|  |  |
| --- | --- |
| 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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | interval |
| Years of Education | Ordinal |

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

Solution:

Sample space of tossing three Coins(s):

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

Number of favourable outcomes = n(two heads and one tail)=3

P(two heads and one tail)=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

Solution:

sample space S of two dice:.

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) }

So, n(S) = 36

P(E) = n(E) / n(S)

1. Solution:

P(Equal to 1)= 0/36=0

b) Solution:

P(Less than or equal to 4)=6/36=1/6

1. Sum is divisible by 2 and 3:

Solution:

Possible outcomes give a Sum is divisible by 2 and 3 : {(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}=6

P(Sum is divisible by 2 and 3)=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?

Solution: Total ball=7

Sample Space=S(n)=7c2

Let event be with no blue balls=n(E)=5c2

P(E)= 5c2/7c2=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

Solution:  0.015+0.8+1.95+0.025+0.06+0.24 = 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**

**Solution:**

print('mean of Points:',data['Points'].mean())

print('mean of Score:',data['Score'].mean())

print('mean of Weigh:',data['Weigh'].mean())

mean of Points: 3.5965625

mean of Score: 3.2172500000000004

mean of Weigh: 17.848750000000003

print('median of Points:',data['Points'].median())

print('median of Score:',data['Score'].median())

print('median of Weigh:',data['Weigh'].median())

median of Points: 3.6950000000000003

median of Score: 3.325

median of Weigh: 17.71

print('mode of Points:',data['Points'].mode())

print('mode of Score:',data['Score'].mode())

print('mode of Weigh:',data['Weigh'].mode())

mode of Points: 0 3.07

1 3.92

Name: Points, dtype: float64

mode of Score: 0 3.44

Name: Score, dtype: float64

mode of Weigh: 0 17.02

1 18.90

Name: Weigh, dtype: float64

print('range of Points:',(data['Points'].max() - data['Points'].min()))

print('range of Score:',data['Score'].max() - data['Score'].min())

print('range of Weigh:',data['Weigh'].max() - data['Weigh'].min())

range of Points: 2.17

range of Score: 3.9110000000000005

range of Weigh: 8.399999999999999

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?

Solution:

Expected Value (Mean)= (108+110+ 123+ 134+ 135+ 145+ 167+ 187+ 199)/9= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

#skweness

car[['speed'  ,'dist']].skew()

speed -0.117510

dist 0.806895

dtype: float64

# Kurtosis

car[['speed'  ,'dist']].kurt()

speed -0.508994

dist 0.405053

dtype: float64

**SP and Weight(WT)**

**Use Q9\_b.csv**

#skweness

SPWT[['SP'  ,'WT']].skew()

SP 1.611450

WT -0.614753

dtype: float64

# Kurtosis

SPWT[['SP'  ,'WT']].kurt()

SP 2.977329

WT 0.950291

dtype: float64

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



Solution: In histogram: Data has right skewness and mean>median>mode.

In Boxplot: It has outliers 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?

Solution: n=2000, x̄=200, σ=30

Confidence Interval = x̄ ± z(σ/√n)

confidence interval of 94%= 200 ± 1.262, or from 198.7 to 201.3

confidence interval of 96%= 200 ± 1.378, or from 198.6 to 201.4

confidence interval of 98%= 200 ± 1.561, or from 198.4 to 201.6

**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.

Solution:

from statistics import mean, median, mode, stdev, variance

test\_scores = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

print(mean(test\_scores))

print(median(test\_scores))

print(mode(test\_scores))

print(variance(test\_scores))

print(stdev(test\_scores))

41

40.5

41

25.529411764705884

5.05266382858645

1. What can we say about the student marks?

Solution: we don’t have outliers and the data is slightly skewed towards right because mean is greater than median

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

Solution: Symmetrical or bell-shaped curve

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

Solution: Positive skewness

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

Solution: Negative skewness

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

Positive kurtosis means the curve is more peaked and have less kurtosis than a normal distribution.

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

Negative Kurtosis means the curve will be flatter and broader and also called platykurtic..

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



What can we say about the distribution of the data?

Solution: Not normally distributed and there is skewness in data

What is nature of skewness of the data?

Solution: Negative Skewness

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

Solution: IQR=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.

Solution: a) In both boxplots there is no outliers.

1. Both normally distributed and there is no skewness in both boxplots
2. Both boxplots median range lies between 250 to 275

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:

#mean of MPG

mean\_of\_MPG= cars['MPG'].mean()

print('mean\_of\_MPG',mean\_of\_MPG)

#standard deviation of MPG

std\_of\_MPG=cars['MPG'].std()

print('std\_of\_MPG',std\_of\_MPG)

mean\_of\_MPG 34.42207572802469

std\_of\_MPG 9.131444731795982

from scipy import stats

#a. P(MPG>38)

a= stats.norm.cdf(38, loc=34.42, scale=9.13)

print('P(MPG>38)=',1-a)

#b. P(MPG<40)

a= stats.norm.cdf(40, loc=34.42, scale=9.13)

print('P(MPG<40))=',a)

#c. P (20<MPG<50)

a= stats.norm.cdf(50, loc=34.42, scale=9.13) - stats.norm.cdf(20, loc=34.42, scale=9.13)

print('P(20<MPG<50)=',a)

P(MPG>38)= 0.34748702501304063

P(MPG<40))= 0.7294571279557076

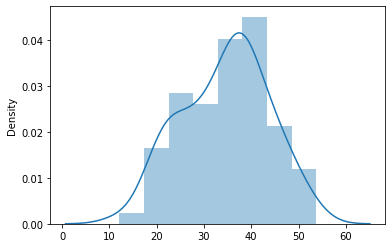
P(20<MPG<50)= 0.8989177824549222

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

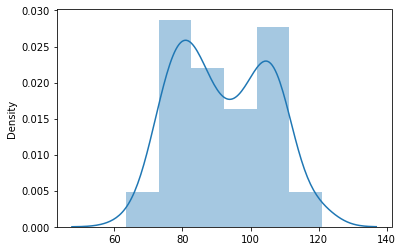
Solution: Yes, MPG of cars Follow normal Distribution

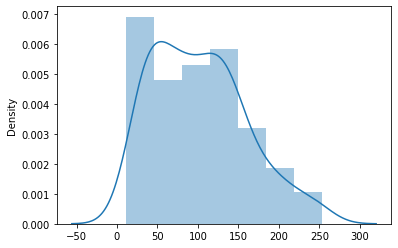


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Solution: Both not follow normally distributed





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

Solution:

For 90% confidence interval:

Significance level at 5 % for two tailed test:

α = 5 % = 0.05, From Z table, Z value will be

z = 1.645.

For 94 % confidence interval:

Significance level at 3 % for two tailed test:

α = 3 % = 0.03, From Z table, Z value will be

z = 1.555.

For 60 % confidence interval, we get:

We have the significance level at 20 % for two tailed test:

α =20 % = 0.2, From Z table, Z value 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

Solution: Sample size=25

Degree freedom = 25-1=24

With the help of t-score table

t scores of 95% confidence interval = 2.064

t scores of 94% confidence interval = 2.053

t scores of 99% confidence interval = 2.797

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

Solution:

Populataion mean=270

Sample mean= 260

Sample std= 90

Sample size=18

import scipy

from scipy import stats

import pandas as pd

import numpy as np

#t= ((sample mean)- Pop Mean)/ Std Error

t= (260-270)/(90/np.sqrt(18))

t

0.32167253567098364

stats.t.cdf(t, df=17)

0.32167253567098364