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

**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 | Nominal |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

**Answer:**

Let S be the sample space then

S={H,T},{H,T},{H,T} then the

event E is {HHT,HTH,THH}

Let X be a random variable denoting the two heads and one tail.

P(X=2)=probability of occurrence of 2 heads and 1 tail.

=P(HHT)+P(HTH)+P(THH)

=12.12.12+12.12.12+12.12.12=38

**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**

**Answer:**

**a) Probability that the sum is equal to 1:**

There is only one way to get a sum of 1, which is by getting a 1 on the first die and a 1 on the second die. There is only one favorable outcome.

Total possible outcomes when rolling two dice = 6 \* 6 = 36

P(Sum is 1) = (Number of favorable outcomes) / (Total possible outcomes) = 1 / 36

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

Favorable outcomes for a sum less than or equal to 4 are:

1 + 1 = 2

1 + 2 = 3

2 + 1 = 3

1 + 3 = 4

3 + 1 = 4

2 + 2 = 4

There are 6 favorable outcomes.

P(Sum is less than or equal to 4) = (Number of favorable outcomes) / (Total possible outcomes) = 6 / 36 = 1 / 6

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

To find the sums that are divisible by both 2 and 3, you are looking for sums that are multiples of 6. The possible sums that meet this criterion are:

6, 12

There are 2 favorable outcomes.

P(Sum is divisible by 2 and 3) = (Number of favorable outcomes) / (Total possible outcomes) = 2 / 36 = 1 / 18

So, the probabilities are:

**a) P(Sum is 1) = 1/36**

**b) P(Sum is less than or equal to 4) = 1/6**

**c) P(Sum is divisible by 2 and 3) = 1/18**

**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 number of balls = (2 + 3 + 2) = 7 Let S be the sample space. Then, n(S) = Number of ways of drawing 2 balls out of 7 =7C2 =(7×6)(2×1) =21 Let E = Event of drawing 2 balls, none of which is blue. ∴n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls. =5C2 =(5×4)(2×1) =10 ∴P(E)=n(E)n(S)=1021

**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 a randomly selected child  = 3.09

Expected number of candies for a randomly selected child

=  1 \* 0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12

= 0.015 + 0.8  + 1.95 + 0.025 + 0.06 + 0.24

=  3.090

=  3.09

**Expected number of candies for a randomly selected child  = 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**

**Answer:**

**Please check the answer in Question-7.ipynb file**

**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:**

Expected Value  =  ∑ ( probability  \* Value )

 ∑ P(x).E(x)

there are 9 patients

Probability of selecting each patient = 1/9

Ex  108, 110, 123, 134, 135, 145, 167, 187, 199

P(x)  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

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 that patient = 145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Answer:**

**Please check the answer in Question-9a-9b.ipynb file**

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



**Answer:**

The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side.



**Answer:**

The boxplot has outliers on the maximum side.

**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 = ´ X ±t(s/√n)

Where ´ X is Mean = 200

s is standard deviation = 30

N is sample size = 2000

**CI94% :α = 1 - (confidence level / 100)**

= 1- (94%/100)

= 0.06

Critical probability (p\*) = 1 - α/2

= 1-0.03

= 0.97

Degrees of freedom =n-1=2000-1=1999

t-score =1.882

**CI94% = 200±1.2620**

**CI98% :α = 1 - (confidence level / 100)**

= 1-(98%/100)=0.02

Critical Probability (p\*) = 1 - α/2

= 0.99

Degrees of freedom =n-1=2000-1=1999

t-score = 2.328

**CI98% = 200± 1.561**

**CI96%:α = 1 - (confidence level / 100)**

= 1 – (96%/100)

=0.04

Critical Probability (p\*) = 1-α/2

= 0.98

Degrees of freedom =n-1=2000-1=1999

t-score = 2.055

**CI98%= 200±1.378**

**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**

**Answer:**

1. **Find mean, median, variance, standard deviation.**

1) To find the mean, median, variance, and standard deviation for the given test scores, follow these steps:

**a) Mean (Average):**

To find the mean, add up all the scores and divide by the total number of scores.

