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

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) | Interval |
| 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 | 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:**

To find the probability of two heads and one tail can be used in Binomial distribution formula.

P(X=k) = (n/k) pk (1-p)n-k

The number of ways to get two heads out of three tosses is 3. You can think of it as choosing which two out of the three tosses will be heads.

The probability of getting heads (H) on a single toss is 1/2, and we want this to happen twice (two heads), so we square it: (1/2) \* (1/2) = 1/4.

The probability of getting tails (T) on a single toss is also 1/2, and we want this to happen once: 1/2.

Now, we multiply these probabilities:

3×(1/4)×(1/2)=3/8

So, the probability of getting exactly two heads and one tail when tossing three coins is 3/8, which is 37.5%.

**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:**

a) The probability of getting a sum of 1 is 0.

b) The probability of getting a sum less than or equal to 4 is 6/36.

c) The probability of getting a sum divisible by both 2 and 3 is 6/36.

**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:**

First, let's calculate the total number of ways to draw 2 balls out of the 7 in the bag:

Total ways to draw 2 balls = Combination (7, 2) = 21 ways

Now, let's calculate the number of ways to draw 2 balls without any of them being blue. Since there are 2 blue balls, we need to choose 2 balls from the 5 non-blue balls (2 red and 3 green):

Ways to draw 2 non-blue balls = Combination (5, 2) = 10 ways

So, the probability of drawing 2 non-blue balls is:

Probability = (Number of Favorable Outcomes) / (Total Possible Outcomes) = 10 / 21

Therefore, the probability that none of the balls drawn is blue 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**

**Answer:**

To calculate the expected number of candies for a randomly selected child:

Multiply the number of candies each child has by their respective probabilities and add them up:

E = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120) = 3.095

So, the expected number of candies for a randomly selected child is approximately 3.095 candies.

**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 Comment about the values/ Draw some inferences.**

**Use Q7.csv file!**

**Answer:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **score** | **Weigh** |
| **Mean** | 3.596563 | 3.21725 | 17.84875 |
| **Median** | 3.695 | 3.325 | 17.71 |
| **Mode** | 3.92 | 3.44 | 17.02 |
| **Variance** | 0.285881 | 0.957379 | 3.193166 |
| **Standard Deviation** | 0.534679 | 0.978457 | 1.786943 |
| **Range** | 2.17 | 3.911 | 8.4 |

Please refer to python notebook for answer

**Comments and Inference:**

1. Points have a relatively low mean and show some variability with a standard deviation of 0.53. The range in Points is from 2.76 to 4.93.
2. Score has a moderate mean and a wider range with a standard deviation of 0.98. The range in Score is from 1.513 to 5.424.
3. Weight has a mean of 18.81 and shows less variability with a standard deviation of 1.97. The range in Weight is from 14.5 to 22.9.
4. The dataset appears to represent various car models, and the differences in Points, Score, and Weight suggest differences in car performance and characteristics.
5. The lack of modes in Points and Score indicates that there are no predominant values, and the data is relatively diverse.
6. The mode in Weight (17.02) suggests that this weight value occurs more frequently than others.
7. Variance and standard deviation provide insights into data spread and dispersion. Points have the lowest variance, indicating less spread compared to Score and Weight.

**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:**

Expected Value (Mean) Calculation:

Given weights of patients: 108, 110, 123, 134, 135, 145, 167, 187, 199

**Expected Value (Mean) = Sum of weights / Total number of patients**

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

Expected Value 145.33 pounds.

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

**Car’s speed and distance**

**Use Q9\_a.csv**

**SP and Weight (WT)**

**Use Q9\_b.csv**

**Answer:**

Please refer to python notebook for answer

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



**Answer:** The histogram’s peak has a 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:**

For 94% confidence interval Range is [ 198.73 – 201.26]

For 98% confidence interval range is [198.43 – 201.56]

For 96% confidence interval range is [198.62 – 201.37]

Please refer to python notebook for answer

**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:**

We don’t have outliers and the data is slightly skewed towards the right because mean is greater than median.

Please refer to python notebook for answer

**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.

Please refer to python notebook for answer

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

**Answer:**

Skewness and tail is towards Right

Please refer to python notebook for answer

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

**Answer:**

Skewness and tail is towards left

Please refer to python notebook for answer

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

**Answer:**

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

Please refer to python notebook for answer

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

**Answer:**

Negative Kurtosis means the curve will be flatter and broader.

Please refer to python notebook for answer

**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)?**

**Answer:**

**Distribution of the Data:** The data appears to be positively skewed. This means that the tail of the data distribution extends to the right, and the majority of the data points are concentrated on the lower end of the range.

**Nature of Skewness:** As mentioned earlier, the data is positively skewed. This is evident from the fact that Q1 (the first quartile) is closer to the lower end of the data range (Q1 = 10), while Q3 (the third quartile) is further towards the upper end of the data range (Q3 = 18). In a positively skewed distribution, the mean is typically greater than the median.

**Approximate IQR:** The Interquartile Range (IQR) can be calculated as the difference between Q3 and Q1, which is:

IQR = Q3 - Q1 = 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:**

There are no outliers, first. Second, the box plots are both normally distributed, with zero to no skewness at either the minimum or maximum whisker range, and they both share the same median, which is about in the range of 275 to 250.

**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)**

**b. P(MPG<40)**

**c. P (20<MPG<50)**

**Answer:**

1. **P(MPG>38):** Prob\_MPG\_greater\_than\_38 = np. round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3)

print('P(MPG>38) =',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

**b. P(MPG<40):** prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

**c. P (20<MPG<50):** prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20))

p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<50)=',(prob\_MPG\_less\_than\_50))

P(MPG<50)= 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20)

print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))

P(20<MPG<50)= 0.013000000000000012

Please refer to python notebook for answer

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

**a) Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

1. MPG of cars follows normal distribution.

Please refer to python notebook for answer

A red line with blue dots

Description automatically generated

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

**Dataset: wc-at.csv**

Adipose Tissue (AT) and Waist does not follow Normal Distribution

Please refer to python notebook for answer

A line with blue dots

Description automatically generated

A blue dotted line with a red line

Description automatically generated

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

**Answer:**

Please refer to python notebook for answer

# z value for 90% confidence interval

print ('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.05),4))

Z score for 60% Conifidence Intervla = -1.6449

# z value for 94% confidence interval

print ('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.03),4))

Z score for 60% Conifidence Intervla = -1.8808

# z value for 60% confidence interval

print ('Z score for 60% Conifidence Intervla =',np.round(stats.norm.ppf(.2),4))

Z score for 60% Conifidence Intervla = -0.8416

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

**Answer:**

Please refer to python notebook for answer

# t score for 95% confidence interval

print ('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.025,df=24),4))

T score for 95% Confidence Interval = -2.0639

# t value for 94% confidence interval

print ('T score for 94% Confidence Inteval =',np.round(stats.t.ppf(0.03,df=24),4))

T score for 94% Confidence Inteval = -1.974

# t value for 99% Confidence Interval

print ('T score for 95% Confidence Interval =',np.round(stats.t.ppf(0.005,df=24),4))

T score for 95% Confidence Interval = -2.7969

**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 refer to python notebook for answer

import numpy as np

Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

(260-270) /90/np.sqrt(18))

t\_score = -0.471

stats.t.cdf(t\_score, df = 17)

0.32 = 32%