|  |  |
| --- | --- |
| 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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Interval |
| 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 3**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Coin 1 | H | H | H | T | H | T | T | T |
| Coin 2 | H | H | T | H | T | T | H | T |
| Coin 3 | H | T | H | H | T | H | T | T |

Total Outcomes = 8

Favorable Outcomes = 3 (highlighted in yellow)

Probability of two heads and one tail = 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 4**

Total number of outcomes if two dice are rolled = 36

1. Probability of getting sum equal to 1 is Zero (0) as minimum sum if two dice are rolled is 2 (1 in each dice)
2. Favorable outcomes are (1,1), (1,2), (2,1), (1,3), (3,1), (2,2)

Hence the probability of getting less than of equal to 4 = 6/36 = 1/6

1. Sum divisible by 2 and 3 are factors of 6 (6, 12)

There are 5 ways to get 6 = (1,5), (5,1), (2,4), (4,2), (3,3)

There is 1 way to get 12 = (6,6)

Total number of favorable outcomes = 6

Probability that sum is divisible by 2 and 3 = 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 5**

Total number of balls in the bag = 2+3+2 = 7

Probability of first ball not being blue is = 5/7 (5 balls are not blue)

Probability of second ball not being blue is = 4/6 (4 balls are not blue of the remaining six)

Probability of both balls not being blue = 5/7 \* 4/6 = 20/42 = 10/21 (‘and’ condition applied)

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 6**

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count (Y) | Probability (P) | Expected Value (Y\*P) |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.20 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.120 | 0.24 |

The Expected number of candies for a randomly selected child = Sum of expected values = 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 7**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| count | 32 | 32 | 32 |
| mean | 3.596563 | 3.21725 | 17.85 |
| std | 0.534679 | 0.978457 | 1.787 |
| min | 2.76 | 1.513 | 14.5 |
| max | 4.93 | 5.424 | 22.9 |
| Range | 2.17 | 3.911 | 8.4 |
| Variance | 0.28588163 | 0.9573781 | 3.193 |

Insights from the data :

The data is a set of attributes for various models of cars. Units of the attributes are not supplied hence I am making certain assumptions. It appears ‘Points’ to a car is given out of 5. Can’t say the same for ‘Score’ which is a continuous variable. ‘Weight’ appears to be in tons.

1. Points : Mean for points is 3.59 suggesting that the datset is right aligned and the tail is toward the left as the max is 4.93 and min is 2.76. So distance from min to mean is 0.83 and Max to mean is 2.17.
2. Score : Min and max are higher than those for points and range also is wider compared to Points.
3. Weigh : Standard deviation is low in the context of max and min thereby signalling that the data is tightly bunched.

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 8**

|  |  |  |
| --- | --- | --- |
| **Weight** | **Probabiliy of Getting Selected** | **EV** |
| 108 | 0.11 | 12.00 |
| 110 | 0.11 | 12.22 |
| 123 | 0.11 | 13.67 |
| 134 | 0.11 | 14.89 |
| 135 | 0.11 | 15.00 |
| 145 | 0.11 | 16.11 |
| 167 | 0.11 | 18.56 |
| 187 | 0.11 | 20.78 |
| 199 | 0.11 | 22.11 |
|  |  | **145.33** |

|  |  |  |
| --- | --- | --- |
| **Weight** | **Probabiliy of Getting Selected** | **EV** |
| 108 | 0.11 | 12.00 |
| 110 | 0.11 | 12.22 |
| 123 | 0.11 | 13.67 |
| 134 | 0.11 | 14.89 |
| 135 | 0.11 | 15.00 |
| 145 | 0.11 | 16.11 |
| 167 | 0.11 | 18.56 |
| 187 | 0.11 | 20.78 |
| 199 | 0.11 | 22.11 |
|  |  | **145.33** |

Expected Value of the Weight of the patient is 145.33 Pounds.

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

Cars speed and distance (Use Q9\_a.csv)

**Ans 9**

|  |  |
| --- | --- |
| **Cars speed and distance** | |
| **Kurtosis** |  |
| Speed -0.508994 | Wide peak and thin tails |
| Dist 0.405053 | Relatiely sharp peak |
| **Skewness** |  |
| Speed -0.117510 | Negative hence concentrated on the right |
| Dist 0.806895 | Positive hence concentrated on the left |

**SP and Weight(WT)**

|  |  |
| --- | --- |
| **SP and Weight(WT)** |  |
| **Kurtosis** |  |
| SP 2.977329 | Thin peak |
| WT 0.950291 | Wide peak (and missing data) |
| **Skewness** |  |
| SP 1.611450 | Concentrated on the left |
| WT -0.614753 | Concentrated on the right |

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



**Ans 10**

Histogram – Max. no of chicks (~ 100) are in 50 to 100 weight units range. Min. no of chicks (~10) are in 400 wt.units range.

