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

2^3=8

HHH, HHT, HTH, THH, TTH, THT, HTT, TTT

=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

Two disk thrown is =36

1. Answer is =0
2. Answer is =(3/36)

=1/6

1. Answer is =(30/36)

=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 no of balls= 2+3+2=7

(7\*6)/(2\*1)=21 :no of way drawing balls=2

(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

:total= 0.015+0.20+0.65+0.005+0.01+0.120=1

1+4+3+5+6+2=21 (21/6)=3.5

Probability of having candy is=3

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 1**

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

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?

There are 9 patient

Probability of selecting each patient=1/9

EX. 108,110,123,135,145,167,187,199

P(x) 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9

Expected value: (1/9) (108+110+123+134+135+145+167+187+199)

(1/9) (1308)

=145.33

Expected value of the weight=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**

****

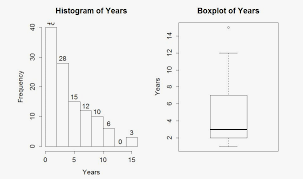
**Q10) Draw inferences about the following boxplot & histogram**



The histogram peak has the right skew and tail is on righ

Mean>median

The boxplot has outlier on the maximum 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?

-CI=94%, 98%, 96%

* Sample meanof https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%3D%20200.
* Sample standard deviation of https://tex.z-dn.net/?f=s%20%3D%2030.
* Sample sizeof https://tex.z-dn.net/?f=n%20%3D%202000.

The interval is:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%5Cpm%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D

Considering a 94% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 1.8916, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%201.8916%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.73

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%201.8916%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.27

The 94% confidence interval is (198.73, 201.27).

Considering a 96% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.0673, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%202.0673%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.61

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%202.0673%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.39

The 96% confidence interval is (198.61, 201.39).

Considering a 98% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.3452, hence:

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20-%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20-%202.3452%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20198.43

https://tex.z-dn.net/?f=%5Coverline%7Bx%7D%20%2B%20t%5Cfrac%7Bs%7D%7B%5Csqrt%7Bn%7D%7D%20%3D%20200%20%2B%202.3452%5Cfrac%7B30%7D%7B%5Csqrt%7B2000%7D%7D%20%3D%20201.57

The 98% confidence interval is (198.43, 201.57).

**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**

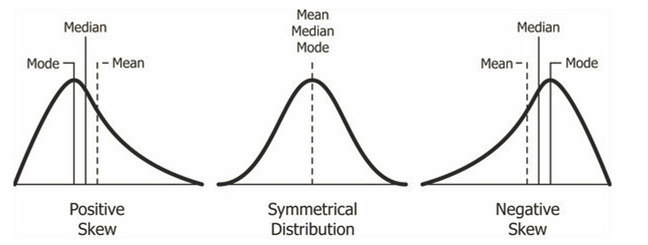
1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | MEAN | DEVIATIN SCORE | SQUARE OF DEV. SCORE | VARIENCE | STANDERD DEVIATION |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| 34 | 41 | 7 | 49 | 21.50815883 | 4.637688953 |
| 36 | 41.41176 | 5.411764706 | 29.28719723 |  |  |
| 36 | 41.75 | 5.75 | 33.0625 |  |  |
| 38 | 42.13333 | 4.133333333 | 17.08444444 |  |  |
| 38 | 42.42857 | 4.428571429 | 19.6122449 |  |  |
| 39 | 42.76923 | 3.769230769 | 14.20710059 |  |  |
| 39 | 43.08333 | 4.083333333 | 16.67361111 |  |  |
| 40 | 43.45455 | 3.454545455 | 11.9338843 |  |  |
| 40 | 43.8 | 3.8 | 14.44 |  |  |
| 41 | 44.22222 | 3.222222222 | 10.38271605 |  |  |
| 41 | 44.625 | 3.625 | 13.140625 |  |  |
| 41 | 45.14286 | 4.142857143 | 17.16326531 |  |  |
| 41 | 45.83333 | 4.833333333 | 23.36111111 |  |  |
| 42 | 46.8 | 4.8 | 23.04 |  |  |
| 42 | 48 | 6 | 36 |  |  |
| 45 | 50 | 5 | 25 |  |  |
| 49 | 52.5 | 3.5 | 12.25 |  |  |
| 56 | 56 | 0 | 0 |  |  |
|  |  |  |  |  |  |
|  |  |  | 365.6387 |  |  |

There are 18 student in class & the performance of student is very average. The mean of student is approximately 41

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

If the distribution is symmetric, then the mean is equal to the median, And the distribution are zero skewness.



