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

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 | Interval |
| Celsius Temperature | Interval |
| Weight | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| 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:**

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

P(X)=3/8=0.375

the probability that two heads and one tail are obtained is 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. Equal to 1

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

1. Less than or equal to 4

P(X<=4) = 6/36 =0.1667

1. Sum is divisible by 2 and 3

P(X)=(2,3,4,5,6,7

3,4,5,6,7,8

4,5,6,7,8,9

5,6,7,8,9,10

6,7,8,9,10,11

7,8,9,10,11,12)

P(X) = 6/36 = 0.1667

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:**

Total number of balls in the bag = 7

Two balls drawn at random = 21

Two balls drawn at random is not blue = 10

P(X) = 10/21 = 0.476

The probability that none of the balls drawn is blue = 0.476.

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 number 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\*0.120

= 3.09

So, the expected number 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:**

import pandas as pd

df=pd.read\_csv("Q7.csv",index\_col=0)

df

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **Mazda RX4** | 3.90 | 2.620 | 16.46 |
| **Mazda RX4 Wag** | 3.90 | 2.875 | 17.02 |
| **Datsun 710** | 3.85 | 2.320 | 18.61 |
| **Hornet 4 Drive** | 3.08 | 3.215 | 19.44 |
| **Hornet Sportabout** | 3.15 | 3.440 | 17.02 |
| **Valiant** | 2.76 | 3.460 | 20.22 |
| **Duster 360** | 3.21 | 3.570 | 15.84 |
| **Merc 240D** | 3.69 | 3.190 | 20.00 |
| **Merc 230** | 3.92 | 3.150 | 22.90 |
| **Merc 280** | 3.92 | 3.440 | 18.30 |
| **Merc 280C** | 3.92 | 3.440 | 18.90 |
| **Merc 450SE** | 3.07 | 4.070 | 17.40 |
| **Merc 450SL** | 3.07 | 3.730 | 17.60 |
| **Merc 450SLC** | 3.07 | 3.780 | 18.00 |
| **Cadillac Fleetwood** | 2.93 | 5.250 | 17.98 |
| **Lincoln Continental** | 3.00 | 5.424 | 17.82 |
| **Chrysler Imperial** | 3.23 | 5.345 | 17.42 |
| **Fiat 128** | 4.08 | 2.200 | 19.47 |
| **Honda Civic** | 4.93 | 1.615 | 18.52 |
| **Toyota Corolla** | 4.22 | 1.835 | 19.90 |
| **Toyota Corona** | 3.70 | 2.465 | 20.01 |
| **Dodge Challenger** | 2.76 | 3.520 | 16.87 |
| **AMC Javelin** | 3.15 | 3.435 | 17.30 |
| **Camaro Z28** | 3.73 | 3.840 | 15.41 |
| **Pontiac Firebird** | 3.08 | 3.845 | 17.05 |
| **Fiat X1-9** | 4.08 | 1.935 | 18.90 |
| **Porsche 914-2** | 4.43 | 2.140 | 16.70 |
| **Lotus Europa** | 3.77 | 1.513 | 16.90 |
| **Ford Pantera L** | 4.22 | 3.170 | 14.50 |
| **Ferrari Dino** | 3.62 | 2.770 | 15.50 |
| **Maserati Bora** | 3.54 | 3.570 | 14.60 |
| **Volvo 142E** | 4.11 | 2.780 | 18.60 |

df.describe()

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **count** | 32.000000 | 32.000000 | 32.000000 |
| **mean** | 3.596563 | 3.217250 | 17.848750 |
| **std** | 0.534679 | 0.978457 | 1.786943 |
| **min** | 2.760000 | 1.513000 | 14.500000 |
| **25%** | 3.080000 | 2.581250 | 16.892500 |
| **50%** | 3.695000 | 3.325000 | 17.710000 |
| **75%** | 3.920000 | 3.610000 | 18.900000 |
| **max** | 4.930000 | 5.424000 | 22.900000 |

**Range = Max-Min**

Points = 4.93-2.76 = 2.17

Score = 5.424-1.513 = 3.911

Weigh = 22.90-14.50 = 8.4

df.mode()

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **0** | 3.07 | 3.44 | 17.02 |
| **1** | 3.92 | NaN | 18.90 |

df.var()

Points 0.285881

Score 0.957379

Weigh 3.193166

dtype: float64

**Conclusion:**

Here the Standard Deviation in the data is less. There is a huge gap of max and min value in weigh as compare to other data.

