|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continuous Data |
| Weight of Gold | Continuous Data |
| Distance between two places | Continuous Data |
| Length of a leaf | Continuous Data |
| Dog's weight | Continuous Data |
| Blue Color | Nominal Data |
| Number of kids | Discrete Data |
| Number of tickets in Indian railways | Discrete Data |
| Number of times married | Discrete Data |
| Gender (Male or Female) | Nominal Data |

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 | Interval |
| Years of Education | Ratio |

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

* Taken question in consideration.

Total outcomes are: 23 = 8

HHH, **HHT**, **HTH**, **THH,** TTH, THT, HTT, TTT so we can see

Answer = 3/8 = 0.375

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

* When two dice are rolled total outcomes = 6 x 6 = 36

1. There is no combination which will give sum 1 = **0/36**
2. Out of all outcomes less than or equal to 4 are: (1,1) (1,2) (1,3) (2,2) (2,1) (3,1)

So the probability is = 6/36 = **1/6**

1. Sum is divisible by 2 and 3

(1,5)(2,4)(3,3)(4,2)(5,1)(6,6)

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?

* All possible outcomes are

2C7 = 21

Number of ways to draw non blue ball

5C2 = 10

Therefore => answer = **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

* (1 \* 0.015 )+ (4\*0.20)  +( 3 \*0.65)  +( 5\*0.005)  + (6 \*0.01 ) +( 2 \* 0.12)

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= **3.090**

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**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.59 | 3.21 | 17.84 |
| Median | 3.7 | 2.82 | 20 |
| Mode | 3.92 | 3.44 | 17.02 |
| Varience | 0.2769 | 0.9274 | 3.0933 |
| Std | 0.526 | 0.963 | 1.7588 |
| range | 2.17 | 3.911 | 8.3999 |
|  |  |  |  |

Points:

The mean score of approximately 3.59 indicates that the average points earned is around this value.

The median score of approximately 3.7 suggests that the middle value of the points distribution is close to this value.

The mode of approximately 3.92 indicates that this value is the most common among the points earned.

Score:

The mean score of approximately 3.21 suggests that the average score is around this value.

The median score of approximately 2.82 indicates that the middle value of the scores is around this value.

The mode score of approximately 3.44 implies that this score appears most frequently.

Weight:

The mean weight of approximately 17.84 suggests that the average weight is around this value.

The median weight of 20 suggests that the middle weight value is relatively higher.

The mode weight of approximately 17.02 indicates that this weight is the most common.

Variability and Spread:

The variance of the points is relatively low (0.2769), indicating that the data points are clustered closer to the mean.

The higher variance of the score (0.9274) suggests more variability in scores compared to the points.

The variance of the weight (3.0933) indicates relatively higher variability in weight values.

Spread and Dispersion:

The standard deviation of the points (0.526) suggests relatively low spread from the mean.

The higher standard deviation of the score (0.963) indicates greater spread compared to the points.

The standard deviation of the weight (1.7588) indicates higher spread in weight values.

Range:

The range of the points (2.17) indicates the difference between the maximum and minimum points.

The range of the score (3.911) indicates a wider spread in score values.

The range of the weight (8.3999) suggests a substantial range of weight values.

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?

* The probability of selecting one patient is = 1/9

Then there are 9 patients.

If we multiply weights with 1/9(probability)

=(1/9)(108) + (1/9)110  + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

=**145.33**

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

**Cars speed and distance**

**Use Q9\_a.csv**

* To calculate skewness we need to calculate mean

data = df["speed"]

mean = np.mean(data)

print("Mean:", mean)

variance = np.var(data, ddof=0)

print("Variance:", variance)

std\_deviation = np.std(data, ddof=0)

print("Standard Deviation:", std\_deviation)

skew = stats.skew(data)

print("Skewness:", skew)

kurt = stats.kurtosis(data)

print("kurtosis:", kurt)