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

**Mean = 698 / 18 = 38.78 (rounded to two decimal places)**

**b) Median:**

To find the median, first arrange the scores in ascending order and then find the middle value. Since there are 18 scores, the median will be the average of the 9th and 10th values when sorted.

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

**Median = (40 + 41) / 2 = 40.5**

**c) Variance:**

To find the variance, you'll need to calculate the average of the squared differences between each score and the mean. Here's the formula for variance:

Variance = Σ(xi - mean)² / (n - 1)

Where Σ represents the summation, xi is each score, mean is the calculated mean from part (a), and n is the number of scores.

Variance = [(34 - 38.78)² + (36 - 38.78)² + ... + (56 - 38.78)²] / (18 - 1)

Variance = [72.32 + 7.84 + ... + 287.68] / 17

**Variance ≈ 81.13 (rounded to two decimal places)**

**d) Standard Deviation:**

The standard deviation is the square root of the variance.

**Standard Deviation ≈ √(81.13) ≈ 9.00 (rounded to two decimal places)**

1. **What can we say about the student's marks?**

**Answer:**

- The mean score of the student is approximately 38.78.

- The median score is 40.5.

- The data has a variance of approximately 81.13.

- The standard deviation is approximately 9.00.

The data appears to be relatively centered around the mean, with the median close to the mean, which suggests that the distribution of scores is not heavily skewed. The variance and standard deviation indicate the spread or dispersion of the scores around the mean. A higher standard deviation indicates greater variability in the scores, while a lower standard deviation suggests less variability. In this case, the standard deviation of 9.00 is relatively moderate, indicating that the student's scores do not have very high variability.

Overall, the student's marks appear to be fairly consistent, with the majority of scores clustering around the mean, and the data does not exhibit extreme outliers or significant skewness.

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

**Answer:**

No skewness is present we have a perfect symmetrical distribution.

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

**Answer:**

Skewness and tail is towards Right.

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

**Answer:**

Skewness and tail is towards left.

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

**Answer:**

Positive kurtosis means the curve is more peaked and it is Leptokurtic.

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

**Answer:**

Negative Kurtosis means the curve will be flatter and broader.

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



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

**Ans:** The above Boxplot is not normally distributed the median is towards the higher value.

**What is nature of skewness of the data?**

**Ans:** The data is a skewed towards left. The whisker range of minimum value is greater than maximum.

**What wil be the IQR of the data (approximately)?**

**Ans:** The Inter Quantile Range = Q3 Upper quartile – Q1 Lower Quartile = 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:**

 First there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

**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)**

**Answer:**

**Please check the answer in Question-20.ipynb file**

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

**Please check the answer in Question-21a.ipynb file**

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

**Dataset: wc-at.csv**

**Please check the answer in Question-21b.ipynb file**

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

**Answer:**

**For 90% confidence interval:**

We have the significance level at 5 % ( as it is a two tailed test)

that is:

α = 5 % = 0.05

z at α = 0.05 from the z table will be:

z = 1.645.

**For 94 % confidence interval:**

We have the significance level at 3 % ( as it is a two tailed test)

that is:

α = 3 % = 0.03

z at α = 0.03 from the z table will be:

z = 1.555.

**For 60 % confidence interval:**

We have the significance level at 20 % ( as it is a two tailed test)

that is:

α =20 % = 0.2

z at α = 0.2 from the z table will be:

z = 0.253

**Therefore, we get that the z score at 90 % confidence interval is 1.645, at 94 % confidence interval is 1.555 and at 60 % confidence interval is 0.253.**

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

**Answer:**

To compute the 95% confidence interval, start by computing the mean and standard error: M = (2 + 3 + 5 + 6 + 9)/5 = 5. σM = = 1.118. Z.95 can be found using the normal distribution calculator and specifying that the shaded area is 0.95 and indicating that you want the area to be between the cutoff points

Confidence Level z

0.90 1.645

0.92 1.75

0.95 1.96

0.96 2.05

With a 90 percent confidence interval, you have a 10 percent chance of being wrong. A 99 percent confidence interval would be wider than a 95 percent confidence interval (for example, plus or minus 4.5 percent instead of 3.5 percent).

**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:**

**Please check the answer in Question-21b.ipynb file**