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

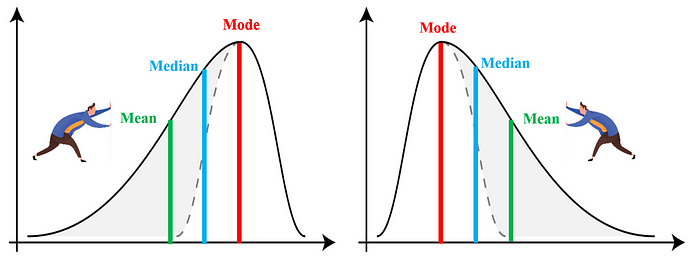
The mean of positively skewed data will be greater than the median. In a negatively skewed distribution, the exact opposite is the case: the mean of negatively skewed data will be less than the median.

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

The mean of positively skewed data will be greater than the median. In a negatively skewed distribution, the exact opposite is the case: the mean of negatively skewed data will be less than the median. If the data graph is symmetrical, the distribution has zero skewness, regardless of how long or fat the tails are.

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

Positive value of kurtosis indicate that distribution is peaked and possesses thick tail. An extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.



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

A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution.

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



What can we say about the distribution of the?

if the distance from the median to the maximum is greater than the distance from the median the to the minimum, then the box plot is positively skewed. The showing boxplot is positively skewed.

What is nature of skewness of the data?

The median is about 10, and the distribution is most likely skewed left.

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

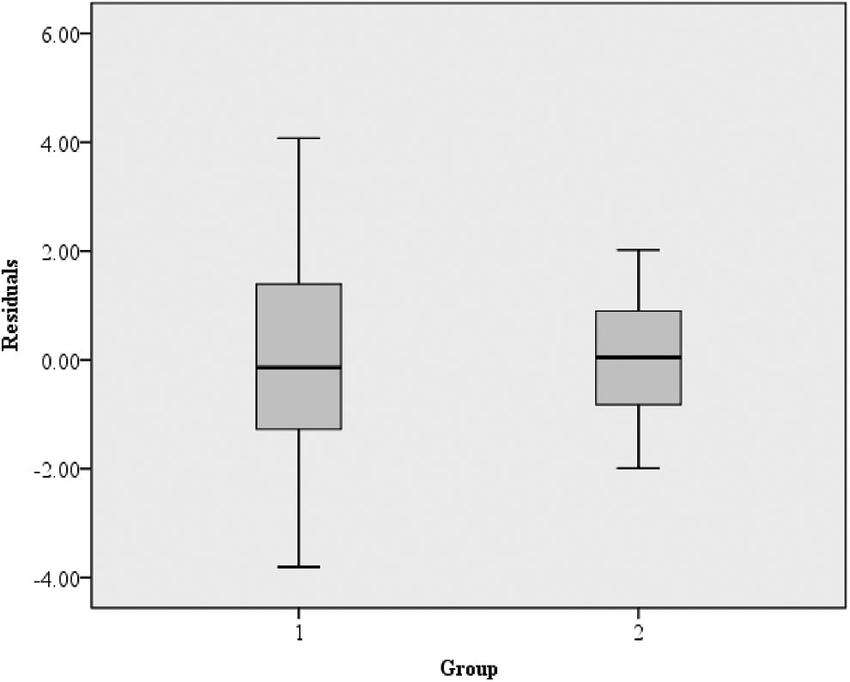
The middle 50% of values when ordered from lowest to highest. in the box plot the box slightly right so its positive skewed.

Q19) Comment on the below Boxplot visualizations?



First there is no outliers. second both the box plot share the same median that is approximately in the range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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



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)

 Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3) print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

b. P(MPG<40) Ans: prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

c. P (20<MPG<50) Ans: prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20)) p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3) print('P(MPG<50)=',(prob\_MPG\_less\_than\_50)) P(MPG<50)= 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20) print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50)) P(20<MPG<50)= 0.013000000000000012

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

cars=pd.read\_csv('/content/Cars (1).csv')

cars.head()

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

cars['MPG'].mean()

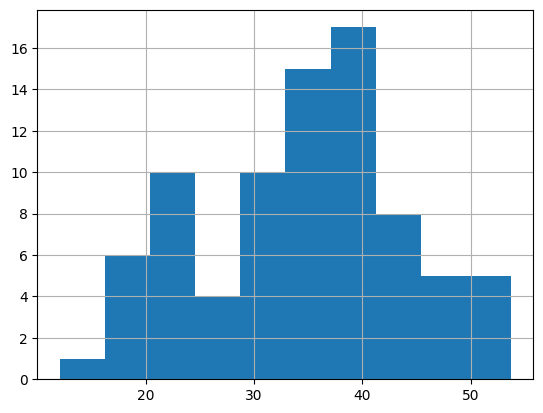
34.42207572802469

cars['MPG'].median()

