|  |  |
| --- | --- |
| 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 | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Ratio |
| 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?

Possible Outcomes : {HHH,HHT,HTT,THT,TTH,HTH,THH,TTT} = 8

Favourable chances : {HHT,HTH,THH} = 3

Probability = 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

Possible outcomes = 36

1. Favourable chances = 0

Probability = 0/36 = 0

1. Favourable outcomes : {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)} = 6

Probability = 6/36 = 1/6

1. Favourable outcomes : {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)} = 6

Probability = 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?

Total possible outcomes = 7C2 = 21

Favourable outcomes = 5C2 = 10

Required Probability = 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

Expected number of candies =

(1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120)

= 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.240

= 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**

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **count** | 32.000000 | 32.000000 | 32.000000 |
| **mean** | 3.596563 | 3.217250 | 17.848750 |
| **(Standard-Deviation)std** | 0.534679 | 0.978457 | 1.786943 |
| **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 |
| **max** | 4.930000 | 5.424000 | 22.900000 |

**Variance: Range = Max – Min Mode:**

**Points** 0.285881 **Points** 2.170000 **Points** 3.07

**Score** 0.957379 **Score** 3.911000 **Score** 3.44

**Weigh** 3.193166 **Weigh** 8.400000 **Weigh** 17.02

**Inferences:**

Using the describe() function in pandas, got the **Mean, Median and Standard Deviation** for the Q7 data.

**Variance** = (Standarad Deviation)2

**Range** = Max – Min

Mode() function in pandas gives the **Mode**. [Q7.Mode()]

\*\*The values of Mean, Median and Mode for each of the columns are almost similar and this seems to have a normal distribution(Symmetric Curve)

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?

Expected weight = (108+110+123+134+135+145+167+187+199)/9

= 145.34

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

**Cars speed and distance**

**Use Q9\_a.csv**

import pandas as pd

from matplotlib import pyplot as plt

%matplotlib inline

Q9\_a = pd.read\_csv('C:/Users/Ravi Kiran/Basic Statistics Level-1/Q9\_a.csv')

Q9\_a.skew()

Index 0.000000

speed -0.117510

dist 0.806895

dtype: float64

Q9\_a.kurt()

Index -1.200000

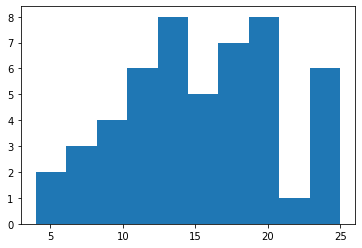
speed -0.508994

dist 0.405053

dtype: float64

plt.hist(Q9\_a['speed'])

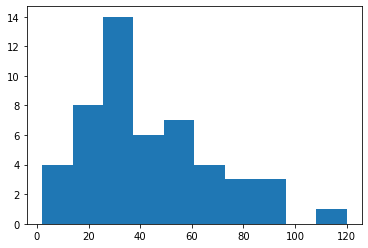
plt.show()



plt.hist(Q9\_a['dist'])

plt.show()

​



**Inferences**

***For Speed***

Left - skewed

Mean < Median

***For Dist***

Right - skewed

Mean > Median

**SP and Weight(WT)**

**Use Q9\_b.csv**

import pandas as pd

from matplotlib import pyplot as plt

%matplotlib inline

Q9\_b = pd.read\_csv('C:/Users/Ravi Kiran/Basic Statistics Level-1/Q9\_b.csv’)

Q9\_b.skew()

Unnamed: 0 0.000000

SP 1.611450

WT -0.614753

dtype: float64

Q9\_b.kurt()

Out[24]:

Unnamed: 0 -1.200000

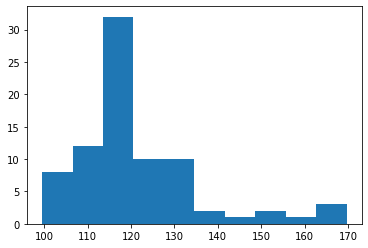
SP 2.977329

WT 0.950291

dtype: float64

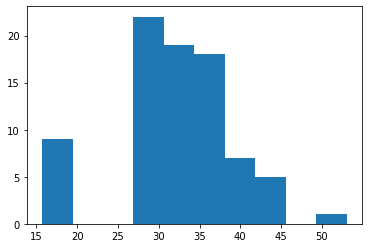
plt.hist(Q9\_b['SP'])

plt.show()

****

plt.hist(Q9\_b['WT'])

plt.show()

****

**Inferences**

***For SP***

Right - skewed

Mean > Median

***For WT***

Left - skewed

Mean < Median

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

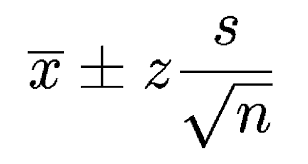


**Inference:** The Distribution is Right – skewed i.e., Mean of the distribution is greater than the Median.



**Inference:** The distribution has more number of outliers at 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?

**Confidence interval: **

Given, mean = 200

Standard Deviation (sd) = 30

Sample Size (n) = 2000

Confidence Interval at 94% = 200 ± (1.88) 30/√2000

= 198.74 – 201.26

Confidence Interval at 98% = 200 ± (2.33) 30/√2000

= 198.44 – 201.56

Confidence Interval at 96% = 200 ± (2.04) 30/√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?
3. import numpy as np

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

np.mean(marks)

41.0

np.median(marks)

40.5

np.var(marks)

24.11111111111111

np.std(marks)

4.910306620885412

2)

Mean(41.0) and Median(40.5) are very close to each other and the distribution is almost normal but slightly right-skewed as mean>median.

