|  |  |
| --- | --- |
| 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 | ORDINAL |
| Time on a Clock with Hands | INTERVAL |
| Number of Children | RATIO |
| 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?

Ans) probability of 3 coins= TTT,TTH,THT,THH,HTT,HTH,HHT,HHH

probability p(a)=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

**ANS)**

1. **p(a) = 0**

because there is no sum of getting 1 when two dice rolled.

1. **p(b) = 5/36**

because sample space (S) = 36,

num of times i.e, less than or equal to 4 = 5

so P(B) = 5/36

1. **p(c) = 6/36 =1/6**

sample space (s) = 36

number that sum is divisible by 2 and 3 are = 6

therefore, probability P( C ) = 6/36

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?

**ANS)**

**Total num of balls = 7**

**Num of ways of drawing 2 balls out of 7 is = 21**

**Num of ways of drawing 2 out of 5 is = 10**

**Probability = num of favourable outcomes/total number of outcomes**

**i.e. P(E ) = 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

**ANS)**

**EXPECTED NUMBER OF CANDIES FOR A RANDOMLY SELECTED CHILD ARE = P(E)**

**P(E ) = 1/21 = 0.0476**

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**

**ANS)** data1=pd.read\_csv(r"C:\Users\91863\Downloads\Q7.csv")

data1

data1.describe()

**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**

**range : 2.17 3.911 8.4**

* the above descriptive stats says that standard deviation is almost low.
* So there is no very large variation among the values of Points, Score and Weigh.
* The data is simple with less variation among datapoints for all columns.
* median value is close to mean for all the columns. So, the distribution could be symmetric for them.

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?

**ANS)**

**EXPECTED VALUE OF THE WEIGHT OF THAT PATIENT IS P(E)**

**P(E ) = 1308/9 = 145.34**

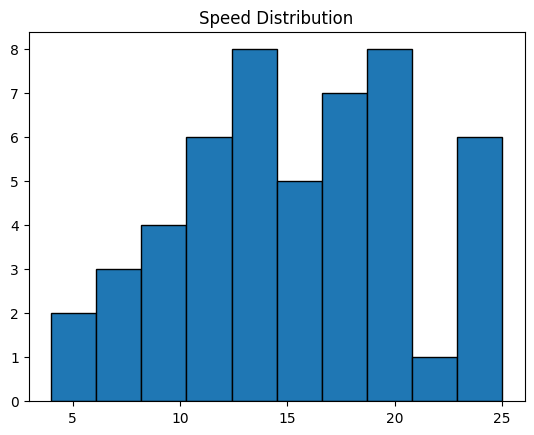
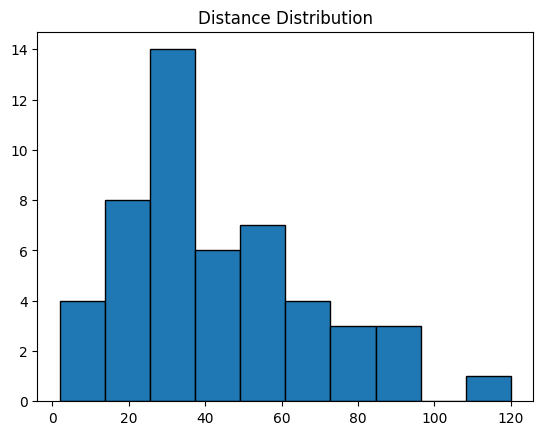
**P( E ) = Sum of all the values(x) / total number of values**

**= 1308/9**

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

**Cars speed and distance**

**Use Q9\_a.csv**

** Ans)**

**SKEWNESS FOR SPEED AND DISTANCE IS “ -0.11395477, 0.78248352”**

**i.e, 1)** NEGATIVE SKEWNESS is observed from the above graph and point

[-0.11395577] in case of speed.

2) POSITIVE SKEWNESS for distance with [0.78248352]

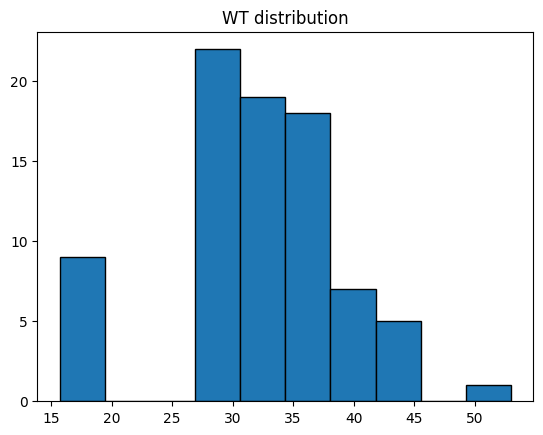
**KURTOSIS FOR SPEED AND DISTANCE IS “ -0.57714742, 0.24801866”**

