|  |  |
| --- | --- |
| 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 |

**Assignment 1**

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

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

Ans: Sample Space = {HHH, HHT, HTH, THH, HTT, TTH, THT, TTT}

P(X)= No. of desired events / Total no. of events

No. of desired events= 3 i.e. (HHT, HTH, THH)

P (X=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

Ans: Sample Space:

(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)

1. P (X= sum is equal to 1) = No. of events with sum equal to 1/ Total no. of events

= 0/36 = 0

1. Events with sum less than or equal to 4 are (1,1) (1,2) (1,3) (2,1) (2,2) (3,1).

Hence probability is,

P (X= sum is less than or equal to 4) = No. of events with sum less than or equal to

4/ Total no. of events= 6/36=1/6=0.166

1. Events with sum divisible by 2 and 3 are (1,5) (2,4) (3,3) (4,2) (5,1), (6,6).

Hence probability is,

P (X= Sum is divisible by 2 and 3) = No. of events with sum divisible by 2 and 3/ Total no. of events = 6/36= 1/6=0.166

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: The combination formula is used to find the number of ways of selecting items from a collection, such that the order of selection does not matter.

No. of ways to randomly draw 2 balls from collection of 7 balls=7C2=21(sample space)

We want to draw 2 balls such that none of them is blue. So we have to draw balls from rest 5 balls. That is 2 red and 3 green.

No. of ways to randomly draw 2 balls from collection of 5 balls= 5C2=10(desired events)

P (X= drawing 2 balls which are not blue) = No. of desired events / Total no. of events

= 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: = 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120 = 3.09

Hence, the Expected number of candies for a randomly selected child are 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:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | Variance | Std. Deviation | Range | CV |
| Points | 3.596563 | 3.695 | 3.92 | 0.285881 | 0.534679 | (4.93,2.76) | 0.14 |
| Scores | 3.217250 | 3.325 | 3.44 | 0.957379 | 0.978457 | (5.424,1.513) | 0.30 |
| Weigh | 17.848750 | 17.710 | 17.02 | 3.193166 | 1.786943 | (22.9,14.5) | 0.10 |

As we know that, Mean is susceptible to outliers. Hence we use median to get a true picture of data instead of distorted one provided by mean.

But here, as we can see, there is not much difference between Mean values and Median values of Points, Scores and Weigh column. It shows that there are not many outliers in dataset, else there would have been a conspicuous difference between mean and median values.

Points and Scores data is left skewed since Median>Mean

Whereas, Weigh data is Right skewed since Mean>Median

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean.

Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out

Coefficient of Variation = Std Deviation/ Mean

CV value greater than 1 is often considered high

Here, for all three features, CV value is less than 1, means data are clustered around the mean.

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: Expected value = ∑ (X \* Probability of 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

Hence, Expected Value of the Weight of randomly chosen patient is 145.33 pounds.

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

**Use Q9\_a.csv -> Cars speed and distance**

**Use Q9\_b.csv ->SP and Weight (WT)**

Ans: Q9\_a

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| Speed | -0.117510 | -0.508994 |
| Distance | 0.806895 | 0.405053 |



Skewness comments on symmetricity of distribution.

Perfectly symmetrical distribution has zero skewness.

Right skewed distribution has positive skewness value. Distribution of “distance” is right skewed.

Left skewed distribution has negative skewness value. Distribution of “speed” is left skewed.

Kurtosis measures sharpness of distribution.

Normal distribution has kurtosis value = 3

Flat distribution has kurtosis value < 3. Hence both speed and distance have less sharp peaks and fat tails (most of observations at ends)

Sharp distribution has kurtosis value >3. Thin tails indicate very less data present at ends.

Ans: Q9\_b

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| SP | 1.611450 | 2.977329 |
| WT | -0.614753 | 0.950291 |

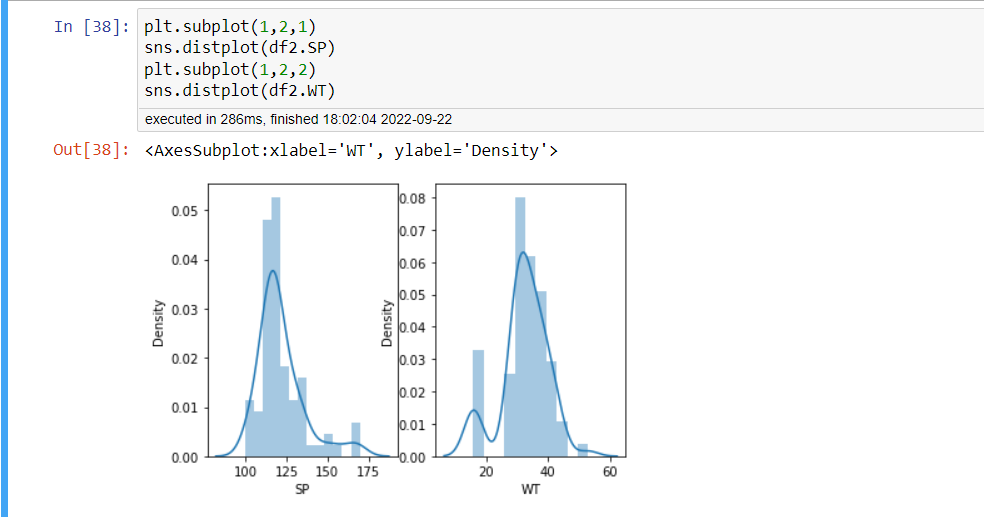


SP has positive skewness valuemeans it has Right skewed distribution.

