* 1
* 2

|  |  |
| --- | --- |
|  | Data Type |
| Number of beatings from Wife | Discreate data |
| Results of rolling a dice | Discreate 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 | Continuous data |
| Number of kids | Discreate data |
| Number of tickets in Indian railways | Discreate data |
| Number of times married | Discreate data |
| Gender (Male or Female) | Discreate data |

|  |  |
| --- | --- |
| 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 | Continuous data |
| Type of living accommodation | Ratio data |
| Level of Agreement | Ordinal data |
| IQ(Intelligence Scale) | Interval 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 | Ratio data |

* 3)

(HHH,HHT,HTH,THH,TTH,THT,HTT,TTT)=8

(HHT,HTH,TTH)=3

i.e, 3/8

* 4)

probability that sum is

1. Equal to 1 = 0 possiblity
2. Less than or equal to 4 = 6/36 or 0.16 possiblity
3. Sum is divisible by 2 and 3 = 6/36 or 0.16 possibility

* 5) the probability that none of the balls drawn is blue?

5/7\*4/6 = 20/42 or 10/21

* 6) the 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\*0.120

0.015+0.8+1.95+0.025+0.06+0.24 = 3.090

* 7)

df.mean()

Points 3.596563 Score 3.217250 Weigh 17.848750 dtype: float64

df.median()

Points 3.695 Score 3.325 Weigh 17.710 dtype: float64

df.mode()

| **Unnamed: 0** | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **0** | AMC Javelin | 3.07 | 3.44 | 17.02 |
| **1** | Cadillac Fleetwood | 3.92 | NaN | 18.90 |
| **2** | Camaro Z28 | NaN | NaN | NaN |
| **3** | Chrysler Imperial | NaN | NaN | NaN |
| **4** | Datsun 710 | NaN | NaN | NaN |
| **5** | Dodge Challenger | NaN | NaN | NaN |
| **6** | Duster 360 | NaN | NaN | NaN |
| **7** | Ferrari Dino | NaN | NaN | NaN |
| **8** | Fiat 128 | NaN | NaN | NaN |
| **9** | Fiat X1-9 | NaN | NaN | NaN |
| **10** | Ford Pantera L | NaN | NaN | NaN |
| **11** | Honda Civic | NaN | NaN | NaN |
| **12** | Hornet 4 Drive | NaN | NaN | NaN |
| **13** | Hornet Sportabout | NaN | NaN | NaN |
| **14** | Lincoln Continental | NaN | NaN | NaN |
| **15** | Lotus Europa | NaN | NaN | NaN |
| **16** | Maserati Bora | NaN | NaN | NaN |
| **17** | Mazda RX4 | NaN | NaN | NaN |
| **18** | Mazda RX4 Wag | NaN | NaN | NaN |
| **19** | Merc 230 | NaN | NaN | NaN |
| **20** | Merc 240D | NaN | NaN | NaN |
| **21** | Merc 280 | NaN | NaN | NaN |
| **22** | Merc 280C | NaN | NaN | NaN |
| **23** | Merc 450SE | NaN | NaN | NaN |
| **24** | Merc 450SL | NaN | NaN | NaN |
| **25** | Merc 450SLC | NaN | NaN | NaN |
| **26** | Pontiac Firebird | NaN | NaN | NaN |
| **27** | Porsche 914-2 | NaN | NaN | NaN |
| **28** | Toyota Corolla | NaN | NaN | NaN |
| **29** | Toyota Corona | NaN | NaN | NaN |
| **30** | Valiant | NaN | NaN | NaN |
| **31** | Volvo 142E | NaN | NaN | Na |

df.values

array([['Mazda RX4', 3.9, 2.62, 16.46], ['Mazda RX4 Wag', 3.9, 2.875, 17.02], ['Datsun 710', 3.85, 2.32, 18.61], ['Hornet 4 Drive', 3.08, 3.215, 19.44], ['Hornet Sportabout', 3.15, 3.44, 17.02], ['Valiant', 2.76, 3.46, 20.22], ['Duster 360', 3.21, 3.57, 15.84], ['Merc 240D', 3.69, 3.19, 20.0], ['Merc 230', 3.92, 3.15, 22.9], ['Merc 280', 3.92, 3.44, 18.3], ['Merc 280C', 3.92, 3.44, 18.9], ['Merc 450SE', 3.07, 4.07, 17.4], ['Merc 450SL', 3.07, 3.73, 17.6], ['Merc 450SLC', 3.07, 3.78, 18.0], ['Cadillac Fleetwood', 2.93, 5.25, 17.98], ['Lincoln Continental', 3.0, 5.424, 17.82], ['Chrysler Imperial', 3.23, 5.345, 17.42], ['Fiat 128', 4.08, 2.2, 19.47], ['Honda Civic', 4.93, 1.615, 18.52], ['Toyota Corolla', 4.22, 1.835, 19.9], ['Toyota Corona', 3.7, 2.465, 20.01], ['Dodge Challenger', 2.76, 3.52, 16.87], ['AMC Javelin', 3.15, 3.435, 17.3], ['Camaro Z28', 3.73, 3.84, 15.41], ['Pontiac Firebird', 3.08, 3.845, 17.05], ['Fiat X1-9', 4.08, 1.935, 18.9], ['Porsche 914-2', 4.43, 2.14, 16.7], ['Lotus Europa', 3.77, 1.513, 16.9], ['Ford Pantera L', 4.22, 3.17, 14.5], ['Ferrari Dino', 3.62, 2.77, 15.5], ['Maserati Bora', 3.54, 3.57, 14.6], ['Volvo 142E', 4.11, 2.78, 18.6]], dtype=object)

df.std()

