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

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 | Ratio |

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

Ans:

No. of possible outcomes = 2^3=8

Possible outcomes = {HHH,HTH,THT,HHT,TTH,HTT,THH,TTT}

No. of 2 heads occurred in sample = 3

Total no. of sample = 8

P ( getting 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 obtained here : total no. of outcomes=36

{ (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 (sum =1) : 0 = no probabilty

in the above sample sum of two outcomes never becomes 1,sum starts from 2

1. p ( sum < = 4) :6/36 = 1/6 = 0.16666 =16.66% probability

from above sample we have sum of 2 outcomes less than or equals to 4 = 6 .

1. p (Sum is divisible by 2 and 3) : = 0.16666 =16.66% probability

sum of outcomes from above sample divisible by both 2 and 3 = 24 .

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:

2 balls taken except blue coloured balls , so no. of balls grabbed = 5

Balls garbbed = 2

P( taking none of the balls drawn is blue) = 2/5 = 0.4=40%

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: Expected no. of candies for a randomly selected child

E(X)=

= (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

= 3.090

= 3.09

So, expected no. of candies for a randomly selected child is 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: To** Calculate Mean, Median, Mode, Variance, Standard Deviation, Range we can use python scripts using pandas

#script

import pandas as pd

data=pd.read\_csv("Q7.csv")

data.describe()

data.mode()

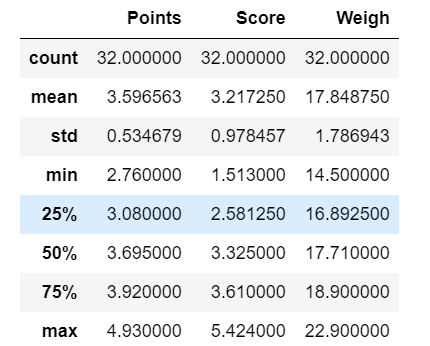
data.std()

data.var()

range = data[['Points','Score','Weigh']].max()-data[['Points','Score','Weigh']].min()

range

this is a screensnap of data.describe() --- which describes the dataset



Here, above data having different types of cars with its points, score and weigh. Average in the points of the cars is around 3.59, average score is 3.21 and the average weigh of the listed cars are around 17.84.

Here the median value of the points in the cars data is 3.69 ,median of the score of the data set is around 3.32 and the median of the weigh of the data set is 17.710.

Mode of the dataset for points is 3.92, mode of the score is 3.44 and mode of the weigh in the dataset is 17.02.

Standard deviation of points is around .534, for score is around .9784 and for weigh is around 1. 786.

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=X1.P1+X2.P2+X3.P3+…………+Xn.Pn

Here, X={108, 110, 123, 134, 135, 145, 167, 187, 199}

Number of total patients =9

P ( one of the patients is chosen at random) = 1/9

Expected value =(108\* 0.1111)+ (110\*0.1111)+( 123\*0.1111)+(134\*0.1111) +(135\*0.1111)+(145\*0.1111)+(167\*0.1111)+(187\*0.1111)+(199\*0.1111)

=145.33 .

So, Expected value =145.33 .

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data Cars speed and distance , Use Q9\_a.csv AND SP and Weight(WT) Use Q9\_b.csv**

**Ans:**

1. for calculating **Skewness, Kurtosis ,here using python scripts ,**

matplotlib and seaborn are used for data visualization and beautification of data.

#script

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

time = pd.read\_csv('Q9\_a.csv)

time\_new=time[['speed','dist']]

sns.distplot(time.speed) # by using distplot we can visualize data.

sns.distplot(time.dist) # by using distplot we can visualize data.time\_new.skew()

time\_new.kurt()

**inference:**

skewness of speed = -0.117510

skewness of dist = 0.806895

kurtosis of speed = -0.508994

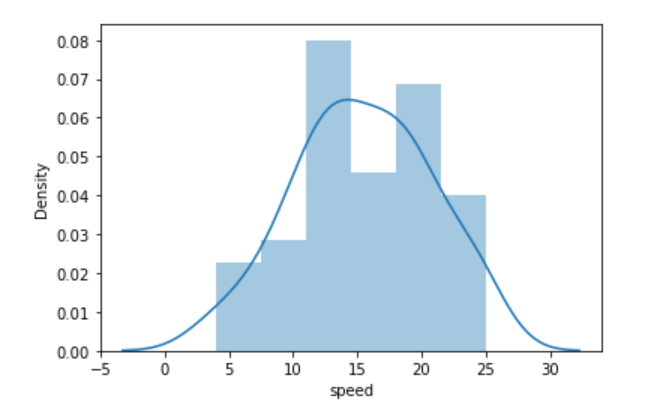
kurtosis of dist = 0.405053

skewness of speed is negative means it having negative skewness ,that tail on left side of distribution is longer.that is data is seen more in right side of distribution .

