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

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ratio |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ordinal |
| Years of Education | Interval |

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

HHT, THH, THT are 3 possible outcomes of 2 Heads and 1 Tail out of 8 possible outcomes. Hence probability is given by (3/8)\*100 = 37.5%

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
4. Since two dice are rolled the minimum possible outcome is (1,1). Hence probability that sum is equal to 1 is 0.
5. (1,1),(1,2),(2,1),(3,1),(1,3),(2,2) are the possible outcomes. Hence probability of sum is less than or equal to 4 is 6/36=16.66%
6. (3,3),(2,4),(4,2),(5,1),(1,5),(6,6) are the possible outcomes. Hence probability of sum is divisible by 2 and 3 is 6/36=16.66%

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?

n(S)=7C2

n(S)=(7\*6)/(2\*1) = 21

n(E)= 5C2 = (5\*4)/(2\*1) = 10

Hence 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

Expected number of candies for a randomly selected child is:



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**

import pandas as pd

a=pd.read\_csv("C:\\Users\\SHARAN\\Desktop\\ExcelR\\Assignment-1\\Q7.csv")

print(a.describe())

a.var()

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

Points 0.285881

Score 0.957379

Weigh 3.193166

dtype: float64

INFERENCE:

Standard deviation of points and scores are quite low hence more number of data is close to the mean that is very less possibilities outliers. But in case of weight, since standard deviation is very high data is spread across the entire range.

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?



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

Cars speed and distance

Use Q9\_a.csv

import pandas as pd

a=pd.read\_csv("C:\\Users\\SHARAN\\Desktop\\ExcelR\\Assignment-1\\Q9\_a.csv")

print(a.skew(axis=0,skipna=True))

a.kurt(axis=0,skipna=True)

Index 0.000000

speed -0.117510

dist 0.806895

dtype: float64

Out[9]:

Index -1.200000

speed -0.508994

dist 0.405053

dtype: float64

Inference:

\* Here both Skewness and kurtosis of car speed is negative which shows that distribution lie towards left. Hence negative skewed. As taking kurtosis into consideration it shows that the distribution has broad peak and thin tail.

\* For distance Skewness are Kurtosis both are positive, hence distribution is towards right, hence positive skewing. As taking kurtosis into consideration it shows that the distribution has pointed peak and wide tail.

SP and Weight(WT)

Use Q9\_b.csv

import pandas as pd

a=pd.read\_csv("C:\\Users\\SHARAN\\Desktop\\ExcelR\\Assignment-1\\Q9\_b.csv")

print(a.skew(axis = 0, skipna = True))

a.kurt(axis = 0, skipna = True)

Unnamed: 0 0.000000

SP 1.611450

WT -0.614753

dtype: float64

Out[2]:

Unnamed: 0 -1.200000

SP 2.977329

WT 0.950291

dtype: float64

INFERENCE

* For SP both Skewness and Kurtosis are positive hence distribution lie towards right. Hence positive Skewing. As taking kurtosis into consideration it shows that the distribution has pointed peak and wide tail.
* For WT Skewness is negative and Kurtosis is positive. Hence distribution is towards left and hence negative Skewing.

Q10) Draw inferences about the following boxplot & histogram



From Histogram and boxplot it is clear that most of the values lie in left and outliers are present in right end. Hence it is negative skewing.

**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?

Sample=2000

Mean=200

Std.D=30

1. For 94%

import numpy as np

import scipy

from scipy import stats

a=stats.norm.interval(0.94,loc=200,scale=30/np.sqrt(2000))

print("Gain at 94% confidence: ",np.round(a,4))

Gain at 94% confidence: [198.7383 201.2617]

1. For 98%

import numpy as np

import scipy

from scipy import stats

a=stats.norm.interval(0.98,loc=200,scale=30/np.sqrt(2000))

print("Gain at 98% confidence: ",np.round(a,4))

Gain at 98% confidence: [198.4394 201.5606]

1. For 96%

import numpy as np

import scipy

from scipy import stats

a=stats.norm.interval(0.96,loc=200,scale=30/np.sqrt(2000))

print("Gain at 96% confidence: ",np.round(a,4))

Gain at 96% confidence: [198.6223 201.3777]

**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

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

print("Mean: ",marks.mean())

print("Median: ",marks.median())

print("Variance: ",marks.var())

print("Standard deviation: ",marks.std())

Mean: 41.0

Median: 40.5

Variance: 25.529411764705884

Standard deviation: 5.05266382858645

1. What can we say about the student marks?

Student’s marks are quite consistent as mean and median are close to each other and deviation of 5 marks from the mean.

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

Normal distribution is exhibited.

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

Negative skewing is exhibited.

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

Positive skewing is exhibited.

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

Positive kurtosis indicates that the distribution is peaked and possess thick tails. It means most of the data located on the tail side.

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

ANS: Negative kurtosis value for a data 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 data?

ANS: Most of the data seems to be distributed in the range 10:18

What is nature of skewness of the data?

ANS: Negative skewing.

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

ANS: IQR= 8

Q19) Comment on the below Boxplot visualizations?



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

ANS: In both the cases the Mean seems to at 262.5, and also boxplots follow normal distribution. And also data is equally distributed.

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)

import pandas as pd

a=pd.read\_csv("C:\\Users\\SHARAN\\Desktop\\ExcelR\\Assignment-1\\Cars.csv")

print(a)

print("MEAN: ",round(a.MPG.mean(),4))

print("STANDARD DEVIATION: ",round(a.MPG.std(),4))

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

.. ... ... ... ... ...