Mean of speed = 746/50 = 14.92

Varience = 33.75

Std = 5.81

**Skewness = -0.11395**

**Kurtosis = -0.577147**

**As the skewness came out to be negative it have long left tail and the data is concentrated at right side.**

**This is Platykurtic as its negative this all means it have flatter center peak and lighter tails**

Distance

from scipy import stats

data = df["dist"]

mean = np.mean(data)

print("Mean:", mean)

variance = np.var(data, ddof=0)

print("Variance:", variance)

std\_deviation = np.std(data, ddof=0)

print("Standard Deviation:", std\_deviation)

skew = stats.skew(data)

print("Skewness:", skew)

kurt = stats.kurtosis(data)

print("kurtosis:", kurt)

Mean: 42.98

: 650.7796000000001

Standard Deviation: 25.51038220019449

Skewness: 0.782483517311496

kurtosis: 0.24801865717051808

**As the skewness comes out positive the data frame of distance have long right tail and the data is concentrated at left side.**

**This is Leptokurtic. It have peak at its center and heavier tails.**

**SP and Weight(WT)**

**Use Q9\_b.csv**

* **SP**
* data = df["SP"]
* mean = np.mean(data)
* print("Mean:", mean)
* variance = np.var(data, ddof=0)
* print("Variance:", variance)
* std\_deviation = np.std(data, ddof=0)
* print("Standard Deviation:", std\_deviation)
* skew = stats.skew(data)
* print("Skewness:", skew)
* kurt = stats.kurtosis(data)
* print("kurtosis:", kurt)

Mean: 121.54027218037035

Variance: 198.63012494127113

Standard Deviation: 14.093620008403487

Skewness: 1.5814536794423764

kurtosis: 2.7235214865269244

**The positive skewness (greater than 0) suggests that the data is skewed to the right (right-tailed). This means that there is a longer tail on the right side of the distribution, and the data may have some outliers or extremely high values.**

**The majority of the data points are likely clustered on the left side of the distribution, with a few larger values dragging the mean to the right.**

**WT**

data = df["WT"]

mean = np.mean(data)

print("Mean:", mean)

variance = np.var(data, ddof=0)

print("Variance:", variance)

std\_deviation = np.std(data, ddof=0)

print("Standard Deviation:", std\_deviation)

skew = stats.skew(data)

print("Skewness:", skew)

kurt = stats.kurtosis(data)

print("kurtosis:", kurt)

Mean: 32.412576910246905

Variance: 55.44913245817723

Standard Deviation: 7.446417424384509

Skewness: -0.6033099322115126

kurtosis: 0.8194658792266849

**A skewness value of approximately -0.6033 indicates that the dataset is negatively skewed. In a negatively skewed distribution:**

**The tail on the left-hand side (the lower values) is longer or more spread out than the tail on the right-hand side (the higher values).**

**The majority of the data points tend to be concentrated on the right side of the distribution, while there are fewer extreme values on the left side.**

**A kurtosis value of approximately 0.8194 suggests that the dataset has positive kurtosis. Positive kurtosis indicates that the dataset has slightly heavier tails and a slightly more peaked central peak compared to a normal distribution.**

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



Here we can see that the major Chick weights fall in the catogory of 50-100g(measures in x) as the maximum which is 200.The minimum weights have afrequency if less than or equal to 5.

The plot is Right sqewed which show that there is lesser concentration ofchick weights in the 300-400gram category .

The expected value should be above 46.45

The data is right skewed.There areoutliers at upper side.



The data is right skewed.,There areoutliers at upper side

Median is less than mean right skewed and we have outlier on the upperside ofbox plot and there is less data points between Q1 and bottom point.

**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?

* X (+/-) Z(1-alpha) \* standard\_dev/sqrt(n)

Here

X = sample average or mean

Z(1-alpha) \* Standard\_dev/sqrt(n) = Margin of error

Standard\_dev/sqrt(n) = 0.6708

Z values

94 = 1.882

98 = 2.326

96 = 2.053

a) 200 (+/-) 1.882 \* 0.6708

= 200 (+/-) 1.262

b) 200 (+/-) 2.326 \* 0.6708

= 200 (+/-) 1.560c) 200 (+/-) 2.053 \* 0.6708

= 200 (+/-) 1.377

**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?
3. data = np.array([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])
4. mean = np.mean(data)
5. print("Mean:", mean)
6. variance = np.var(data, ddof=0)
7. print("Variance:", variance)
8. std\_deviation = np.std(data, ddof=0)
9. print("Standard Deviation:", std\_deviation)

Mean: 41.0

Variance: 24.11111111111111

Standard Deviation: 4.910306620885412

The data appears to be somewhat positively skewed, as there are more students with marks above the mean (41.78) than below it. However, it's not extremely skewed.

