|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete Data |
| Results of rolling a dice | Discrete Data |
| Weight of a person | Continous Data |
| Weight of Gold | Continous Data |
| Distance between two places | Continous Data |
| Length of a leaf | Continous Data |
| Dog's weight | Continous Data |
| Blue Color | Continous Data |
| Number of kids | Discrete Data |
| Number of tickets in Indian railways | Discrete Data |
| Number of times married | Discrete Data |
| Gender (Male or Female) | Continous Data |

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

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

ANSWER:-

There are three possible outcomes when you toss a single coin: H (head) or T (tail). Since we want to find the probability of getting two heads and one tail in any order, we can calculate the probability of each possible outcome and then add them up.

The possible outcomes for getting two heads and one tail are:

H, H, T

H, T, H

T, H, H

Now, let's calculate the probability of each of these outcomes:

1:-Probability of H, H, T:

* Probability of getting a head on the first toss: 1/2
* Probability of getting a head on the second toss: 1/2
* Probability of getting a tail on the third toss: 1/2
* Probability of this outcome = (1/2) \* (1/2) \* (1/2) = 1/8

2:-Probability of H, T, H:

* Probability of getting a head on the first toss: 1/2
* Probability of getting a tail on the second toss: 1/2
* Probability of getting a head on the third toss: 1/2
* Probability of this outcome = (1/2) \* (1/2) \* (1/2) = 1/8

3:-Probability of T, H, H:

* Probability of getting a tail on the first toss: 1/2
* Probability of getting a head on the second toss: 1/2
* Probability of getting a head on the third toss: 1/2
* Probability of this outcome = (1/2) \* (1/2) \* (1/2) = 1/8

Total Probability = (1/8) + (1/8) + (1/8) = 3/8 (ANS)

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

* Equal to 1
* Less than or equal to 4
* Sum is divisible by 2 and 3

ANSWER:-

To find the probability of different sums when two dice are rolled, you can first list all the possible outcomes and then count the favorable outcomes for each condition. Each die has 6 sides, so there are a total of 6 \* 6 = 36 possible outcomes when rolling two dice.

a) Probability that the sum is equal to 1:

Equal to 1= 0

b) Probability that the sum is less than or equal to 4:

The possible sums less than or equal to 4 are 2, 3, and 4. Let's count the favorable outcomes for each:

- For a sum of 2, there is only one way to achieve it: rolling a 1 on both dice (1, 1). So, the probability is 1/36.

- For a sum of 3, you can roll (1, 2) or (2, 1). There are two favorable outcomes. So, the probability is 2/36 = 1/18.

- For a sum of 4, you can roll (1, 3), (2, 2), or (3, 1). There are three favorable outcomes. So, the probability is 3/36 = 1/12.

Now, add the probabilities for each case:

Total Probability = (1/36) + (1/18) + (1/12) = (2/72) + (4/72) + (6/72) = 12/72 = 1/6

c) Probability that the sum is divisible by both 2 and 3:

There are 5 rolls that produce 6, i.e., 1–5, 2–4, 3–3, 4–2, 5–1. This we have 6 of the 36 possible rolls that produce sums that are divisible by both 2 and 3. That is 16.67%.

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?

ANSWER:-

Total no of red balls= 2

Total no of green balls=3

Total no of blue balls= 2

Total no of balls = 2+3+2 = 7

Probability of getting no blue balls when random picked:

P= (5\*4/2\*1) / (7\*6/2\*1) =10/21

P= 0.476

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:-

To calculate the expected number of candies for a randomly selected child, you can multiply each child's candy count by their respective probabilities and then sum these values.

Expected number of candies (E) = (Probability of A \* Candy count of A) + (Probability of B \* Candy count of B) + (Probability of C \* Candy count of C) + (Probability of D \* Candy count of D) + (Probability of E \* Candy count of E) + (Probability of F \* Candy count of F)

Expected number of candies = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

Expected number of candies = 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24

Expected number of candies = 3.14

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**

| **Points** |  | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **count** |  | 32.000000 | 32.000000 | 32.000000 |
| **mean** |  | 3.596563 | 3.217250 | 17.848750 |
| **std** |  | 0.534679 | 0.978457 | 1.786943 |
| **min** |  | 2.760000 | 1.513000 | 14.500000 |
| **25%** |  | 3.080000 | 2.581250 | 16.892500 |
| **50%** |  | 3.695000 | 3.325000 | 17.710000 |
| **75%** |  | 3.920000 | 3.610000 | 18.900000 |
| **max** |  | 4.930000 | 5.424000 | 22.900000 |

**calculate Median**

Points 3.695

Score 3.325

Weigh 17.710

**calculate Mode**

**Points Score Weigh**

**0 3.07 3.44 17.02**

**1 3.92 NaN 18.90**

**calculate Variance**

Points 0.285881

Score 0.957379

Weigh 3.193166

calculate Range

Points 2.170

Score 3.911

Weigh 8.400

Q8) Calculate Expected Value for the problem below

* 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:-

To calculate the expected value (also known as the mean)

