|  |  |
| --- | --- |
| 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) | **Categorical** |

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) | **ratio** |
| Sales Figures | **ratio** |
| Blood Group | **nominal** |
| Time Of Day | **interval** |
| Time on a Clock with Hands | **Interval** |
| Number of Children | **nominal** |
| Religious Preference | **nominal** |
| Barometer Pressure | **Interval** |
| SAT Scores | **Interval** |
| Years of Education | **Interval** |

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

**Answer: sample space=(HHT,HTH.THH,TTT,THT,HHH,TTH,HTT)=8**

**OUTCOME**

**Required outcome=(HHT,HTH,THH)**

**P=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

**Answer : a) equal to 1=0**

**b) less than or equal to 4=6/36=16.66%**

**c)sum is divisible by 2 and 3=6/36=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 : 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 | 0.12 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

**Answer: expected number of candies for a randomly selected child**

**= (1\*0.05)+(4\*0.20)+(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**

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:**

**code**

**import pandas as pd**

**df=pd.read\_csv("Q7 (1).csv")**

**df**

**df.mean()**

**df.median()**

**df.mode()**

**df.var()**

**df.std()**

**df['Points'].max()-df['Points'].min()**

**df['Score'].max()-df['Score'].min()**

**df['Weigh'].max()-df['Weigh'].min()**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weight** |
| **Mean** | **3.596563** | **3.217250** | **17.848750** |
| **Median** | **3.695000** | **3.325000** | **17.710000** |
| **mode** | **3.92** | **3.44** | **17.02** |
| **Variance** | **0.285881** | **0.957379** | **3.193166** |
| **Standard deviation** | **0.534679** | **0.978457** | **1.786943** |
| **Range** | **2.17** | **3.911** | **8.4** |
|  | | | |

**INFERENCES:**

* **The mean is useful for identify trends in the data because by using means value we can compare over a time to trends.**
* **The median is a central value of data and also median is a useful for some extreme data cases. The values of means in above case are almost equal to median.so,there is no outliers.Sometime its differs because of outliers.Median is trustable.**
* **Standard deviation and Variance are shows how data spread across the mean.In this above case standard deviation of point is low, so data distribution is low and standard deviation of score is high,so data spread is high.**

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  =  ∑ ( probability  \* Value )**

**∑ P(x).E(x)**

**there are 9 patients**

**Probability of selecting each patient = 1/9**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **E(x)** | **108** | **110** | **123** | **134** | **135** | **145** | **167** | **187** | **199** |
| **P(x)** | **1/9** | **1/9** | **1/9** | **1/9** | **1/9** | **1/9** | **1/9** | **1/9** | **1/9** |

**Expected Value  =  (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**

**Expected Value of the Weight of that patient = 145.33**

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

Cars speed and distance

Use Q9\_a.csv

**Answer :**

**Code**

**import pandas as pd**

**df=pd.read\_csv("Q9\_a.csv")**

**df**

**df.skew(axis=0,skipna=True)**

**df.kurtosis(axis=0,skipna=True)**

|  |  |  |
| --- | --- | --- |
|  | **Speed** | **Distance** |
| **Skew** | **-0..117509861** | **0.8689496** |
| **kurtosis** | **-0.50899442** | **0.40505282** |

**Inferences**

**Skewness is negative, that tells us that the distribution is skewed towards left. Mean of distribution is less than the Median. Kurtosis Value is less than 3, that tells us that the distribution has broad peak and thin tails.**

**Skewness is positive, that tells us that the distribution is skewed towards right. Mean of distribution is more than the Median. Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Code**

**import pandas as pd**

**df=pd.read\_csv("Q9\_b.csv")**

**df**

**df.skew(axis=0,skipna=True)**

**df.kurtosis(axis=0,skipna=True)**

|  |  |  |
| --- | --- | --- |
|  | **Sp** | **Wt** |
| **skew** | **1.1611450** | **-0.614753** |
| **kurtosis** | **2.977329** | **0.950291** |

**Inference**

**Skewness is positive, that tells us that the distribution is skewed towards right. Mean of distribution is more than the Median. Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram.**

**Skewness is negative, that tells us that the distribution is skewed towards left. Mean of distribution is less than the Median.Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram**

