|  |  |
| --- | --- |
| **Activity** | **Data Type** |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continuous data |
| Weight of Gold | Continuous data |
| Distance between two places | Continuous data |
| Length of a leaf | Continuous data |
| Dog's weight | Continuous data |
| Blue Color | Nominal data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Nominal 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 data |
| High School Class Ranking | Ordinal data |
| Celsius Temperature | Interval data |
| Weight | Ratio data |
| Hair Color | Nominal data |
| Socioeconomic Status | Ordinal data |
| Fahrenheit Temperature | Interval data |
| Height | Ratio data |
| Type of living accommodation | Nominal data |
| Level of Agreement | Ordinal data |
| IQ(Intelligence Scale) | Ordinal data |
| Sales Figures | Ratio data |
| Blood Group | Nominal data |
| Time Of Day | Interval data |
| Time on a Clock with Hands | Interval data |
| Number of Children | Ratio data |
| Religious Preference | Nominal data |
| Barometer Pressure | Interval data |
| SAT Scores | Interval data |
| Years of Education | Ordinal data |

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

**Ans:** Three coin are tossed is sample space is

**S = {HHH,HHT,HTH,THH,HTT,TTH,THT,TTT} , n(S) = 8**

Probability of three coins tossed is : 1/8

The two heads and one tail is sample space is

S1= {HHT,HTH,THH}, n(S1) = 3

Probability of Two heads and One tail is P(S1) = 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:-** When two dice are rolled simultaneously the possible no of outcomes are 36 Probability that rolling two dice is 1/36.

The sample space of two dice rolled is

S= {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),

(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),

(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),

(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6),

(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),

(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

n(S)= 36

1. Probability that sum is equal to 1 is 0
2. Less than or equal to 4 The Sample space S = { (1,1), (1,2), (1,3), (2,1), (2,2), (3,1) } , n(s)= 6

The probability of less then or equal to 4 is P(S) = 6/36 = 1/6

1. Sum is divisible by 2 and 3

its possible values and Sample space is

S = {(1,5), (3,3), (2,4), (4,2), (5,1), (6,6)}

n(S) = 6

probability of sum is divisible by 2 and 3 is P(S) = 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:**

Total number of balls = (2 + 3 + 2) = 7. Let S be the sample space.

Then, n(S) = Number of ways of drawing 2 balls out of 7 = 7C2​ = 21

Let E = Event of drawing 2 balls, none of which is blue.

n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls = 5C2​ = 10

∴P(E)= n(E)​/n(S) **=10/21**

∴The probability that none of the balls drawn is blue 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

**Ans:** Expected number of candies for a randomly selected child = ∑XiPi

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

= 3.09

Expected number of candies for a randomly selected child is **3.09**

**Q7)** Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weight

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**Ans:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Point** | **Score** | **Weight** | | **Mean** | 3.596563 | 3.21725 | 17.84875 | | **Median** | 3.695 | 3.325 | 17.71 | | **Mode** | 3.92 | 3.44 | 17.02 | | **Variance** | 0.285881 | 0.957379 | 3.193166 | | **SD** | 0.534679 | 0.978457 | 1.786943 | | **Range** | 2.17 | 3.911 | 8.4 | |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

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:**

There is total 9 patients at clinic. We have to assume one of the patients is chosen at random so its probability is 1/9.

The Expected Value of the Weight of that patient

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

= 144.3333

The Expected Value of the Weight of that patient is **145.3333**

**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

**Ans:**

Skewness: **-0.1140**(Car’s speed) and **0.7825**(distance)

Kurtosis: **-0.5771**(Car’s speed) and **0.2480**(distance)

Skewness: **1.5814**(SP) and **-0.6033**(Weight)

Kurtosis: **2.7235**(SP) and **0.8194** (Weight)

**Interpretation:**

In cars speed and distance data, cars speed is negatively skewed and the value of kurtosis is negative i.e. it is less peaked than normal curve therefore the curve is platykurtic. while distance is positively skewed and it is more peaked than normal curve hence it is leptokurtic curve.

For SP and Weight data, SP is positively skewed and the value of kurtosis is negative i.e. it is less peaked than normal curve therefore the curve is platykurtic. while WT is positively skewed and its kurtosis value is positive therefore it is more peaked than normal curve hence it is leptokurtic curve.

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



**Interpretation:**

From the given histogram of chickWeight$weight it is clear that the most of the data falls to the right side of the graphs peak thus, the histogram skews in such way that its tail on the right side of distribution is longer so it is positively skewed. In positively skewed distribution mean is always greater than median. Also, from boxplot it is reveals that the distribution is positively skewed and there are some outliers are present in the data.

