Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Categorical |
| Results of rolling a dice | Categorical |
| 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 |
| Number of kids | Categorical |
| Number of tickets in Indian railways | Categorical |
| Number of times married | Categorical |
| Gender (Male or Female) | Categorical |

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 | Ordinal |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Nominal |
| SAT Scores | Ordinal |
| Years of Education | Ratio |

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

**Ans:**

Since three coins are tossed, the sample space will be 23 = 8

Event space of two heads and one tail will be {HHT, HTH, THH}

So, the probability of getting two heads and one tail will be 3/8 = **0.357**

**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:**

Two dice are rolled so Sample space will be 62 = 36

1. No matter how unlucky we are, no way we can get a sum less than 2. So, the probability (sum Equals to 1) = 0.
2. The event having a sum less than or equal to 4 will be {(1,1), (1,2), (1,3),

(2,1), (2,2),

(3,1)}

P(sum less than or equal to 4) = |E| / |S| = 6/36 = **1.67**

1. We found 6 such cases where the sum will be divisible by 2 and 3 i.e., sum is divisible by 6.

Those are: {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}

So probability (Sum is divisible by 2 and 3) = 6/36 = **1.67**

**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 no of balls = 2 (red) + 3 (green) + 2 (blue) = 7

Two balls are drawn randomly. So Sample space will be 7C2 = 7!/(5!2!) = 21

Since we don't want blue color, so all we left is 2 red and 3 green color.

And we have to choose 2 balls out of 5 balls i.e., not a blue color.

So event space will be 5C2 = 5!/(3!2!) = 10

Probability that none of the balls drawn is blue = 10/21 = **0.476**

**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:**

Expected values = ∑( x \* P(x) )

= count candies of A \* P(A) + count candies of \* p(B) + ……

= 1\*0.015 + 4\*0.2 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.12

= **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>**

**Use Q7.csv file**

**Ans:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Mode** | **Variance** | **Standard**  **Deviation** | **Range** |
| **Points** | 3.6 | 3.7 | 3.92, 3.07 | 0.28 | 0.53 | 2.17 |
| **Score** | 3.22 | 3.32 | 3.44 | 0.93 | 0.96 | 3.91 |
| **Weigh** | 17.85 | 17.71 | 18.09, 17.02 | 3.09 | 1.76 | 8.4 |

Weight column has a high variance, also the range 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?**

**Ans:**

Total no of patients are 9. Probability of choosing any one of them will be 1/9

so Expected value will be :

(108+110+123+134+135+145+167+187+199)/9 = **145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:**

|  |  |  |
| --- | --- | --- |
|  | **Speed** | **Distance** |
| **Skewness** | -0.11 | 0.78 |
| **Kurtosis** | -0.58 | 0.25 |

Both of the column does not form a normal distribution.

**Speed** column has negative skewness i.e, the data set has a larger tail towards the left side. And kurtosis is negative which tells it is a platykurtic data. This tells us that the data is more spread from the mean.

Its opposite in case of **Distance** column. Where both skewness and kurtosis are positive I.e, a right-skewed leptokurtic data. Here we have more outliers lies around the mean.

**SP and Weight(WT)**

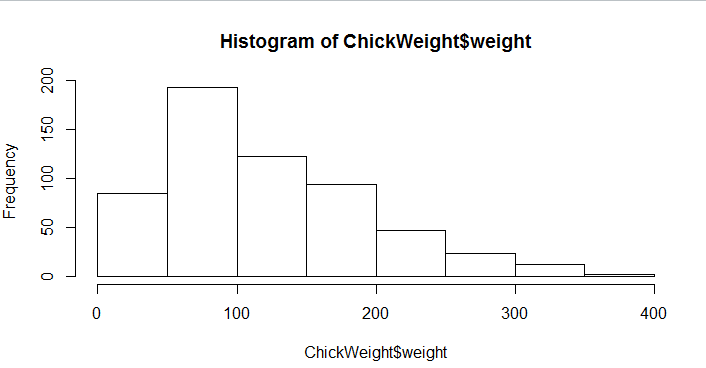
**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| **Skewness** | 1.58 | -0.6 |
| **Kurtosis** | -0.58 | 0.25 |

SP column has +ve skewness and -ve kurtosis So it’s a positively-skewed platykurtic distribution. Most of the data points are lies far from the mean.

WT column has -ve skewness and +ve kurtosis so it’s a negatively skewed leptokurtic distribution. Here most of the data stay around the mean.

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





**Ans:**

Histogram shows the distribution is right-skewed. About 200 chicks weighs between 50-100. A few chicks weigh between 350-400.

The box plot shows there are some outliers present inside the dataset. And the outliers are mostly present when chicks weight is higher say 350-400.

