|  |  |
| --- | --- |
| 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 | Nomial |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| Religious Preference | Nomial |
| Barometer Pressure | Interval |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

If three coins are tossed, Outcome will be S={HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}.

Number of outcome =2^3=8

Outcome of two heads and one tail = {HHT,HTH,THH}=3

Therefore,Probability of two head and one tail are = 3/8

Q4) Two Dice are rolled, find the probability that sum is

If two dice are rolled,Outcome will be

S={(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)}

Number of outcome = 36

1. Equal to 1

Outcome that sum of 2 dice is 1 = 0

Therefore,Probability = 0

1. Less than or equal to 4

Outcome that sum of 2 dice is Less than or equal to 4 =

{(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)} = 6

Therefore,Probability = 6/36=1/6

1. Sum is divisible by 2 and 3

Outcome that sum of 2 dice is divisible by 2 and 3 ={(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}=6

Therefore,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?

A bag contains 2 red, 3 green & 2 blue ball,

If two balls are drawn, Total no. of outcome = nCr = 7C2 = = 21.

Number of outcome that none of the balls drawn is blue

2C2+3C2 = 1+3

= 4

Therefore, Probability P(A) = 4/7

=0.571

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 E(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.

import pandas as pd

df = pd.read\_csv("Q7 (1).csv",index\_col=0)

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 |

df.median()

Points 3.695

Score 3.325

Weigh 17.710

df.var()

Points 0.285881

Score 0.957379

Weigh 3.193166

df.mode()

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

**Use Q7.csv file**

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 Value = =

= 145.33

Expected value of the Weight of that patient = **145.33** pounds

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

**Cars speed and distance**

**import pandas as pd**

**Use Q9\_a.csv**

**SP and Weight(WT)**

df = pd.read\_csv("Q9\_a (1).csv")

**1)Skessness**

df.skew()

Index 0.000000

speed -0.117510

dist 0.806895

**2)Kurtosis**

df.kurt()

Index -1.200000

speed -0.508994

dist 0.405053

**Use Q9\_b.csv**

df1=pd.read\_csv("Q9\_b (1).csv")

**1)Skessness**

df.skew()

Unnamed: 0 0.000000

SP 1.611450

WT -0.614753

**2)Kurtosis**

df.kurt()

Unnamed: 0 -1.200000

SP 2.977329

WT 0.950291

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



1)In above Histogram mass of distribution is concentrated on left and tail on right side.

2)Therefore,this is **positive skewness** .



In above Box plot or whisker plot **outlier** are in upper side

**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?

N=3000000 n=2000 =200 s=30

Confidence interval Estimate = = 200

94% of confidence interval

Z=(( )+0.94)=(0.97)=1.89

Therefore 200 = 2002.56

Lower limit =197.44

Upper limit =202.56

98% of confidence interval

Z=(( )+0.99)=(0.99)=2.33

Therefore 200 = 2003

Lower limit =197

Upper limit =203

96% of confidence interval

Z=(( )+0.96)=(0.98)=2.06

Therefore 200 = 2002.73

Lower limit =197.27

Upper limit =202.73

**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.

import pandas as pd

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

df.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

df.var()

25.529411764705884

In [13]:

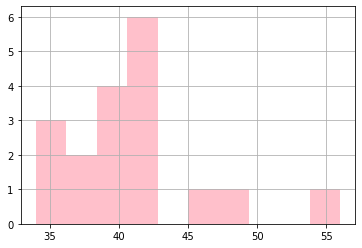
df.mode()

0 41

1. What can we say about the student marks?

plt.hist(df,color='pink')

plt.grid()



Conclusion: In above graph we can see that large amount in near to mean.Therefore most of the student mark is in between 34 to 42.

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

When mean,median of data are equal then symmetric distribution is form and nature of skewness is zero skewness

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

When mean > median then nature of skewness is Right skewness distribution.

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

When mean > median then nature of skewness is Left skewness distribution.

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

For a data positive kurtosis value indicates that middle peack will be hight kurtosis and possess thick tail

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

For a data positive kurtosis value indicates that middle peack will be flat kurtosis and possess thin tail.

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



What can we say about the distribution of the data?

