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

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

Ans: Total no of outcomes =8 (HHH,TTT,HHT,TTH,HTH,THH,THT,HTT)

No of favorable outcomes=3

P= No of favorable outcomes/Total no of outcomes =**3/8**

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:

Total no of possible outcomes = 36

i.--> Equal to 1 = **0**

ii-> Less than or equal to 4 = (1,1), (1,2),(1,3),(2,2) => 4

favorable outcomes/possible outcomes=**4/36**

iii-> Sum is divisible by both 2 and 3 is =

(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)

No of favorable outcomes = 6

Total number of outcomes = 36

No of favorable outcomes/ Total number of outcomes=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?

Ans: Total number of balls 2+3+2=7

Total no of outcomes = 7

Probability of getting 2 balls out of 7 is 7C2.

* n(s) = (7\*6)/(2\*1)=21

Number of ways of getting 2 balls out of 5 balls none of the ball is blue is 5C2

* n(e) = (5\*4)/(2\*1)=10

**P(E) =n(e)/n(s) = 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: (1\*0.015)+(4\*020)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120)

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

= **3.09**

The Expected number of candies for a randomly selected child is = **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:**

**Mean median mode Variance Standard Deviation Range**

**Points**: 3.5965 3.695 for 0:3.07 0.28588 0.534679 2.17

For 1:3.92

**Mean median mode Variance Standard Deviation Range**

**Score:** 3.217250, 3.325 3.44 3.1931 0.978457 3.91100

**Weigh:** 17.848 17.17 for 0:17.02 3.1931 1.786943 8.39999

For 1:18.90

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  =  Σ ( probability  \* Value )

Total no of patients = 9

Let us assume one of the patient chosen at random=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
* 12+12.22+13.66+14.88+15+16.11+18.55+20.77+22.11=145.3

==145.3

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

**Car’s speed and distance**

**Use Q9\_a.csv**

**Ans:** Skewness of Speed and Distance is

speed : -0.117510

distance: 0.806895

Kurtosis of Speed and distance

speed : -0.508994

distance: 0.405053

The skewness of the "**speed**" column is approximately **-0.1175**. Since this value is negative but close to zero, it suggests a very slight left-skewness.so it also known as negative skewness.

Kurtosis value is **-0.509,** which is negative. This indicates that the distribution is platykurtic, meaning it has thinner tails and is flatter compared to a normal distribution.

The "**Distance**" data is right-skewed, so it called as positive skewness.

The positive kurtosis suggests that the "distance" data has fatter tails and is more peaked at the center compared to a normal distribution. Positive kurtosis is also called "**leptokurtic**" kurtosis.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans:** Skewness of Sp and Weight

SP 1.611450

WT -0.614753

Kurtosis of Sp and Weight

SP 2.977329

WT 0.950291

The skewness of a **SP** value is approximately **1.611**, which is positive. Positive skewness indicates that the data is right-skewed, meaning it has a tail on the right side.

The kurtosis of a **SP** value is approximately **2.977**, which is positive. Positive kurtosis indicates that the distribution has fatter tails and is more peaked at the center compared to a normal distribution.

The skewness of a **Weight** value is approximately -0.615, which is negative. Negative skewness indicates that the data is left-skewed, meaning it has a tail on the left side.

The kurtosis of **Weight** value is approximately 0.950, which is positive. Positive kurtosis indicates that the distribution has fatter tails and is more peaked at the center compared to a normal distribution.

**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.

* The box plot contains outliers on higher 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:** Sample mean (X) = 200 pounds

Standard deviation (σ) = 30 pounds

Sample Size (n) = 2000

* For 94% confidence interval = **200 +1.261** = **201.261** or

**200-1.261**= **198.739**

* For 98% confidence interval = **200+1.563** = **201.563**  or

**200-1.563 = 198.437**

* For 96% confidence interval =  **200+1.375= 201.375**  or

**200-1.375 = 198.625**

**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?

Ans: Mean =

34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56/18=**41.0**

Median= **40.5**

Variance= **24.11111111**

Standard deviation = **5.052664**

**2 Ans :**   **Some of the students got less than mean and some of the students got greater than mean.**

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

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

**Ans:** when the mean and median of given data set are equal indicates that the data is symmetrically distributed. The skewness is **perfectly Symmetric**

this distribution is perfectly balanced on both sides on the mean.

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

**Ans:** when the mean is greater than median of given data set that the data is right-sided skew. The skewness is **Positively Skewed** the tail on the right side is longer than other tails.

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

**Ans:** when the median is greater than mean of a given data set is left sided skew. Then the skewness is **Negatively Skewed**  the tail on the left side is longer than other tails.

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

**Ans:** A positive kurtosis indicates that the data has thick tails and is more peaked at the center when compared to a normal distribution. These distributions are

called **“ Leptokurtic. ”**

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

**Ans:** A negative kurtosis value indicates that the data has thin tails and is flatter at the center when compared to a normal distribution. These distributions are called as **“ playkurtic. “**

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 nature of the skewness of the above data is left sided skew. That means the nature of the skew is “ **Negatively skewed “**

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

Ans: Inter Quartile Range (IQR)

Upper whisker length 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.

**Ans:** The above box plots shows two different plots . **1st** box plot shows no outliers, and the range is in between 248-280 .  **2nd**  plot the range is in between 190-325 above and there are no outliers. Both the plots are said to be Normal distribution.

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)

c. P (20<MPG<50)

**Ans:** A 🡪 P(MPG>38) = **0.34759392515827137**

B 🡪P(MPG<40) = **0.7293498762151609**

C 🡪P(20<MPG<50) = **0.8988689169682047**

The above A,B,C are the probabilities of MPG>38, MPG>40, 20<MPG<50.

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:** The MPG of cars follows the 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 Circumference (Wc) does not follow Normal distribution.

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

**Ans:**  To calculate Z-score.

**Z-score = (1-confidenceinterval)/2**

* For 90% confidence interval **(1-0.90)/2 = 0.05**
* For 94% confidence interval  **(1-0.94)/2 = 0.03**
* For 60% confidence interval **(1-0.60)/2 = 0.20**

The confidence intervals for 90% , 94% , 60% are **0.05** , **0.03** , **0.20.**

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

Ans: To calculate T-score.

T-score = (1-confidence level)/2 \* Sample size

Where Sample size = n-1

🡺 For 95% confidence interval = (1-0.95)/2 \* (25-1)

= **0.025\*24 = 0.6**

**🡺**  For 96% confidence interval = (1-0.96)/2 \* (25-1)

= **0.02\*24 = 0.48**

🡺 For 99% confidence interval = (1-0.99)/2 \* (25-1)

= **0.005 \* 24 = 0.12**

The confidence intervals for 95%, 96% , 99% are **0.6** , **0.48**, **0.12** .

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(t\_score, df )

df 🡪 degrees of freedom

**Ans:** The T score of the above problem is  **-0.4714045207910317**

The probability of 18 randomly selected bulbs are **=0.32167253567098364**

from scipy.stats import t, norm

sm=260 # sample mean

pm=270 # population mean

std=90 # standard devation

ss=18 # sample size

df=17 # degrees of freedom

# (sample size - 1)

t\_score =(sm-pm)/(std/(ss\*\*0.5))

t\_score

# probability for 18 randomly selected bulbs

p= t.cdf(t\_score,df)

p