|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data - Ratio |
| Results of rolling a dice | Discrete data - Ordinal |
| Weight of a person | Continuous data - Ratio |
| Weight of Gold | Continuous data - Ratio |
| Distance between two places | Continuous data - Ratio |
| Length of a leaf | Continuous data - Ratio |
| Dog's weight | Continuous data - Ratio |
| Blue Color | Discrete data - Nominal |
| Number of kids | Discrete data - Ratio |
| Number of tickets in Indian railways | Discrete data - Ratio |
| Number of times married | Discrete data - Ratio |
| Gender (Male or Female) | Discrete data - 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 | 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) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans:

P(2 heads and 1 tail) = Number of favorable outcomes/Total number of outcomes

P(2 heads and 1 tail) = 3/8

Therefore, the probability is 0.375.

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:

1. Equal to 1:

P(Sum=1) = Number of favorable outcomes/Total number of outcomes

= 0/36 = 0

1. Less than or equal to 4:

The possible combinations for sums less than or equal to 4 are (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1).

P(Sum≤4) = Total number of outcomes / Number of favorable outcomes

= 6/36 = 1/6 = 0.167

1. Sum is divisible by 2 and 3:

The sums that are divisible by both 2 and 3 are 6 and 12.

Total no. of possible combinations which sums up to 6 are 5:

(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)

Total no. of possible combinations which sums up to 12 is 1:

(6, 6)

So, the total number of favorable outcomes is 6.

P(Sum is divisible by 2 and 3) = Number of favorable outcomes/Total number of outcomes

= 6/36 = 1/6 = 0.167.

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:

Given, a bag contains 2 red, 3 green and 2 blue balls.

We need to find the probability that none of the drawn balls is blue.

Total number of balls = 2 + 3 + 2 = 7

Two balls can be drawn in 7C2 ways = 21 ways

Number of ways of drawing 2 balls such that none is blue = Number of ways of drawing 2 balls from 2 red and 3 green balls = 5C2 = 10 ways

Probability of drawing 2 balls such that none is blue = 10/21

Therefore, the correct answer 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 number of candies for a randomly selected child, we need to multiply each candy count by its corresponding probability and then sum up these values. The formula for the expected value (E) is given by:

E = Sum(Xi\*Pi)

where:

- Xi is the candy count for child i,

- Pi is the probability of child i having that candy count,

- The summation is taken over all children.

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count | Probability | Candies count \* Probability |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.2 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.12 | 0.24 |

E (Total summation) = 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

E = 3.09

Therefore, the expected number of candies for a randomly selected child is 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.

**Use Q7.csv file**

**Ans: The solution file has been attached (Attached file - Assignment 1 Q7 solution).**

1. **Points**

**Mean - 3.5965625**

**Median - 3.695**

**Mode - 3.92**

**Variation - 0.276947559**

**Standard deviation - 0.526258072**

**Inferences –**

* **The mean and median are relatively close, suggesting that the data is roughly symmetrically distributed.**
* **The standard deviation (0.526258072) indicates moderate variability in the dataset.**

1. **Score**

**Mean - 3.21725**

**Median - 3.325**

**Mode - 3.44**

**Variation - 0.927460875**

**Standard deviation - 0.963047701**

**Inferences –**

* **The mean and median are relatively close, suggesting that the data is roughly symmetrically distributed.**
* **The standard deviation is around 0.963047701, representing the spread of the data.**

1. **Weigh**

**Mean - 17.84875**

**Median - 17.71**

**Mode - 17.02**

**Variation - 3.093379688**

**Standard deviation - 1.758800639**

**Inferences –**

* **The mean and median are relatively close, suggesting that the data is roughly symmetrically distributed.**
* **The standard deviation is approximately 1.75, representing the spread of the data.**

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?

Sol:

**The solution file has been attached (Attached file - Assignment 1 Q8 solution)**

Expected Value = (1/9)(108) + (1/9)110 + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

= (1/9) ( 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9) ( 1308)

= 145.33

Expected Value of the Weight of a random patient = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Sol:**

**(Please check the files Assignment 1 Q9\_a solution and Assignment Q9\_b solution for calculation details.)**

1. **Cars speed and distance.**

**Skewness for speed = -0.117509861**

**Skewness for distance = 0.80689496**

**Kurtosis for speed = -0.50899442**

**Kurtosis for distance = 0.405052582**

**Inferences:**

**For "speed," the skewness is negative, indicating a left-skewed distribution.**

**For " distance," the skewness is positive, indicating a right-skewed distribution.**

**For "speed," the kurtosis is negative, indicating a Platykurtic distribution (Tails are lighter and distribution is less peaked).**

**For " distance," the kurtosis is positive, indicating a slightly leptokurtic distribution (more peaked than a normal distribution).**

1. **SP and Weight**

**Skewness for SP = 1.611450196**

**Skewness for Weight = -0.614753326**

**Kurtosis for SP = 2.977328944**

**Kurtosis for Weight = 0.950291491**

**Inferences:**

**For "SP," the skewness is positive, indicating a right-skewed distribution.**

**For "Weight," the skewness is negative, indicating a left-skewed distribution.**

**For " SP," the kurtosis is positive, indicating a leptokurtic distribution (more peaked than a normal distribution).**

**For "Weight," the kurtosis is positive, indicating a slightly leptokurtic distribution (more peaked than a normal distribution).**

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



Sol:

1. Histogram

Range of 50-100 has the highest frequency of 200.

