|  |  |
| --- | --- |
| 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 | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Ordinal |
| Number of Children | Nominal |
| 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?

Ans :

(Number of favorable outcomes) / (total number of possible outcomes)=3/8=0.375=37.5%

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:

Number of possible outcomes for the above event is

N (Event (Two dice rolled)) = 6^2 = 36

1. P (sum is Equal to 1) = ‘0’ zero null nada none.
2. P (Sum is less than or equal to 4) = N (Event (Sum is less than or equal to

4)) / N (Event (Two dice rolled))

= 6 / 36 = 1/6 = 0.166 = 16.66%

1. P (Sum is divisible by 2 and 3) = N (Event (Sum is divisible by 2 and 3)) / N

(Event (Two dice rolled))

= 6 / 36 = 1/6 = 0.16 = 16.66%

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:

Total number of balls =7 balls

N (Event (2 balls are drawn randomly from bag) = 7! / 2! \* 5!

= (7\*6\*5\*4\*3\*2\*1) / (2\*1) \* (5\*4\*3\*2\*1)

N (Event (2 balls are drawn randomly from bag) = (7\*6)/ (2\*1) = 21

If none of them drawn 2 balls are blue = 7 – 2 = 5

N (Event (None of the balls drawn is blue) = 5! / 2! \* 3! = (5\*4) / (2\*1)

= 10

P (None of the balls drawn is blue) = N (Event (None of the balls drawn is blue) /

N (Event (2 balls are drawn randomly from

bag)

= 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 :

Expected number of candies for a randomly selected child is=

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,Weigh >

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

**Ans :**

Mean for the Points 3.5965625

Mean for the Score 3.2172500000000004

Mean for the Weigh 17.848750000000003

Median for the Points 3.6950000000000003

Median for the Score 3.325

Median for the Weigh 17.71

Mode for the Points 3.07

Mode for the Score 3.44

Mode for the Weigh 17.02

Variance for the Points 0.28588135080645166

Variance for the Score 0.9573789677419356

Variance for the Weigh 0.9573789677419356

Standard Deviation for the Points 0.5346787360709716

Standard Deviation for the Score 0.9784574429896967

Standard Deviation for the Weigh 1.7869432360968431

Range for the Points 2.17

Range for the Score 3.9110000000000005

Range for the Weigh 8.399999999999999

**Draw Inferences:**







**Use Q7.csv 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?

Ans :

Expected Value = (sum of all weights) / (total number of weights)

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

Expected Value = 1318 / 9= 145.33

The Expected Value of the Weight of that patient is 144.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Answers are given in the Jupyter Notebook**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Answers are given in the Jupyter Notebook**

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



Ans :- The histograms peak has right skew and tail is on right. Mean > Median. We have outliers on the higher side.(Postive Skew)



Ans:- 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?

Ans :

Confidence Interval = Sample Mean ± (Critical Value × Standard Error)

Critical value for 94% is 1.88 for 98% is 2.33 for 96% is 1.75

Sample mean=200

Standard Error (SE) = Standard Deviation / sqrt(Sample Size)=0.671

94% Confidence interval is in between (198.74,201.26)

98% Confidence interval is in between (198.44,201.56)

96% Confidence interval is in between (198.83,201.17)

**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?

1)Ans:

Mean ≈ 39.94

Median=40.5

Variance=67.9

Standard Deviation=8.24

2)Ans :

The average score obtained by the student is approximately 39.94

The middle value of the score is 40.5

The variance of the scores is approximately 67.9. It measures the spread or dispersion of the scores around the mean. A higher variance indicates greater variability in the scores.

Standard deviation is 8.24. It is a measure of the average distance between each score and the mean. A higher standard deviation indicates more spread or variability in the scores.

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

Ans :

When the mean and median are equal, it suggests that the data is centered and balanced around the middle point. This implies that the distribution is likely to be symmetric, and therefore the skewness would be close to zero or negligible.

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

Ans :

When the mean is greater than the median, it indicates a positive skew or a right-skewed distribution, where the majority of the data is concentrated towards the left and there are some larger values that extend the right tail.

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

Ans:

Negative skewness means that the tail of the distribution is longer on the left side, and the majority of the data points are concentrated towards the right. The mean is being pulled to the left by the smaller values in the distribution, resulting in it being smaller than the median.

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

Ans :

A positive kurtosis value indicates that a dataset has leptokurtic or heavy-tailed distribution. Leptokurtic distributions with positive kurtosis values have a higher peak or more concentrated data in the center compared to a normal distribution. The heavier tails indicate that there is an increased probability of extreme values occurring in the dataset.

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

Ans:

A negative kurtosis value indicates that a dataset has platykurtic or light-tailed distribution. Platykurtic distributions with negative kurtosis values have flatter peaks and less concentration of data in the center compared to a normal distribution. The lighter tails indicate a decreased probability of extreme values occurring in the dataset.

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



What can we say about the distribution of the data?

Ans :

By looking the box plot we can say left-skewed distribution, also known as a negative-skewed distribution.

What is nature of skewness of the data?

Ans:

the tail on the left side of the distribution is longer than the tail on the right side. This means that there are more values that extend towards the lower end of the data range, resulting in a longer tail on the left side.

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

Ans:

IQR=Q3(upper quartile) - Q1(lower quartile) =18.2-10=8.2

Q19) Comment on the below Boxplot visualizations?



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

Ans :

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)

Ans:

By doing calculation in jupyter notebook we get the probabilities

1. P(MPG>38) = 0.348
2. P(MPG<40) = 0.729
3. P(20<MPG<50)= 0.013000000000000012

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

Ans:

1. MPG of cars follows normal distribution.



1. Ans :

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





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

Ans :

Z\_score =np.round(stats.norm.ppf((1 + confidence\_level=0.90) / 2),4)

Z score for 90% Confidence Interval = 1.6449

Z score for 94% Confidence Interval = 1.8808

Z score for 60% Confidence Interval = 0.8416

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

Ans :

t\_score = np.round(stats.t.ppf((1 + confidence\_level=0.90) / 2, degrees\_of\_freedom=25),4)

T score for 95% Confidence Interval = 2.0595

T score for 96% Confidence Interval = 2.1666

T score for 99% Confidence Interval = 2.7874

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 :

import scipy.stats as stats

# Given information

mu = 270 # Claimed average lifespan

n = 18 # Sample size

x\_bar = 260 # Sample mean

s = 90 # Sample standard deviation

# Calculating the t-score

t\_score = (x\_bar - mu) / (s / (n \*\* 0.5))

# Calculating the probability using the t-distribution

df = n - 1

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

probability is 0.32167253567098364= 32%