|  |  |
| --- | --- |
| 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 | 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) | 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,HTH,THH,TTH,THT,HTT,TTT)

Interested Events =(HHT,HTH,THH)

Probability(P) = I.E/Total no. of Events = 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 : Two Dice are Rolled

Sample Space = {11,12,13,14,15,16,

21,22,23,24,25,26,

31,32,33,34,35,36,

41,42,43,44,45,46,

51,52,53,54,55,56,

61,62,63,64,65,66}

Total no. of events = 36

1. P(a) = 0
2. I.E = {11,12,13,21,22,31}

P(b) = 6/36 = 1/6

1. I.E = {15,24,33,42,51,66}

P(c) = 6/36 = 1/6

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 Red, 3 Green, 2 Blue = 2+3+2 = 7 = no of balls (n)

Let Sample space for drawing 2 balls out of 7 = n(S)

= 7C2 = 7!/2!\*5! = (7\*6)/(2\*1) = 21

Let event of drawing 2 balls from 5 balls, none of the balls are blue = n(E)

= 5C2 = 5!/2!\*3! = (5\*4)/(2\*1) = 10

P(E) = n(E)/n(S) = 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 : Expected number of candies for a randomly selected child = **∑** x\*P(x)

= 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120 = 3.090

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

import numpy as np

Q7 = pd.read\_csv("C:/Users/HARSHA/Desktop/Data Science/Data Science/Data Science Assignments/Basic statistics level 1/Q7.csv")

Q7.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 |
| **(Range)min** | 2.760000 | 1.513000 | 14.500000 |
| **25%** | 3.080000 | 2.581250 | 16.892500 |
| **(Median) 50%** | 3.695000 | 3.325000 | 17.710000 |
| **75%** | 3.920000 | 3.610000 | 18.900000 |
| **(Range)max** | 4.930000 | 5.424000 | 22.900000 |

Q7.var() (Variance)

Points 0.285881

Score 0.957379

Weigh 3.19316

Inferences :

Mode is a way of expressing, in a single number, class

Ex: Numeric.

To get mode, first we have to write a method Q7.mode(Name).

Q7.mode()

Unnamed: 0 Points Score Weigh

0 AMC Javelin 3.07 3.44 17.02

1 Cadillac Fleetwood 3.92 NaN 18.90

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  =  ∑ ( probability  \* Value )

 ∑ P(x).E(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

= 1308/9

= 145.33

**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 : Q9\_a.csv**

import pandas as pd

import numpy as np

Q9\_a= pd.read\_csv("C:/Users/HARSHA/Desktop/Data Science/Data Science/Data Science Assignments/Basic statistics level 1/Q9\_a.csv")

import matplotlib.pyplot as plt

%matplotlib inline

Q9\_a.skew()

Index 0.000000

speed -0.117510 (Left Skewness,‘-ve’)

dist 0.806895 (Right Skewness,’+ve’)

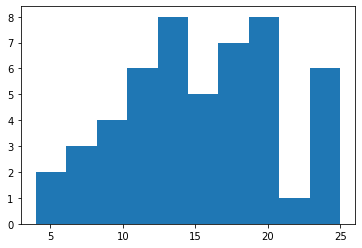
Q9\_a.kurt()

Index -1.200000

speed -0.508994(Negative Krutosis)

dist 0.405053(Positive Krutosis)

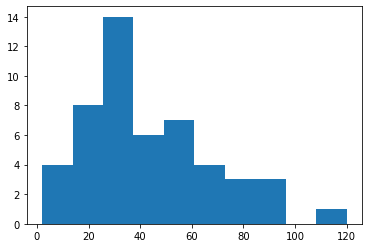
plt.hist(Q9\_a.speed)



Inference :

1. Skewness is negative.
2. Distribution is skewed towards left.
3. Mean of distribution is less than the Median.
4. Distribution has broad peak and thin tails.

plt.hist(Q9\_a.dist)



Inference :

1. Skewness is positive.
2. Distribution is skewed towards right.
3. Mean of distribution is more than the Median.
4. Distribution has sharp peak and wide tails.

**Q9\_b.csv**

import pandas as pd

import numpy as np

Q9\_b= pd.read\_csv("C:/Users/HARSHA/Desktop/Data Science/Data Science/Data Science Assignments/Basic statistics level 1/Q9\_b.csv")

import matplotlib.pyplot as plt

%matplotlib inline

Q9\_b.skew()

Unnamed: 0 0.000000

SP 1.611450(‘+ve’,right skew)

WT -0.614753(‘-ve‘,left skew)

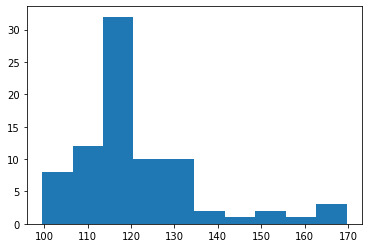
Q9\_b.kurt()

Unnamed: 0 -1.200000

SP 2.977329(+ve kurt)

WT 0.950291(+ve kurt)

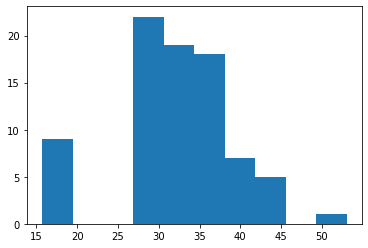
plt.hist(Q9\_b.SP)



**Inference:**

1. Skewness is positive.
2. Distribution is skewed towards right.
3. Mean of distribution is more than the Median.
4. Distribution has sharp peak and wide tails.