WT has negative skewness value means it has Left skewed distribution.

SP has kurtosis value almost equal to 3 means it has Normal distribution. (ref. graph below)

WT has kurtosis value < 3 means distribution has less sharp peak.

****

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



Ans:

It is a unimodal histogram since it has one peak

Histogram of Chicken Weight is right skewed. Hence Mean > Median because the mean is more sensitive to the higher values and is drawn towards the tail of the histogram.

Histogram shows that majority of chicken are having weight below 200gms.

And only a few chickens have weight more than 300gm; which are outliers, which also can be seen in boxplot.

Lower whisker of boxplot representing 25% of chickens is very short. It shows these chickens constitute low weight group with very less variation in weight.

Higher whisker of boxplot representing 25% of chickens is long. It shows these chickens constitute high weight group with high variation in weight.

**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: Population size N= 3,000,000

Sample size n= 2,000

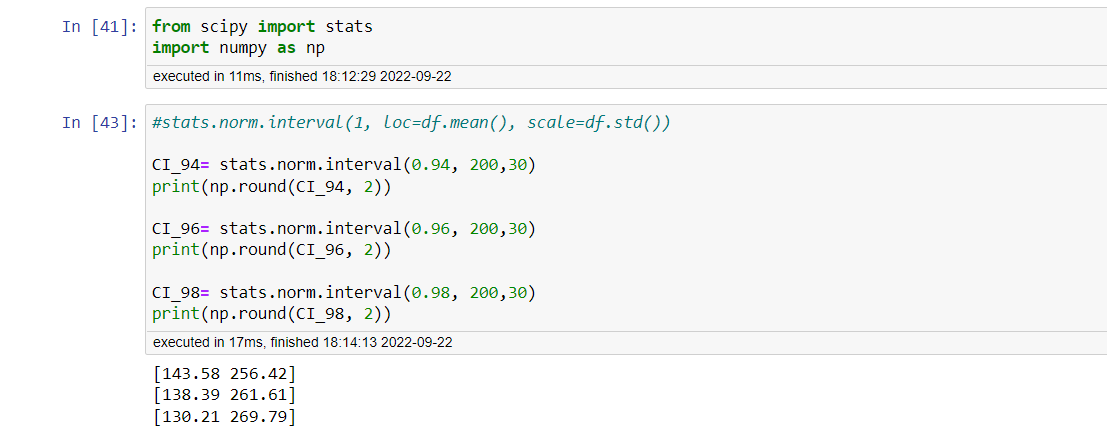
Sample mean x̅ = 200 pounds

Sample std deviation S = 30 pounds

At 94% confidence interval average weight of an adult male in Mexico would be in the range of [143.58, 256.42]

At 96% confidence interval average weight of an adult male in Mexico would be in the range of [138.39, 261.61]

At 98% confidence interval average weight of an adult male in Mexico would be in the range of [130.21, 269.79]

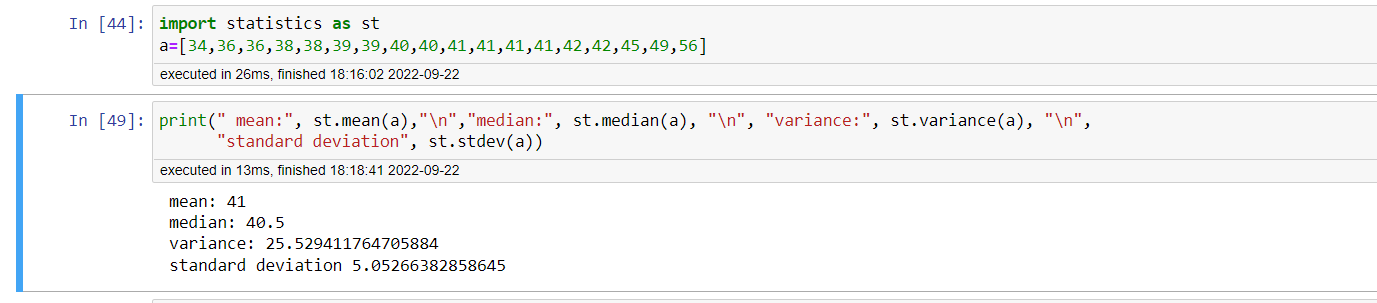


**Q12)** Below are the scores obtained by a student in tests

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



Difference between mean and median is not much which indicates outliers do not have very high values in order to distort mean.