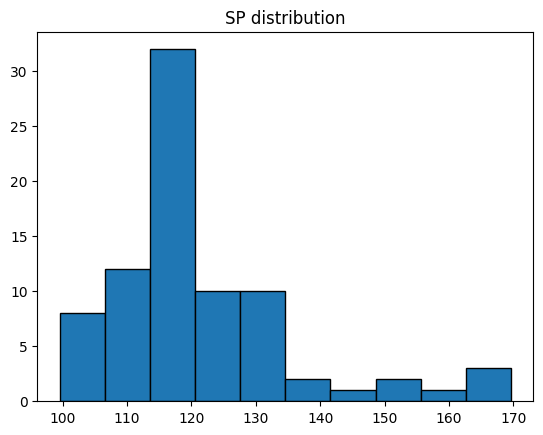
**3)** NEGATIVE KURTOSIS i.e, lower in case of speed .so, thinner tails.

4) for distance, POSITIVE KURTOSIS , so it has heavier tails.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans)**

****

KURTOSIS:

**SP 2.977329 WT 0.950291**

SKEWNESS:

**SP 1.611450** **WT -0.614753**

1. **Both "SP" and "WT" have positive kurtosis values, indicating that have heavier tails than a normal distribution i.e, the presence of outliers**
2. **"SP" has a positive skewness, that its distribution is skewed to the right. This means there may be a concentration of data points on the left side of the mean with a longer tail to the right.**
3. **"WT" has a negative skewness, that its distribution is skewed to the left. This suggests a concentration of data points on the right side of the mean with a longer tail to the left.**

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





Ans)

a) From the histogram we can say that , the distribution is right-skewed with its right tail since more data points are on the left. **(positive skewed)**

1. In the boxplot, the outliers are present on the top side.
2. And the more datapoints are also on the top whiskers. Because it extended very long and spread of the data.
3. **The longer top whisker suggests that there is more variability or spread in the upper part of the dataset. This may indicate a positively skewed distribution i.e, Positive skewness**

**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?

Ans)

da=pd.DataFrame({"scores":[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]})

da.mean()

da.median()

da.var()

da.std()

1. Mean : scores 41.0

Median : scores 40.5

Variance : scores 25.529412

Standard variation : scores 5.052664

1. Since the mean and median are close, the distribution is symmetric roughly.

* The standard deviation is moderate value , the data is moderately spread around the mean.

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

Ans)

* The distribution is approximately symmetrical.
* The skewness is close to zero tells that the tails on both sides of the distribution are roughly equal in length.

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

Ans) **The distribution is right-skewed (positively skewed).The tail on the right side is longer or fatter than the left side.The skewness is positive.**

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

Ans) **The distribution is left-skewed (negatively skewed).The tail on the left side is longer or fatter than the right side. The skewness is negative**.

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

Ans) **Positive kurtosis (greater than 3) indicates that the distribution has heavier tails and a more peaked central region than a normal distribution.**

**The data has more outliers or extreme values than a normal distribution.**

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

Ans) **Negative kurtosis indicates that the distribution has lighter tails than a normal distribution. The data has fewer outliers or extreme values than a normal distribution.**

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



a)What can we say about the distribution of the data?

b)What is nature of skewness of the data?

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

Ans)

a) The distribution of the data in the above boxplot tells that maximum of the data is on the left whiskers while the right whiskers data spread is less and median is close to the right whiskers.

b) the nature of skewness of the data is “negatively skewed”. Since the left whiskers spreads the data more than the right.

c) IQR OF THE DATA is the difference between the first quartile(Q1) and third quartile(Q3)

IQR = Q3 – Q1

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.

Ans)

* Boxplot 1 and Boxplot 2 has same median (Q2).
* The IQR for boxplot1 is approximately 25. Since the data is less spread. Whereas boxplot spreads more data. hence, the IQR is approximately 75.
* Boxplot1 has less whiskers so less variability. Besides, boxplot2 has more variability.
* There are no outliers present in the both boxplot visualizations.