The range (22) and the standard deviation (approximately 6.67, rounded to two decimal places) indicate moderate variability in the dataset. Some students scored quite high (e.g., 56), while others scored relatively low (e.g., 34).

The range (22) and the standard deviation (approximately 6.67, rounded to two decimal places) indicate moderate variability in the dataset. Some students scored quite high (e.g., 56), while others scored relatively low (e.g., 34).

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

* The skewness will be symmetrical. Hence both the sides of the plot must be equal in proportion for the data should be normally distributed.

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

* For data which produces mean > median the skewness will be a +ve skewness or the data will be right skewed. Most of the data will be lying on the left side of the plot. Mean always tends to go towards the most skewed part since skewness influences the mean.

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

* For data which produces median > mean the skewness will be a -ve skewness or the data will be left skewed. Most of the data will be lying on the right side of the plot. Mean always tends to go towards the most skewed part since skewness influences the mean.

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

* A distribution with a positive kurtosis value indicates that the distributionhas heavier tails than the normal distribution. For example, data that follow a tdistribution have a positive kurtosis value.

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 data?

The data is not actually equally distributed across the plane. There might be outliers influencing the data. Median of the data is 14.7(app x)

25 percent of the data lies between 0-10

50 percent of the data lies between 10-18

25 percent of the data lies after 18-20 appx

What is nature of skewness of the data?

The data will be left skewed since whisker length on the upper quadrant is higher than the data on the lower quadrant. Median will be greater than the mean since data is left skewed

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

IQR is the inter quartile range.Here Q1 = 10

Q2 = 14.7

Q3 = 18

IQR = Q3

Q1 = 8(approx)

Q19) Comment on the below Boxplot visualizations?



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

* Here when we compare box plot 1 with box plot 2 we can say that the data in box plot 1 is widely spread. Here the main inference is that since the data range varies high in box plot 2 it is hard to make a prediction in box plot 2. The median in the 2box plots are equal. And the data spread in both of them are symmetrical

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)

P(MPG>38)

= mean(MPG)=34.42208

= sd(MPG)=9.131445

= 1–pnorm(38,mean(MPG),sd(MPG))

= 0.330

= 33%

P(MPG<40)

=pnorm(40,mean(MPG),sd(MPG))

=0.7293499

=72.3%

P (20<MPG<50)

=pnorm(50,mean(MPG),sd(MPG))–pnorm(20,mean(MPG),sd(MPG))

