INDEX

1. Applications of Various Types of Analysis
2. An Introduction to ANOVA
3. What does the Analysis of Variance Reveal
4. One Way ANOVA
5. Full Factorial ANOVA (two way ANOVA)
6. Why does ANOVA Work
7. Utility of ANOVA
8. ANOVA Terminology
9. Correlation Analysis
10. Regression Analysis
11. Difference b/w Correlation and Regression
12. Univariate Data
13. Bivariate Data
14. Multivariate Data
15. Difference b/w Univariate, Bivariate & Multivariate Descriptive Statistics

1. Applications of various types of analysis

The four types of data analysis are

a. Descriptive Analysis
b. Diagnostic Analysis
c. Predictive Analysis
d. Prescriptive Analysis

a. Descriptive Analysis: Its Applications include
    i. KPI dashboards
    ii. Monthly revenue reports
    iii. Sales leads overview

Applications of various types of analysis

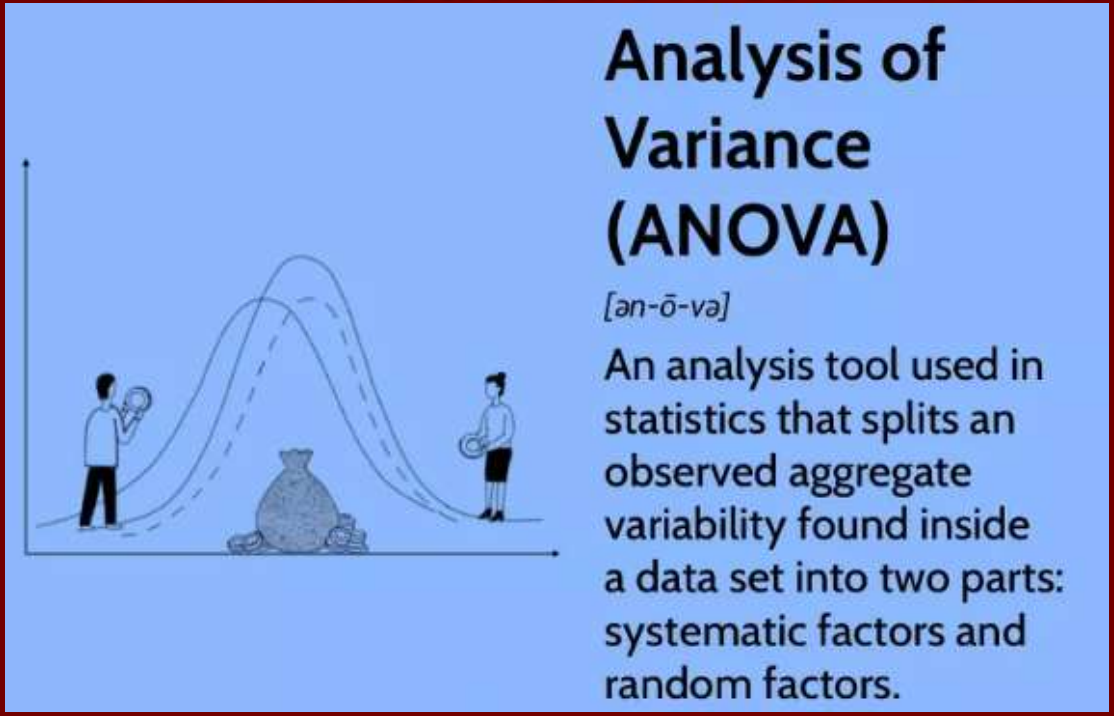
b. Diagnostic Analysis: Its Applications include

i. A freight company investigating the cause of slow shipments in a certain region
ii. A SaaS company drilling down to determine which marketing activities increased trials

c. Predictive Analysis: Its Applications include

i. Risk assessment
ii. Sales forecasting
iii. Using customer segmentation to determine which leads have the best chance of converting
iv. Predictive analytics in customer success teams

2. An introduction to ANOVA



Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

An introduction to ANOVA

The t-test and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method. **ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests.** The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers." It was employed in experimental psychology and later expanded to subjects that were more complex.

