|  |  |
| --- | --- |
| 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 | **Categorical / Discrete** |
| Number of kids | **Discrete** |
| Number of tickets in Indian railways | **Discrete** |
| Number of times married | **Discrete** |
| Gender (Male or Female) | **Categorical / 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(Morning, Afternoon, Eve)? | **Nominal** |
| 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?

P(HHT) = (½)^3 \* (3 choose 2) = **3/8**

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

1. Equal to 1

Die 1 = 1 to 6(6 possibilities)

Die 2 = 1 to 6(6 possibilities)

So we have 36 possibilities when 2 dice are rolled.

P(X=1) = Probability of getting the sum equal to 1 by rolling 2 dice

P(X=1) = 0/36 = **0**

1. Less than or equal to 4

P(X<=4) = (1,1) (1,2) (1,3) (2,1) (2,2) (3,1)

= 6/36 = **1/6**

1. Sum is divisible by 2 and 3

P(X/2 and X/3) = (1,5)(5,1)(2,4)(4,2)(3,3)(6,6)

= 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?

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

Probability of picking the first ball not being blue is (7-2)/7 = 5/7

Probability of picking the second ball considering the first ball is picked and not blue is (6-2)/6 = 4/6 =2/3

So the Probability that none of the balls drawn is blue = 5/7\*2/3 = **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

Expected number of candies for a randomly selected child

= 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= **3.09**

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

* For Points(P), Score(S), Weight(W)>

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

**Use Q7.csv file**

Mean = **3.596(P) 3.217(S) 17.848(W)** ; Median = **3.695(P) 3.325(S) 17.710(W)**;

Mode = **3.07(P) 3.44(S) 17.02 & 18.9(W)** ; Variance = **0.285(P) 0.957(S) 3.193(W)** ;

Standard Deviation = **0.534(P) 0.978(S) 1.786(W)** ; Range = **2.17(P) 3.911(S) 8.4(W)**

All data sets follow Normal Distribution, as the Mean and Median values are closer to each other.

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?

The Expected value is equivalent to the mean of the population.

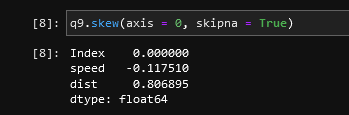
E(X) = (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)/9 = **145.33**

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

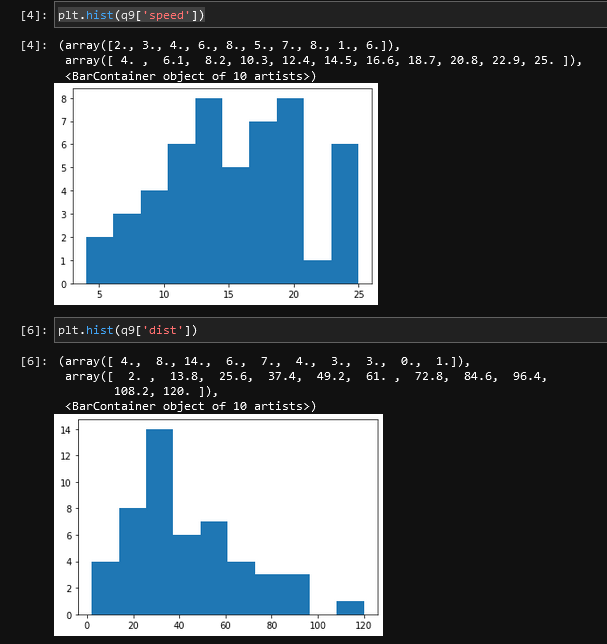
Cars speed and distance

Use Q9\_a.csv

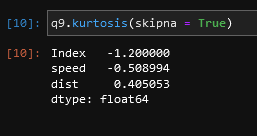
**Skewness:**



1. Speed has a skewness of -0.11, meaning the mass of the data is slightly on the right side of the distribution (In this example, since the skewness is negative and also closer to zero, mass is almost in the center)
2. Distance has a skewness of 0.80, meaning the mass of the data is on the left side of the distribution.