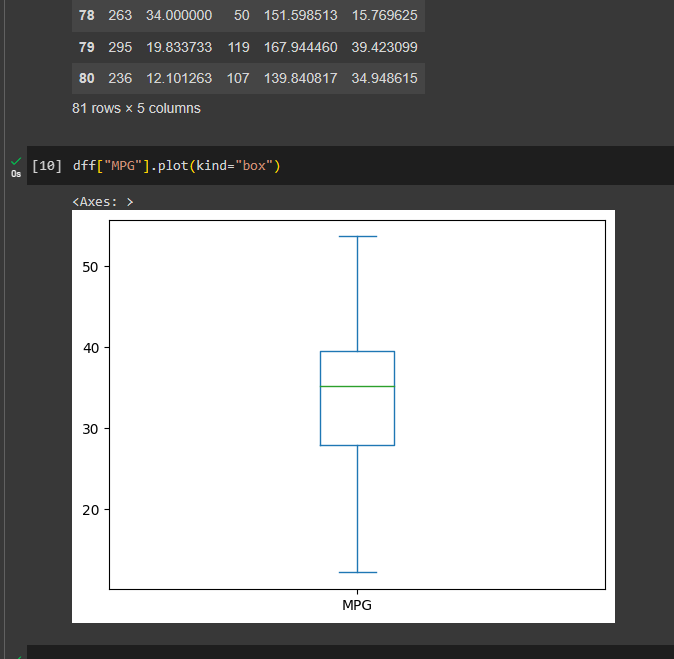
=0.955 -0.057

=0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

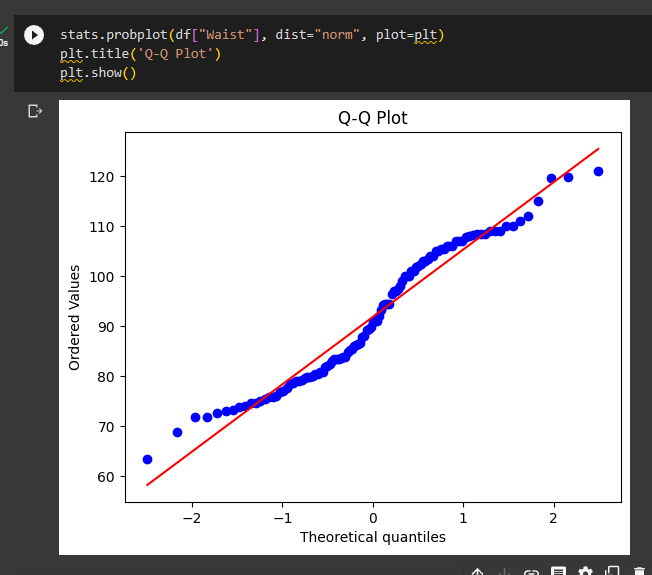
Dataset: Cars.csv



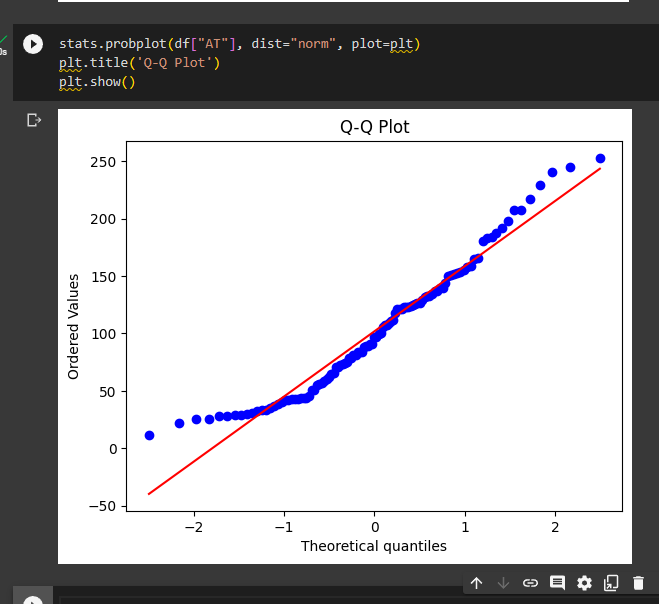
As we can see the box plot follows a near normal distribution.

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

Dataset: wc-at.csv

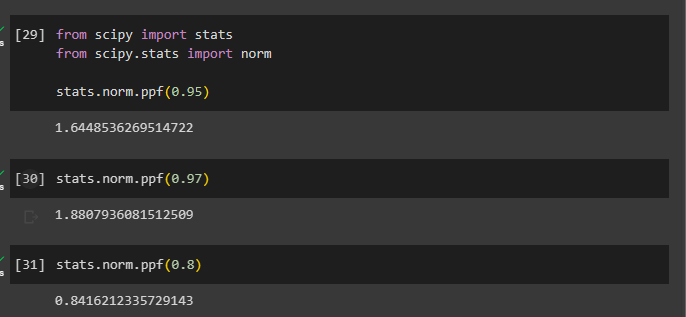


This shows the data follows a abnormal curve thus not normal distribution.

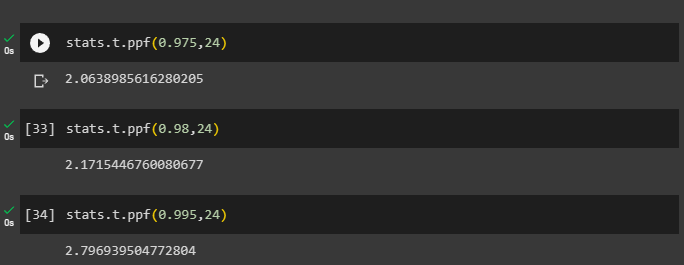


This dataset is somewhat following a normal distribution.

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



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



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