49 and 56 are outlier scores which makes data right skewed.

Coefficient of variance (Sd/mean) is 0.123 which is low. It indicates data has low variance and is closely distributed around the mean.

Most of the times he/she scored marks in range of 34 to 45.

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

Ans: When mean & median are equal, then data is distributed symmetrically around mean. Skewness is zero.

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

Ans: When mean > median, data trails off to the right and has a longer right tail then data is said to be skewed on the right side. Higher values are infrequent. Skewness is positive.

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

Ans: When mean < median, data trails off to the left and has a longer left tail then data is said to be skewed on the left side. Lower values are infrequent. Skewness is negative.

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

Ans: Kurtosis measures sharpness of distribution. Positive kurtosis value indicates sharp peaks and thin tail of distribution. Most of the data is in center and very less at ends.

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

Ans: Kurtosis measures sharpness of distribution. Negative kurtosis value indicates flat peaks and fat tail of distribution. Many observations are at ends.

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



What can we say about the distribution of the data?

Ans: Length of lower whisker and lower box is more than length of higher whisker and higher box. It shows that lower 50% of data is sparsely distributed and has high variation. Whereas higher 50% of data is closely distributed and has low variation.

What is nature of skewness of the data?

Ans: Data trails off to the left and has a longer lower whisker, hence data is skewed on the left side

What will be the IQR of the data (approximately)?   
Ans: The interquartile range (IQR), is the difference between the third quartile and the first quartile.

This tells us how spread out the middle 50% of values are in a given dataset.

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:

Median value for both datasets is same.

Length of boxplot 1 is short w.r.t boxplot 2 which indicates data points of 1st boxplot are closely spaced and have less variation among themselves, Whereas data points of 2nd boxplot are sparsely placed, and have higher variation among themselves.

Skewness for both boxplots is zero, means they are perfectly symmetrical around mean.

There are no outliers in both boxplots.

Q 20) Calculate probability from the given dataset for the below cases

Calculate the probability of MPG of Cars for the below cases. (Data \_set: Cars.csv.)

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

Ans: P(MPG>38) = 0.3475

P(MPG<40) = 0.7293

P(20<MPG<50) = 0.899



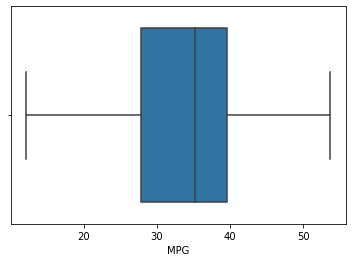
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: MPG of Cars does not follow “perfect” normal distribution since skewness is -0.17.

It is left skewed distribution. But it is near normal distribution.



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

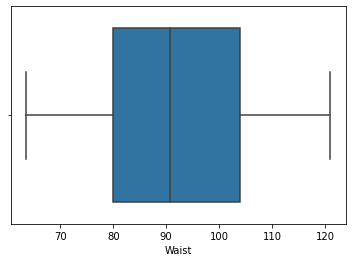
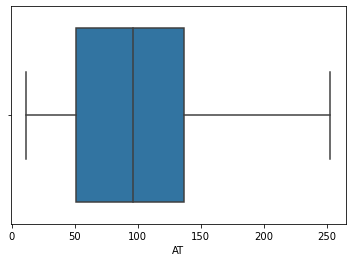
Dataset: wc-at.csv

Ans:

Waist skewness= 0.1340 AT skewness= 0.5848

Waist circumference does not follow “perfect” normal distribution since skewness is 0.134. It is little right skewed. But it is near normal distribution.

Adipose Tissue does not follow normal distribution since it is skewed towards right.



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

Ans: By referring Z table, we get

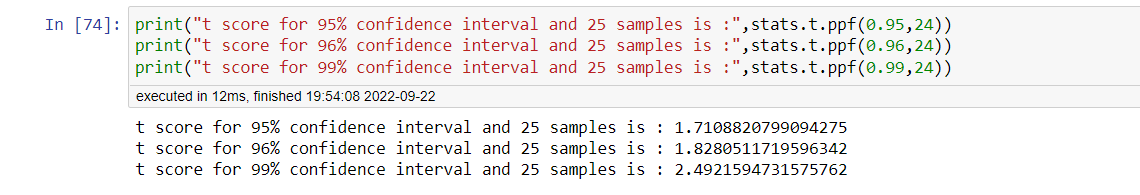
Z scores of 90% confidence interval is 1.645

Z scores of 94% confidence interval is 1.88.

Z scores of 60% confidence interval is 0.253

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

Ans:



(Scroll down…)

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 less than 260 days

Ans: 

*Thank You*