Weight:-108, 110, 123, 134, 135, 145, 167, 187, 199

Expected Value = (Sum of Weights) / (Number of Patients)

Expected Value = 1,308 / 9 ≈ 145.33 pounds

So, the expected value of the weight of a randomly chosen patient is approximately 145.33 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

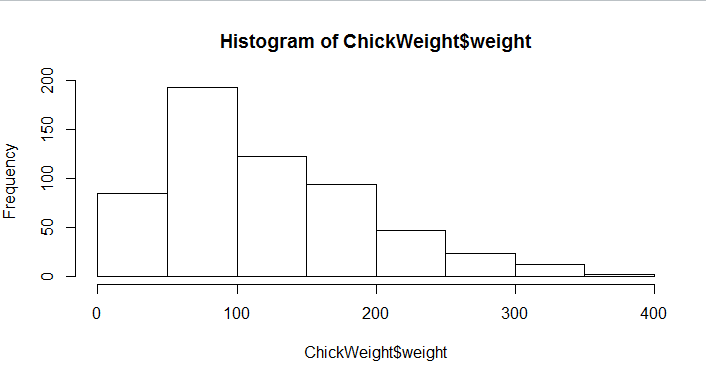
|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| speed | -0.11 | -0.508994 |
| distance | 0.806895 | 0.405053 |

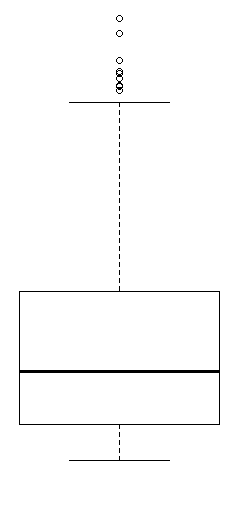
**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| speed | 1.611450 | 2.977329 |
| weight | -0.614753 | 0.950291 |

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





75%(300)(Upper quartile)

25%(100)(Lower Quartial)

1.5

Median (200)

IQR=UQ-LQ=300-100=200

LE(Lower Extreme)=LQ-1.5\*IQR=100-(1.5\*200)=-200

UE(Upper Extreme)=UQ-1.5\*IQR=300-(1.5\*200)=0

* The above histogram has right tail
* The data is positive Skewed
* Most number of chick weigh from the range of 50-100 and followed by 100-200

**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:-

Confidence Interval = X̄ ± Z \* (σ/√n)

Let:

1. X̄ is the sample mean=200 pounds
2. N is the sample size =2,000
3. σ is the population standard deviation =30 pounds
4. Z is the Z-score corresponding to the desired confidence level (e.g., 94%, 98%, or 96% confidence). You can find these Z-scores in a standard normal distribution table.

Calculate 94% confidence intervals:-

CI= X̄ ± Z \* (σ/√n)

CI= 200 ±1.88\*(30/√2000)

CI= 200 ± 2.65

Calculate the confidence interval:(197.35, 202.65)

Calculate 98% confidence intervals:-

CI= X̄ ± Z \* (σ/√n)

CI= 200 ±2.33\*(30/√2000)

CI= 200 ± 3.28

Calculate the confidence interval:( 196.72, 203.28)

Calculate 96% confidence intervals:-

CI= X̄ ± Z \* (σ/√n)

CI= 200 ±1.75\*(30/√2000)

CI= 200 ± 2.47

Calculate the confidence interval:( 197.53, 202.47)

**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**

* Find mean, median, variance, standard deviation.
* What can we say about the student marks?

ANSWER:-

Scores: 34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

Mean (Average):

Mean = (34 + 36 + 36 + 38 + 38 + 39 + 39 + 40 + 40 + 41 + 41 + 41 + 41 + 42 + 42 + 45 + 49 + 56) / 18

Mean =41.83

Median:

The median is the middle value when the scores are arranged in ascending order. Since there are 18 scores, the median will be the average of the 9th and 10th values.

Median = (40 + 40) / 2 = 40

Variance:

Variance = Σ(xi - x̄)² / (n - 1)

x̄ (mean)=41.83

xi= each individual score

n = 18

Variance = [(34 - 41.83)² + (36 - 41.83)² +(36 - 41.83)²+ (38 - 41.83)²+ (38 - 41.83)²+ (39 - 41.83)²+ (39 - 41.83)²+ (40 - 41.83)²+ (40 - 41.83)²+ (41 - 41.83)²+ (41 - 41.83)²+ (41 - 41.83)²+ (41 - 41.83)²+ (42 - 41.83)²+ (42 - 41.83)²+ (45 - 41.83)²+ (49 - 41.83)²+ (56 - 41.83)²] / (18-1)

Variance ≈ 61.71 (rounded to two decimal places)

Standard Deviation:

Standard Deviation = √Σ(xi - x̄)² / (n - 1)

(Σ(xi - x̄)² / (n - 1)) it al so know as variance

So .