76 322 36.900000 50 169.598513 16.132947

77 238 19.197888 115 150.576579 37.923113

78 263 34.000000 50 151.598513 15.769625

79 295 19.833733 119 167.944460 39.423099

80 236 12.101263 107 139.840817 34.948615

[81 rows x 5 columns]

MEAN: 34.4221

STANDARD DEVIATION: 9.1314

import scipy

from scipy import stats

print('probability for >38 is:',round(1-stats.norm.cdf(38,loc=a.mean,scale=a.std),4))

probability for >38 is: 0.3476

import scipy

from scipy import stats

print('probability for <40 is:',round(stats.norm.cdf(40,loc=a.mean,scale=a.std),4))

probability for <40 is: 0.7293

import scipy

from scipy import stats

print('probability for 20<x<50 is:',round(stats.norm.cdf(50,loc=a.mean,scale=a.std)-stats.norm.cdf(20,loc=a.mean,scale=a.std),4))

probability for 20<x<50 is: 0.8989

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

a=pd.read\_csv("C:\\Users\\SHARAN\\Desktop\\ExcelR\\Assignment-1\\Cars.csv")

import seaborn as sn

import numpy as np

import matplotlib.pyplot as plt

a["gain"]=a.MPG.pct\_change(periods=1)

a=a.dropna()

a=a[["MPG","gain"]]

print(a)

sn.distplot(a.gain,label="cars")

plt.xlabel('gain')

plt.ylabel('density')

plt.legend();

MPG gain

1 50.013401 -0.068664

2 50.013401 0.000000

3 45.696322 -0.086318

4 50.504232 0.105214

5 45.696322 -0.095198

.. ... ...

76 36.900000 0.966654

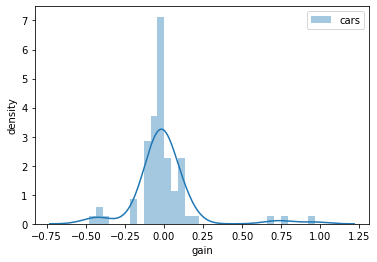
77 19.197888 -0.479732

78 34.000000 0.771028

79 19.833733 -0.416655

80 12.101263 -0.389865

[80 rows x 2 columns]



It follows 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

For AT:

import pandas as pd

a=pd.read\_csv("C:\\Users\\SHARAN\\Desktop\\ExcelR\\Assignment-1\\wc-at.csv")

import seaborn as sn

import numpy as np

import matplotlib.pyplot as plt

a["gain"]=a.AT.pct\_change(periods=1)

a=a.dropna()

a=a[["AT","gain"]]

print(a)

sn.distplot(a.gain,label="AT")

plt.xlabel('gain')

plt.ylabel('density')

plt.legend();

AT gain

1 25.89 0.006610

2 42.60 0.645423

3 42.80 0.004695

4 29.84 -0.302804

5 21.68 -0.273458

.. ... ...

104 124.00 -0.340426

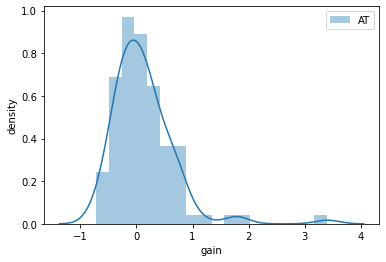
105 62.20 -0.498387

106 133.00 1.138264

107 208.00 0.563910

108 208.00 0.000000

[108 rows x 2 columns]



It follows Positive skewing

For WC:

import pandas as pd

a=pd.read\_csv("C:\\Users\\SHARAN\\Desktop\\ExcelR\\Assignment-1\\wc-at.csv")

import seaborn as sn

import numpy as np

import matplotlib.pyplot as plt

a["gain"]=a.Waist.pct\_change(periods=1)

a=a.dropna()

a=a[["Waist","gain"]]

print(a)

sn.distplot(a.gain,label="Waist")

plt.xlabel('gain')

plt.ylabel('density')

plt.legend();

Waist gain

1 72.60 -0.028763

2 81.80 0.126722

3 83.95 0.026284

4 74.65 -0.110780

5 71.85 -0.037508

.. ... ...

104 100.10 -0.008911

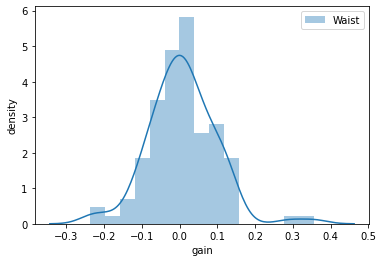
105 93.30 -0.067932

106 101.80 0.091104

107 107.90 0.059921

108 108.50 0.005561

[108 rows x 2 columns]



It follows Normal Distribution

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

1. For 90% :- import scipy

from scipy import stats

round(stats.norm.ppf(0.95),4)

1.6449

1. For 94% :- import scipy

from scipy import stats

round(stats.norm.ppf(0.97),4)

1.8808

1. For 60% :- import scipy

from scipy import stats

round(stats.norm.ppf(0.8),4

0.8416

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

1. For 95% :- import scipy

from scipy import stats

round(stats.t.ppf(0.975,df=24),4)

2.0639

1. For 96% :- import scipy

from scipy import stats

round(stats.t.ppf(0.98,df=24),4)

2.1715

1. For 99% :- import scipy

from scipy import stats

round(stats.t.ppf(0.995,df=24),4)

2.7969

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

Mean=270

N=18

X=260

Std.D=90

import pandas as pd

import scipy

from scipy import stats

import numpy as np

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

round(stats.t.cdf(t,df=17),4)

0.3217