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

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

3/8

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

1. Equal to 1 -> 0
2. Less than or equal to 4 – 6/36 = 1/6
3. Sum is divisible by 2 and 3 -> 6/36 = 1/6

**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?**

Total number of ways 2 balls are drawn – 7C2 = 21

No. of ways in which no blue ball is drawn - 5C2 = 10

Probability = 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

Expected number of candies for a randomly selected child are:

  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 = 3.09

**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.**

For points -> Mean = 3.596; Median = 3.695; Mode = 3.07, 3.92; Variance = 0.2858; Standard Deviation = 0.5346; Range = 2.17; Honda Civic has the maximum points and Dodge Challenger has the minimum points.

For Score -> Mean = 3.217; Median = 3.325; Mode = 3.44; Variance = 0.9572; Standard Deviation = 0.9784; Range = 3.911; Lincoln Continental has the maximum points and Lotus Europa has the minimum points.

For Weigh -> Mean = 17.848; Median = 17.71; Mode = 17.02,18.9; Variance = 3.193; Standard Deviation = 1.7869; Range = 8.4; Merc 230 has the maximum Weigh and Ford Pantera L has the minimum points.

**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?

Each person has a unique weight. Probability of selecting a person is 1/9.

Then expected weight = Sum(P\*W)

* (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

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

**Cars speed and distance**

**Use Q9\_a.csv**

Speed -> Skew = -0.1175; Kurtosis = -0.5089. Has a negative skew and kurtosis. Most of the speeds are to the right of the median.

Distance -> Skew = 0.8068; Kurtosis = 0.405. Has a positive skew and kurtosis.

As speed rises distance also rises.

**SP and Weight(WT)**

**Use Q9\_b.csv**

SP -> Skew = 1.611; Kurtosis = 2.977. Has a positive skew and kurtosis. Most of the data points are to the left of the median.

WT -> Skew = -0.6147; Kurtosis =0.95. Has a negative skew and a positive kurtosis. Data in some ranges is absent.

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



It is a positively skewed data. Most of the data is concentrated on the left of the median. Means a lot of weights are in the 50 to 100 range and as the weight increases above 200, the number of entities with that weight reduces. Very few have a weight above 300.



Outliers are present above the upper whisker. And the most of the data is distributed in the upper whisker region.

**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?**

Solution:

Confidence - 94%

from scipy.stats import t, norm

norm.interval(0.94, loc = 200, scale = 30)

Output - (143.57619175546247, 256.42380824453755)

Confidence – 96%

from scipy.stats import t, norm

norm.interval(0.96, loc = 200, scale = 30)

Output - (138.38753268104531, 261.61246731895466)

Confidence – 98%

from scipy.stats import t, norm

norm.interval(0.98, loc = 200, scale = 30)

Output - (130.2095637787748, 269.7904362212252)

**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.**

Mean – 41; Median – 40.5; Variance – 24.108; Standard Deviation – 4.91

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

Most of the students got 41 marks. Mean = Mode.

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

No skew will be there. Bell curve or normal distribution.

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

Positive skewness

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

Negative skewness

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

The peak is very high and the tails are thick. Means all the data is concentrated close to the center.

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

The peak is flat and the tails are thin. The variation in data is very high and very less concentration towards the center.

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



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

Data is not uniformly distributed. 25% of the data is between 0 to 10, 50% of data is between 10 to 18, and the remaining 25% is between 18 to 20.

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

Negative skew as the distribution is concentrated closely after the median.

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

IQR = 18 – 10 = 8

Q19) Comment on the below Boxplot visualizations?



The distributions are uniform in both. But plot 1 has much lesser variance in the data. The medians are same for both. But the variance is more in plot 2.

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

Since the variance is more plot 2 it could be harder to make a prediction in it whereas in plot 1 all the data is close together so predictions are easier compared to plot 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) – 0.3475
  2. P(MPG<40) – 0.7293

c. P (20<MPG<50) – 0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Almost normally distributed with a slight skew of -0.177

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

Dataset: wc-at.csv

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

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

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

Solution:

import numpy as np

from scipy.stats import t

t\_stat = (260 - 270)/(90/np.sqrt(18))

print(t\_stat)

if (t\_stat > 0):

print(str(np.round(2\*(t.cdf(-1\*t\_stat, 17)\*100),2))+'%')

else:

print(str(np.round(2\*(t.cdf(t\_stat, 49)\*100),2))+'%')

Output -> t\_stat is -0.471. Probability is 0.6394