|  |  |
| --- | --- |
| 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) | Nominal |

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 | Continuous |
| Weight | Continuous |
| Hair Color | Nominal |
| Socioeconomic Status | Nominal (Lower economical class, Middle, Higher etc ) |
| Fahrenheit Temperature | Continuous |
| Height | Continuous |
| Type of living accommodation | Nominal (Ex: Individual house, Row house, Apartment etc) |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Continuous |
| Blood Group | Nominal |
| Time Of Day | Continuous |
| Time on a Clock with Hands | Continuous |
| Number of Children | Discrete |
| Religious Preference | Nominal |
| Barometer Pressure | Continuous |
| SAT Scores | Discrete |
| Years of Education | Discrete |

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

Ans:

Single coin’s total possible outcomes: 2 (head / tail)

Three coin’s total possible outcomes: (Outcomes Number of coins ) = 23 = 8 (HHH, HHT, HTH, HTT, THH, THT, TTH, TTT)

Favorable outcomes (2 heads, 1 tail): 3 (HHT, HTH, THH)

Probability of getting two heads and one tail where three coins are tossed: 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:

Total number of outcomes of single dice: 6

Total number of outcomes of two dice: 62 = 36

1. Probability that is equal to 1

Total number of favorable outcomes combinations: 0

Probability: 0/36 = 0

1. Less than or equal to 4

Total number of favorable outcomes: 6 (1,1; 1,2; 1,3; 2,1; 2,2; 3,1;)

Probability: 6/36 = 0.167

1. Sum is divisible by 2 and 3

Two dice sum range: 2 to 12

Numbers which are divisible by both 2 and 3 are: 6, 12

Hence, total number of favorable outcomes: 2

Probability: 2/36 = 1/18 = 0.0555

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:

nCr =  n!/(n-r)! \* r!

Total number of balls: 7

Total number of combinations to draw 2 balls out of 7 balls: 7C2 =

= (7! / (7-2)! \* 2!

= 7 \* 6 / 2

= 42/2

= 21

Total number of balls without blue balls: 5

Total number of combinations to draw 2 balls out of 5 balls: 5C2 =

= (5! /(5- 2)! \* 2!

=5\*4 / 2

= 20 / 2

= 10

Probability that none of the balls drawn is blue: 10/21 = 0.4761

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

Answer:

Expected number of candies for 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,Weight>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Points Analysis |  |  | Score Analysis | |  | Weight Analusis | |
|  |  |  |  |  |  |  |  |
| Mean | 3.596563 |  | Mean | 3.21725 |  | Mean | 17.84875 |
| Median | 3.695 |  | Median | 3.325 |  | Median | 17.71 |
| Mode | 3.92 |  | Mode | 3.44 |  | Mode | 17.02 |
|  |  |  |  |  |  |  |  |
| Variance | 0.285881 |  | Variance | 0.957379 |  | Variance | 3.193166 |
| Standard Deviation | 0.534679 |  | Standard Deviation | 0.978457 |  | Standard Deviation | 1.786943 |
| Range | 2.17 |  | Range | 3.911 |  | Range | 8.4 |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Minimum value | 2.76 |  |  | 1.513 |  |  | 14.5 |
| Maximum value | 4.93 |  |  | 5.424 |  |  | 22.9 |
|  |  |  |  |  |  |  |  |
| IQR (Inter Quartile Range) |  |  | IQR Calculations | |  | IQR Calculations | |
| Q1 | 3.08 |  | Q1 | 2.58125 |  | Q1 | 16.8925 |
| Q3 | 3.92 |  | Q3 | 3.61 |  | Q3 | 18.9 |
|  |  |  |  |  |  |  |  |
| IQR | 0.84 |  | IQR | 1.02875 |  | IQR | 2.0075 |
|  |  |  |  |  |  |  |  |
| Outlier Min | 1.82 |  | Outlier Min | 1.038125 |  | Outlier Min | 13.88125 |
| Outlier Max | 5.18 |  | Outlier Max | 5.153125 |  | Outlier Max | 21.91125 |
|  |  |  |  |  |  |  |  |
| Inferences |  |  |  |  |  |  |  |
| 1. No big difference between mean and median indicates that data distribution is normal. May not have any extreme values. | | | | | | | |
| 2. No outliers exists in points column. |  |  |  |  |  |  |  |
| 3. Honda Civic got highest points |  |  |  |  |  |  |  |
| 4. Cadillac Fleetwood, Lincoln Continental, Chrysler Imperial have Scores as outliers. | | | | | |  |  |
| 5. Merc230 has maximum weight and can be stated as outlier. | | | |  |  |  |  |

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?