**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 size = 2000

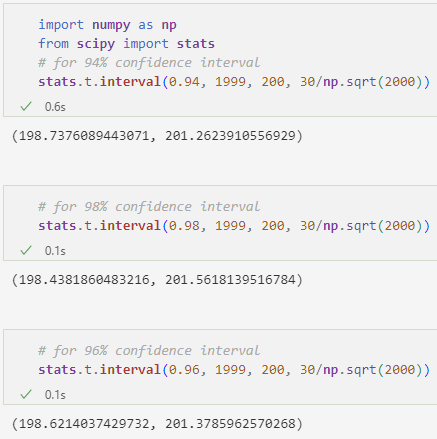
Therefore degrees of freedom will be 1999

population is 3000000

Sample mean = 200

Standard deviation of sample is 30

Here we don’t know population standard deviation so we use students t distribution.



**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**

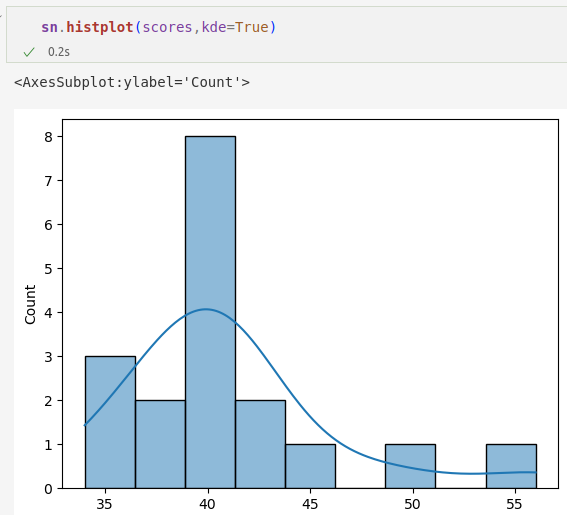
1. **Find mean, median, variance, standard deviation.**
2. **What can we say about the student marks?**

**Ans:**



Average mark of the students is 41.

Drawing histogram of marks.



Here we can see most of the students i.e 8 no of students’ scored 40 marks.

The graph shows its right-skewed data. This means there are a few no. of students who score more than 50.

3 students scored 35 which is a poor score compared to the average.

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

**Ans.**

The skewness value will be 0, Which will form a symmetric graph and also it forms a normal distribution considering it is not a multi-modal.

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

**Ans:**

It's a **right**-skewed distribution. The value of skewness is **greater than** 0.

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

**Ans:**

It's a **left**-skewed distribution. The value of skewness is **less than** 1.

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

**Ans:**

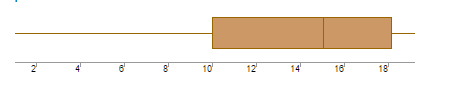
Data comes from a **Leptokurtic** distribution. Where most of the values lie close to the mean.

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

**Ans:**

Data comes from a **Platykurtic** distribution. Where most of the values lie far from the mean.

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



**What can we say about the distribution of the data?**

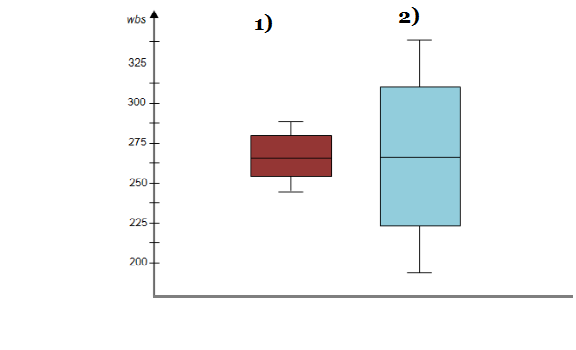
The box plot is not symmetrical so we can say it's not a normal distribution. The lower whisker is too large, this shows the data may be stretched out toward the left of the mean.

**What is nature of skewness of the data?**

The skewness is negative. Data comes from a left-skewed distribution.

**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.**

**Ans:**

Both box plots have the same median i.e, 262.5

Both of them are symmetrical which means these are coming from a normal distribution.

The lower quantile and upper quantile of these plots are not the same. So we can say the range of these two differs. The range of 2nd data is more than the first one.

Also, the variance of the 2nd data is more compared to the 1st data.

**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: [ Please check: stats 1.ipynb ]

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

**Ans:**

[ Please check: stats 1.ipynb ]

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

**Dataset: wc-at.csv**

**Ans:** [ Please check: stats 1.ipynb ]

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

**Ans:** [ Please check: stats 1.ipynb ]

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

**Ans:** [ Please check: stats 1.ipynb ]

**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:**

Population mean is 270

Sample size = 18

Sample mean = 260

Sample standard deviation = 90

Degrees of freedom = 17

[ Please check: stats 1.ipynb ]