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 of the weight of the patient=

108+110+123+134+135+145+167+187+199/9 = 145.33

So, the expected value of the patient chosen at random is 145.33.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANS:**

import pandas as pd

df=pd.read\_csv("Q9\_a.csv",index\_col=0)

df

|  |  |  |
| --- | --- | --- |
| **Index** | **speed** | **dist** |
| **1** | 4 | 2 |
| **2** | 4 | 10 |
| **3** | 7 | 4 |
| **4** | 7 | 22 |
| **5** | 8 | 16 |
| **6** | 9 | 10 |
| **7** | 10 | 18 |
| **8** | 10 | 26 |
| **9** | 10 | 34 |
| **10** | 11 | 17 |
| **11** | 11 | 28 |
| **12** | 12 | 14 |
| **13** | 12 | 20 |
| **14** | 12 | 24 |
| **15** | 12 | 28 |
| **16** | 13 | 26 |
| **17** | 13 | 34 |
| **18** | 13 | 34 |
| **19** | 13 | 46 |
| **20** | 14 | 26 |
| **21** | 14 | 36 |
| **22** | 14 | 60 |
| **23** | 14 | 80 |
| **24** | 15 | 20 |
| **25** | 15 | 26 |
| **26** | 15 | 54 |
| **27** | 16 | 32 |
| **28** | 16 | 40 |
| **29** | 17 | 32 |
| **30** | 17 | 40 |
| **31** | 17 | 50 |
| **32** | 18 | 42 |
| **33** | 18 | 56 |
| **34** | 18 | 76 |
| **35** | 18 | 84 |
| **36** | 19 | 36 |
| **37** | 19 | 46 |
| **38** | 19 | 68 |
| **39** | 20 | 32 |
| **40** | 20 | 48 |
| **41** | 20 | 52 |
| **42** | 20 | 56 |
| **43** | 20 | 64 |
| **44** | 22 | 66 |
| **45** | 23 | 54 |
| **46** | 24 | 70 |
| **47** | 24 | 92 |
| **48** | 24 | 93 |
| **49** | 24 | 120 |
| **50** | 25 | 85 |

df.skew()

Index 0.000000

speed -0.117510

dist 0.806895

dtype: float64

df.kurt()

Index -1.200000

speed -0.508994

dist 0.405053

dtype: float64

**Conclusion:**

Here the skewness of speed is negative hence, the tail is flatter on the left side and the distance is positively skewed hence, it is flatter on the right side.

Here the speed, distance less than 3 the curve is platykurtic.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANS:**

import pandas as pd

df1=pd.read\_csv("Q9\_b.csv",index\_col=0)

df1

|  | **SP** | **WT** |
| --- | --- | --- |
| **1** | 104.185353 | 28.762059 |
| **2** | 105.461264 | 30.466833 |
| **3** | 105.461264 | 30.193597 |
| **4** | 113.461264 | 30.632114 |
| **5** | 104.461264 | 29.889149 |
| **...** | ... | ... |
| **77** | 169.598513 | 16.132947 |
| **78** | 150.576579 | 37.923113 |
| **79** | 151.598513 | 15.769625 |
| **80** | 167.944460 | 39.423099 |
| **81** | 139.840817 | 34.948615 |

81 rows × 2 columns

df1.skew()

SP 1.611450

WT -0.614753

dtype: float64

df1.kurt()

SP 2.977329

WT 0.950291

dtype: float64

**Conclusion:**

Here the skewness of SP is positive hence, the tail is more flat on the right side and the WT is negatively skewed hence, it is more flat on the left side.

Here the SP,WT less than 3 the curve is platykurtic .

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



**ANS:**

**HISTOGRAM:**

Here the histogram shows that the data is right skewed that is it is positively skewed.

**BOXPOT:**

Here the boxplot shows that the data has outliers.

**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:**

N = 3000000, n=2000, s = 30

Confidence Interval = = 200+Z

**For 94% confidence interval**

From Z table value for 94% =1.88

200+1.88\*30/sqrt(2000)=201.261

200-1.88\*30/sqrt(2000)=198.739

**CI (198.739,201.261)**

**For 98% confidence interval**

From Z table value for 98% =2.32

200+2.32\*30/sqrt(2000)=201.556

200-2.32\*30/sqrt(2000)=198.444

**CI (198.444,201.556)**

**For 96% confidence interval**

From Z table value for 96% =2.05

200+2.05\*30/sqrt(2000)=201.375

200-2.05\*30/sqrt(2000)=198.625

**CI (198.625,201.375)**

**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:**

1)

import pandas as pd

df2=pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])

df2.describe()

**count 18.000000**

**mean 41.000000**

**std 5.052664**

**min 34.000000**

**25% 38.250000**

**50% 40.500000**

**75% 41.750000**

**max 56.000000**

**dtype: float64**

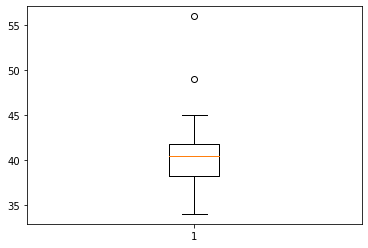
df2.var()

**25.529411764705884**

2)

plt.boxplot(df2)

plt.show()



By plotting a box plot we can see that there is 2 outliers present in the data.