**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 = n = 2000

Population size = N = 30000000

Sample mean = X = 200

Sample SD = 30

Now we calculate CI for 94%,98% and 98%.

For 94% Confidence Interval is (198.72, 201.26) and Z score value = 1.88

For 96% Confidence Interval is (198.62, 201.37) and Z score value = 2.05

For 98% Confidence Interval is (198.44, 201.56) and Z score value = 2.32

**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:**

From the above scores of students,

1. Mean = 41, Median = 40.5, Variance = 24.1111 and Standard Deviation = 4.9103
2. Average of student’s marks are 41. By calculating 95% confidence interval for given data it is clear that the students marks are lies in the interval (38, 43).

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

**Ans:**

If the mean is equal to the median then distribution is symmetric and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median = mode.

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

**Ans:** If the mean is greater than the median, the distribution is positively skewed.

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

**Ans:** If the mean is less than the median, the distribution is negatively skewed.

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

**Ans:** A distribution with a positive kurtosis value indicates that the distribution has heavier tails than the normal distribution. This distribution is known as leptokurtic distribution. Leptokurtic distribution is more peaked than normal curve.

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

**Ans:-** A distribution with a negative kurtosis value indicates that the distribution is flat and has thin tails than the normal distribution. This distribution is known as platykurtic distribution. Platykurtic distribution is less peaked than normal curve.

**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:**

1. From the above boxplot it is clear that the distribution is negatively skewed because, the median is closer to the third quartile of the data which is tend towards right so it is negatively skewed distribution.
2. The nature of given boxplot is negatively skewed.
3. From the given boxplot, first quartile is approx. 10 and third quartile is 18.

So, the inter quartile range is (Q3-Q1) i.e. (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:** From the above boxplot 1 and 2, boxplot 1 is positively skewed because its median value is tended towards Q1 and the mean is always greater than median and the median is always greater than the mode. Boxplot 2 is normal distribution it is approximately symmetric distribution. i.e mean = median = mode.

**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)

**Ans:**

* 1. P(MPG>38) = 1-stats.norm.cdf (38, loc = data.MPG.mean(), scale= data.MPG.std ()) =**0.3475**
  2. P(MPG<40) = stats.norm.cdf (38, loc = data.MPG.mean(), scale= data.MPG.std ()) = **0.7293**
  3. P (20<MPG<50) = stats.norm.cdf (20, loc = data.MPG.mean(), scale= data.MPG.std ()) - (1-stats.norm.cdf (50, loc data.MPG.mean(), scale= data.MPG.std ())) = **0.01311**

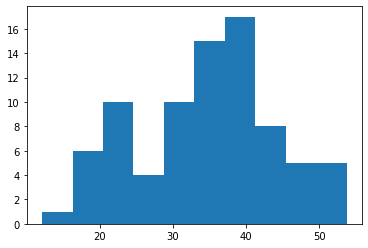
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:**

The MPG of cars does not follow Normal Distribution it has negatively skewed distribution. (From histogram)



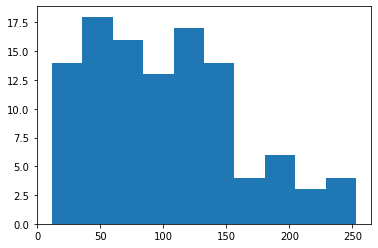
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

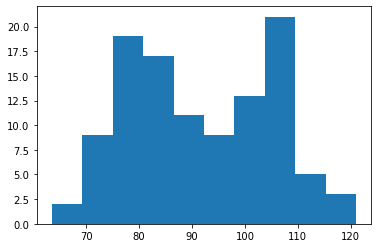
**Ans:**

The AT and Waist from wc-at data set both have positively skewed distribution.

**Histogram of AT:**



**Histogram of Waist:**



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

**Ans**: stats.norm.ppf (1-alpha/2)

Z score of 90% is 1.6448

Z score of 94% is 1.8807

Z score of 60% is 0.8416

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

**Ans**: stats.t.ppf(P%, df = 24)

t score of 95% is 1.7108

t score of 96% is 1.8280

t score of 99% is 2.4921

**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:**

t – statistic is

t =

x = sample mean = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = sample size = 18

t =

t =

t =

t =

t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here we need the t-distribution with 17 degrees of freedom.

(Using stats.t.cdf (-0.471, 17))

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is **0.3218**.