|  |  |
| --- | --- |
| 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 | Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Nominal |

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 | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| 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?

Soln:

Sample space,S=2^3=8

No. of events,E={hth,thh,hht}=3

Required 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

Soln:

Total no. of outcomes=6^2=36

(a). Probability=0

(b). Possible events=(2,2),(3,1),(1,3),(1,1),(1,2),(2,1)=6

Probability=6/36=1/6

(c). Possible events=(3,3),(2,4),(4,2),(6,6),(1,5),(5,1)=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?

Soln:

Total number of events= ===21

Interested events===10

Probability that none of the balls is blue =10/21= 0.47

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

Soln:

Expected number = E(x) \*P(x)=1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120= **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.** df\_7.describe()

index, Points, Score, Weigh

count, 32.0,32.0,32.0

mean, 3.5965625,3.2172500000000004,17.848750000000003

std, 0.5346787360709715,0.9784574429896966,1.7869432360968431

min, 2.76,1.513,14.5

25%, 3.08,2.58125,16.8925

50%, 3.6950000000000003,3.325,17.71

75%, 3.92,3.61,18.9

max, 4.93,5.424,22.9

df\_7.mode()

index, Points, Score, Weigh

0,3.07,3.44,17.02

1,3.92,NaN,18.9

df\_7['Points'].max()-df\_7['Points'].min()= 2.17#range\_points

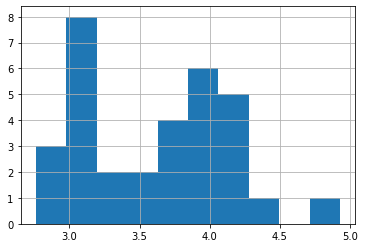
df\_7['Score'].max()-df\_7['Score'].min()=3.9110000000000005#range\_score

df\_7['Weigh'].max()-df\_7['Weigh'].min()=8.399999999999999 #range\_weigh

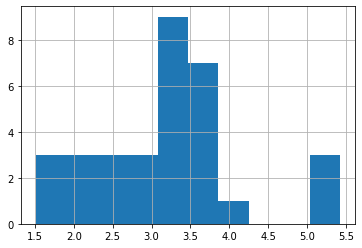
df\_7.var()

Points: 0.285881 ,Score: 0.957379 ,Weigh: 3.193166 ,dtype: float64

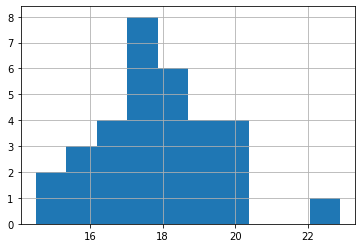
df\_7['Points'].hist()



df\_7['Score'].hist()



df\_7['Weigh'].hist()



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

weight\_mean=np.mean(weight)

print(weight\_mean)

145.33333333333334

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans.**

df\_9a=pd.read\_csv('Q9\_a.csv')

df\_9a=df\_9a.set\_index('Index')

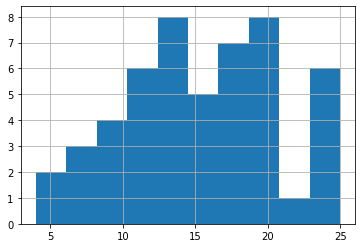
df\_9a['speed'].describe()

count 50.000000 ,mean 15.400000, std 5.287644 ,min 4.000000 ,25% 12.000000 50% 15.000000, 75% 19.000000 ,max 25.000000 ,Name: speed, dtype: float64

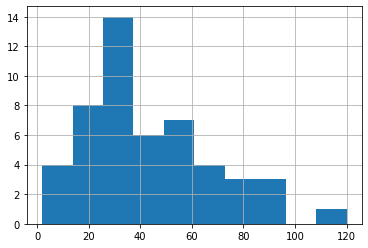
df\_9a['dist'].describe()

count 50.000000, mean 42.980000 ,std 25.769377 ,min 2.000000 ,25% 26.000000 ,50% 36.000000 ,75% 56.000000 ,max 120.000000 ,Name: dist, dtype: float64

df\_9a['speed'].hist()

****

df\_9a['dist'].hist()

****

from scipy.stats import skew

from scipy.stats import kurtosis

df\_9a['speed'].skew()= -0.11750986144663393

df\_9a['dist'].skew()= 0.8068949601674215

**Observation**

"dist" is positively skewed where as “speed” is negatively skewed. Thus, dist has distribution of data concentrated on the left whereas speed has distribution on the right. As seen in the graph

df\_9a['speed'].kurtosis()= -0.5089944204057617

df\_9a['dist'].kurtosis()= 0.4050525816795765

**Observation** : "speed" has negative kurtosis while "dist" has positive kurtosis

**SP and Weight(WT)**

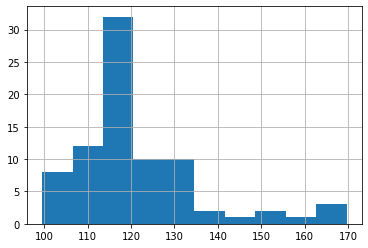
**Use Q9\_b.csv**

df\_9b=pd.read\_csv('Q9\_b.csv')

