# DATA ANALYTICS THEORY

**LAB 01**  Date: 11/03/2024

**1.data analysis principles**

Data analysis is the systematic process of inspecting, cleaning, transforming, and modeling data to discover meaningful patterns, relationships, and insights. It involves using statistical techniques, mathematical algorithms, and computational methods to extract valuable information from raw data. The goal of data analysis is to uncover hidden patterns, make predictions, and support decision-making based on evidence and empirical findings. It plays a crucial role in various fields, including scientific research, business intelligence, market research, and data-driven decision-making.

There are some principles of data analytics

1. Accuracy: Ensuring that the data you're analyzing is reliable and free from errors.

2. Relevance: Focusing on the data that is most relevant to your analysis goals.

3. Objectivity: Approaching the analysis without bias or preconceived notions.

4. Interpretation: Analyzing the data in a meaningful way to draw insights and conclusions.

5. Communication: Presenting the analysis findings clearly and effectively.

The principles help guide the process of data analysis and ensure that it is accurate, meaningful, and useful.

**2.Statistical analytics concept**

Statistical analytics in data analytics refers to the use of statistical methods and techniques to analyze and interpret data and draw meaningful conclusions.It's a fundamental concept in data analytics that helps in extracting valuable insights from data. It involves applying statistical models, hypothesis testing, regression analysis, and other statistical tools to uncover patterns, relationships, and insights within the data.

By using statistical analytics, analysts can make data-driven decisions, identify trends, and predict future outcomes based on the analysis of numerical data. It's a powerful tool in many fields, including business, finance, healthcare, and social sciences.

Some common statistical models used in data analytics include linear regression, logistic regression, decision trees, random forests, support vector machines (SVM), k-means clustering, and time series analysis. These models help in understanding relationships between variables, making predictions, identifying patterns, and grouping similar data points. Each model has its own strengths and is used based on the specific analysis goals and characteristics of the data.

**3. hypothesis**

A hypothesis is a statement or assumption that is made about the relationship between variables or the characteristics of a population. It serves as a starting point for analysis and is tested using statistical methods. The hypothesis can be either null (no relationship or difference) or alternative (there is a relationship or difference).

By testing the hypothesis with data, analysts can determine if there is enough evidence to support or reject the hypothesis. Hypothesis testing is an important part of data analytics as it helps in making informed decisions and drawing conclusions from data.

There are two types of hypotheses: the null hypothesis and the alternative hypothesis.

1. The null hypothesis, denoted as H0, states that there is no significant relationship or difference between variables or populations. It assumes that any observed differences are due to chance or random variation.

2. The alternative hypothesis, denoted as Ha or H1, proposes that there is a significant relationship or difference between variables or populations. It suggests that the observed differences are not due to chance alone. Hypothesis testing involves collecting and analyzing data to determine whether there is enough evidence to support or reject the null hypothesis in favor of the alternative hypothesis.

**4. Regression and its types**

Regression is a statistical technique used in data analytics to understand the relationship between a dependent variable and one or more independent variables. It helps in predicting the value of the dependent variable based on the values of the independent variables. There are several types of regression, including:

1. Linear Regression: This is the most common type of regression, where the relationship between the dependent variable and independent variable(s) is assumed to be linear. It aims to find the best-fit line that minimizes the difference between the observed and predicted values.

Let's say we want to predict a student's exam score based on the number of hours they studied. Linear regression would help us find the line that best fits the data points, allowing us to predict the exam score for a given number of study hours.

2. Logistic Regression: Logistic regression is used when the dependent variable is categorical or binary. It predicts the probability of an event occurring based on the values of the independent variables.

Suppose we want to predict whether a customer will churn or not based on their demographic information. Logistic regression can help us calculate the probability of churn based on variables like age, gender, and income.

3. Polynomial Regression: Polynomial regression is an extension of linear regression where the relationship between the dependent variable and independent variable(s) is modeled using higher-degree polynomial equations.

Imagine we have data on the number of years of experience and corresponding salary for a group of employees. Polynomial regression can help us model a curve that fits the data points, allowing us to predict salary based on years of experience.

4. Ridge Regression: Ridge regression is a regularization technique used to handle multicollinearity (high correlation) among independent variables. It adds a penalty term to the regression equation to reduce the impact of multicollinearity.

Let's say we have a dataset with highly correlated independent variables, such as height and weight. Ridge regression can help us handle this multicollinearity issue by adding a penalty term, allowing us to make more accurate predictions.

5. Lasso Regression: Lasso regression is another regularization technique that not only handles multicollinearity but also performs feature selection by shrinking the coefficients of less important variables to zero.

Suppose we want to predict the price of a house based on various features like square footage, number of bedrooms, and location. Lasso regression can help us select the most important features and shrink the coefficients of less important variables, improving the model's predictive power.

**5. Correlation**

Correlation in data analytics is like a way to see if two things are connected or related. It helps us understand if there's a pattern between two or more things. It refers to the statistical relationship between two or more variables. It helps us understand how changes in one variable are associated with changes in another variable. Correlation is measured using a correlation coefficient, which ranges from -1 to 1.

A correlation coefficient of 0 suggests no linear relationship between the variables.

When we say there's a positive correlation, it means that when one thing goes up, the other thing also tends to go up. For example, if you study more, your exam scores might also go up.A positive correlation (between 0 and 1) indicates that as one variable increases, the other variable tends to increase as well. For example, there might be a positive correlation between the amount of studying done and exam scores.

A negative correlation means that when one thing goes up, the other thing tends to go down. For instance, if you spend more time watching TV, your physical fitness level might go down.A negative correlation (between -1 and 0) indicates that as one variable increases, the other variable tends to decrease. For instance, there might be a negative correlation between the number of hours spent watching TV and physical fitness level.

But correlation doesn't always mean that one thing causes the other. It just shows a relationship between them. We need more analysis to figure out if there's a cause-and-effect relationship.

Correlation is helpful in many areas, like figuring out trends in economics, understanding relationships in social sciences, or making predictions in marketing. It's a cool tool to see how things are connected!

**6. Anova**

ANOVA, or analysis of variance, is a statistical technique used to compare the means of three or more groups. It helps us determine if there are any significant differences between the groups.

Imagine you have different groups of people and you want to know if there is a difference in their average heights. ANOVA can tell you if the differences you observe are statistically significant or just due to chance.

It does this by partitioning the total variation in the data into two components: the variation between the groups and the variation within the groups. The variation between the groups is compared to the variation within the groups to determine if there is a significant difference in the means.

By analyzing the variance between groups and within groups, ANOVA helps us understand if there is a significant difference in the means of the groups. It's commonly used in research, psychology, and other fields to compare multiple groups and draw conclusions.

ANOVA calculates an F-statistic, which is the ratio of the between-group variation to the within-group variation. If the F-statistic is large enough, it indicates that the group means are significantly different.

ANOVA is commonly used in experimental studies and research to compare the effects of different treatments or interventions on a dependent variable. It helps researchers determine if there is a significant effect of the independent variable(s) on the outcome variable.

The basic formula for calculating the F-statistic in ANOVA is:

F = (Between-group variation / Within-group variation)

To calculate the between-group variation, sum of squares between (SSB), which measures the differences between the group means. The within-group variation is calculated using the sum of squares within (SSW), which measures the variability within each group.

The sum of squares is obtained by summing the squared differences between each observation and the group mean. Then, these sums of squares are divided by their respective degrees of freedom (DFB and DFW) to calculate the mean squares.

Finally, the F-statistic is calculated by dividing the mean square between by the mean square within.