Range of 350 to 400 has the least frequency of around 0-10.

Skewness – It has a long tail towards the right. So it is heavily right skewed.

Mean is greater than the median.

1. Boxplot

It is skewed towards right.

Median is less than mean.

There are outliers on the upper side of the box plot.

The majority of the data is concentrated towards lower values.

**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?

Sol:

To calculate confidence intervals for the average weight of adult males in Mexico, we can use the formula for the below formula:

Confidence Interval = Sample Mean ± (Critical Value \* Standard Deviation/Sqrt of Sample Size)

The critical value is determined by the desired confidence level and the distribution of the data.

For a normal distribution (which is often assumed for large sample sizes), we can use the Z-table.

For a 94% confidence level, the critical value is Z = 1.88

For a 96% confidence level, the critical value is Z = 2.05

For a 98% confidence level, the critical value is Z = 2.33

Now, let's calculate the confidence intervals:

1. 94% Confidence Interval:

CI for 94% = 200 ± (1.88 × 30/Square root of 2000)

CI for 94% = 200 ± 1.261142339

So, the 94% confidence interval is approximately (198.74, 201.26)

2. 96% Confidence Interval:

CI for 96% = 200 ± (2.05 × 30/Square root of 2000)

CI for 96% = 200 ± 1.375181806

So, the 96% confidence interval is approximately (198.62, 201.38)

3. 98% Confidence Interval:

CI for 98% = 200 ± (2.33 × 30/Square root of 2000)

CI for 98% = 200 ± 1.563011516

So, the 98% confidence interval is approximately (198.44, 201.56)

94% Confidence interval is approximately (198.74, 201.26)

94% Confidence interval is approximately (198.62, 201.38)

94% Confidence interval is approximately (198.44, 201.56)

**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?

Sol:

Mean is 41

Median is 40.5

Variance is 24.11

Standard deviation is 4.91

The student's average performance is around 41, with a relatively symmetric distribution of scores. However, there is some variability in the scores, as indicated by the variance and standard deviation.

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

Ans: Symmetrical

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

Ans: Positively skewed or right-skewed

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

Ans: Negatively skewed or left-skewed

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

Ans:

Positive kurtosis indicates that a probability distribution has heavier tails and a more peaked or sharp central peak than the normal distribution.

There is an increased probability of extreme values in the tails of the distribution.

The tails of the distribution have more mass, and the distribution has a higher peak.

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

Ans:

A negative kurtosis value indicates that a probability distribution has lighter tails and a flatter central peak than the normal distribution.

There is a decreased probability of extreme values in the tails of the distribution.

The tails of the distribution have less mass, and the distribution has a flatter central peak.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

1) 50% of the population is above 10 and the remaining are less.

2) It is Negatively skewed or left-skewed as the Median is greater than the Mean.

3) IQR = Q3 - Q1

= 18 – 10

= 8

The IQR of the data is approximately 8.

Q19) Comment on the below Boxplot visualizations?



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

Ans:

Mean and median in both of the boxplots are approximately equal.

The range and the IQR is less in boxplot 1 compared to boxplot 2.

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:

1. P(MPG>38)

= 0.3475939251582705

1. P(MPG<40)

= 0.7293498762151616

1. P(20<MPG<50)

= 1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

Sol:

1. The Kurtosis of -0.611 shows that the MPG of Cars doesn’t follow Normal Distribution as it’s not 0. It has a Negative Kurtosis (Platykurtic). However it’s very close to normal distribution.
2. The Adipose Tissue (AT) doesn’t follow normal distribution as its Kurtosis is -1.102 (not zero).

The Waist Circumference(Waist)’s Kurtosis is -0.28 which shows that it’s very close to Normal distribution although it doesn’t exactly follow Normal distribution.

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

Ans:

For a 90% confidence interval, the Z score is approximately 1.645.

For a 94% confidence interval, the Z score is approximately 1.88.

For a 60% confidence interval, the Z score is approximately 0.84.

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

Ans:

95% confidence interval = ±2.064,

96% confidence interval = ±2.177,

99% confidence interval = ±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

Ans:

To calculate the probability that 18 randomly selected bulbs would have an average life of no more than 260 days if the CEO's claim were true, we will perform a one-sample t-test:

Hypotheses:

- Null Hypothesis (H0): The average life of the bulbs is 270 days (mu = 270).

- Alternative Hypothesis (H1): The average life of the bulbs is less than 270 days (mu < 270).

Standard Error of the Mean (SE):

SE = s/sqrt(n)

where s is the sample standard deviation, and n is the sample size.

SE = 90/sqrt(18)

t-Score (t):

t = (X – mu)/SE

where X is the sample mean.

t = (260-270)/(90/sqrt(18))

t = -0.4714

Degrees of Freedom (df):

Degrees of freedom (df) for a one-sample t-test is n - 1, where n is the sample size.

df = 18 - 1 = 17

Calculate the Probability:

Using R’s pt function, we found that the probability is 0.3216741

Hence the probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.3216741.