ASSIGNMENT

**STATISTIC-1**

Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| **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/Categorical*** |
| Number of kids | ***Discrete*** |
| Number of tickets in Indian railways | ***Discrete*** |
| Number of times married | ***Discrete*** |
| Gender (Male or Female) | ***Discrete/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 | ***Nominal*** |
| Level of Agreement | ***Ordinal*** |
| IQ(Intelligence Scale) | ***Interval*** |
| Sales Figures | ***Interval*** |
| Blood Group | ***Nominal*** |
| Time Of Day | ***Ratio*** |
| Time on a Clock with Hands | ***Interval*** |
| Number of Children | ***Ratio*** |
| Religious Preference | ***Nominal*** |

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

Ans: Since three coins are tossed so total possible outcomes are:

{HHH, HHT, HTT, TTT, TTH, THH, THT, HTH}= **8**

Interested outcomes (two heads and one tail):

{HHT, THH, HTH}= **3**

So the probability that two heads and one tail obtained is:

P(A)=outcome of interest

Total number of outcomes

P(A)**= 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: The total possible outcomes when two dice rolled are:

{ (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)} = **36**

The probability that sum is:

1. Equal to 1:

P(A)= **0**(zero). Because two dice are rolled and assuming that both dice got least value i.e. (1, 1) so when we sum up we get value 2, were finding sum is equal to 1 is not possible hence probability is zero.

1. Less than or equal to 4:

Here the possible outcomes are:

{ (1, 1)(1, 2)(1,3)(2,1)(2,2)(3,1)= **6**

P(A)**= 6**

**36**

1. Sum is divisible by 2 and 3:

The possible outcomes for divisible by 2 are:

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

(5, 1) (5, 3) (5, 5) (6, 2) (6, 4) (6, 6)}= **18**

P(A)= **18**

**36**

The possible outcomes for divisible by 3 are:

{ (1, 2)(1, 5)(2, 1)(2, 4)(3, 3)(3, 6)(4, 2)(4, 5)(5, 1)(5, 4)(6, 3)(6, 6)}= **12**

P(A)= **12**

**36**

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 total number of balls in the bag are:

2+3+2= **7balls** where 2 balls are blue.

So the probability of the first ball not being blue is = **5**

**7**

So now we are left with the 6 balls with 2 blue.

The probability of the second ball not being blue assuming that the first

wasn’t is = **4**

**6**

The probability that neither ball drawn was blue is:  **5** X **4 = 10**

**7 6 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**:** Here the Expected value is denoted as:

EV = Ʃ P(Xi)(Xi)

Where,

P(Xi): the probability of the event

(Xi) : the event

Hence the expected number of candies for the given probabilities is:

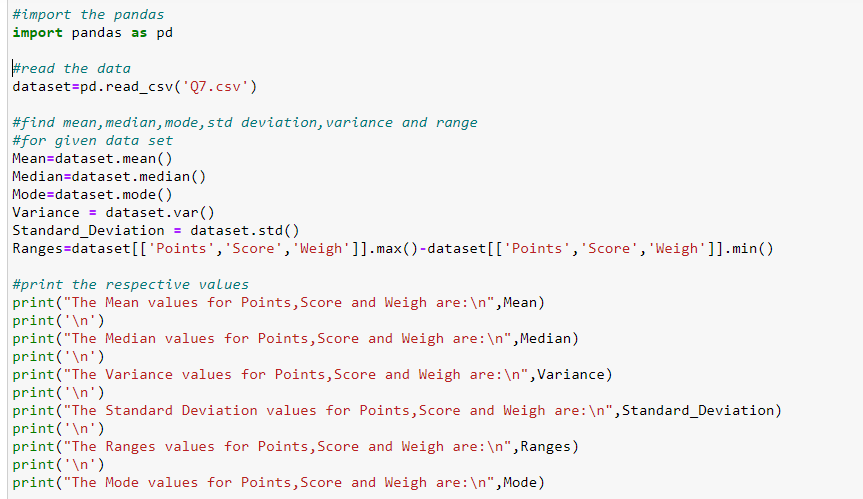
EV= (0.015x1)+(0.20x4)+(0.64x3)+(0.005x5)+(0.01x6)+(0.120x2)