The formula for ANOVA is

$$F = \frac{MST}{MSE}$$

where:

$F = \text{ANOVA coefficient}$

$MST = \text{Mean sum of squares due to treatment}$

$MSE = \text{Mean sum of squares due to error}$

3. What does the analysis of variance reveal?

The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

4. One way ANOVA

The one-way analysis of variance is also known as single-factor ANOVA or simple ANOVA. The one-way ANOVA is suitable for experiments with only one independent variable (factor) with two or more levels. For instance a dependent variable may be what month of the year there are more flowers in the garden. There will be 12 levels.

Assumptions

- **Independence:** The value of the dependent variable for one observation is independent of the value of any other observations.
- **Normalcy:** The value of the dependent variable is normally distributed
- **Variance:** The variance is comparable in different experiment groups.
- **Continuous:** The dependent variable (number of flowers) is continuous and can be measured on a scale which can be subdivided.

5. Full factorial ANOVA (two-way ANOVA)

It is used when there are two or more independent variables. Each of these factors can have multiple levels. Full-factorial ANOVA can only be used in the case of a full factorial experiment, where there is use of every possible permutation of factors and their levels. This might be the month of the year, when there are more flowers in the garden, and then the number of sunshine hours. This two-way ANOVA not only measures the independent vs the independent variable, but if the two factors affect each other.

Assumptions

- **Continuous:** The same as a one-way ANOVA, the dependent variable should be continuous.
- **Independence:** Each sample is independent of other samples, with no crossover.
- **Variance:** The variance in data across the different groups is the same.
- **Normalcy:** The samples are representative of a normal population.
- **Categories:** The independent variables should be in separate categories or groups.

6. Why does ANOVA work

Some people question the need for ANOVA; after all, mean values can be assessed just by looking at them. But ANOVA does more than only comparing means. Even though the mean values of various groups appear to be different, this could be due to a sampling error rather than the effect of the independent variable on the dependent variable. If it is due to sampling error, the difference between the group means is meaningless. ANOVA helps to find out if the difference in the mean values is statistically significant.

ANOVA indirectly reveals if an independent variable is influencing the dependent variable. For example, in the blood sugar level experiment, suppose ANOVA finds that group means are not statistically significant, and the difference between group means is only due to sampling error. This result infers that the type of medication (independent variable) is not a significant factor that influences the blood sugar level.

Limitations

ANOVA can only tell if there is a significant difference between the means of at least two groups, but it can't explain which pair differs in their means. If there is a requirement for granular data, deploying further follow up statistical processes will assist in finding out which groups differ in mean value. Typically, ANOVA is used in combination with other methods.

ANOVA also makes assumptions that the dataset is uniformly distributed, as it compares means only. If the data is not distributed across a normal curve and there are outliers, then ANOVA is not the right process to interpret the data.

ANOVA assumes the standard deviations are the same or similar across groups. If there is a big difference in standard deviations, the conclusion of the test may be inaccurate.

7. Utility of ANOVA

ANOVA can help to

- **Compare the yield of two different wheat varieties under three different fertilizer brands.**
- **Compare the effectiveness of various social media advertisements on the sales of a particular product.**
- **Compare the effectiveness of different lubricants in different types of vehicles.**
- **ANOVA use in data science is in email spam detection. Because of the massive number of emails and email features, it has become very difficult & resource-intensive to identify and reject all spam emails. ANOVA and f-tests are deployed to identify features that were important to correctly identify which emails were spam and which not.**

Example of how to use ANOVA

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

Example of how to use ANOVA

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

8. ANOVA Terminology

Dependent variable: **This is the item being measured that is theorized to be affected by the independent variables.**

Independent variable/s: **These are the items being measured that may have an effect on the dependent variable.**