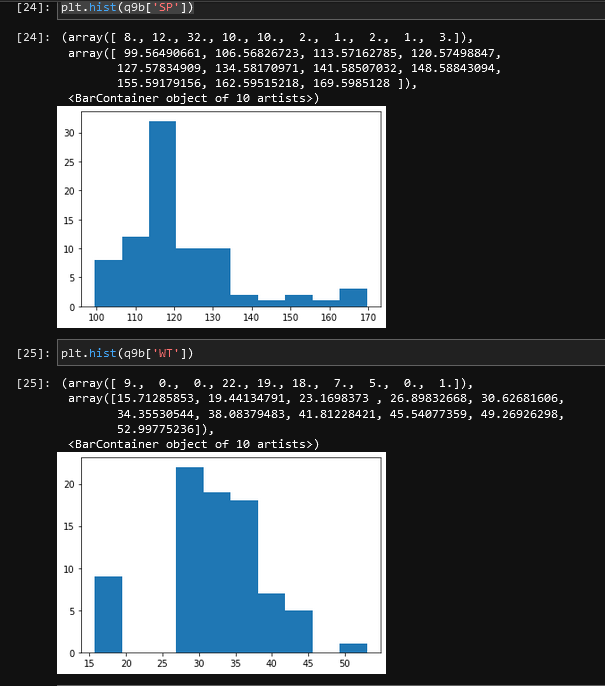


**Kurtosis:**

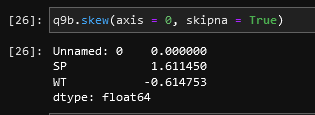


1. Speed has a negative Kurtosis value of -0.50, indicating the distribution having a wider peak and thinner tails.
2. Distance has a Kurtosis value of 0.40, indicating the distribution having a thinner peak.

SP and Weight(WT) {Use Q9\_b.csv}

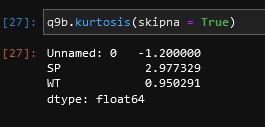


**Skewness:**



1. SP has a skewness of 1.61, meaning the mass of the data is concentrated on the left side of the distribution
2. Weight(WT) has a skewness of -0.60, meaning the mass of the data is slightly on the right side of the distribution (Yet close to the center as the value is closer to 0).

**Kurtosis:**



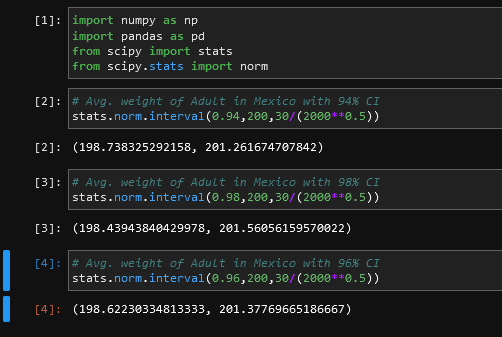
1. SP has a Kurtosis value of 2.97, indicating the distribution having very thin peak (as the value is on the extreme positive side).
2. WT has a Kurtosis value of 0.95, indicating the distribution having a thinner peak.

Q10) Draw inferences about the following boxplot & histogram



1. Histogram is Right Skewed (Positive Skewness), meaning the mass of the data is on the left of the distribution and the tail is on the right side.
2. Box plot has data (~75%) concentrated towards the left of the distribution (Skewed right) and there are multiplier outliers towards the extreme right.