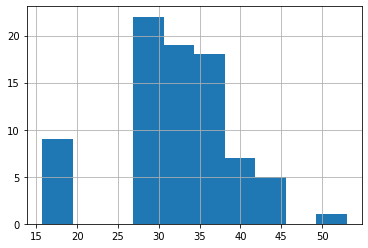
df\_9b.describe()

| **SP** | **WT** |
| --- | --- |
| **count** | 81.000000 | 81.000000 |
| **mean** | 121.540272 | 32.412577 |
| **std** | 14.181432 | 7.492813 |
| **min** | 99.564907 | 15.712859 |
| **25%** | 113.829145 | 29.591768 |
| **50%** | 118.208698 | 32.734518 |
| **75%** | 126.404312 | 37.392524 |
| **max** | 169.598513 | 52.997752 |

df\_9b['SP'].hist()



df\_9b['WT'].hist()

****

df\_9b['SP'].skew()=1.6114501961773586

df\_9b['WT'].skew()= -0.6147533255357768

**Observation**: "SP" is positively skewed while "WT" is negatively skewed

df\_9b['SP'].kurtosis()=2.9773289437871835

df\_9b['WT'].kurtosis()=0.9502914910300326

**Observation**: both "SP" and "WT" have positive kurtosis

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



ANS :

* Majority of the Chicks has weight in range 50 – 100, followed by 100 -150 and 150 – 200
* The data is positively Skewed



ANS:

* Data 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?

Soln:

n=2000

= 200

s= 30

Confidence Interval Estimate= Z => 200 Z

#z-value at 94% CL

np.round(stats.norm.interval(0.94),2)[1]=1.88

moe\_94=1.88\*(30/np.sqrt(2000))

ci\_94=(200-moe\_94,200+moe\_94)

print("Confidence interval at 94% is:",ci\_94)

Confidence interval at 94% is: (198.73885766069012, 201.26114233930988)

#z-value at 98% CL

np.round(stats.norm.interval(0.98),2)[1]=2.33

moe\_98=2.33\*(30/np.sqrt(2000))

ci\_98=(200-moe\_98,200+moe\_98)

print("confidence interval at 98% is :",ci\_98)

confidence interval at 98% is : (198.43698848372765, 201.56301151627235)

#z-value at 96% CL

np.round(stats.norm.interval(0.96),2)[1]=2.05

moe\_96=2.05\*(30/np.sqrt(2000))

ci\_96=(200-moe\_96,200+moe\_96)

print("confidence interval at 96% is:",ci\_96)

confidence interval at 96% is: (198.62481819383763, 201.37518180616237)

**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.

Ans:

* Mean=41
* Median=40.5
* Variance=25.53
* Std dev=5.0526

1. What can we say about the student marks?

Ans: **Mean > Median, This implies that 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. Perfectly symmetric bell-shaped curve

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

Ans: Skewness = Positive. Data is distributed more on left

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

ANS: Skewness = Negative. Data is distributed more on right

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

Ans: peakedness (sharp peak) and less variation.

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

Ans: Less peakness (Broad peak) and more variation.

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 symmetric. Data is more concentrated towards right side

What is nature of skewness of the data?

ANS : Skewness = Negative

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

ANS: IQR data is 8 (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. Data is Normally Distributed. No Outliers. Center around 262.5. Comparatively, first graph has less range
2. Data is Normally Distributed. No Outliers. Center around 262.5

Comparatively, second graph has more range

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)

**Solution:**

df\_20=pd.read\_csv('Cars.csv')

df\_20.shape=(81,5)

df\_20.describe()

| **HP** | **MPG** | **VOL** | **SP** | **WT** |
| --- | --- | --- | --- | --- |
| **count** | 81.000000 | 81.000000 | 81.000000 | 81.000000 | 81.000000 |
| **mean** | 117.469136 | 34.422076 | 98.765432 | 121.540272 | 32.412577 |
| **std** | 57.113502 | 9.131445 | 22.301497 | 14.181432 | 7.492813 |
| **min** | 49.000000 | 12.101263 | 50.000000 | 99.564907 | 15.712859 |
| **25%** | 84.000000 | 27.856252 | 89.000000 | 113.829145 | 29.591768 |
| **50%** | 100.000000 | 35.152727 | 101.000000 | 118.208698 | 32.734518 |
| **75%** | 140.000000 | 39.531633 | 113.000000 | 126.404312 | 37.392524 |
| **max** | 322.000000 | 53.700681 | 160.000000 | 169.598513 | 52.997752 |