35.15272697

cars['MPG'].mode()

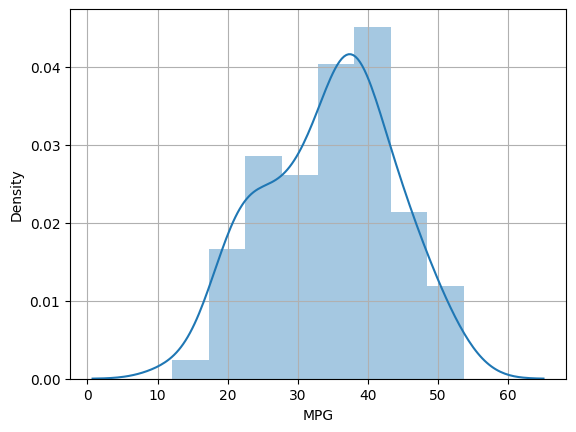
0 29.629936 Name: MPG, dtype: float64

cars['MPG']**.**hist()

sns.distplot(cars['MPG'])

plt.grid(True)

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

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

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

df.head()

| **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 |

df.mean()

Waist 91.901835 AT 101.894037 dtype: float64

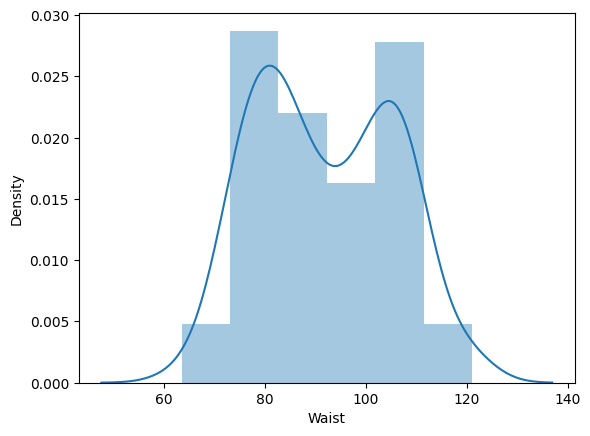
df.median()

Waist 90.80 AT 96.54 dtype: float64

df.mode()  
WaistAT094.5121.01106.0123.02108.5NaN

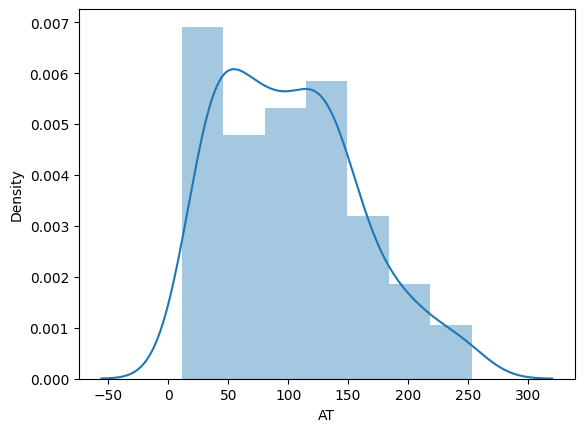
sns.distplot(df['Waist'])

plt.show()



sns.distplot(df['AT'])

plt.show()



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

For 90% confidence interval:

We have the significance level at 5 % ( as it is a two tailed test)

that is:

α = 5 % = 0.05

z at α = 0.05 from the z table will be:

z = 1.645.

For 94 % confidence interval, we get:

We have the significance level at 3 % ( as it is a two tailed test)

that is:

α = 3 % = 0.03

z at α = 0.03 from the z table will be:

z = 1.555.

For 60 % confidence interval, we get:

We have the significance level at 20 % ( as it is a two tailed test)

that is:

α =20 % = 0.2

z at α = 0.2 from the z table will be:

z = 0.253

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

from scipy.stats import norm

# t scores of 95% confidence interval for sample size of 25

stats.t.ppf(0.975,24)  # df = n-1 = 24

2.0638985616280205

# t scores of 96% confidence interval for sample size of 25

stats.t.ppf(0.98,24)

2.1715446760080677

# t scores of 99% confidence interval for sample size of 25

stats.t.ppf(0.995,24)

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

from scipy import stats

from scipy.stats import norm

# Assume Null Hypothesis is: Ho = Avg life of Bulb >= 260 days

# Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days

# find t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))

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

t

-0.4714045207910317

# Find P(X>=260) for null hypothesis

# p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function

p\_value=1-stats.t.cdf(abs(-0.4714),df=17)

p\_value

0.32167411684460556

#  OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function

p\_value=stats.t.sf(abs(-0.4714),df=17)

p\_value

0.32167411684460556