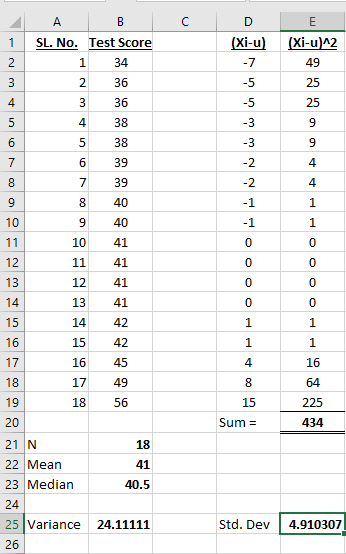
Q11)Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% 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



1. What can we say about the student marks?

On an average, students have scored 41 in the tests. Majority of the data is concentrated between 34 and 45. Hence the distribution is positive skewed or right skewed. Also the width of the peak is thinner.

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

Skewness will be zero and the distribution will have a perfect bell shape

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

Distribution will be Right Skewed

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

Distribution will be Left Skewed

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

Curve will have a thin peak

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

Curve will have a wider peak and thinner tails

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



What can we say about the distribution of the data?

75% of the data is between 10 and 18

What is nature of skewness of the data?

Data distribution is Left Skewed

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 has an IQR of 25 compared to Boxplot 2 which has an IQR of 87.5. So, the Kurtosis of Boxplot 1 is on the high positive side, indicating the width of the peak (of data distribution) is thinner. Boxplot 2 has a negative Kurtosis value, indicating the peak is wider.

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

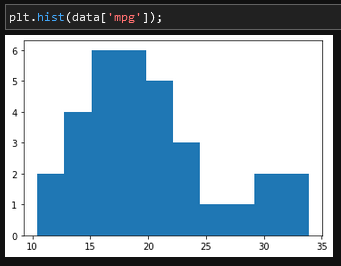
* 1. P(MPG>38) = **33/81**
  2. P(MPG<40) = **61/81**

c. P (20<MPG<50) = **69/81**

Note: Total count of MPG is 81 and hence the probability is calculated as per the values given

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution



|  |  |
| --- | --- |
| Mean(MPG)= | 34.42208 |
| Median(MPG)= | 35.15273 |

Mean is almost closer to the Median. And as per the empirical rule of Normal Distribution, more than 95% of the data is within the range of (µ± 2\*SD). Hence the MPG data follows Normal Distribution.

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

Fig(a)

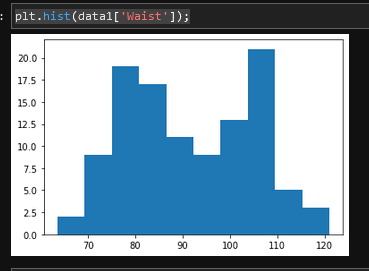
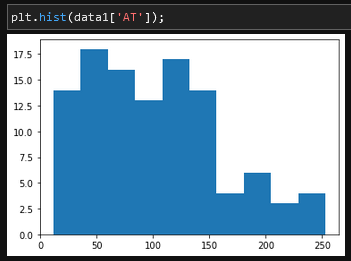


Fig. (b)



1. Data distribution of “Waist”, as shown in the Fig (a), has the following measures of central tendency:

|  |  |
| --- | --- |
| Mean(Waist)= | 91.90183 |
| Median(Waist)= | 90.8 |

As the mean and median are almost close to each other, the distribution is normal.

1. Data distribution of “AT”, as shown in the Fig(b), has the following measures of central tendency:

|  |  |
| --- | --- |
| Mean(AT) = | 101.894 |
| Median(AT)= | 96.54 |

As the mean and median values are quite far apart, it doesn’t follow the normal distribution.

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

Z score of 90% =1.65

Z score of 94% = 1.56

Z score of 60% = 0.26

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

t-score of 95% = 2.060

t-score of 96% = 1.828

t-score of 99% = 2.787

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

Mean(µ) = 270, Sample size(n)=18, Sample mean(×)=260,

Std. Dev. of Sample(s) = 90

t=(x-μ)/(s/sqrt(n))

t = (260-270)/(90/sqrt(18)) = **-0.471**

The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is **32.16%**