A null hypothesis (H0): **This is when there is no difference between the groups or means. Depending on the result of the ANOVA test, the null hypothesis will either be accepted or rejected.**

An alternative hypothesis (H1): **When it is theorized that there is a difference between groups and means.**

Factors and levels: **In ANOVA terminology, an independent variable is called a factor which affects the dependent variable. Level denotes the different values of the independent variable that are used in an experiment.**

Fixed-factor model: **Some experiments use only a discrete set of levels for factors. For example, a fixed-factor test would be testing three different dosages of a drug and not looking at any other dosages.**

Random-factor model: **This model draws a random value of level from all the possible values of the independent variable.**

9. Correlation Analysis

Economic and business variable are related. For instance, demand and supply of a commodity is related to its price. Demand for a commodity decreases as its price rises. We say demand and price are inversely related or negatively correlated. But sellers supply more of a commodity when its price rises. Supply of the commodity decreases when its price falls. We say supply and price are directly related or positively co-related. Thus, correlation indicates the relationship between two such variables in which categories in the value of one variable is accompanies with a change in the value of other variable.

According to **LR Connore**, "if two or more quantities very in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other(s) they are said to be correlated."

Correlation Analysis

WI King defined " Correlation means that between two series or groups of data, there exists some casual connection."

The definitions make it clear that the term Correlation refers to the study of relationship between two or more variables. **Correlation is a statistical device, which studies the relationship between two variables. If two variables are said to be correlated, change in the value of one variable result in a corresponding change in the value of other variable.** Heights and weights of a group of people, age of husbands and wives etc., are examples of bi-variant data that change together.

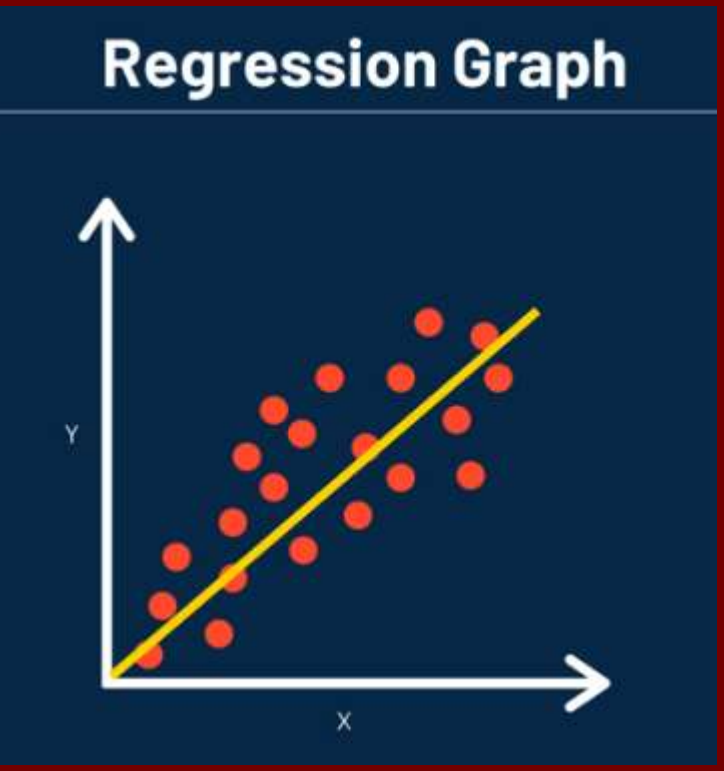
10. Regression Analysis

Regression is how one variable affects another or changes in a variable that trigger changes in another, essentially cause and effect. It implies that the outcome is dependent on one or more variables. For instance, while correlation can be defined as the relationship between two variables, regression is how they affect each other. An example of this would be how an increase in rainfall would then cause various crops to grow, just like a drought would cause crops to wither or not grow at all.

