|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continous |
| Weight of Gold | Continous |
| Distance between two places | Continous |
| Length of a leaf | Continous |
| Dog's weight | Continous |
| 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 | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratios |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratios |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratios |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratios |
| Religious Preference | Nominal |
| Barometer Pressure | Ratios |
| SAT Scores | Interval |
| Years of Education | Ratios |

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

Sol.) Number of possible outcomes = 23 = 8

The possible outcomes will be HHH, TTT, HTT, THT, TTH, THH, HTH, HHT.

We have to find the probability of two heads and one tail.

So favourable outcomes = {THH, HTH, HHT}

Number of favourable outcomes = 3

so the req. probability is : =3/8

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

* Equal to 1 = 0
* Less than or equal to 4 = 1/6
* Sum is divisible by 2 and 3 = 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?

Sol.)

Total number of balls = (2 + 3 + 2) = 7

Then, Number of ways of drawing 2 balls out of 7 =7C2 = 21

Number of ways of drawing 2 balls out of (2 + 3) balls =5C2 = 10

Therefore, 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

Sol.) the mean or expectated mean is given by = summation of X\*P(E)

mean=1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

=3.0515

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.

**Sol.) using csv file**







**ConClusion:**  The CSV Dataset have 4 Columns, in which ‘Points’, ‘Score’ and ‘Weigh’ are Numeric Data type and ‘Unnamed’ is String(object) Data Type.

There are 0 Null values.

and Mean, Mode, Median Standard deviation and variance are calculated for numeric columns as shown in above SSs.

Q8) Calculate Expected Value for the problem below

* 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?

**So, The Expected Value is : 145.33333**

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Solution(9A) :-**

**Here, we get the skewness of cars speed and distance**

speed -0.117510

dist 0.806895

conclusion:

1. the data of Cars speed is almost or near to symmertrical distribution.

2. the data of Cars disatance is positively skewed i.e. most of the data points falls before the median.(becoz mode<median<mean)

**Here, we get the Kurtosis of cars speed and distance**

speed -0.508994

dist 0.405053

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Solution(9B):-**

**Here, we get the skewness of cars SP and WT**

SP 1.611450

WT -0.614753

conclusion:

1. the data of Cars SP is positively skewed.

2. the data of Cars disatance is negatively skewed i.e. most of the data points falls after the median.(becoz mode>median>mean)

**Here, we get the Kurtosis of cars SP and WT**

SP 2.977329 (distribution is not symmetric with low variance called

leptokurtic)

WT 0.950291

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





**Inference about histogram** :-

1. data points between 50 to 100 have highest frequency.

2. it is positively skewed (i.e mean>median>mode)

3. distribution is skewed to the right side.

**Inference about Boxplot** :-

1. there are some outliers presents in data.

2. median is very close to lower whisker.

3. data distribution is negatively skewed.

**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?

**Sol** :- C.I.= sample mean + Z(1-alpha)\*population S.D./sqrt(sample size)

where (1-alpha) is confidence interval.

given, sample mean 200 pounds

S.D 30 pounds

**1. for (1-alpha)=94%**

from scipy import stats

import numpy as np

n=3000

sd=30

mean=200

scale=sd/np.sqrt(n) #scale is standard error of mean

#now calculate the C.I.

stats.norm.interval(0.94, loc=mean, scale=scale)

**Confidence Interval is: (198.96984691480404, 201.03015308519596)**

**2. for (1-alpha)=98%**

from scipy import stats

import numpy as np

n=3000

sd=30

mean=200

scale=sd/np.sqrt(n) #scale is standard error of mean

#now calculate the C.I.

stats.norm.interval(0.98, loc=mean, scale=scale)

**Confidence Interval is: (198.72580679278366, 201.27419320721634)**

**3. for (1-alpha)=96%**

from scipy import stats

import numpy as np

n=3000

sd=30

mean=200

scale=sd/np.sqrt(n) #scale is standard error of mean

#now calculate the C.I.

stats.norm.interval(0.96, loc=mean, scale=scale)

**Confidence Interval is: (198.8751153941953, 201.1248846058047)**

**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**

* Find mean, median, variance, standard deviation.