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)

Ans) mpg\_data = cars['MPG']

a) P(mpg > 38)

prob\_a = 1 - norm.cdf(38, loc=mpg\_data.mean(), scale=mpg\_data.std())

c) P(20 < mpg < 50)

prob\_b = norm.cdf(50, loc=mpg\_data.mean(), scale=mpg\_data.std()) - norm.cdf(20, loc=mpg\_data.mean(), scale=mpg\_data.std())

1. P(MPG<40)

prob\_less\_than\_40 = norm.cdf(40, loc=mpg\_data.mean(), scale=mpg\_data.std())

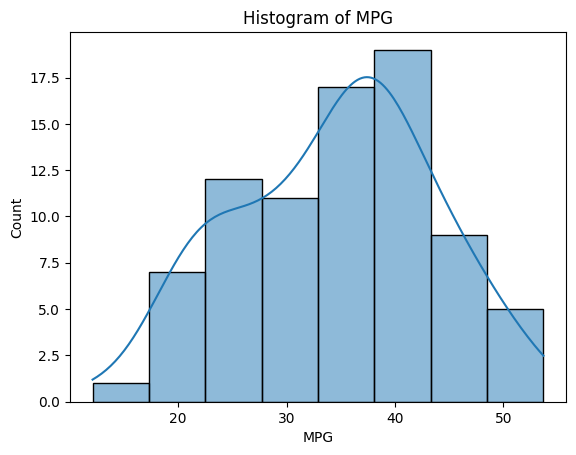
1. P(MPG > 38): 0.3476
2. P(MPG < 40): 0.7293
3. P(20 < MPG < 50): 0.8989

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

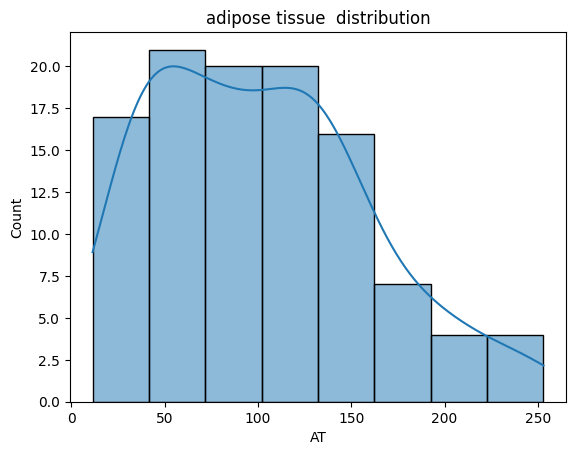
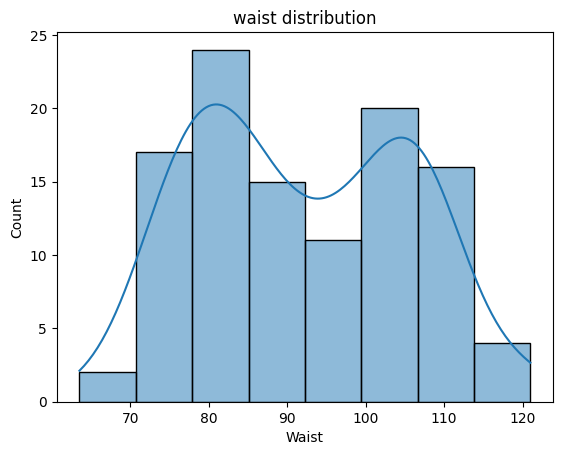
Ans)



MPG gives normal distribution , but not symmetric.

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

Dataset: wc-at.csv

 Ans)

* Waist distribution does not follow normal distribution while adipose tissue follows roughly normal distribution, But not symmetric.

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

Ans)

confidence\_interval\_90 = 0.90

z\_score\_90 = norm.ppf((1 + confidence\_interval\_90) / 2)

confidence\_interval\_94 = 0.94

z\_score\_94 = norm.ppf((1 + confidence\_interval\_94) / 2)

confidence\_interval\_60 = 0.60

z\_score\_60 = norm.ppf((1 + confidence\_interval\_60) / 2)

Z-score for 90% Confidence Interval: 1.6449

Z-score for 94% Confidence Interval: 1.8808

Z-score for 60% Confidence Interval: 0.8416

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

sample\_size = 25

df = sample\_size – 1 # 24

confidence\_interval\_95=0.95

t\_score\_95 = stats.t.ppf((1 + confidence\_interval\_95) / 2,24)

confidence\_interval\_96=0.96

t\_score\_96 = stats.t.ppf((1 + confidence\_interval\_96) / 2,24)

confidence\_interval\_99=0.99

t\_score\_99 = stats.t.ppf((1 + confidence\_interval\_90) / 2,24)

Ans) T-score for 95% Confidence Interval: 2.0639

T-score for 96% Confidence Interval: 2.1715

T-score for 99% Confidence Interval: 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

ans)

population\_mean = 270

sample\_mean = 260

sample\_std\_dev = 90

sample\_size = 18

# Degrees of freedom

df = sample\_size - 1

# Calculate the t-score

t\_score = (sample\_mean - population\_mean) / (sample\_std\_dev / math.sqrt(sample\_size))

# Calculate the probability using the cumulative distribution function (CDF)

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

print(t\_score,probability)

T-score: -0.4714

Probability that the average life is no more than 260 days: 0.3217