The distribution is skewed towards left implying the average weight of the chicks could be between 100 to 200.

Box Plot : IQR is not very wide hence the SD of the distribution will be not very high. Q3 is wide and there are many outliers beyond upper extreme. Median will be lower than the Mean.

**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 11**

Sample size n = 2000

Population mean µ = Not known

Sample mean x̄ = 200 pounds

Population SD σ = Not known

Sample SD s = 30 pounds

To calculate confidence intervals at 94%, 98%, 96% confidence

Standard Error of the Mean = s/sqrt n = 30/√2000 = 0.67

Df = 2000-1 =1999

Confidence interval = x̄ ± t 1-α, n-1 s/√n

Code for 94% = stats.t.ppf(.97,df=1999) => 200 ± 1.88\*0.67 => 201.25 to 198.74

Code for 96% = stats.t.ppf(.98,df=1999) => 200 ± 2.05\*0.67 => 201.37 to 198.62

Code for 98% = stats.t.ppf(.99,df=1999) => 200 ± 2.32\*0.67 => 201.5 to 198.44

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 12**

Mean = 41

Mode = 41

Median = 40.5

SD = 5.05

Variance = 25.52

Average score is 40.5 which is close to median 41. 17 out of 18 scores i.e 95% of the data scores lie between two standard deviation from the mean.

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

**Ans 13**

When mean and median of data are equal, the skewness is zero.

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

**Ans 14**

When mean > median the distribution is right tailed (Mostly but not necessarily).

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

**Ans 15**

When median > mean the distribution is left tailed (Mostly but not necessarily).

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

**Ans 16**

Positive kurtosis mean that the distributio has sharp peak and has fat tails.

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

**Ans 17**

Negative kurtosis mean that the distribution is wide and has flat peak with thin tails.

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)?   
  
**Ans 18**

The data is not grouped tightly.

Q1 = 10

Q2 = 15.5

Q3 = 18

IQR = Q3-Q1 = 18-15.5 = 2.5

The data is skewed towards left. Median is > mean.

Q19) Comment on the below Boxplot visualizations?



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

**Ans 19**

Data is tightly grouped in Boxplot 1 compared to Boxplot 2.Both have same median but different distribution. As IQR for Boxplot 1 is way too smaller compared to that of Boxplot 2. Overall range between maximum and minimum values if Boxplot 2 is larger. This indicates that the SD ib Boxplot 1 should be lower compared to 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)

c. P (20<MPG<50)

**Ans 20**

1. P(MPG>38) = 0.407
2. P(MPG<40) = 0.246

c. P (20<MPG<50) = 0.851

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

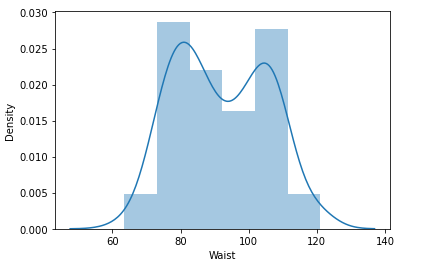
**Ans 21**

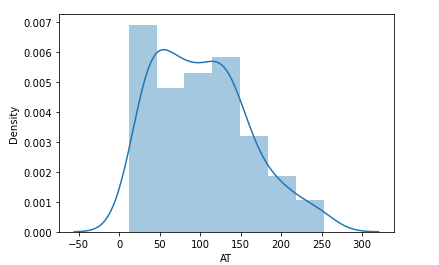
No. A simple histogram demonstrates that the distribution is not normal. Also mean (34.42) and median (35.15) are not equal for MPG.

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

Dataset: wc-at.csv

No. Both histograms appear to be skewed. Plot for Waist and AT both have two peaks as shown below.





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

**Ans 22**

Z Scores

|  |  |  |
| --- | --- | --- |
| **CI** | **Z Score** | **Code** |
| 90% | 1.64 | stats.norm.ppf(0.95) |
| 94% | 1.88 | stats.norm.ppf(0.97) |
| 60% | 0.841 | stats.norm.ppf(0.80) |

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

**Ans 23**

|  |  |  |
| --- | --- | --- |
| **CI** | **t Score** | **Code** |
| 95% | 2.06 | stats.t.ppf(0.975, df=24) |
| 96% | 2.17 | stats.t.ppf(0.98, df=24) |
| 99% | 2.79 | stats.t.ppf(0.995, df=24) |

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 24**

Sample size n = 18

Population mean µ = 270 days

Sample mean x̄ = 260 days

Population SD σ = Not known

Sample SD s = 90 days

t score = (x̄ - µ)/ (s/√n)

t = (260-270)/(90/√18)

t = - 0.471

Df = 18-1 =17

Code: .t.cdf(-0.471, df=17)

Probability p = 0.3218.