Answer: Average of patients weight = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Speed Analysis |  |  | |  |  |
|  |  |  | |  |  |
| Skewness | -0.11751 |  | |  |  |
| Kurtosis | -0.50899 |  | |  |  |
|  |  |  | |  |  |
|  |  |  | |  |  |
| Inferences for speed column |  |  | |  |  |
| Skewness is between -0.8 and 0.8. So, speed data is fairly symmetrical and data distribution is normal. |  |  | |  |  |
| Kurtosis is between acceptable range i.e. -3 to +3. -ve Kurtosis indicates range is slightly flatter. | |  | |  |  |
|  |  |  | |  |  |
| dist column analysis |  |  | |  |  |
|  |  |  | |  |  |
| Skewness | 0.806895 |  | |  |  |
| Kurtosis | 0.405053 |  | |  |  |
|  |  |  | |  |  |
| Inferences for dist column |  |  | |  |  |
| Skewness is in the range of +/-0.8. So, speed data is also symmetrical and data distribution is normal. | | | | |  |
| Kurtosis is between acceptable range i.e. +/-3. Most of the dist values are in acceptable range without much peakedness. | | |

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |
| --- | --- |
| SP column analysis |  |
|  |  |
| Skewness | 1.61145 |
| Kurtosis | 2.977329 |
|  |  |
| Inteferences: |  |
| Skewness value is outside of acceptable range -0.8 to +0.8. So, data is mostly asymmetrical and distribution is not normal. |  |
| Kurtosis value is within range +/-3. No too much peakedness and it is acceptable. |  |
|  |  |
| WT column analysis |  |
|  |  |
| Skewness | -0.61475 |
| Kurtosis | 0.950291 |
|  |  |
| Inferences of WT column |  |
| Skewness is within acceptable range +/-0.8. So, distribution is symmetrical. |  |
| Kurtosis value is within acceptable range +/-3. So, no peakedness and no outliers mostly. |  |

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



Inferences:

Histogram indicates data is asymmetrical and right skewed data.

Boxplot indicates that there are some outliers which are above upper maximum range (Q3 + 1.5 \* IQR). Since median is near to lower range, the data is right skewed or positively skewed.

**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?

Answer:

Given parameters are as follows.

Population (N) = 3,000,000

Sample size (n) = 2000

Sample mean ( xbar) = 200 pounds

Sample Standard Deviation (s) = 30 pounds

Confidence Interval (CI) = xbar (+or-) z\*s/sqrt(n)

1. **94% Confidence Interval:**

z scord for 94% confidence interval = 1.88

Confidence Intervals:

1. 200 + 1.88 \* 30/sqrt(2000) = **201.26**
2. 200 - 1.88 \* 30/sqrt(2000) = **198.73**
3. **98% Confidence Interval:**

z score for 98% confidence interval: 2.33

Confidence Intervals:

1. 200 + 2.33 \* 30/sqrt(2000) = **201.56**
2. 200 - 2.33 \* 30/sqrt(2000) = **198.43**
3. **96% Confidence Interval:**

z score for 96% confidence interval: 2.06

Confidence Intervals:

1. 200 + 2.06 \* 30/sqrt(2000) = **201.38**
2. 200 - 2.06 \* 30/sqrt(2000) = **198.61**

**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?

Answer:

1. Mean: Average of all scores: 41

Median: Total number of scores: 18

Average of middle numbers: (40+41)/2 = 40.5

Mode: Most repetitive number: 41 (repeated 4 times)

Variance:

S2 = Sigma(xi - xbar)2/n-1

= 434/17 = 25.529

Standard Deviation: Sqrt(Variance) = 5.052

1. Student always maintained consistency in scoring the marks with average marks as 41.

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

Answer: Data is symmetrical and normal distribution.

