|  |  |
| --- | --- |
| 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 | Continuous |
| 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 | Nominal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Nominal |
| 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. Sample – HHH, HTT, HTH, HHT, TTT, THH, THT, TTH

Two Heads and One Tail – HTH, HHT, THH

Probability of two heads and one tail = 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

Ans.

1. Probability that the sum is equal to 1=0
2. Sum less than or equal to 4 – (1,1), (1,2), (1,3), (2,1), (2,2), (3,1)

Probability that sum is less than or equal to 4 = 6/36 = 0.167

1. Sum is divisible by 2 and 3 – (1,5), (2,4), (3,3), (4,2), (5,1), (6,6)

Probability that the sum is divisible by 2 and 3 = 6/36 = 0.167

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. Probability that first ball drawn is not blue = (No of green + No of red balls)/ Total balls

= (2+3)/ (2+3+2) = 5/7

Probability that second ball drawn is not blue = (No of green + No of red balls) - 1/ Total balls left after first ball drawn

= 4/6 = 2/3

Probability that none of the balls drawn is blue = 5/7 \* 2/3 = 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 |

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

**Points –**

**Ans.**

mean 3.596

median 3.695

mode 3.07

var 0.286

std 0.535

min 2.76

max 4.93

**Score-**

mean 3.217

median 3.325

mode 3.44

var 0.957

std 0.978

min 1.513

max 5.424

**Weigh-**

mean 17.849

median 17.71

mode 17.02

var 3.193

std 1.787

min 14.5

max 22.9

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

Ans. Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

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

= 145.34

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

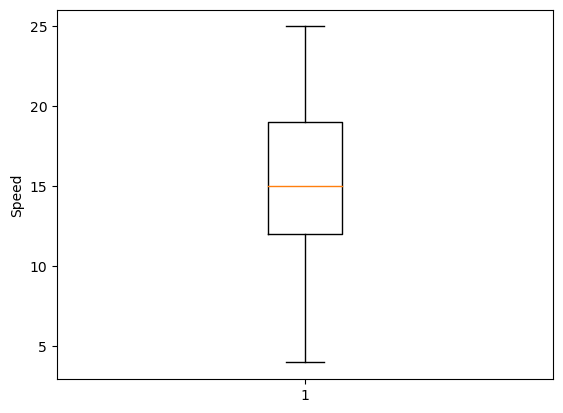
**Cars speed and distance**

**Use Q9\_a.csv**

Ans. Speed –

Skewness = -0.11750986144663393

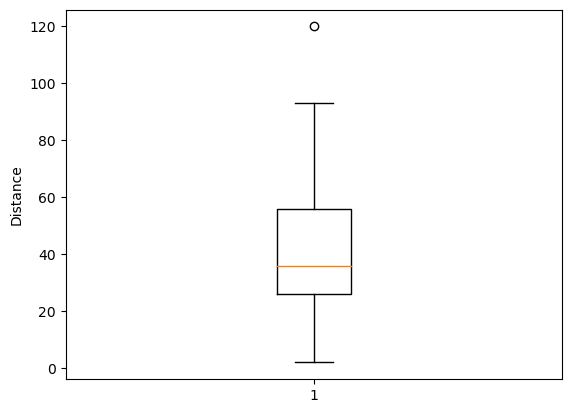
Kurtosis = -0.5089944204057617



Dist –

Skewness = 0.8068949601674215

Kurtosis = 0.4050525816795765

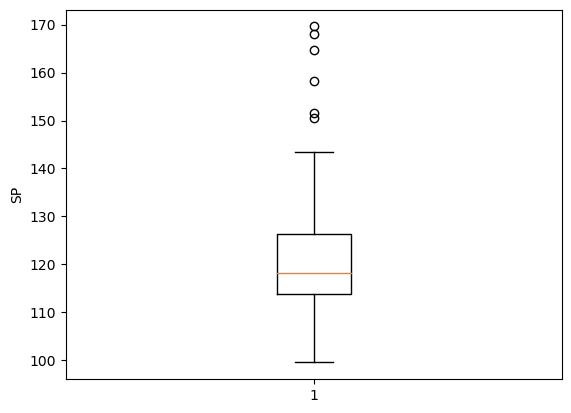


**Use Q9\_b.csv**

Ans. SP –

Skewness = 1.6114501961773586

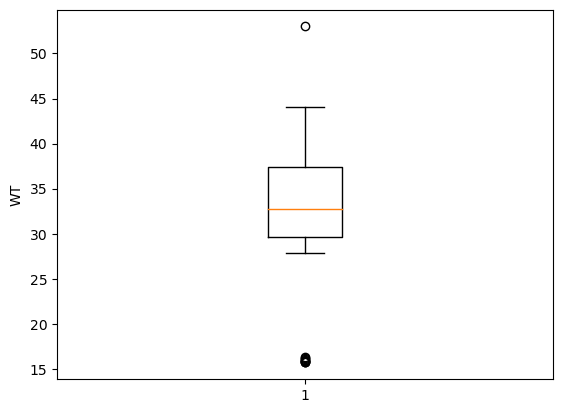
Kurtosis = 2.9773289437871835



WT-

Skewness = -0.6147533255357768

Kurtosis = 0.9502914910300326



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



Ans. From the histogram, we can see that the data is positively skewed. This means that most of the data is less than the mean of the data. It also means mean is more than the median. It also have positive kurtosis.

**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?

Ans. As the population standard variation is not given, we will use tstat to calculate the confidence interval.

Sample size =2000

Degree of freedom = 1999

Standard deviation of sample = 30

Mean of Sample = 200

For 94% confidence interval,

stat.t.interval(.94, 1999, loc=200, scale=30 / (2000 \*\* 0.5))

Interval -> (198.7376089443071, 201.2623910556929)

For 98% confidence interval,

stat.t.interval(.98, 1999, loc=200, scale=30 / (2000 \*\* 0.5))

Interval -> (198.4381860483216, 201.5618139516784)

For 96% confidence interval,

stat.t.interval(.96, 1999, loc=200, scale=30 / (2000 \*\* 0.5))

Interval -> (198.6214037429732, 201.3785962570268)

**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.

Ans. Mean=(34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56)/18

=41

Median=(40+41)/2 = 40.5

Variance = ∑(xi-mean)2/n-1 = 25.52

Standard deviation = = 5.05

1. What can we say about the student marks?

Ans. As mean is very close to median, we can say the marks are mostly normally distributed. The marks are mostly consistent with little increase towards the end.

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

Ans. When the mean and median are equal, the concentration of the data is centered around a single point, and there isn't a significant tail pulling the mean away from the median. In other words, there is no skewness or the skewness is very close to zero.

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

Ans. When the mean is greater than the median, it indicates that the distribution of the data is positively skewed. Positively skewed distributions are also known as right-skewed distributions. The majority of the data points are concentrated on the left-hand side (lower values) of the distribution. The median is typically closer to the lower end of the data, while the mean is influenced by the presence of the longer tail on the right.

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

Ans. When the median is greater than the mean, it indicates that the distribution of the data is negatively skewed. Negatively skewed distributions are also known as left-skewed distributions. The majority of the data points are concentrated on the right-hand side (higher values) of the distribution. The median is typically closer to the higher end of the data, while the mean is influenced by the presence of the longer tail on the left.

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

Ans. Positive kurtosis indicates that the distribution has more outliers and is more heavy-tailed and more peaked compared to a normal distribution. This can happen in datasets that have clusters of extreme values or in cases where the distribution is not as spread out around the mean as in a normal distribution.

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

Ans. Negative kurtosis indicates that the distribution has fewer outliers and is more light-tailed and flatter compared to a normal distribution. This can happen in datasets that are relatively spread out and lack prominent clusters of extreme values.

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



What can we say about the distribution of the data?

Ans. The data is negatively skewed. Most of the data is more than the mean and median is more than the mean.

What is nature of skewness of the data?

Ans. The data is negatively skewed.

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

Ans. IQR = Q3 – Q1 = 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 has positive kurtosis as the data is not that much widely spread compared to Boxplot2 where the data is widely spread and can be considered as having negative kurtosis.