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

**ANS:**

When mean = median there is no skewness.

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

**ANS:**

When mean > median the distribution is positively skewed.

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

**ANS:**

When median > mean the distribution is negatively skewed.

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

**ANS:**

Positive kurtosis value indicates peakness for a data.

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

**ANS:**

Negative kurtosis value indicates less peakness for the data.

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



What can we say about the distribution of the data?

**ANS:**

The data is not normally distributed.

What is nature of skewness of the data?

**ANS:**

The data is left skewed.

What will be the IQR of the data (approximately)?   
**ANS:**

IQR=Q3-Q1

IQR=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:**

Here in both the boxplot we can see that there is no outliers present.

The median of both the box plot is same.

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

from scipy.stats import norm

df=pd.read\_csv("C:\\Users\\HP\\Desktop\\Cars.csv")

1. P(MPG>38)

1-stats.norm.cdf(38,df.MPG.mean(),df.MPG.std())

0.3475939251582705

1. P(MPG<40)

stats.norm.cdf(40,df.MPG.mean(),df.MPG.std())

0.7293498762151616

c) P (20<MPG<50)

stats.norm.cdf(0.50,df.MPG.mean(),df.MPG.std())-stats.norm.cdf(0.20,df.MPG.mean(),df.MPG.std())

1.2430968797327613e-05

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**ANS:**

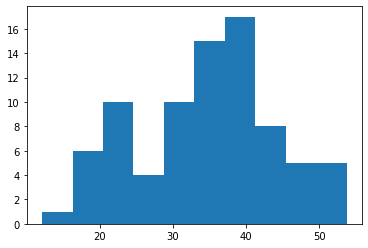
import pandas as pd

import matplotlib.pyplot as plt

df=pd.read\_csv("Cars.csv")

plt.hist(df['MPG'])

plt.show()



**Conclusion:**

From the above histogram we can see that the data of MPG is not normally distributed.

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

Dataset: wc-at.csv

**ANS:**

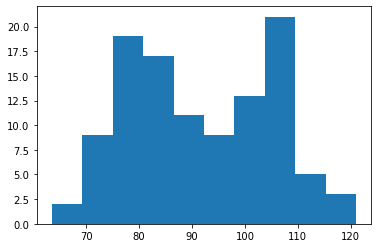
import pandas as pd

import matplotlib.pyplot as plt

df=pd.read\_csv("wc-at.csv")

plt.hist(df['Waist'])

plt.show()

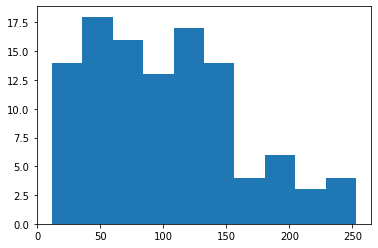


**Conclusion:**

From the above histogram we can see that the data of Waist is not normally distributed.

plt.hist(df['AT'])

plt.show()



**Conclusion:**

From the above histogram we can see that the data of AT is not normally distributed.

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

**ANS:**

**90% confidence interval:**

df\_ci=stats.norm.interval(0.90,

loc=df.mean(),

scale=df.std())

print( 'Mean at 90% confidence interval is:', np.round(df\_ci, 4))

Mean at 90% confidence interval is:

Waist - [[ 69.5991 7.6525]

AT - [114.2046 196.1355]]

**94% confidence interval:**

df\_ci=stats.norm.interval(0.94,

loc=df.mean(),

scale=df.std())

print( 'Mean at 94% confidence interval is:', np.round(df\_ci, 4))

Mean at 94% confidence interval is:

Waist - [[ 66.3999 -5.8656]

AT - [117.4037 209.6537]]

**60% confidence interval:**

df\_ci=stats.norm.interval(0.60,

loc=df.mean(),

scale=df.std())

print( 'Mean at 60% confidence interval is:', np.round(df\_ci, 4))

Mean at 60% confidence interval is:

Waist - [[ 80.4902 53.6735]

AT - [103.3135 150.1145]]

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

ANS:

**95% confidence interval**

stats.t.cdf(0.975,24)

0.8303570471638759

**96% confidence interval**

stats.t.cdf(0.98,24)

0.8315688116127068

**99% confidence interval**

stats.t.cdf(0.995,24)

0.8351685156761681

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:**

H0: Average life of is 260 days.

H1: Average life is more than 260 days.

Here μ = 270, n = 18, = 260, s = 90

from scipy import stats

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

t

-0.4714045207910317

p\_value=stats.t.cdf(t,17)

p\_value

0.32167253567098364

**Conclusion:**

Here p-value > 0.05 so we do not reject Ho and conclude that the average life of the bulb is 260 days.