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

Answer: Right skewed or Positive skewed distribution data

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

Answer: Left skewed or Negatively skewed data distribution.

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

Answer: Positive Kurtosis value indicates peakedness. It indicates heavy or long tails. Data range is high with extreme high and low values. High risk data with large data variations.

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

Answer: Negative Kurtosis indicates almost flattened data with thin tails. Small range of data. Peakedness is less compared to normal distribution also. This is also low risky data with less data variation.

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



What can we say about the distribution of the data?

Answer: Data distribution is not normal and asymmetrical.

What is nature of skewness of the data?

Answer: Since median is near to Q3 and away from Q1, expect that data is negatively skewed or left skewed.

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

Answer: IQR = 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.

Answer:

Brown box plot data is of narrow range and blue boxplot has wide range of data comparatively.

Median is almost at the center of the data and whiskers are almost equal for both of the boxplots indicates symmetrical data. No outliers in both the plots.

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)

Answer:

1. Number of rows with MPG value greater than 38 is : 33

Total number of rows: 81

Probability: 33/81 = 0.407

1. P(MPG < 40)

Number of rows with MPG value less than 40: 61

Total number of rows: 81

Probability = 61/81 = 0.7530

1. P(20<MPG<50)

Number of rows with MPG values above 20 and below 50: 69

Total number of rows: 81

Probability: 69/81 = 0.8518

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer: Skewness of MPG of Cars is -0.17795. This is within acceptable range -0.8 to +0.8 of normal distribution. So, data is 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

Answer: Skewness of AT is 0.584 which is within acceptable range +/-0.8. Hence, can be considered data is symmetrical and follows normal distribution.

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

**Answers:**

**Calculation of Z-score for 90% confidence interval:**

Area = (1 + CI)/2 = (1+0.90)/2 = 0.95

z score for above value using z-table: (1.64 + 1.65)/2 = 1.645

**Calculation of Z-score for 94% confidence interval:**

Area = (1 + CI)/2 = (1+0.94)/2 = 0.97

z score for above value using z-table: (1.88 + 1.89)/2 = 1.885

**Calculation of Z-score for 60% confidence interval:**

Area = (1 + CI)/2 = (1+0.60)/2 = 0.8

z score for above value using z-table: (0.84 + 0.85)/2 = 0.845

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

Answers:

**Calculation of t score for 95% confidence interval:**

Significance level (alpha) = 1-0.95 = 0.05

Degree of freedom = n-1 = 24

t score for above value using t table (two tails): 2.064

**Calculation of t score for 96% confidence interval:**

Significance level (alpha) = 1-0.96 = 0.04

Degree of freedom = n-1 = 24

t score for above value using t table (two tails): 2.20

(95% Confidence and for df 24=2.064; 98% confidence and for df 24: 2.492;

i.e. for 1% confidence between 95 and 98: (2.492 – 2.064)/3 = 0.142

96% confidence and for df24 = 2.064+0.142 = 2.206)

**Calculation of t score for 99% confidence interval:**

Significance level (alpha) = 1-0.99 = 0.01

Degree of freedom = n-1 = 24

t score for above value using t table(two tails): 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

Answer:

Null Hypothesis and alternate hypothesis definitions for this issue:

H0 = Bulb lasts for no more than 260 days i.e. Mu <=260

H1 = Mu > 260

Given values:

Population mean (Mu) = 270 days

Sample size (n) = 18

Sample mean (xbar) = 260 days

Sample standard deviation (s) = 90 days

Since the sample size is less than 30 and population standard deviation is not known, need to go for t-score.

t = (xbar – mu)/(s/sqrt(n))

= (260 – 270)/(90/sqrt(18))

= -0.4714

This is single tail as per hypothesis definition.

Significance value alpha = 0.05 (95% confidence interval)

Probability value for above t-score: stats.t.cdf(-0.474, df=18) = 0.3219

Probability value (0.3219) is greater than significance alpha value (0.05).

Hence accepts the null hypothesis

**i.e. the bulbs average life is <= 260 days**