Negative distribution of the data ie. Left side distribution of data

What is nature of skewness of the data?

Negative Skewness is nature of data

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

Inter Quartile Range(IQR) = Upper Quartile – Lower Quartile

= 18 – 10 = 8  
Inter Quartile Range(IQR) of the data is 8.

Q19) Comment on the below Boxplot visualizations?



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

1. Both Boxplot shows normal distribution.
2. Both Boxplot have same median.
3. No outlier found both of a boxplot.
4. Range of both boxplot are different.

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

import pandas as pd

import seaborn as sns

import scipy as smf

from scipy.stats import norm

from scipy import stats

df=pd.read\_csv("Cars (1).csv")

* 1. P(MPG>38)

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

0.3475939251582705

* 1. P(MPG<40)

stats.norm.cdf(38,loc=df.MPG.mean(),scale=df.MPG.std())

* 1. P (20<MPG<50)

df.MPG.mean()

34.422075728024666

df.MPG.std()

9.131444731795982

In [ ]:

Q 21) Check whether the data follows normal distribution

import pandas as pd

import matplotlib.pyplot as plt

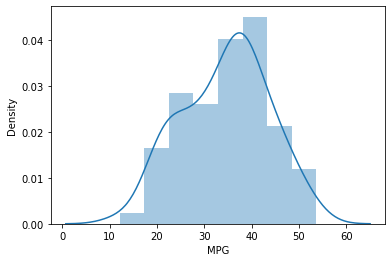
import seaborn as sns

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

df=pd.read\_csv("Cars (1).csv")

sns.distplot(df["MPG"])



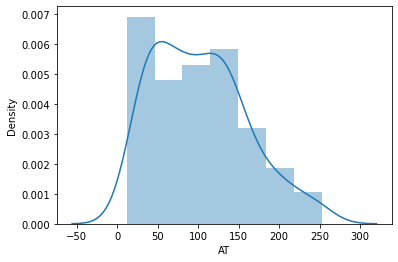
Conclusion: Therefore,The MPG of car does not follow normal distribution.

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

Dataset: wc-at.csv

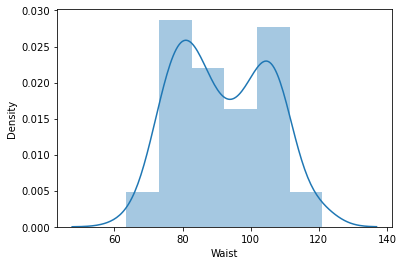
df1=pd.read\_csv("wc-at (1).csv")

sns.distplot(df1["AT"])



conclusion:therefore,at Adipose Tissue(AT) data does not show normal distribution.

sns.distplot(df1["Waist"])



Conclusion : therefore,at Waist data does not show normal distribution

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

from scipy import stats

from scipy.stats import norm

#z-score of 90% of confidence interval

z=((1-0.90)/2)+0.90

stats.norm.ppf(z)

1.6448536269514722

#z-score of 94% of confidence interval

z=((1-0.94)/2)+0.94

stats.norm.ppf(z)

1.8807936081512509

#z-score of 60% of confidence interval

z=((1-0.60)/2)+0.60

stats.norm.ppf(z)

0.8416212335729143

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

n = 25, n-1 = 25-1 =24

import numpy as np

from scipy import stats

from scipy.stats import norm

#t scores of 95% confidence interval

tscore = ((1-0.95)/2)+0.95

stats.t.ppf(tscore,24)

2.0638985616280205

#t scores of 96% confidence interval

tscore = ((1-0.96)/2)+0.96

stats.t.ppf(tscore,24)

2.1715446760080677

#t scores of 99% confidence interval

tscore = ((1-0.99)/2)+0.99

stats.t.ppf(tscore,24)

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

µ=270 =260 n=18 s=90

= Average life of the bulb is 270 days

=Average life of the bulb is not a 270 days

tscore =

import numpy as np

from scipy import stats

from scipy.stats import norm

import math

tscore = (260-270)/(90/math.sqrt(18))

tscore

-0.4714045207910317

stats.t.cdf(tscore,17)

0.32167253567098364

Therefore,the probability that 18 randomly selected bulbs would have an average life of no more than 260 days =32%