**Solution** :-

import pandas as pd

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

data=pd.Series(ls)

print(data.count())

# mean

print("the mean is :",data.mean())

#median

print("the median is:", data.median())

#standard deviation

print("the SD of data is:",data.std())

#variance

print("the var is:",data.std()\*\*2)

#mode

print("the mode is",data.mode())

**18**

**the mean is : 41.0**

**the median is: 40.5**

**the SD of data is: 5.05266382858645**

**the var is: 25.529411764705884**

**the mode is : 41**

* What can we say about the student marks?

**Solution**:-

1. about skewness

pearson's cofficient of skewness =(mean - mode)/SD

=(41-41)/5.05

=0

i.e data is symmetrically distributed.

2. data follows the normal distribution.

3. data points 41 has highest frequency in among all points.

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

**Solution** :- if mean = median

**then the distribution is called as symmetric distribution**.

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

**Solution** :- if mean > median

**then the distribution is called as positively skewed**.

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

**Solution** :- if mean > median

**then the distribution is called as negatively skewed**.

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

**Solution** :-

A distribution with a positive kurtosis value indicates that the distribution has heavier tails and a sharper peak than the normal distribution.

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

**Solution** :-

A distribution with a positive kurtosis value indicates that the distribution has lighter tails and shape nearer to the normal distribution.

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



What can we say about the distribution of the data?

**data is not normally distributed.**

What is nature of skewness of the data?

**data is positively skewed.**

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

**IQR = Q3-Q1**

**= 18-10**

**= 6**

Q19) Comment on the below Boxplot visualizations?



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

**Solution** :-

1. median value for both boxplot are same, i.e 262.5

2.IQR values

for boxplot 1st = 275-250=25

for boxplot 2nd =300-225=75

3. both boxplots are repersenting normal distribution.

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** :-

import pandas as pd

import numpy as np

from scipy import stats

data=pd.read\_csv("Cars.csv")

**#for P(MPG>38)**

p1=1-stats.norm.cdf(38, loc=data.MPG.mean(), scale=data.MPG.std())

print("the probability for MPG>38: ", p1)

**#for P(MPG<40)**

p2=stats.norm.cdf(40, loc=data.MPG.mean(), scale=data.MPG.std())

print("the probability for MPG<40: ", p2)

**#for P(20<MPG<50)**

p31=stats.norm.cdf(20, loc=data.MPG.mean(), scale=data.MPG.std())

p32=stats.norm.cdf(50, loc=data.MPG.mean(), scale=data.MPG.std())

p3=p32-p31

print("the probability for 20<MPG<50: ", p3)

**the probability for MPG>38: 0.3475939251582705**

**the probability for MPG<40: 0.7293498762151616**

**the probability for 20<MPG<50: 0.8988689169682046**

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Solution** :- we can check by using histogram



**So, according to above histogram of MPG data , the data is not following the normal distribution.**

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

Dataset: wc-at.csv

**Solution** :-



**According to above histogram for (Waist and AT) both do not follow the normal distribution of data.**

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

**Solution** :-

from scipy import stats

print('90% confidence interval z-score: ',round(stats.norm.ppf(0.95),2))

print('94% confidence interval z-score: ',round(stats.norm.ppf(0.97),2))

print('60% confidence interval z-score: ',round(stats.norm.ppf(0.80),2))

**90% confidence interval z-score: 1.64**

**94% confidence interval z-score: 1.88**

**60% confidence interval z-score: 0.84**

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

**Solution** :-

from scipy import stats

print('95% confidence interval t-score: ',round(stats.t.ppf(0.975,24),2))

print('96% confidence interval t-score: ',round(stats.t.ppf(0.98,24),2))

print('99% confidence interval t-score: ',round(stats.t.ppf(0.995,24),2))

**95% confidence interval t-score: 2.06**

**96% confidence interval t-score: 2.17**

**99% confidence interval t-score: 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

**Solution** :-

from scipy import stats

import numpy as np

sample\_mean=260

sample\_SD=90

n=18

df=n-1

population\_mean=270

t\_score=(sample\_mean - population\_mean)/(sample\_SD/np.sqrt(n))

res=stats.t.cdf(t\_score, df)

print("the probability that 18 randomly selected bulbs would have an average life of no more than 260 days:",np.round(res,2))

**the probability that 18 randomly selected bulbs would have an average life of no more than 260 days: 0.32**