Q10) Draw inferences about the following boxplot & histogram



**Inferences :**

**\*THE DISTRIBUTION IS RIGHT SKEWED.**

**\*MEAN VALUE IS GREATER THAN MEDIAN.**



**Inferences :**

**\*ABOVE BOXPLOT SHOWS DISTRIBUTION HAVE MANY OUTLIERS ON UPPER EXTREME.**

**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**

**94% Confidence:**

**X-bar = 200**

**Sd = 30**

**n = 2000**

**Interval Estimate = X-bar ± Z\*Sd/sqrt(n)**

**=200 ± 1.88\*30/sqrt(2000)**

**=198.74 – 201.26**

**>98% Confidence:**

**X-bar = 200**

**Sd = 30**

**n = 2000**

**Interval Estimate = X-bar ± Z\*Sd/sqrt(n)**

**=200 ± 2.33\*30/sqrt (2000)**

**=198.44-201.56**

**96% Confidence:**

**X-bar = 200**

**Sd = 30**

**n = 2000**

**Interval Estimate = X-bar ± Z\*Sd/sqrt(n)**

**=200 ± 2.05\*30/sqrt (2000)**

**=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.

**Answer:**

**Mean = 41**

**Median = 40.5**

**Variance = 25.5**

**Standard Deviation = 5.05**

1. What can we say about the student marks?

**Answer :**

**TWO THINGS ABOUT STUDENT MARKS:**

1. **There is no outliers in the student marks.**
2. **Mean is greater than median. So ,data little-bit Right skewed.**

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

**Answer:**

**Skewness is equal to zero(Skewness = 0).So,it is Symmetric**.

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

**Answer:**

**Right Skewed.**

**Skewness is positive .So,it is Asymmetric.**

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

**Answer :**

**Left Skewed.**

**Skewness is negative.So,it is Asymmetric**.

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

**Answer :**

**It indicates given case is asymmetric with Sharp Peak, Thick Tails.**

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

**Answer :**

**It indicates given case is asymmetric with Broad Peak, Thin Tails.**

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



What can we say about the distribution of the data?

**Answer :**

**According to above boxplot the data is Not a Normal Distribution**

What is nature of skewness of the data?

**Answer :**

**It is asymmetric and left Skewed**

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

**Answer :**

**IQR=18-10=8  
 IQR=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 Boxplot 1 and Boxplot 2 are follows 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)
  3. P (20<MPG<50)

**Answer:**

1. **P(MPG>38)**

**P(MPG>38) = 0.35**

1. **P(MPG<40)**

**P(MPG<40) = 0.73**

1. **P (20<MPG<50)**

**P(20<MPG<50) = 0.898**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Answer:**

**Code :**

**import pandas as pd**

**from scipy import stats**

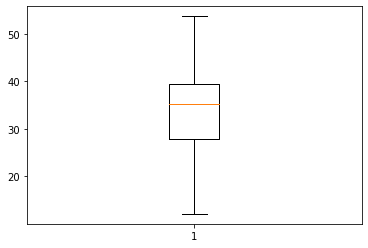
**import matplotlib.pyplot as plt**

**cars=pd.read\_csv('cars.csv')**

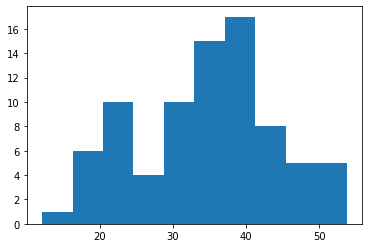
**cars**

**plt.boxplot(cars['MPG'])**

**plt.hist(cars['MPG'])**

****

**Box plot**

** Histogram**

**From this above boxplot and histogram we can say the 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

**Answer:**

**Code**

**import pandas as pd**

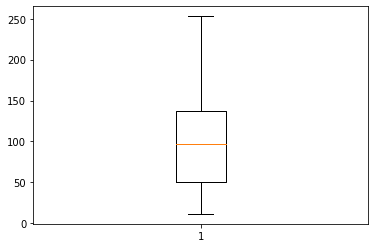
**from scipy import stats**

**import matplotlib.pyplot as plt**

**wc-at=pd.read\_csv('wc-at.csv')**

**cars**

**plt.boxplot(cars['AT'])**

**plt.hist(cars['AT'])**

**Boxplot**

****

**Histogram**

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

**Answer:**

**Code**

**from scipy import stats**

**stats.norm.ppf(0.90)**

**stats.norm.ppf(0.94)**

**stats.norm.ppf(0.60)**

**Answer:**

**90% Confidence:**

**1.644854**

**94% Confidence:**

**1.880794**

**60% Confidence:**

**0.84162**

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

**Answer:**

**95% Confidence:**

**2.063899**

**96% Confidence:**

**2.171545**

**99% Confidence:**

**2.79694**

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:

**Given:**

**µ = 270, n= 18, x bar = 260, s = 90**

**t score = (x bar - µ)/(s/sqrt(n))**

**=(260 – 270)/(90/sqrt(18))**

**=-10/21.23=-0.47**

**=0.322163**

**So, Required Probability = 0.32**