1. **What is Univariate Analysis? Give an example.**

- The main purpose of univariate analysis is to take data,
  - **summarize that data**, and **find patterns** among the values.
- Example
  1. The salaries of workers in a specific industry; the variable in this example is workers' salaries.
  2. The heights of ten students in a class are measured; the variable here is the students' heights.
  3. A veterinarian wants to weigh 20 puppies; the variable, in this case, is the weight of the puppies

2. **What are the two techniques for reducing the number of digits?**

- There are two techniques for reducing the number of digits.
- The first is known as rounding.
- Values from zero to four are **rounded down,**
- A second method of losing digits is simply **cutting off or 'truncating'** the ones that we do not want.

3. **State the important aspects of histograms during inspection of data.**

Histograms allow inspection of four important aspects of any distribution:

level	What are typical values in the distribution?
spread	How widely dispersed are the values? Do they differ very much from one another?
shape	Is the distribution flat or peaked? Symmetrical or skewed?
outliers	Are there any particularly unusual values?

4. **Define standardized variable.**

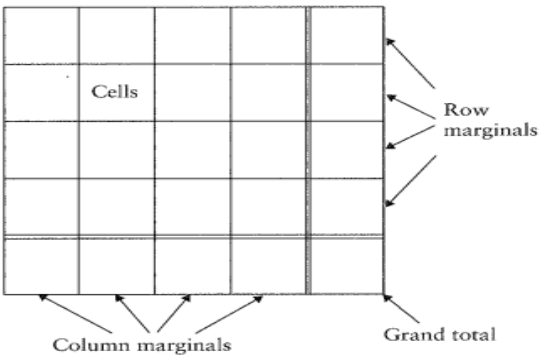
Subtracting a constant from every data value altered the level of the distribution, and dividing by a constant scaled the values by a factor.

5. **Write the formula for Gini coefficient?**

A measure that summarizes what is happening across all the distribution is the Gini coefficient.

$$\text{Gini coefficient} = \frac{2}{\overline{Y}N^2}\sum iY_i - \frac{N+1}{N}$$

6. **Show the anatomy of a contingency table.**



Anatomy of a contingency table.

7. **List out types of graphs are used to depict the bivariate analysis?**

Bivariate information is investigated utilising the scatterplot of Y against X, giving a visual image of

the information's relationship.

8. **Show some examples of bivariate analysis?**
   Information for two factors (normally two sorts of related information). Model: Ice cream deals versus the temperature on that day. The two factors are Ice Cream Sales and Temperature.

9. **Define the degree of freedom.**
   It does numerically what the three-dimensional bar chart does graphically. True only under existing or specified conditions.

$$\text{Degrees of freedom (Df)} = (r-1) \times (c-1)$$

10. **List out the use of scatterplots**.
- Is the relationship monotonic?
- Are the variables positively or negatively related?
- Can the relationship be summarized as a straight line or will it need a curve?
- Hhow much does Y increase (or decrease) for every unit increase of X?
- Are there any gaps in the plot?
- Are there any obvious outliers

1. Compare Univariate , Bivariate and Multivariate Analysis.

Univariate	Bivariate	Multivariate
It only summarize single variable at a time.	It only summarize two variables	It only summarize more than 2 variables.
It does not deal with causes and relationships.	It does deal with causes and relationships and analysis is done.	It does not deal with causes and relationships and analysis is done.
It does not contain any dependent variable.	It does contain only one dependent variable.	It is similar to bivariate but it contains more than 2 variables.
The main purpose is to describe.	The main purpose is to explain.	The main purpose is to study the relationship among them.
The example of a univariate can be height.	The example of bivariate can be temperature and ice sales in summer vacation.	Example, Suppose an advertiser wants to compare the popularity of four advertisements on a website. Then their click rates could be measured for both men and women and relationships between variable can be examined

2.

The dataset below shows the gross earnings in pounds per week of twenty men and twenty women drawn randomly from the 1979 New Earnings Survey (see appendix to this chapter on the accompanying website). The respondents are all full-time adult workers. Men are deemed to be adult when they reach age 21, women when they reach age 18.

Men		*Women*	
150	58	90	39
55	122	76	47
82	120	87	80
107	83	58	42
102	115	50	40
78	69	46	99
154	99	63	77
85	94	68	67
123	144	116	49
66	55	60	54

Calculate the mean and standard deviation of the male earnings of the data. Compare them with the median and midspread you calculated. Why do they differ?

mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

mean = 98.05(men), 65.40(Female)

median

Odd Number of Observations

If the total number of observations given is odd, then the formula to calculate the median is:

$$Median = (\frac{n+1}{2})^{th}\ term$$

where n is the number of observations

Even Number of Observations

If the total number of observation is even, then the median formula is:

$$Median = \frac{(\frac{n}{2})^{th}\ term + (\frac{n}{2}+1)^{th}\ term}{2}$$

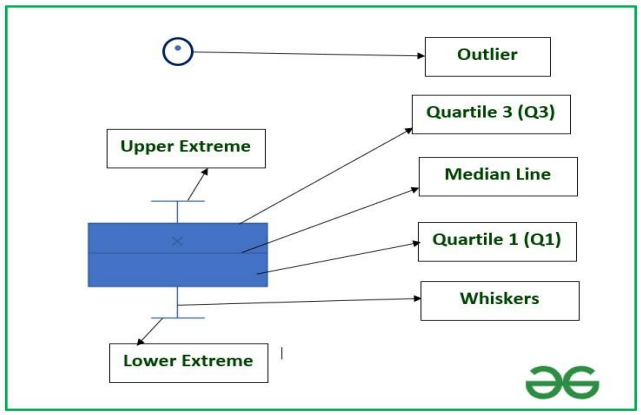
where n is the number of observations

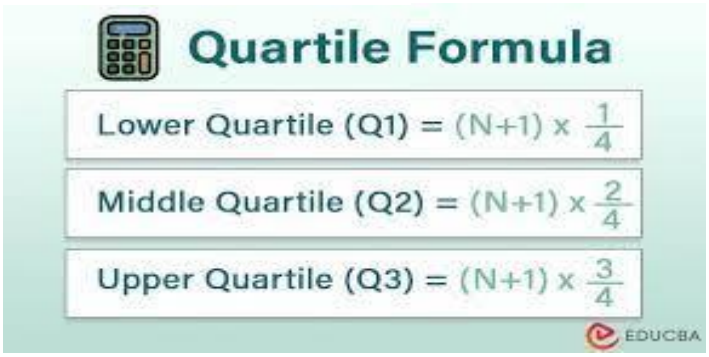
. the *standard deviation* to be

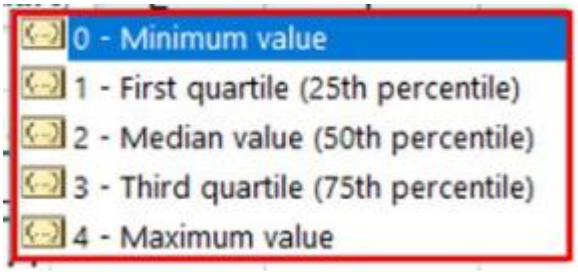
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

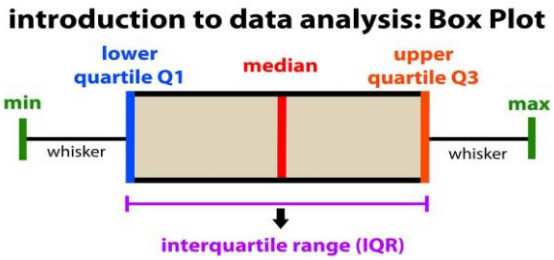
s.d. = 77.85,65.74

Inter Quartile Range









Q_i= [i * (n + 1) /4] th observation

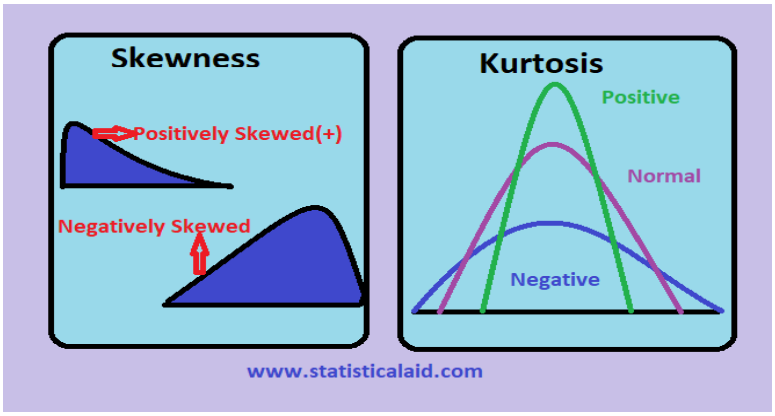
	men	women
Q0	55	39
Q1	75.75	48.5
Q2	96.5	61.5
Q3	120.5	77.75
Q4	154	116
IQR(Q3-Q1)	44.75	29.25

Midspread

Midspread =The difference between Q_L and Q_U

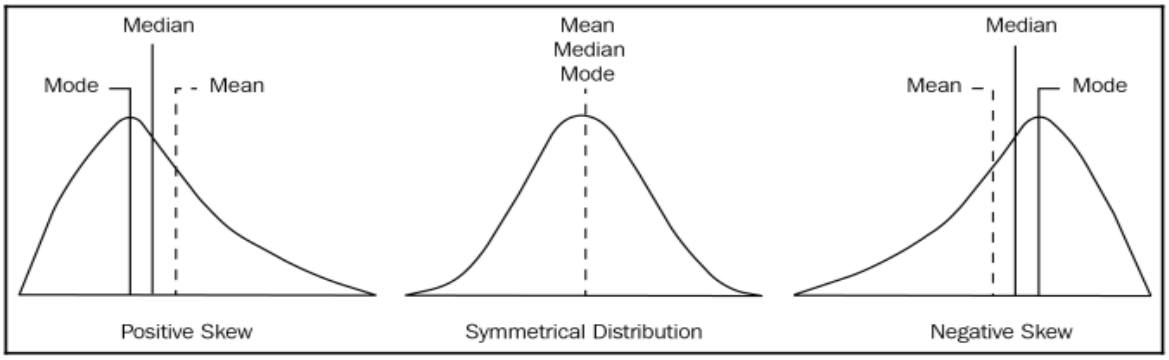
3. **What is measure of dispersion? Explain the following**

i) Skewness
ii) Kurtosis
iii) Calculating percentiles
iv) Quartiles



Skewness

In probability theory and statistics, skewness is a measure of the asymmetry of the variable in the dataset about its mean. The skewness value can be positive or negative, or undefined. The skewness value tells us whether the data is skewed or symmetric. Here's an illustration of a positively skewed dataset, symmetrical data, and some negatively skewed data:



Note the following observations from the preceding diagram:

- The graph on the right-hand side has a tail that is longer than the tail on the right-hand side. This indicates that the distribution of the data is skewed to the left. If you select any point in the left-hand longer tail, the mean is less than the mode. This condition is referred to as **negative skewness**.
- The graph on the left-hand side has a tail that is longer on the right-hand side. If you select any point on the right-hand tail, the mean value is greater than the mode. This condition is referred to as **positive skewness**.
- The graph in the middle has a right-hand tail that is the same as the left-hand tail. This condition is referred to as a **symmetrical condition**.

Kurtosis

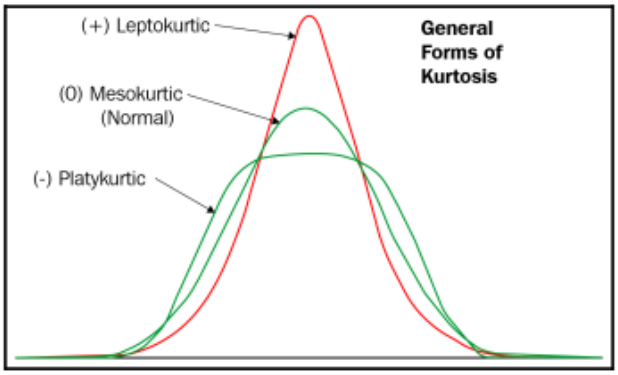
kurtosis is a statistical measure that illustrates how heavily the tails of distribution differ from those of a normal distribution. This technique can identify whether a given distribution contains extreme values.

Types of kurtosis

There are three types of kurtosis—mesokurtic, leptokurtic, and platykurtic. Let's look at these one by one:

- **Mesokurtic**: If any dataset follows a normal distribution, it follows a mesokurtic distribution. It has kurtosis around 0.
- **Leptokurtic**: In this case, the distribution has kurtosis greater than 3 and the fat tails indicate that the distribution produces more outliers.
- **Platykurtic:** In this case, the distribution has negative kurtosis and the tails are very thin compared to the normal distribution.

All three types of kurtosis are shown in the following diagram:



Calculating percentiles

Percentiles measure the percentage of values in any dataset that lie below a certain value. In order to calculate percentiles, we need to make sure our list is sorted. An example would be if you were to say that the 80th percentile of data is 130: then what does that mean? Well, it simply means that 80% of the values lie below 130. Pretty easy, right? We will use the following formula for this:

$$\text{The formula for calculating percentile of X} = \frac{\text{Number of values less than X}}{\text{Total number of observations}} * 100$$

Suppose we have the given data: 1, 2, 2, 3, 4, 5, 6, 7, 7, 8, 9, 10. Then the percentile value of 4 = (4/12) * 100 = 33.33%.

This simply means that 33.33% of the data is less than 4.

Quartiles

Given a dataset sorted in ascending order, quartiles are the values that split the given dataset into quarters. Quartiles refer to the three data points that divide the given dataset into four equal parts, such that each split makes 25% of the dataset. In terms of percentiles, the 25th percentile is referred to as the first quartile (Q1), the 50th percentile is referred to as the second quartile (Q2), and the 75th percentile is referred to as the third quartile (Q3).

Based on the quartile, there is another measure called inter-quartile range that also measures the variability in the dataset. It is defined as follows:

$$IQR = Q3 - Q1$$

Quartile Formula

The Quartile Formula For Q1 $= \frac{1}{4}(n+1)^{th}$ term

The Quartile Formula For Q3 $= \frac{3}{4}(n+1)^{th}$ term

The Quartile Formula For Q2 $=$ Q3 – Q1 (Equivalent to Median)

4. From the following given table find

i) Total percentage table

ii) Row percentage table

iii) Column percentage table

	a	b	c	d	Total
x	5	20	5	40	**70**
y	15	15	5	20	**55**
z	15	10	15	20	**60**
Total	**35**	**45**	**25**	**80**	**185**

total

	a	b	c	d	Total
x	2.70	10.81	2.70	21.62	37.84
y	8.11	8.11	2.70	10.81	29.73
z	8.11	5.41	8.11	10.81	32.43
Total	18.92	24.32	13.51	43.24	100.00

Row

	a	b	c	d	Total
x	7.14	28.57	7.14	57.14	100.00
y	27.27	21.43	9.09	36.36	100.00
z	25.00	14.29	25.00	33.33	100.00
Total	18.92	64.29	13.51	43.24	100.00

Coloumn

	a	b	c	d	Total
x	14.29	44.44	20.00	50.00	37.84
y	42.86	33.33	20.00	25.00	29.73
z	42.86	22.22	60.00	25.00	32.43
Total	100.00	100.00	100.00	100.00	100.00

5. Explain the various types and importance of bivariate analysis.

Bivariate analysis is an important statistical method because it lets researchers look at the relationship between two variables and determine their relationship. This can be helpful in many different kinds of research, such as social science, medicine, marketing, and more.

Here are some reasons why bivariate analysis is important:

- **Bivariate analysis helps identify trends and patterns:** It can reveal hidden data trends and patterns by evaluating the relationship between two variables.
- **Bivariate analysis helps identify cause and effect relationships:** It can assess if two variables are statistically associated, assisting researchers in establishing which variable causes the other.
- **It helps researchers make predictions:** It allows researchers to predict future results by modeling the link between two variables.
- **It helps inform decision-making:** Business, public policy, and healthcare decision-making can benefit from bivariate analysis.

The ability to analyze the correlation between two variables is crucial for making sound judgments, and this analysis serves this purpose admirably.

Types of bivariate analysis

Many kinds of bivariate analysis can be used to determine how two variables are related. Here are some of the most common types.

1. Scatterplots

A scatterplot is a graph that shows how two variables are related to each other. It shows the values of one variable on the x-axis and the values of the other variable on the y-axis.

The pattern shows what kind of relationship there is between the two variables and how strong it is.

2. Correlation

Correlation is a statistical measure that shows how strong and in what direction two variables are linked.

A positive correlation means that when one variable goes up, so does the other. A negative correlation shows that when one variable goes up, the other one goes down.

3. Regression

This kind of analysis gives you access to all terms for various instruments that can be used to identify potential relationships between your data points.

The equation for that curve or line can also be provided to you using regression analysis. Additionally, it may show you the correlation coefficient.

4. Chi-square test

The **chi-square test** is a statistical method for identifying disparities in one or more categories between what was expected and what was observed. The test's primary premise is to assess the actual data values to see what would be expected if the null hypothesis was valid.

Researchers use this statistical test to compare categorical variables within the same sample group. It also helps to validate or offer context for frequency counts.

5. T-test

A t-test is a statistical test that compares the means of two groups to see if they have a big difference. This analysis is appropriate when comparing the averages of two categories of a categorical variable.