Standard Deviation = √Variance ≈ √61.71 ≈ 7.85 (rounded to two decimal places)

Based on these statistics, we can say that the student's marks tend to cluster around the mean score of 41.83, but there is some variability in the scores. The majority of scores are between 34 and 56, with the middle score (median) being 40. The standard deviation suggests that the spread of scores is moderate, with most scores falling within approximately one standard deviation of the mean.

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

ANSWER:-

When the mean and median of a data distribution are equal, the distribution has zero skewness. This means that the observations are distributed similarly on the left and right sides of the peak.

Skewness can be positive, negative, or zero. When the mean is greater than the median, the distribution is positively skewed. When the mean is less than the median, the distribution is negatively skewed.

In a symmetrical distribution, the mean and median are approximately equal in value. However, in a symmetrical distribution with two modes (bimodal), the two modes would be different from the mean and median.

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

ANSWER:-

Skewness is a measure of how asymmetrical a distribution is. When the mean is greater than the median, the distribution is positively skewed.

For distributions that have outliers or are skewed, the median is often the preferred measure of central tendency because the median is more resistant to outliers than the mean

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

ANSWER:-

Skewness is a measure of how asymmetrical a distribution is. When the median is greater than the mean, the distribution is negative skewed.

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

ANSWER:-

Kurtosis is a measure of how peaked a distribution is. A positive kurtosis value indicates that a distribution is more peaked than normal. Positive kurtosis values also indicate that a distribution has thick tails.

A distribution with a positive kurtosis value is called a leptokurtic distribution. A leptokurtic distribution has a higher peak and taller tails than a normal distribution.

A distribution with a positive kurtosis value indicates that the data values are more concentrated around the mean than those in a normal distribution.

A distribution with a negative kurtosis value indicates a shape flatter than normal.

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

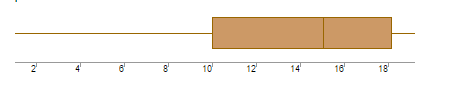
ANSWER:-

A negative kurtosis value indicates that a distribution has a flatter peak and thinner tails than a normal distribution. This type of distribution is called platykurtic.

Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. A kurtosis value near zero indicates a shape close to normal. A negative value indicates a distribution which is more peaked than normal. A positive kurtosis indicates a shape flatter than normal.

Data sets with high kurtosis have heavy tails and more outliers. Data sets with low kurtosis tend to have light tails and fewer outliers.

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



What can we say about the distribution of the data?\

Answer: The given data lies on right side, so it not symmetrical in nature and the distribution is left hand tail

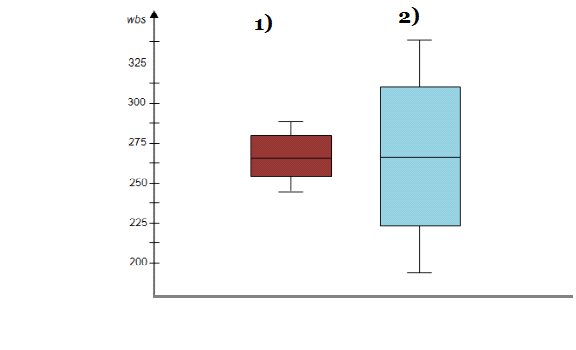
What is nature of skewness of the data?

Answer: Negatively Skewed

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

Answer: IQR=UQ-LQ=18-10=8(app)

Q19) Comment on the below Boxplot visualizations?



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

Answer:

* Both the Box Plots are Normally Distributed
* Both have Same Median around 262(app)
* Comparatively Second BoxPlot has large range compared to 1st Box plot

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)

Ans:-

0.34759394041453007

1. P(MPG<40)

Ans:-

0.7293498604157946

c. P (20<MPG<50)

Ans:-

0.8988689076273199

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer: MPG has normal Distribution

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

Dataset: wc-at.csv

Answer: wc-at.csv dataset has not normal Distribution

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

Ans:-

|  |  |  |
| --- | --- | --- |
| Confidence interval | Alpha a | Z score |
| 94% | 0.03 | 1.880 |
| 90% | 0.05 | 1.644 |
| 60% | 0.02 | 0.841 |

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

Ans:-

|  |  |
| --- | --- |
| 95% | ±2.060 |
| 96% | ±2.167 |
| 99% | ±2.787 |

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

Define your hypotheses:

Ans:-

1>Null Hypothesis (H0): The CEO's claim is true, and the population mean (μ) is 270 days.

Alternative Hypothesis (Ha): The CEO's claim is not true, and the population mean (μ) is not 270 days (two-tailed test).

2>Calculate the standard error of the sample mean (standard error of the mean, SEM):

SEM = Standard Deviation (σ) / √n

SEM = 90 / √18 ≈ 21.21

3>Calculate the t-score:

t = (Sample Mean - Population Mean) / SEM

t = (260 - 270) / 21.21 ≈ -0.471

4>Determine the degrees of freedom (df):

df = n - 1 = 18 - 1 = 17

5>Calculate the probability (p-value) associated with the t-score. You'll use a t-distribution table, calculator, or software for this step.

stats.t.cdf(-0.471,17)

Ans:-

0.3218140331685075

P>α

0.321>0.05(Reject Null Hypothesis)