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)

c. P (20<MPG<50)

Ans. a. P(MPG>38) = 1- P(MPG<38)

import scipy.stats as stat

car = pd.read\_csv(r'C:\Users\Simran Kaur\Downloads\Cars.csv')

mean\_car = car.MPG.mean()

std\_car = car.MPG.std()

1 - stat.norm.cdf(38, mean\_car, std\_car )

Answer: 0.34759392515827137

* 1. P(MPG<40)

stat.norm.cdf(40, mean\_car, std\_car)

Answer: 0.729349876215160

* 1. P(20<MPG<50)

=P(MPG<50) – P(MPG<20)

stat.norm.cdf(50, mean\_car, std\_car) - stat.norm.cdf(20, mean\_car, std\_car)

Answer: 0.8988689169682047

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans. mean\_car = car.MPG.mean()

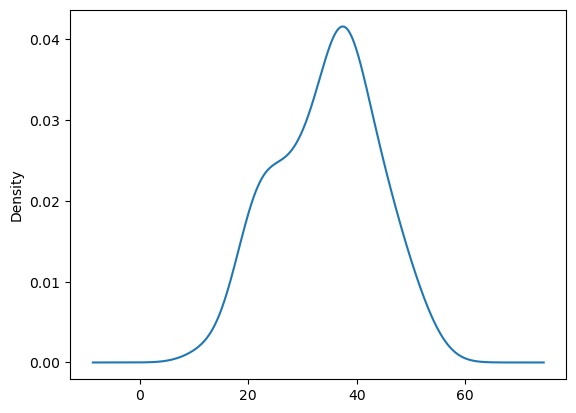
= 34.422

median\_car = car.MPG.median()

=35.15

skew\_car = car.MPG.skew()

= -0.177



Therefore, it is not a normal distribution as mean is not equal to median. Also, we are not getting a smooth bell shaped curve.

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

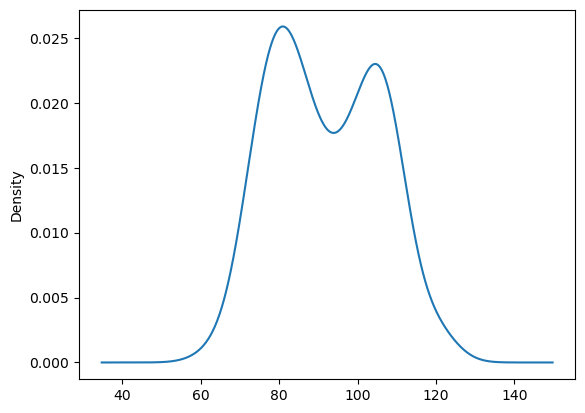
Dataset: wc-at.csv

Ans. Waist Circumference:

mean\_wc = wc.Waist.mean() = 91.90183486238531

median\_wc = wc.Waist.median() = 90.8

skewness\_wc = wc.Waist.skew() = 0.1340560824786468

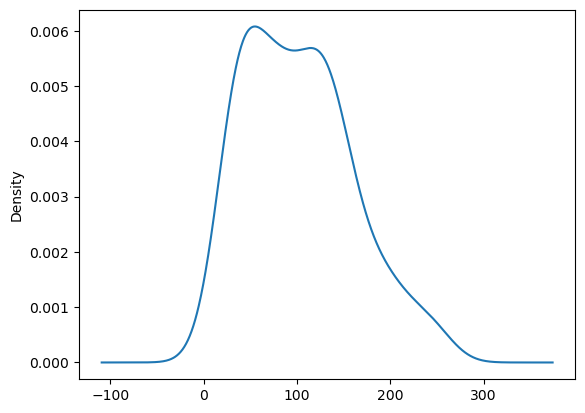


Adipose Tissue:

mean\_at = wc.AT.mean() = 101.89403669724771

meadian\_at = wc.AT.median() = 96.54

skewness\_at = wc.AT.skew() = 0.584869324127853



Thus, both are not normally distributed

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

Ans.

Z score for 90%-

stat.norm.ppf(1-.10/2) -> 1.6448536269514722

Z score for 94%-

stat.norm.ppf(1-.06/2) -> 1.8807936081512509

Z score for 60%

stat.norm.ppf(1-.40/2) -> 0.8416212335729143

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

Ans.

T score for 95%-

stat.t.ppf(q=1-.05/2,df=25) -> 2.059538552753294

T score for 96%-

stat.t.ppf(q=1-.04/2,df=25) -> 2.1665866344527562

T score for 99%-

stat.t.ppf(q=1-.01/2,df=25) -> 2.787435813675851

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. = 260

µ =270

s = 90

n = 18

tstat = = = -0.47

Probability that 18 random bulb selected would have an average life of no more that 260 days

stat.t.cdf(-0.47,17) = 0.32216