Regression analysis helps to determine the functional relationship between two variables (x and y) so that you're able to estimate the unknown variable to make future projections on events and goals.

Regression Analysis

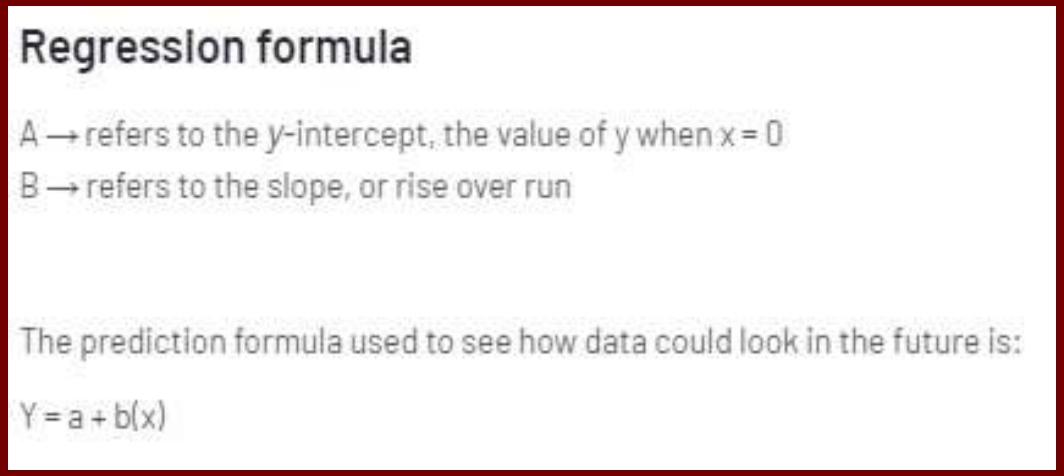
The main objective of regression analysis is to estimate the values of a random variable (z) based on the values of your known (or fixed) variables (x and y). Linear regression analysis is considered to be the best fitting line through the data points.



The main advantage in using regression within your analysis is that it provides you with a detailed look of your data (more detailed than correlation alone) and includes an equation that can be used for predicting and optimizing your data in the future.

Regression Analysis

When the line is drawn using regression, we can see two pieces of information



Example of regression: When it comes to using regression, we at G2 utilize regression to predict certain trends, like how our traffic is expected to grow over the coming months.

One person in particular who uses regression is our SEO and Data Analyst, Sarah Harenberg. Being able to visualize our data, analyze it, see trends, and predict what the data could look like in the future is a big part of her job. Many teams at G2 rely on Sarah when they set our team goals and to understand how our traffic could look in the coming months.