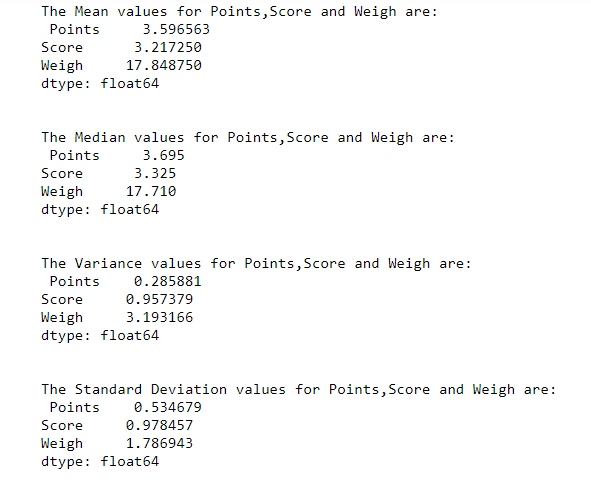
EV= **3.06**

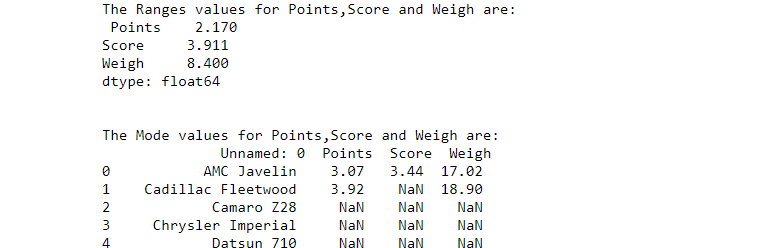
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weigh>

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

Ans: 





|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.596563 | 3.217250 | 17.848750 |
| Median | 3.695 | 3.325 | 17.710 |
| Mode | 3.07 and 3.92 | 3.44 | 17.02 and 18.90 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Standard Deviation | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.170 | 3.911 | 8.400 |

* From the above table we can say that mean and median of Points, Score and Weigh have no much difference so the data almost follows the normal distribution that means most of the data lies in the center.
* Also the variance and standard deviation of Weigh is high compared to Points and Score’s variance and standard deviation.
* The most occurred values in Points are 3.07 and 3.92, in Score is 3.44 and in Weigh are 17.02 and 18.9.So here the Points and Weigh are bimodal because they have two modes and they appear at the same number of times (Points=3 times and Weigh =2 times).
* The Weigh’s has high variability than Points and score because Weigh has higher range than Points and Score, also the Points has low variability compared to Score and Weigh.

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: Sometimes the Expected Value is also known as mean or average.

So we know that mean or average are calculated by adding all the values

dividing by the number of values.

Hence,

E(X) = 108+110+123+134+135+145+167+187+199

9

E(X) = **145.333**

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

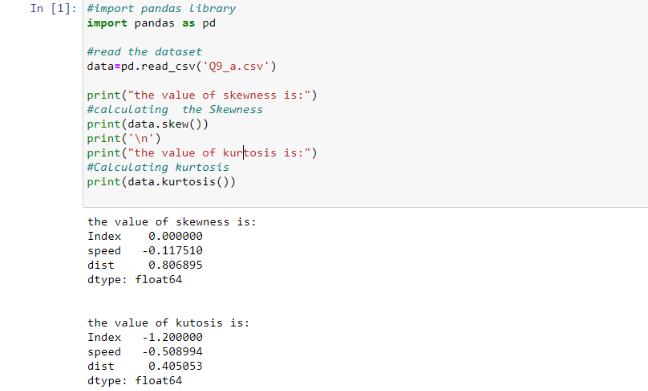
Cars speed and distance

Use Q9\_a.csv

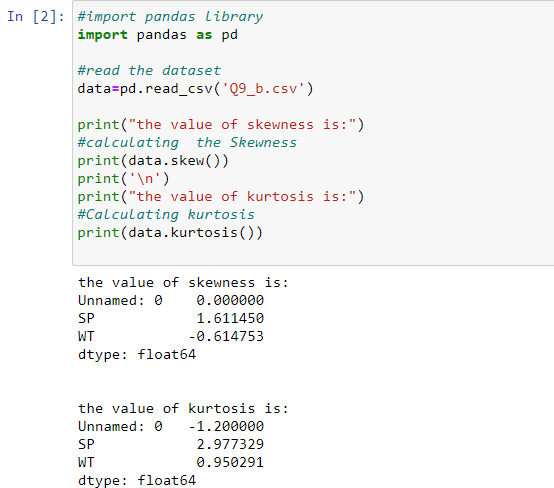
SP and Weight (WT)