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

Skewness is zero, Symmetric curve exists

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

It is Right-tailed

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

It is left-tailed

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

Sharp Peak

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

Broad peak

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



What can we say about the distribution of the data?

Not a Normal Distribution

What is nature of skewness of the data?

Left-skewed

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

IQR = 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.

**Inferences:**

1. No outliers
2. Median of both the plots is 260(approximately)
3. 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)

c. P (20<MPG<50)

import pandas as pd

import numpy as np

from scipy import stats

cars = pd.read\_csv('C:/Users/Ravi Kiran/Basic Statistics Level-1/Cars.csv')

cars

| **HP** | **MPG** | **VOL** | **SP** | **WT** |
| --- | --- | --- | --- | --- |
| **0** | 49 | 53.700681 | 89 | 104.185353 | 28.762059 |
| **1** | 55 | 50.013401 | 92 | 105.461264 | 30.466833 |
| **2** | 55 | 50.013401 | 92 | 105.461264 | 30.193597 |
| **3** | 70 | 45.696322 | 92 | 113.461264 | 30.632114 |
| **4** | 53 | 50.504232 | 92 | 104.461264 | 29.889149 |
| **...** | ... | ... | ... | ... | ... |
| **76** | 322 | 36.900000 | 50 | 169.598513 | 16.132947 |
| **77** | 238 | 19.197888 | 115 | 150.576579 | 37.923113 |
| **78** | 263 | 34.000000 | 50 | 151.598513 | 15.769625 |
| **79** | 295 | 19.833733 | 119 | 167.944460 | 39.423099 |
| **80** | 236 | 12.101263 | 107 | 139.840817 | 34.948615 |

81 rows × 5 columns

cars['MPG'].mean()

34.422075728024666

cars['MPG'].std()

9.131444731795982

**a)P(X>38)**

P **=** stats.norm.cdf(x **=** 38, loc **=** 34.422075728024666 ,scale **=** 9.131444731795982) *#p(x<38)*

P

0.6524060748417295

print('Required Probability : '**, 1-P**)*#p(x>38)*

Requird Probability : 0.3475939251582705

​

**b)P(X<40)**

P = stats.norm.cdf(x = 40, loc = 34.422075728024666 ,scale = 9.131444731795982) #p(x<40)

print('Requird Probability : ', P)#p(x<40)

Requird Probability : 0.7293498762151616

**c)P(20<X<50)**

P = (1-(stats.norm.cdf(x = 20, loc = 34.422075728024666 ,scale = 9.131444731795982))) - stats.norm.cdf(x = 50, loc = 34.422075728024666 ,scale = 9.131444731795982)

print('Requird Probability : ', P)#p(20<x<50)

Requird Probability : -0.013116469610523374

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

cars = pd.read\_csv('C:/Users/Ravi Kiran/Basic Statistics Level-1/Cars.csv')

cars

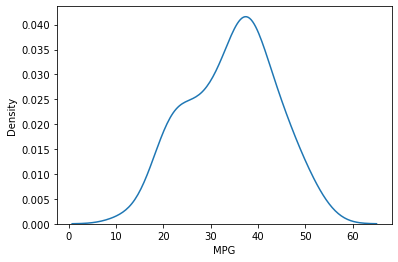
| **HP** | **MPG** | **VOL** | **SP** | **WT** |
| --- | --- | --- | --- | --- |
| **0** | 49 | 53.700681 | 89 | 104.185353 | 28.762059 |
| **1** | 55 | 50.013401 | 92 | 105.461264 | 30.466833 |
| **2** | 55 | 50.013401 | 92 | 105.461264 | 30.193597 |
| **3** | 70 | 45.696322 | 92 | 113.461264 | 30.632114 |
| **4** | 53 | 50.504232 | 92 | 104.461264 | 29.889149 |
| **...** | ... | ... | ... | ... | ... |
| **76** | 322 | 36.900000 | 50 | 169.598513 | 16.132947 |
| **77** | 238 | 19.197888 | 115 | 150.576579 | 37.923113 |
| **78** | 263 | 34.000000 | 50 | 151.598513 | 15.769625 |
| **79** | 295 | 19.833733 | 119 | 167.944460 | 39.423099 |
| **80** | 236 | 12.101263 | 107 | 139.840817 | 34.948615 |

81 rows × 5 columns

In [12]:

sns.distplot(a**=**cars['MPG'], hist**=False**)

plt.show()



cars['MPG'].skew()

-0.17794674747025727

cars['MPG'].kurtosis()

-0.6116786559430913

From the plot, and with the value of kurtosis<3, we can say that the MPG of cars follows Normal Distribution

cars['MPG'].mean()

34.422075728024666

cars['MPG'].median()

35.15272697

cars['MPG'].mode()

0 29.629936

dtype: float64

With the mean, median and mode almost near to each other, the distribution is normal.