skewness is positive says its positive skewed .and tail at light side of distribution

kurtosis is negative says ,it is its having negatived kurtosis , that says the data has light tail and broad peak.

sns.distplot(time.speed)



b)

mass=pd.read\_csv('Q9\_b.csv')

mass\_new=mass[['SP','WT']]

sns.distplot(mass\_new.SP)

sns.distplot(mass\_new.WT)

mass\_new.skew()

mass\_new.kurt()

**inference:**

skewness of SP 1.611450

skewness of WT -0.614753

kurtosis of SP 2.977329

kurtosis of WT 0.950291

The skewness of the SP is positive it means that the tail on the right side of the distribution is longer. And about kurtosis of the SP, it is also positive means that the data has heavy tail and pointed peak.Wt is negatively skewed and kurtosis is positive says distribution is more toward right .

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



Ans: The histogram speaks about distribution that ,it is positively skewed. Most of the values are clustered around the left tail of the distribution while the right tail of the distribution is longer. So the most data arranged in range of 50 to 100 in chickWeight$weight (based on x- axis).





Ans:

Here the box plot speaks about the positively skewness of the distribution because the median is closer to the lower quartile and also there are some outliers in the upper extreme.

**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=3,000,000 sample size(n)=2,000

Sample mean=200 sample standard deviation(s)=30

For calculating the t value:

From scipy import stats

From math import sqrt

stats.t.ppf(CI,df)

t value for 94%= 1.8818614764780113

t value for 96%=2.055089962825778

t value for 98%= 2.328214776106972

For calculating the confidence interval:

Stats.norm.interval(x,mean,scale=())

Confidence interval of 94% is [198.738325292158, 201.261674707842]

Confidence interval of 98% is [198.62230334813333, 201.37769665186667]

Confidence interval of 96% is [198.43943840429978, 201.56056159570022]

**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:

Import pandas as pd

data=pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

data\_new=pd.DataFrame(data,columns=['daily'])

data\_new.mean()

data\_new.median()

data\_new.mode()

data\_new.std()

data\_new.var()

mean = 41.0

median = 40.5

mode = 41

standard deviation = 5.052664

variance = 25.529412

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

Ans: Skewness will be 0

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

Ans: positive skewness

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

Ans: negative skewness

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

Ans: a distribution is peaked and possess thick tails.

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

Ans: 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?

What is nature of skewness of the data?

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

Ans: Most of the observations are on the high end of the scale, so the distribution is [skewed left](https://stattrek.com/statistics/dictionary.aspx?definition=skewness). The [interquartile range](https://stattrek.com/statistics/dictionary.aspx?definition=Interquartile%20range) is indicated by the length of the box, which is 18 - 10. And the median is indicated by the vertical line running through the middle of the box, which is roughly centred over 15. So the median is about 15.

So it is positively skewed

Q19) Comment on the below Boxplot visualizations?



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

Ans:

Both are normally distributed

No ouliers

Mean is around 263 .

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)

c. P (20<MPG<50)

Ans:

import pandas as pd

From scipy import stats

data=pd.read\_csv(“Cars.csv”)

**To calculate mean and standard deviation**

data.describe()

**For P(MPG>38):**

1-stats.norm.cdf(38,loc=34.422076,scale=9.131445)= 0.34759394041453007

**For P(MPG<40):**

stats.norm.cdf(40,loc=34.422076,scale=9.131445)= 0.7293498604157946

**For P (20<MPG<50):**

stats.norm.cdf(50,loc=34.422076,scale=9.131445)- stats.norm.cdf(20,loc=34.422076,scale=9.131445)= 0.8988689076273199

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

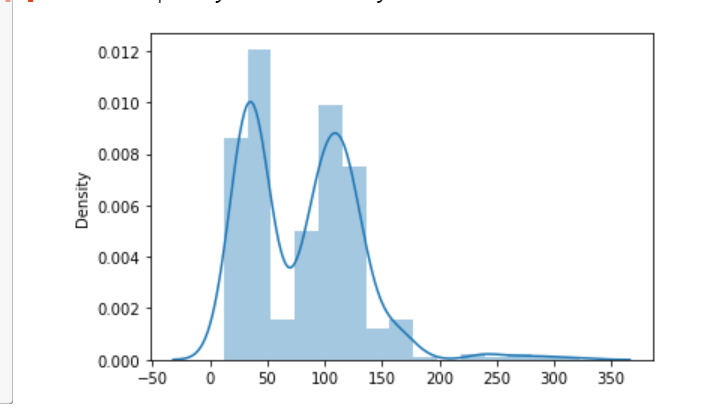
Dataset: Cars.csv

Ans: Import pandas as pd

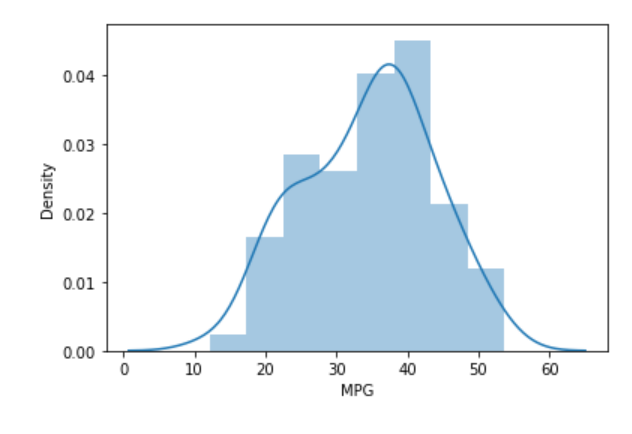
Import seaborn as sns

data=pd.read\_csv('cars.csv')

sns.distplot(data)



The above screenshot is about the whiole data ,here it is not normally distributed ,that it does not show a bell shaped curve

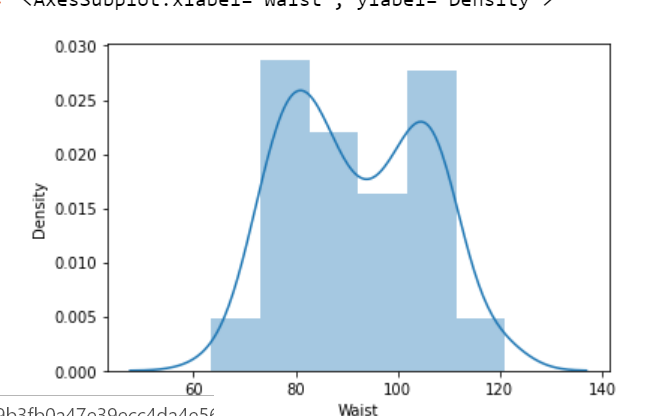
a) sns.distplot(data.MPG)

The above screenshot shows the distribution of the MPG , It show a normal distribution

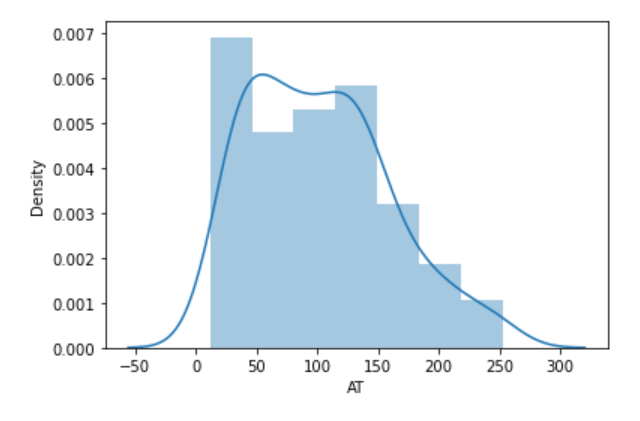
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Ans : sns.distplot(df.Waist)



sns.distplot(df.AT)



Here both Waist and AT shows normall distribution ,even it shows some variations

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

Ans: z-score of

90% - 1.64

94% - 1.88

60% - 0.84

by python scrpts we can easily calculate ,

For 90% confidence interval

from scipy import stats

stats.norm.ppf(.95)

1.6448536269514722

=> 1.64

For 94% confidence interval

from scipy import stats

stats.norm.ppf(.97)

1.8807936081512509

=> 1.88

For 60% confidence interval

from scipy import stats

stats.norm.ppf(.80)

0.8416212335729143

=>0.84

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

t score for 95% confidence interval:

from scipy import stats

stats.t.ppf(.975,df=24)

output: 2.0638985616280205

= 2.06

t score for 96% confidence interval:

from scipy import stats

stats.t.ppf(.98,df=24)

output: 2.1715446760080677

= 2.17

t score for 99% confidence interval:

from scipy import stats

stats.t.ppf(.995,df=24)

output: 2.796939504772804

=2.79

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:

population mean(µ)=270

Sample mean (x̅) =260

Sample size(n)=18

Sample standard deviation(s)=90

t score = (x̅- µ)/(s/sqrt(n))

=(260-270)/(90/sqrt(18))

=-.471

We need to impose this t score to below script

From scipy import stats

stats.t.ppf(-.471,17)

Probability of 18 randomly selected bulbs=.3498