Points 0.534679 Score 0.978457 Weigh 1.786943 dtype: float64

df.range()

points 2.17 score 3.224 weigh 7.4

* 8)

aList = [108,110,123,134,135,145,167,187,199]

sampled\_list = random.sample(aList, 1)

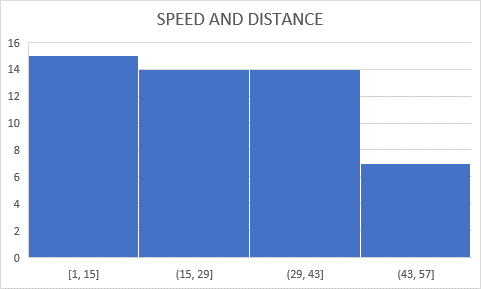
print(sampled\_list)

output=[123]

| ***x*** | ***P*(*x*)** | ***x*\**P*(*x*)** |
| --- | --- | --- |
| 1 | 123 | (1)(123) = 123 |

Expected value of weight is the patient is 123

9)



CALCULATION OF SKEWNESS

import scipy

from scipy.stats import skew

df

print(skew(df,axis=0,bias=true))

output= [0. -0.11395477 0.78248352]

CALCULATION OF KURTOSIS

From scipy.stats import kurtosis

df

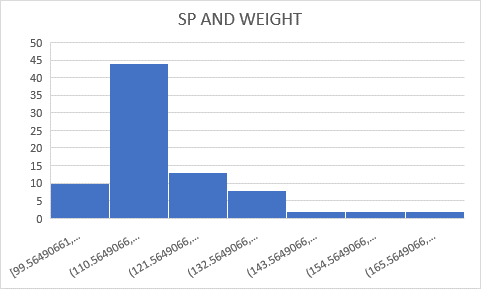
calculate the kurtosis

print(kurtosis(df, axis=o ,bias=true))

OUTPUT=[-1.2 2.97732894 0.95029149]

In conclusion, we can say that the symmetrical distribution of the skewness data is positive tail as the right tail is more in skewness.

And whereas, the kurtosis measurement of peak of a data has wider peak it’s a negative kurtosis.



CALCULATION OF SKEWNESS

import scipy

from scipy.stats import skew

df

print(skew(df,axis=0,bias=true))

output= [0.158145368 -0.60330993]

CALCULATION OF KURTOSIS

From scipy.stats import kurtosis

df

calculate the kurtosis

print(kurtosis(df, axis=o ,bias=true))

OUTPUT=[-1.20036585 2.72352149 0.81946588]

In conclusion, we can say that the symmetrical distribution of the skewness data is positive tail as the right tail is more.

And whereas, the kurtosis measurement of peak of a data has thin peak it’s a positive kurtosis.

* 10)



From the above histogram chart we can say that, a histogram is skewed to the right, if most of the data values are on the left side of the histogram and a histogram tail is skewed to right.

When the data are skewed to the right, that is from 250-400 people are of more weight with less frequency the mean value is more than the median of the data set as the median is the middle value and mode is always the highest value.

Along with the skewness, kurtosis is an important descriptive statistic distribution that the measurement of peak of the data is thin with positive kurtosis.



The boxplot represents, lower (minimum) extreme and upper (maximum) extreme the boxplot indicates middle of 50% of data lies with upper extreme and lower end of box is first quartile i,e Q1 and upper end is Q3 in between Q1 and Q3 the interquartile range. And thus the solid line indicates median. The t-shaped whiskers go to last point, which is 1.5 times interquartile range point further consideration as outliers.

**Interquartile Range (IQR):**The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1

**Outlier:**The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile.

(i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 . IQR).

* 11)

Given=

population of the Men = 30,00,000

Mean= 200 pounds

Random Sample= 2000 men

Standard deviation = 30 pounds

**For 94% confidence interval is 1.88**

* **Confidence Interval**= (200 – 1.88\* 30 / √2000) to (200 + 1.88 \* 30 / √2000)
* **Confidence Interval**=**135.42 to 132.90**

Therefore, the Confidence Interval at a 94% confidence level is 135.42 to 132.90.

**For 96% confidence interval is 2.054**

* **Confidence Interval**= (200 – 2.054\* 30 / √2000) to (200 + 2.054\* 30 / √2000)
* **Confidence Interval**=**135.72 to 132.60**

Therefore, the Confidence Interval at a 96% confidence level is 135.72 to 132.60

**For 98% confidence interval is 2.33**

* **Confidence Interval**= (200 – 2.33\* 30 / √2000) to (200 + 2.33 \* 30 / √2000)
* **Confidence Interval**=**135.5 to 132.78**

Therefore, the Confidence Interval at a 94% confidence level is 135.5 to 132.78.