Use Q9\_b.csv

Ans: The Skewness and kurtosis value for cars speed and distance is given below



The skewness and Kurtosis values for SP and Weight (WT) are:



|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| Speed | -0.117510 | -0.508994 |
| Distance | 0.806895 | 0.405053 |
| SP | 1.611450 | 2.977329 |
| Weight(WT) | -0.614753 | 0.950291 |

Inferences:

**Speed:**

* **Skewness:** A negative skewness value in the table above indicates an asymmetry in the distribution and the tail is larger towards the left hand side of the distribution.
* **Kurtosis:** Since the Kurtosis value is -0.508 that means it is slightly flatter than the normal distribution.

**Distance:**

* **Skewness:** A positive skewness value in the table above indicates an asymmetry in the distribution and the tail is larger towards the right hand side of the distribution.
* **Kurtosis:** Since the kurtosis value is 0.405 that means it is slightly pointy to the normal distribution.

**SP:**

* **Skewness:** A positive skewness value in the table above indicates an asymmetry in the distribution and the tail is larger towards the right hand side of the distribution.
* **Kurtosis:** Since the kurtosis value is 2.977 that means it is almost equal to the normal distribution.

**Weight (WT):**

* **Skewness:** A negative skewness value in the table above indicates an asymmetry in the distribution and the tail is larger towards the left hand side of the distribution.
* **Kurtosis:** Since the kurtosis value is 0.95 that means it is slightly pointy to the normal distribution.

Q10) Draw inferences about the following boxplot & histogram





Ans: **Histogram:** From the histogram plot above we can say that most of the data is falling under 50-200 with the frequency of 200.We can also infer that most of data is concentrated on the left side following the right side tail which is Positive skewed in nature and does not follow normal distribution . It has sharper peak that leads to positive Kurtosis.

**Boxplot:** From the box plot we can say that middle box contains the 50% of data. It has outliers at the upper extreme points and most of the data is on the lower extreme side.

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: In this question we have not given population standard deviation, then we will use t distribution for Calculate 94%, 98%, 96% confidence interval.

Given Population (N) = 3,000,000, sample size (n) =2,000, Sample Mean (x̄) =200, sample standard deviation (s) =30, Degree of freedom=n-1=2000-1

1999, Standard Error=s/=30/30/44.7213=0.671

1**. for 94% Confidence interval**

α =1-0.94=0.06

(T-value for 1999 by using t-table)

t 1-α,n-1=t 0.94,1999=,

+ t 0.94,1999\*s/=200+1.881861\*0.671=200+1.26272873=201.262729

- t 0.94,1999\*s/=200- 1.881861\*0.671=200-1.26272873=198.737271

[198.737271, 201.104466]

2**. for 98% confidence interval**

α =1-0.98=0.02

(T-value for 1999 by using t-table at 98%)

t 1-α,n-1=t 0.98,1999= 2.328215

+ t 0.98,1999\*s/=200+2.328215\*0.671=200+1.56223227=201.562232

- t 0.98,1999\*s/=200-2.328215\*0.671=200-1.56223227=198.437768

[198.437768, 201.562232]

**For 96% confidence interval**

α =1-0.96=0.04

(T-value for 1999 by using t-table at 96%)

t 1-α,n-1=t 0.96,1999=2.05509

+ t 0.96,1999\*s/=200+2.05509\*0.671=200+1.37896539=201.378965

- t 0.96,1999\*s/=200-2.05509\*0.671=200-1.37896539=198.621035

[198.621035, 201.378965]

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=24.111, Standard Deviation=4.910

We say that the scores obtained by a student in tests in increasing order and most Occurring value is 41 and the smallest marks is 34 and highest marks is 56.

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

Ans: If our data has equal mean and median the **skewness is zero** in nature and the distribution is symmetric i.e. normal distribution.

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

Ans: If mean is greater than the median then the nature of data is said to be

**Positively skewed.**

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

Ans: If median is greater than the mean the nature of data is said to be **Negatively**

**Skewed.**

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

Ans: The Positive Kurtosis value indicates that a distribution has a higher peak and taller tails than a normal distribution where more of the values are located in the tails of the distribution rather than around the mean.

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

Ans: The Negative Kurtosis value indicates that a distribution has a less peaked and shorter tails than the normal distribution with the fewer values in its shorter tails.

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:

* From the above boxplot we can say that the data is **not normally distributed.**
* The data is **negative skewed** as it is has tail towards left and most of the data are falling on the right side median is greater than the mean here.
* The IQR is given by upper quartile (Q3)-lower quartile (Q1)

IQR=Q3-Q1

= 18-10

IQR = **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 two Boxplot1 and Boxplot2 we can infer that both the boxplot have same Median of 260 approximately.
* There is no skewness in both of the boxplot hence they are symmetrical i.e. normally distributed.
* There are no outliers for both boxplot in upper and lower quartile.

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)= interested events=26/total number of events=81

=26/81=**0.3209**

1. P(MPG<40)=interested events =61/total events=81

=61/81=**0.75308**

1. P (20<MPG<50)=interested events=69/total events=81

=69/81=**0.85185**

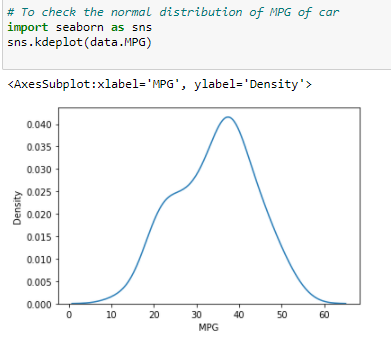
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

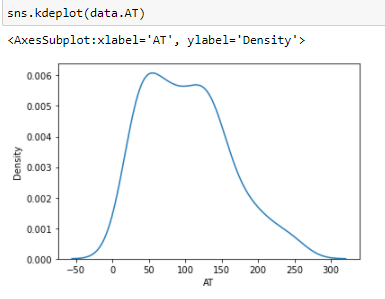
Ans) a)



From the above graph it is evident that the car dataset follows almost

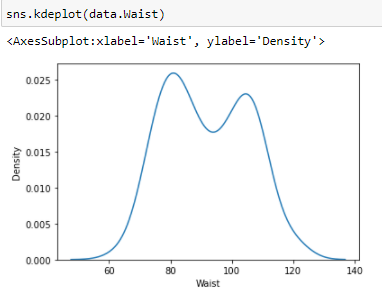
Normal distribution.

b)



From the above graph it is evident that Adipose Tissue (AT) follows almost

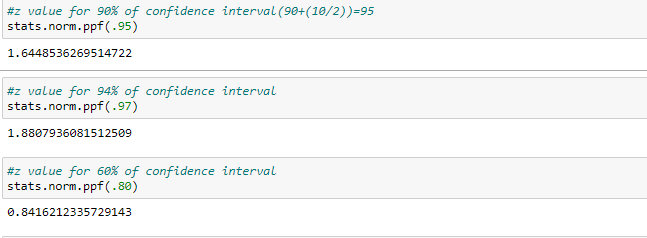
Normal distribution.



From the above distribution it is evident that Waist Circumference (WT) almost follows the normal distribution curve.

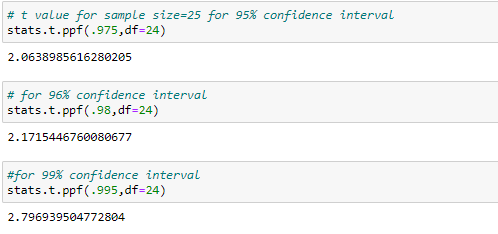
Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

Ans)



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

Ans:



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) µ=270, =260, SD=90, n=18, df=n-1=18-1= 17

tscore= = = -10/21.23= -0.47