df\_20['MPG'].agg(['mean','std'])

mean 34.422076 , std 9.131445 Name: MPG, dtype: float64

1. P(MPG>38)

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

1. P(MPG<40)

stats.norm.cdf(40,34.42076,9.131445)= 0.7293975674106012

1. P (20<MPG<50)

stats.norm.cdf(50,34.422076,9.131445)-stats.norm.cdf(20,34.422076,9.131445)

=0.8988689076273199

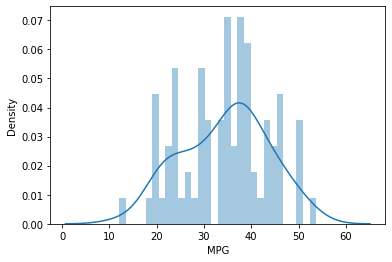
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

import seaborn as sns

sns.distplot(df\_20['MPG'],bins=30)



df\_20['MPG'].mean()=34.42207572802469

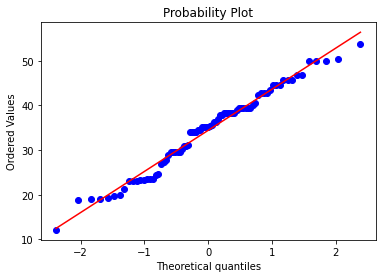
observation: The peakness is centered around the mean(34.422).Hence "MPG" of cars is normally distributed

Display a Q-Q plot to check a normal distribution

Interpretation: If our variable follows a normal distribution, the quantiles of our variable must be perfectly in line with the “theoretical” normal quantiles: a straight line on the QQ Plot tells us we have a normal distribution.

stats.probplot(df\_20['MPG'],dist="norm",plot=plt)

plt.show()



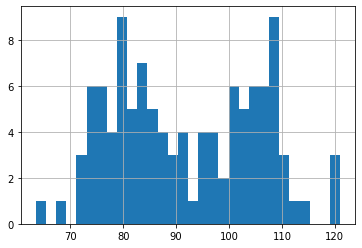
b)Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

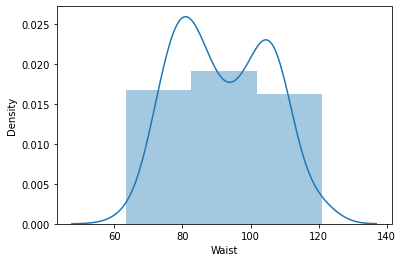
df\_21=pd.read\_csv('wc-at.csv')

df\_21.shape=(109,2)

df\_21['Waist'].hist(bins=30)

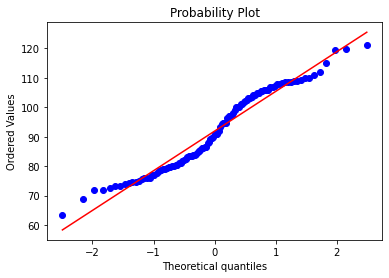


sns.distplot(df\_21['Waist'],bins=3)



stats.probplot(df\_21['Waist'],dist="norm",plot=plt)

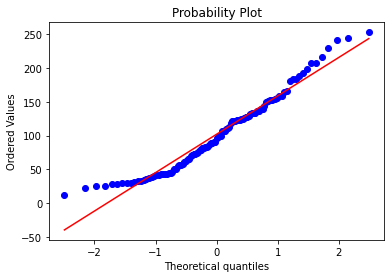
plt.show()



**Waist circumference does not follow normal distribution**

stats.probplot(df\_21['AT'], dist="norm",plot=plt)

plt.show()



**Adipose Tissue does not follow normal distribution**

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

stats.norm.interval(0.9)[1]= 1.6448536269514722

stats.norm.interval(0.94)[1]= 1.8807936081512509

stats.norm.interval(0.60)[1]= 0.8416212335729143

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

stats.t.interval(0.95,24)[1]= 2.0638985616280205

stats.t.interval(0.96,24)[1]= 2.1715446760080677

stats.t.interval(0.99,24)[1]= 2.796939504772804

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

sample size,n=18; dof=17, sample mean=260, population mean=270, sample standard deviation,s=90

t\_val=(260-270)/(90/np.sqrt(18))= -0.4714045207910317

prob=stats.t.cdf(t\_val,17)= 0.32167253567098364=32.2%