|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Numerical-Discrete Data |
| Results of rolling a dice | Numerical-Discrete Data |
| Weight of a person | Numerical-Continuous Data |
| Weight of Gold | Numerical-Continuous Data |
| Distance between two places | Numerical-Continuous Data |
| Length of a leaf | Numerical-Continuous Data |
| Dog's weight | Numerical-Continuous Data |
| Blue Color | Categorical Data |
| Number of kids | Numerical-Discrete Data |
| Number of tickets in Indian railways | Numerical-Discrete Data |
| Number of times married | Numerical-Discrete Data |
| Gender (Male or Female) | Categorical Data |

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 | Normal |
| 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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| 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?

BY Probability Method

Total Number of Events

=Number of Possibilities per Experiments Number of Experiments =23=8

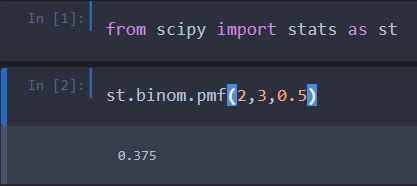
Total Number of Interested Events

= nCr = 3C2 =

Probability of Interested Events

P(x=2H) = =

**By Python**

****

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

1. Equal to 1

When we roll two dice,

Total Number of Events

= Number of possibilities per experimentsnumber of experiments = 62 = 36

We will get minimum sum of 2,Therefore

Total number of interested events = 0

Probability of interested events

P(sum=1) =

1. Less than or equal to 4

When we roll two dice,

Total Number of Events

= number of possibilities per experimentsNumber of Experiments = 62 = 36

Total number of interested events

= Number below that gives sum of 4 or less

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

Probability of interested events

P(sum≤4) = =

1. Sum is divisible by 2 and 3

When we roll two dice,

Total Number of Events

= number of possibilities per experimentsNumber of Experiments = 62 = 36

Number below that divisible by 2 & 3 both = [6,12]

Number of combination that gives sum 6

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

Number of combination that gives sum 12

= (6,6) = 1

Total Number of interested events = 5+1 = 6

Probability of interested events

= =

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 Events

= 7C2 = 7C2 =

Total Number of interested events

= nCr = 5C2 =

Probability of interested events

P(none of the ball is blue) = =

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

Excepted Random value = ∑ p(Xi) \* Xi

= (1\*0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

=3.09

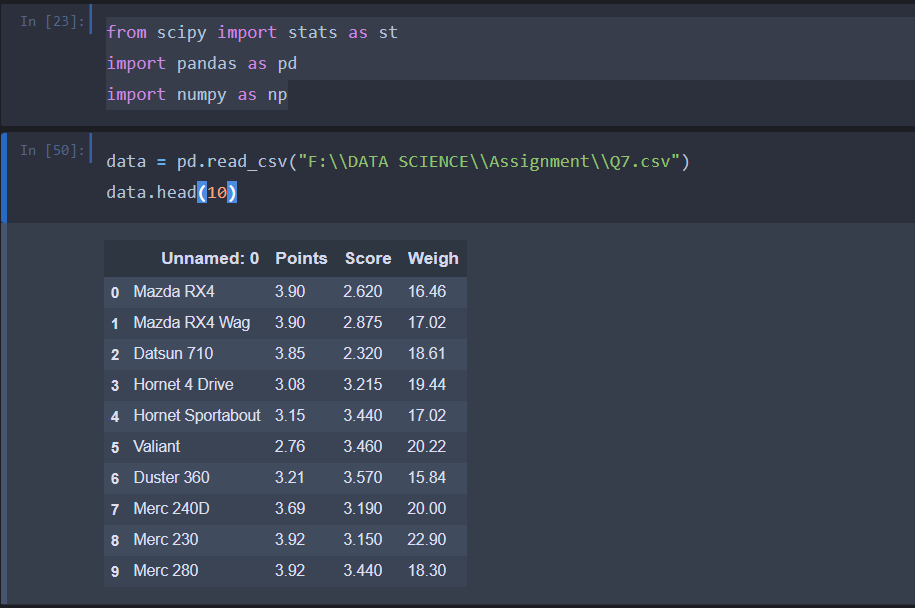
Excepted number of candies for randomly selected child = 3.09

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**

****

|  |  |  |
| --- | --- | --- |
| |  | | --- | |  |      |  | | --- | |  | |

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?

|  |  |  |
| --- | --- | --- |
| Xi | P(Xi) | P(Xi) \* Xi |
| 108 | 0.111111 | 12 |
| 110 | 0.111111 | 12.22222 |
| 123 | 0.111111 | 13.66667 |
| 134 | 0.111111 | 14.88889 |
| 135 | 0.111111 | 15 |
| 145 | 0.111111 | 16.11111 |
| 167 | 0.111111 | 18.55556 |
| 187 | 0.111111 | 20.77778 |
| 199 | 0.111111 | 22.11111 |
|  | ∑(P(Xi) \* (Xi)) | 145.3333 |

Excpected Value of the weight of Randomly chosen patient = 145.33 pounds

**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**

|  |
| --- |
|  |

**Inference:**

**Speed:**

1. **Skewness = -0.117**

* Data is slightly negatively skewed or left skewed data
* Mean data spared is more on left side of the median

1. Kurtosis = -0.508

* Data has platykurtic distribution and has thin tails compared to normal distribution
* The distribution is flat as compared to normal distribution

Distance:

1. Skewness = 0.806

* Data is skewed positively or right skewed data
* Means data spared is more on right side of the median

1. Kurtosis = 0.405

* Data has leptokurtic dist and has thick tails as compared to normal dist
* The distribution is peak as compared to normal distribution.

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



**Inference :**

* Positively skewed data
* Spared of the data on right side of the distribution is more and mass of data is on left side of median
* Frequency of data between 50 – 100 is more



**Inference :**

* Positively skewed data
* Spared of the data on right side of the distribution is more and mass of data is on left side of median
* Positive outliers are there on right side of the distribution

**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?

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | |  |  |  | | --- | |  |  |  | | --- | |  | |

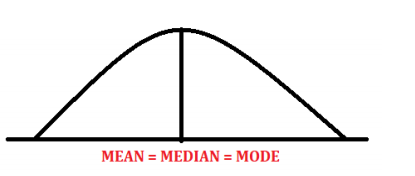
**Q12)** Below are the scores obtained by a student in tests

**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?

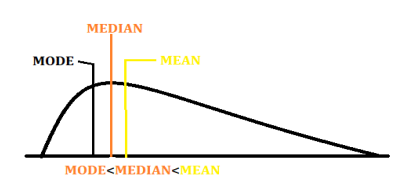
|  |
| --- |
|  |

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



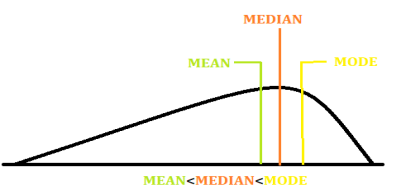
* When mean = median , we can say data is Normally Distributed

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



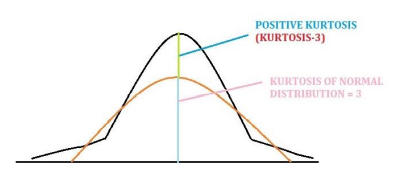
* When mean > median , we can say positively skewed data

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



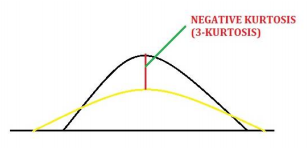
* When mean < median , we can say Negatively skewed data

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



* Positive kurtosis
* Distribution is leptokurtic
* Spared there are more values around mean

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



* Negative kurtosis indicates that distribution is platykurtic
* Spread of the data is more

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



What can we say about the distribution of the data?

* Most of the data lies between 10 – 18
* Q1 = 10
* Q2 = 15 = median = 50th percentile
* Q3 = 18

What is nature of skewness of the data?

* Negatively skewed data

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

* IQR = 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.

|  |  |
| --- | --- |
| Boxplot 1 | Boxplot 2 |
| Data ranges between 240-280 | Data ranges between 190 – 340 |
| Mean = Median = Mode=Qurtile(Q2)=260 | Mean = Median = Mode=Qurtile(Q2)=260 |
| Normally distributed | Normally distributed |
| Q1=255 | Q1=220 |
| Q3 = 280 | Q3 = 310 |
| IQR is less = 280-255=25 | IQR is more = 310 - 220=90 |

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)

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | |  |      |  | | --- | |  |  |  | | --- | |  | |

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

|  |  |
| --- | --- |
|  |  |
|  |  |

MPG OF CARS

1. Mean ≠ Median
2. Skewness = 0.177
3. Kurtosis = 0.6116
4. In box plot Q2 is not at center , whisker is more negative side ,

Midian(Q2) is nearer to Q3 and in bell curve skewed towards negative numbers

**We can say the “MPG” data is slightly right skewed or negatively skewed data .**

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

Dataset: wc-at.csv

|  |  |
| --- | --- |
|  |  |
|  |  |

Since ,

1. Mean = Median = 91.9018,
2. Skewness = 0.134 = 0
3. Kurtosis = -1.01
4. In box plot Q2 is approximately at center We can say that the “Waist” data is Normally Distributed

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

|  |  |  |
| --- | --- | --- |
| Confidence Interval | Alpha = (1-Cl)/2 | Z score |
| 90% | 0.10/2 = 0.05 | ±1.64 |
| 94% | 0.06/2 = 0.03 | ±1.88 |
| 60% | 0.40/2 = 0.20 | ±0.84 |

**By Using Python**

|  |  |  |
| --- | --- | --- |
|  |  |  |

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

|  |  |  |
| --- | --- | --- |
| Confidence Interval | Df | T score |
| 95% | 25 | 2.060 |
| 96% | 2.060 |
| 99% | 2.787 |

**By Using Python**

|  |  |  |
| --- | --- | --- |
|  |  |  |

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

An average light bulb lasts 270 days

= µ = 270

Number of sample bulbs = n = 18

Average days of sample = x = 260 days

S.D of sample = S = 90 days

Then t = -0.4714

Pt = -0.471,df = 17

|  |
| --- |
|  |

18 randomly selected bulbs would have an average life of no more than 260 days = 0.3216 = 32.16 %