**plt.hist(Q9\_b.WT)**

****

**Inference:**

1. Skewness is negative.
2. Distribution is skewed towards left.
3. Mean of distribution is less than the Median.
4. Distribution has sharp peak and wide tails.

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



**inference :** The distribution is right skew(‘+ve’), Mean >Median



**Inference** : The distribution has lots of outliers towards 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:

Select Sample space from total population .

Sample space size n=2000

Standard deviation =30

Mean = 200

**Confidence interval: *Z***

**For 94%:**

=200Za/2\*30/sqrt of 2000

[Za/2= 0.94/2(check z table)=1.88]

=2001.88\*30/sqrt of 2000

=198.74 – 201.26

**For 98%**

=2002.33\*30/sqrt (2000)

=198.44-201.56

**For 96%**

=200 ± 2.05\*30/sqrt (2000)

=198.62-201.38

**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)

Mean = 41

Median = 40.5

Variance = 25.52941

Standard Deviation = 5.0526

2)Mean > Median.

The distribution is slightly skewed towards right.

No outliers are present.

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

Ans : skewness=0, Symmetric

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

Ans : Right skewed(tail on the right side)

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

Ans : Left skewed(tail on the left side)

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

Ans: Sharp Peak, Thick Tails

Q17) What does negative kurtosis value indicates for a data?  
Ans: Broad Peak, Wide Tails

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



What can we say about the distribution of the data?

Ans: It is Not a Normal Distribution

What is nature of skewness of the data?  
Ans: Left Skewed

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

Ans: Inter Quartile Range=Upper Quartile-Lower Quartile=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: 1) The median of the two boxplots are same 260(approx).

2) The boxplots are not skewed in ‘+ve’ or ‘–ve’ direction.

3) Outliers doesn’t exist in both of the boxplots.

4) Both are Normally Distributed.

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)

> mean(Cars$MPG)

[1] 34.42208

> sd(Cars$MPG)

[1] 9.131445

> pnorm(38,34.42,9.13)

[1] 0.652513

P(MPG>38)=1-P(MPG<38)

>1 - 0.65

[1] 0.35

* + 1. P(MPG<40):

pnorm(40,34.42,9.13)

[1] 0.7294571

c) P (20<MPG<50):

> pnorm(50,34.42,9.13)-pnorm(20,34.42,9.13)

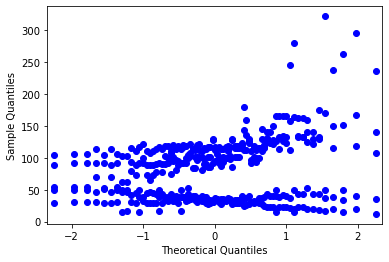
[1] 0.8989178

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:



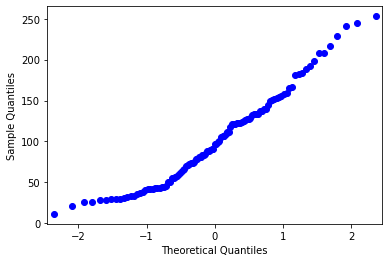
**Normally Distributed**

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

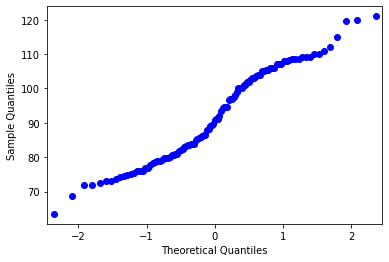
Dataset: wc-at.csv

Ans:

**Normally Distributed(AT)**

****

**Normally Distributed(Waist)**



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

Ans :

90% 🡪> qnorm(0.95)

[1] 1.644854

94% 🡪> qnorm(0.97)

[1] 1.880794

60% 🡪> qnorm(0.8)

[1] 0.8416212

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

Ans:

95% 🡪> qt(0.975,24)

[1] 2.063899

96% 🡪> qt(0.98,24)

[1] 2.171545

99% 🡪 qt(0.995,24)

[1] 2.79694

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: Average=270, x bar= 260, std=90, n=18,

degrees of freedom= n-1=18-1=17

t score = (x bar - µ)/(s/sqrt(n))

= (260 – 270)/(90/sqrt(18))

= -10/21.23

= -0.47

> pt(-0.47,17)

[1] 0.3221639

Req Prob= 0.32= 32%