We can conclude that the average balance of an adult male from the above data from the population of 30,00,000 men weight . And we can say that the chances of 94% 96% 98% is 135.42 135.72 135.5. So we can say that the average weight is of 135 of adult men in Mexico.

* 12)

import statistics

items=[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

statistics.mean(items)

output=41

statistics.median(items)

output=40.5

statistics.stdev(items)

output=5.0526638

from statistics import variance

sample1 = (34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)

print(variance(sample1))

output= 25.529411764705884

We can say that the student average marks scored is 41 and the median is of 40.5 and their standard deviation is 5.05 and variance is of 25.5294.

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

If the mean is equal to median as well as the mode, hence the skewness is zero. If the distribution is symmetric, the mean equals the median, and the skewness of the distribution is zero

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

Right skewed : If the mean is greater then the median, the distribution is postively skewed. It is positively skewed distribution.

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

Left skewed : the mean is less than the median. Then mean underestimates the most common values in a negatively skewed distribution.

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

Positive values of kurtosis indicates the distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more numbers are located in the tails of distribution instead of around the mean.

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

A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution.

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



What can we say about the distribution of the data?

The boxplot indicates the data in which the median is 15 , the density of the data is more at the right side. The outliers in the boxplot is none ,so the distribution of the data is considered as positively skewed.

What is nature of skewness of the data?

For the distribution that is positively skewed, the box plot showing the median closer to the lower quartile.

What will be the IQR of the data (approximately)?

Q3+(1.5 . IQR) 10+(1.5\*2) = 22%(approx)

Q1-(1.5 . IQR) 10-(1.5\*2) = 17%(approx)

19)



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

The line in the middle of the box plot for boxplot1 has median line as same with the line for boxplot2, which indicates that the both boxplot have same median point.

The box plot 2 is much longer than boxplot1, which indicates that the data is much more spread out which used in boxplot2. The line in the middle of the box plot for boxplot 1 is between Q1 & Q3, which indicates that the distribution for used in boxplot1 is normally distributed .

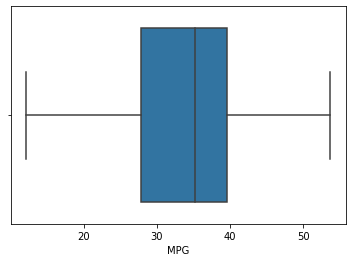
Conversely, the line in the middle of the boxplot2 is near the center of the box, which means the distribution of scores has little skew at all.

Neither of the box plot’s has tiny circles that extend beyond the top or bottom whiskers, which means neither dataset had any clear outliers.

20) Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

sns.boxplot(cars.MPG)



The above boxplot of the cars dataset is positively distributed as the median is at close to the lower quartile.

a.P(MPG>38) 1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())

0.34759392515827137

b.P(MPG<40) 1-stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())

0.2706501237848391

c. P (20<MPG<50) stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std())-stats.norm.cdf(0.20,cars.MPG.std())

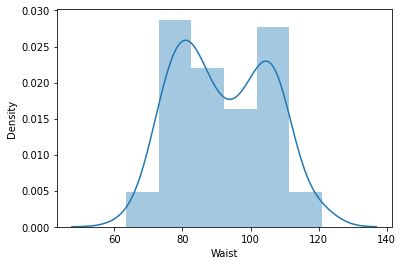
0.00010165648735741544

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

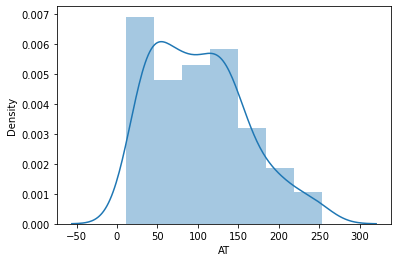
Dataset: wc-at.csv

sns.distplot(df["Waist"])

plt.show()



sns.distplot(df['AT'])

plt.show(

Q22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

For 90% confidence interval:

We have the significance level at 5 % ( as it is a two tailed test)

that is:

α = 5 % = 0.05

z at α = 0.05 from the z table will be:

z = 1.645.

For 94 % confidence interval, we get:

We have the significance level at 3 % ( as it is a two tailed test)

that is:

α = 3 % = 0.03

z at α = 0.03 from the z table will be:

z = 1.555.

For 60 % confidence interval, we get:

We have the significance level at 20 % ( as it is a two tailed test)

that is:

α =20 % = 0.2

z at α = 0.2 from the z table will be:

z = 0.253

**Therefore, we get that the z score at 90 % confidence interval is 1.645, at 94 % confidence interval is 1.555 and at 60 % confidence interval is 0.253.**

Q23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

Q24) 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

t=(260-270)/(90/18\*\*0.5)

t

-0.4714045207910317

p\_value=1-stats.t.cdf(abs(-0.4714),df=17)

p\_value

0.32167411684460556