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

Dataset: wc-at.csv

wc\_at = pd.read\_csv('C:/Users/Ravi Kiran/Basic Statistics Level-1/wc-at.csv')

wc\_at

|  | **Waist** | **AT** |
| --- | --- | --- |
| **0** | 74.75 | 25.72 |
| **1** | 72.60 | 25.89 |
| **2** | 81.80 | 42.60 |
| **3** | 83.95 | 42.80 |
| **4** | 74.65 | 29.84 |
| **...** | ... | ... |
| **104** | 100.10 | 124.00 |
| **105** | 93.30 | 62.20 |
| **106** | 101.80 | 133.00 |
| **107** | 107.90 | 208.00 |
| **108** | 108.50 | 208.00 |

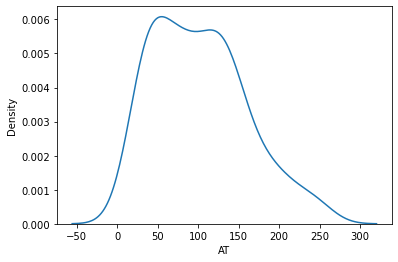
109 rows × 2 columns

109 rows × 2 columns

In [25]:

sns.distplot(a**=**wc\_at['AT'], hist**=False**)

plt.show()



wc\_at['AT'].skew()

0.584869324127853

wc\_at['AT'].kurtosis()

-0.28557567504584425

From the plot, and with the value of kurtosis<3, we can say that the AT of wc-at follows Normal Distribution

wc\_at['AT'].mean()

101.89403669724771

wc\_at['AT'].median()

96.54

wc\_at['AT'].mode()

0 121.0

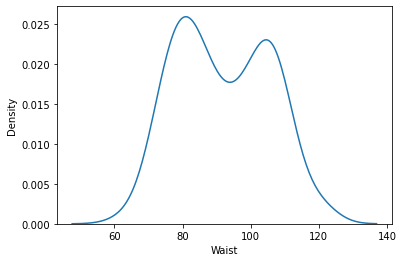
1 123.0

dtype: float64

With the mean, median and mode almost near to each other, the distribution is normal.

sns.distplot(a**=**wc\_at['Waist'], hist**=False**)

plt.show()



wc\_at['Waist'].skew()

0.1340560824786468

wc\_at['Waist'].kurtosis()

-1.1026666011768886

From the plot, and with the value of kurtosis<3, we can say that the Waist of wc-at follows Normal Distribution

wc\_at['Waist'].mean()

91.90183486238533

wc\_at['Waist'].median()

90.8

wc\_at['Waist'].mode()

0 94.5

1 106.0

2 108.5

dtype: float64

With the mean, median and mode almost near to each other, the distribution is normal.

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

**from** scipy **import** stats

print('Z scores at 90% confidence interval is',stats**.**norm**.**ppf(0.95))

Z scores at 90% confidence interval is 1.6448536269514722

print('Z scores at 94% confidence interval is',stats**.**norm**.**ppf(0.97))

Z scores at 94% confidence interval is 1.8807936081512509

print('Z scores at 60% confidence interval is',stats**.**norm**.**ppf(0.80))

Z scores at 60% confidence interval is 0.8416212335729143

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

**from** scipy **import** stats

print(' t scores at 95% confidence interval is', np**.**round(stats**.**t**.**ppf(0.975, df **=** 24), 2))

t scores at 95% confidence interval is 2.06

print(' t scores at 96% confidence interval is', np**.**round(stats**.**t**.**ppf(0.98, df **=** 24), 2))

t scores at 96% confidence interval is 2.17

print(' t scores at 99% confidence interval is', np**.**round(stats**.**t**.**ppf(0.995, df **=** 24), 2))

t scores at 99% confidence interval is 2.8

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

**Worked using Python**

Mu = 270

x̄ = 260

n = 18

s = 90

**Goal :** We have to find the p-value here

t = (x̄ - Mu)/(s/sqrt(n))

from scipy import stats

import numpy as np

t\_value =np.round( (260 - 270)/(90/np.sqrt(18)),2)

t\_value

-0.47

In [ ]:

p\_value = np.round(stats.t.cdf(-0.47, 18),3)

p\_value

0.322

In [ ]: