|  |  |
| --- | --- |
| 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 | Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Nominal |

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 | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

Solution: Probability= Favorable outcomes / Total no. of Outcomes

{(HHH), (TTT), (HTT), (THH), (HTH), (THT), (TTH), (HHT)}

Prob (two heads and one tail) = 3/8 = 0.375

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

Solution: Number of possible outcomes for the above event is 6^2=36

a) Prob (sum is Equal to 1) = 0

b) Prob (sum is Less than or equal to 4)

{(1,1)(1,2)(1,3)(2,1)(2,2)(3,1)}

=6/36= 1/6 =0.1666

1. Prob (Sum is divisible by 2 and 3)

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

= 6/36 = 0.1666

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?

Solution: Total number of balls in a bag = 7

N (Event (2 balls are drawn randomly from bag) = 7! / 2! \* 5!

= (7\*6\*5\*4\*3\*2\*1) /

(2\*1) \* (5\*4\*3\*2\*1)

N (Event (2 balls are drawn randomly from bag) = (7\*6)/ (2\*1) = 21

If none of them drawn 2 balls are blue = 7 – 2 = 5

N (Event (None of the balls drawn is blue) = 5! / 2! \* 3! = (5\*4) / (2\*1)

= 10

P (None of the balls drawn is blue) = N (Event (None of the balls drawn is blue) /

N (Event (2 balls are drawn randomly from

bag)

= 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

Solution: Expected number= sum (X \* Prob(X))

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

= 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**

Solution: Mean for Points = 3.59, Score = 3.21 and Weigh = 17.84

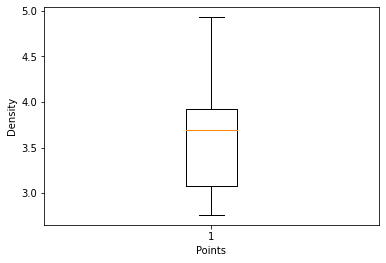
Median for Points = 3.69, Score = 3.32 and Weigh = 17.71

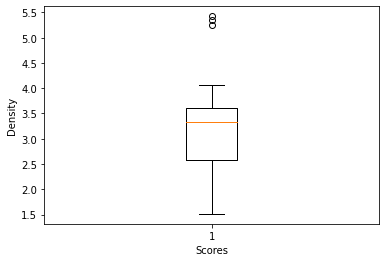
Mode for Points = 3.07, Score = 3.44 and Weigh = 17.02

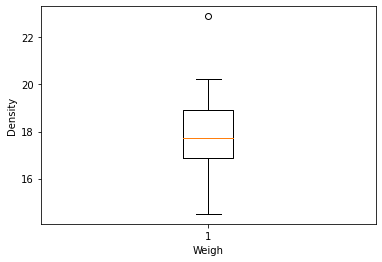
Variance for Points = 0.28, Score = 0.95, Weigh = 3.19

Standard Deviation for Points = 0.53, Score = 0.97, Weigh = 1.78

Range [Min-Max] for Points [3.59 – 4.93], Score [3.21 – 5.42] and Weigh [17.84 – 22.9]







The means of both Points and Score are closer. There are no outliers present for points.

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?

Solution: Expected Value= Sum(X\*Prob(X)) = (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) = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Solution: For cars speed Skewness value= -0.12 The skewness value is negative so it is slightly left skewed

For cars speed Kurtosis value= -0.51 Negative Kurtosis indicates that the distribution is not peaked, the distribution is flat, has thin tails

For cars dist Skewness value= 0.81 The skewness value is positive so it is slightly right skewed

For cars dist Kurtosis value= 0.41 Positive Kurtosis value indicates that the distribution is peaked and has thick tails

**SP and Weight (WT)**

**Use Q9\_b.csv**

For SP Skewness= 1.61 The skewness value is positive so it is right skewed

For SP Kurtosis= 2.98 Positive Kurtosis value indicates that the distribution is peaked and has thick tails

For WT Skewness= -0.61 The skewness value is negative so it is slightly left skewed

For WT Kurtosis= 0.95 Positive Kurtosis value indicates that the distribution is peaked and has thick tails

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



Histogram: The chick weight data is right skewed or positively skewed. The histogram 's peak has right skew and tail is on right. Mean>Median. Most of the chicks’ weight is between the intervals 50 to 100. The least range of weight is 400.



Box Plot: The data is right skewed. Mean> Median. Here outliers exist and are at the upper side of the box plot and there are less data points at the bottom.

**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?

Solution: The confidence intervals for 94% is [198.74 - 201.26]

The confidence intervals for 96% is [198.62 - 201.38]

The confidence intervals for 98% is [198.44 - 201.56]

**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.

Solution: Mean =41, Median =40.5, Variance =25.52 and Standard Deviation =5.05

1. What can we say about the student marks?

Solution: We don’t have outliers and the data is slightly skewed towards right because mean is greater than median

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

Solution: When mean, median of data are equal the distribution is symmetric and the distribution has zero skewness

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

Solution: The nature of skewness when mean > median is the distribution is positively skewed

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

Solution: The nature of skewness when mean > median is the distribution is negatively skewed

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

Solution: Positive Kurtosis value indicates that the distribution is peaked and has thick tails

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

Solution: Negative Kurtosis value indicates that the distribution is not peaked, the distribution is flat and has thin tails

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



What can we say about the distribution of the data?

The data is not actually equally distributed across the plane. There might be outliers influencing the data. Median is not dividing the data in two equal halves. Hence it is not normally distributed.

What is nature of skewness of the data?

The data is left skewed. Hence median will be greater than mean since the data is left skewed.

What will be the IQR of the data (approximately)?   
  
IQR = Upper quartile - Lower quartile   
 =18 – 10   
 =8 (approx.)

Q19) Comment on the below Boxplot visualizations?

Solution: Here in the representation of 2 box plots, plot 2) is highly distributed across the plane and plot 1) is slightly less distributed. Whiskers are distributed. 100% of the data is spread in plot 2) and in plot 1) it is spread in the range 250-290.



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

Both the plots share the same median and they are normally distributed with zero or no skewness. Whisker level is high in plot 2). The data is spread symmetrically in both of the plots.

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)**

np. round (1-(stats.norm.cdf(38, cars['MPG'].mean(), cars['MPG'].std())),3)

0.348

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

np.round(stats.norm.cdf(40,cars['MPG'].mean(),cars['MPG'].std()),3)

0.729

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

prob1=stats.norm.cdf(50, cars['MPG'].mean(),cars['MPG'].std())

prob1

0.9559926932893639

prob2=stats.norm.cdf(20, cars['MPG'].mean(),cars['MPG'].std())

prob2

0.05712377632115912

=> prob1 - prob2

0.9559926932893639-0.05712377632115912

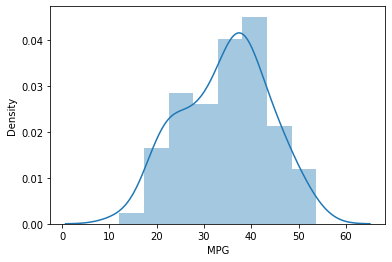
=0.8988689169682047

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

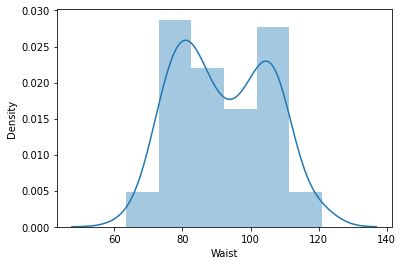
1. MPG of cars follows normal distribution

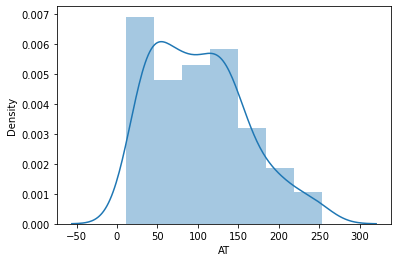


1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc -at data set follows Normal Distribution

Dataset: wc-at.csv

Adipose Tissue (AT) and Waist does not follow Normal Distribution





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

Solution: Z score for 90% confidence interval is [-1.645 - 1.645]

Z score for 94% confidence interval is [-1.881 - .881]

Z score for 90% confidence interval is [-0.842 - 0.842]

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

Solution: T score for 95% confidence interval is [-2.064 - 2.064]

T score for 96% confidence interval is [-2.172 - 2.172]

T score for 99% confidence interval is [-2.797 - 2.797]

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

Solution: t=np.round((260-270)/(90/np.sqrt(18)),3)

-0.471

print('The probability that bulbs have an average life of no more than 260 days is',np.round(stats.t.cdf(-0.471,17),2))

The probability that bulbs have an average life of no more than 260 days is 0.32