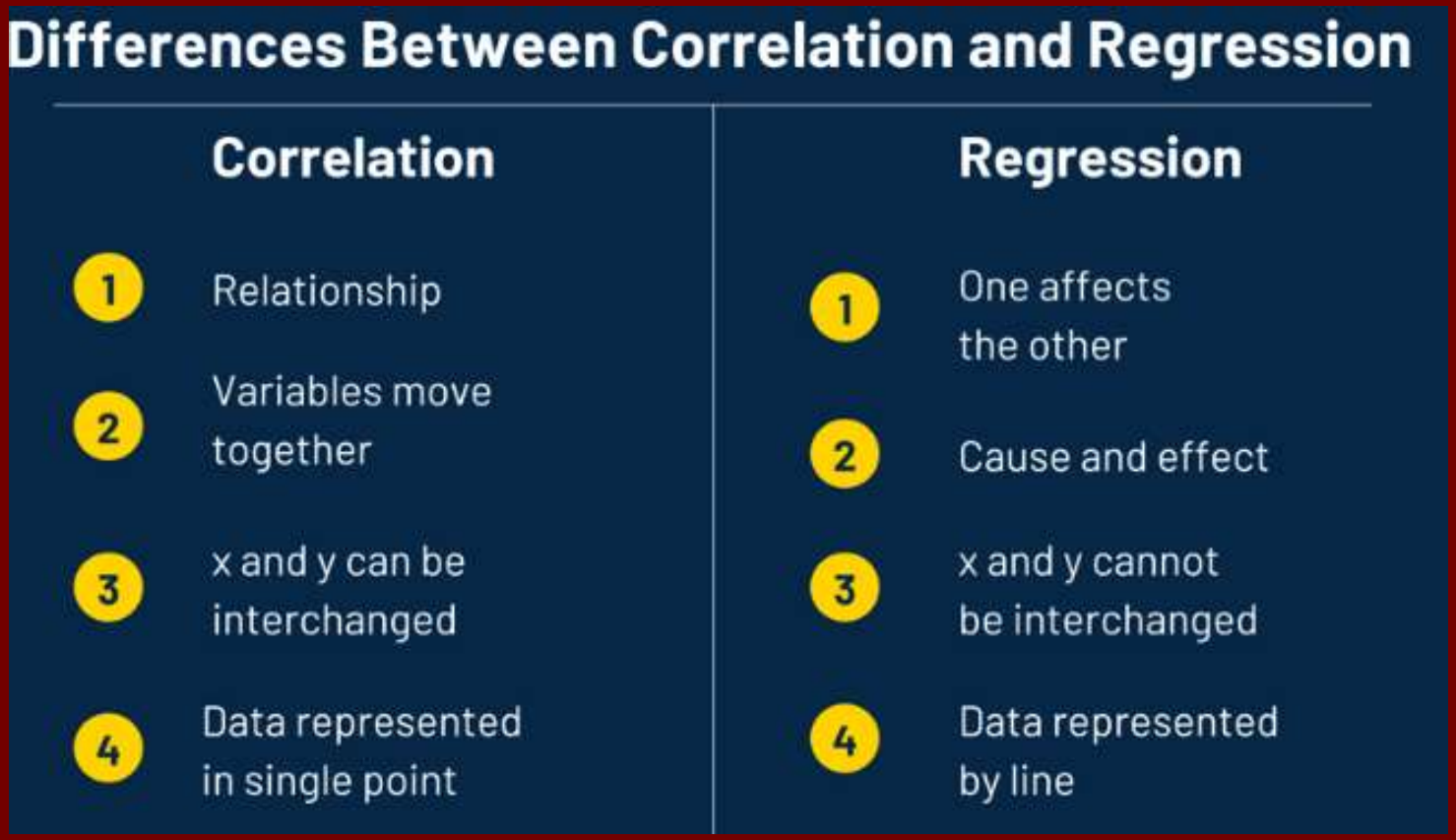
11. Differences between correlation and regression

There are some key differences between correlation and regression that are important in understanding the two.

- Regression establishes how x causes y to change, and the results will change if x and y are swapped. With correlation, x and y are variables that can be interchanged and get the same result.
- Correlation is a single statistic, or data point, whereas regression is the entire equation with all of the data points that are represented with a line.
- Correlation shows the relationship between the two variables, while regression allows us to see how one affects the other.

Differences between correlation and regression

- **The data shown with regression establishes a cause and effect, when one changes, so does the other, and not always in the same direction. With correlation, the variables move together.**



12. Univariate data

This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

Heights (in cm)	164	167.3	170	174.2	178	180	186

Suppose that the heights of seven students of a class is recorded(figure 1),there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

13. Bivariate data

This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables. Example of bivariate data can be temperature and ice cream sales in summer season.

TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

Bivariate data

Suppose the temperature and ice cream sales are the two variables of a bivariate data. Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.

Thus bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

14. Multivariate data

When the data involves three or more variables, it is categorized under multivariate. Example of this type of data is 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 variables can then be examined.

It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

15. Difference between univariate, bivariate and multivariate descriptive statistics

1. Univariate statistics summarize only one variable at a time.

2. Bivariate statistics compare two variables.

3. Multivariate statistics compare more than two variables.