|  |  |
| --- | --- |
| 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- [HHT, HTH, THH]

(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= 0
2. Less than or equal to 4

= [(1,1) (1,2) (2,1), (2,2) (1,3) (3,1)]

= (Event (Sum is less than or equal to 4)) / (Event (Two dice are rolled))

= 6/36= 1/6= 0.166= 16.6%

1. Sum is divisible by 2 and 3=

= [([(1,5) (5,1) (3,3) (6,6) (2,4) (4,2)]

= (Event (Sum is divisible by 2 and 3)) / (Event (Two dice are rolled))

= 6/36= 1/6= 0.166= 6.6%

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?

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 Value

= [(1\*0.015) +(4\*0.2) +(3\*.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.

**Use Q7.csv file**

**ANS-** Calculation is available in python notebook. File is attached with the assignment

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Title | Mean | Median | Mode | Variance | Standard Deviation | Range |
| Points | 3.5965 | 3.6950 | 3.07,3.92 | 0.2858 | 0.5346 | 2.17 |
| Score | 3.2172 | 3.325 | 3.44 | 0.9573 | 0.9784 | 3.9110 |
| Weigh | 17.8487 | 17.71 | 17.02,18.90 | 0.9573 | 1.7869 | 8.3999 |

* The average points of all the vehicles are 3.59.
* The average weight of all vehicles is 3.21.
* The average weight of all vehicles is 17.84.

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 Weights)/ (Total Number of Weights)

= [(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)]

= [12+12.2222+13.6666+14.8888+15+16.1111+18.5555+20.7777+22.1111]

=145.333

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

**Cars speed and distance**

**Use Q9\_a.csv**

Skewness of speed=-0.1139 Kurtosis of speed=-0.5771

Skewness of distance=0.7824 Kurtosis of speed=0.2480

**SP and Weight (WT)**

**Use Q9\_b.csv**

Skewness of SP=1.5814 Kurtosis of SP=2.7235

Skewness of WT=-0.6033 Kurtosis of WT=0.8194

**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-

Total population=3000000

Sample mean=200

Standard deviation of sample=30

Sample size=2000

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

Critical value for 94% is 1.88.

Critical value for 98% is 2.33.

Critical value for 96% is 1.75.

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

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 (198.62,201.37)

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

= 41

Median= 40,41

Variance= 25.52

Standard deviation= 5.05

1. What can we say about the student marks?

ANS-The average student marks are 41.

Also 3 students got marks of 41.

More students got marks between 35 to 45.

The average score obtained by the student is approximately 41.

The middle value of the score is 40,41.

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- The Skewness will be ZERO.

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- The given data is positively skewed.

When the mean is greater than the median, it indicates a positive skew or a right-skewed distribution, where most 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- The given data is negatively skewed.

Negative skewness means that the tail of the distribution is longer on the left side, and most 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-Positive kurtosis indicates more values around tail than mean.

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- Negative Kurtosis indicates wider peaks and thinner tails.

i.e., more values around mean less around tails.

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- The data contains values distributed mostly between 10 to 18.

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

Mean is around 15.

What is nature of skewness of the data?

ANS- It is negatively Skewed data.

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-Q1=7

Q19) Comment on the below Boxplot visualizations?



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

* In box plot 1, 50% of the values lie between260-280.
* In box plot 2, 50% of the values lie between225-325.
* In box plot 1, maximum values lie between285-235.
* In box plot 2, maximum values lie between200-330.
* In box plot 1, outliers lie above 285 and below 235.
* In box plot 2, outliers lie above 200 and below 330.
* 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)

By doing calculation in jupyter notebook we get the probabilities

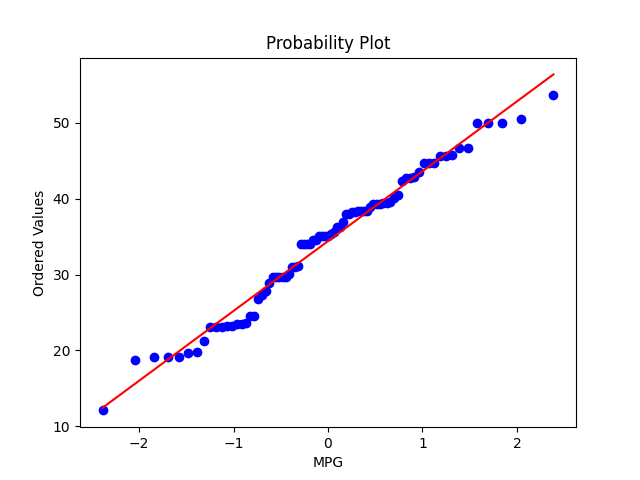
1. P(MPG>38) = 0.348
2. P(MPG<40) = 0.729
3. P(20<MPG<50) = 0.013000000000000012

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans. MPG of cars follows normal distribution.



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

Dataset: wc-at.csv

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

A graph with blue dots and red lines

Description automatically generated

A graph with blue dots

Description automatically generated

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

ANS-Z score at 90% confidence interval

α=0.05

Z score=1.645

Z score at 94% confidence interval

α=0.03

Z score=1.880

Z score at 60% confidence interval

α=0.2

Z score=0.841

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 at 95% confidence interval-2.0638

t score at 96% confidence interval-2.1715

t score at 99% 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

ANS-

Mean of the sample of bulbs=260=x

Total bulbs’ mean=270=µ

Standard deviation of sample=90=s

Number of items in sample=18=n

Degree of freedom=n-1=17

Calculating the t-score

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

Calculating the probability using the t-distribution

df = n - 1

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

t=(260-270)/(90/srqt (18))=-0.471

p-value=0.3216