|  |  |
| --- | --- |
| 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 | Categorical/Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Categorical/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 | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

**Answer**: The Probability of getting two heads and one tail in the toss of three coins is 3/8 or 0.375.

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

**Answer:** The number of possible outcomes for the above event is 6^2 = 36

a.) P (sum is Equal to 1) = 0

b.) P (Sum is less than or equal to 4) = 6/ 36 = 1/6 = 0.166 = 16.66%

c.) P (Sum is divisible by 2 and 3) = 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?

**Answer:** Probability = Favourable outcome/Total outcome

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

\*Number of ways of drawing 2 balls out of 7 = 7C2 = (7x6) / (2x1) = 21

Number of balls other than blue = 5

\* Number of ways of drawing 2 balls out of 5 = 5C2 = (5x4) / (2x1) = 10

Therefore the required probability 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:** The expected number of candies = (1x0.015) + (4x0.2) + (3x0.65) + (5x0.005) +(6x0.01) + (2x0.120) = 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**

**Answer:** Mean for Points = 3.59, Score = 3.21 and Weigh = 17.84

Median for Points = 3.69, Score = 3.32 and Weigh = 17.71

Mode for Points = 3.07, Score = 3.44 and Weigh = 17.02

Variance for Points = 0.28, Score = 0.95, Weigh = 3.19

Standard Deviation for Points = 0.53, Score = 0.97, Weigh = 1.78

Range for points = (4.93-2.76) = 2.17

Range for Score = (5.424-1.51) = 3.91

Range for Weigh = (22.9-14.5) = 8.39

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?

**Answer:** Expected value = Sum (X \* Probability of X)

Expected value = (1/9) (108)+ (1/9)(110)+ (1/9)(123)+ (1/9)(134)+

(1/9) (145)+ (1/9)(167)+ (1/9)(187)+ (1/9)(199)

Expected value = 145.33

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

Cars speed and distance

Use Q9\_a.csv

SP and Weight(WT)

Use Q9\_b.csv

**Answer:** 1. Cars speed and distance

Skewness for speed = -0.117510,

Kurtosis for speed = -0.508994

Skewness for distance = 0.806895

Kurtosis for distance = 0.405053

**Inference:** The skewness value is negative so it is left-skewed.The mean is less than the median. Speed and distance is strongly correlated to each other positively.

2. SP and Weight (WT)

Skewness for SP = 1.611450

Kurtosis for SP = 2.977329

Skewness for WT = -0.614753

Kurtosis for WT = 0.950291

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



**Answer**: The most of the data points are concentrated in the range of 50-100 with a frequency of 200. And the least range of weight is 400 somewhere around 0-10. So the expected value of the above distribution is 75. Skewness- we can notice a long tail towards the right so it is heavily right-skewed.



Answer: The median is less than the mean right skewed and we have an outlier on the upper side of the box plot and there are fewer data points between Q1 and the bottom point.

**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:** import pandas as pd

from scipy import stats

conf\_interval1=stats.t.interval(0.94,1999,200,(30/np.sqrt(2000)))

print(np.round(conf\_interval1,2))

* confidence interval for 94% is [ 198.74, 201.26 ]

conf\_interval2=stats.t.interval(0.94,1999,200,(30/np.sqrt(2000)))

print(np.round(conf\_interval2,2))

* confidence interval for 96% is [ 198.44, 201.56 ]

conf\_interval3=stats.t.interval(0.94,1999,200,(30/np.sqrt(2000)))

print(np.round(conf\_interval3,2))

* confidence interval for 94% is [ 198.62, 201.38 ]

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:** (1) Mean=(34+36+36+38+38+39+39+40+40+41+41+41+41+42+45+

49+56)/18

Mean= 738/18=41

 Median = (9th+ 10th Term)/2 = (40+41)/2 = 40.5

Variance=25.47

Standard deviation = (Variance)^(1/2)

Standard deviation= (24.05) ^ (1/2) = 5.05

(2) The above marks don’t have outliers and the data is slightly skewed towards the right because the mean is greater than the median. And more than 75% of students scored 36 – 42 marks and only one student scored more than 50

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

Ans: No skewness is present we have a perfect symmetrical distribution.

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

Ans: Right skewed

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

Ans: Left skewed

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

Ans: Positive kurtosis means the curve is more peaked and it is Leptokurtic. Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.

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

Ans: Negative Kurtosis means the curve will be flatter and broader and it is Platykurtic. Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean.

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:**

1. The above Boxplot is not normally distributed the median is towards the higher value.
2. The data is skewed towards the left. The whisker range of minimum value is greater than maximum.
3. The Inter Quantile Range = Q3 Upper quartile – Q1 Lower Quartile = 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:**

* Both boxplots are normally distributed.
* The mean and median for both boxplots are the same that is approximately in a range between 275 to 250 .
* there are no outliers.
* By observing both the plots whisker’s level is high in boxplot 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)
  2. P(MPG<40)
  3. P (20<MPG<50)

**Answer:** code- from scipy import stats

import pandas as pd

data3=pd.read\_csv('Cars.csv')

data3.describe()

1. stats.norm.cdf(40,34.422,9.13)

P(MPG>38)=0.347

1. 1-stats.norm.cdf(38,34.422,9.13)

P(MPG<40)=0.729

1. stats.norm.cdf(50,34.422,9.13)-(1-stats.norm.cdf(20,34.422,9.13))

P(20<MPG<50=0.013

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

**Answer:** a) MPG of cars follows normal distribution.

b)  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

**Answer:**  Z score for 90% Confidence Interval = -1.6449

 Z score for 94% Confidence Interval = -1.880

 Z score for 60% Confidence Interval = -0.841

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

**Answer:**  T score for 95% Confidence Interval = -2.0639

T score for 94% Confidence Inteval = -1.974

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:** Sample mean=260

Population mean=270

Sample standard deviation=90

Sample size=18

*t = (s\_mean-P\_mean)/(s\_SD/sqrt(n))*

t **=** (260**-**270 **/** (90**/**18**\*\***0.5)

t = -0.471

from scipy import stats

from scipy.stats import norm

p\_value=1-stats.t.cdf(abs(-0.4714),df=17)

p\_value = 0.321