ANalysis Of Variance “ANOVA” Statistical Test

A guide for choosing the most common statistical tests

Q1 Bivariate / Multivariable	Q2 Difference / Correlation	Q3 Independent / Paired	Q4 Type of outcome (and Normality)	Q5 No of groups (conditions)	Statistical test
Bivariate	Difference	Independent (un-paired)	Continuous (Normal)	2	Independent samples t-test (Student)
				>2	One-way ANOVA
			Continuous (Non-normal)/ Ordinal	2	Mann-Whitney U test
				>2	Kruskal-Wallis H test
			Nominal	2	Chi-square test/ Fisher's exact test
				>2	Chi-square test
			Time to Event (survival)		Log-Rank test (Kaplan-Meier plot)
		Dependent (paired)	Continuous (Normal)	2	Paired t-test
				>2	Repeated measured ANOVA
			Continuous (Non-normal)/ Ordinal	2	Wilcoxon signed-rank test
				>2	Friedman test
			Nominal	2	McNemar's test
	Correlation		Continuous (Normal)		Pearson's correlation (r)
			Continuous (Non-normal)/ Ordinal		Spearman's correlation (ρ)/ Kendall's Tau
Multivariable/ Prediction			Continuous		Linear Regression
			Ordinal		Ordered Logistic Regression
			Nominal	(2 levels)	Binary Logistic Regression
			Nominal	(>2 levels)	Multinomial Logistic Regression
			Time to Event (survival)		Cox Regression
			Count variable/rate		Poisson regression

ANOVA	
ONEWAY	REPEATED MEASURED
Bivariat	
Difference	
Independent Variables	Dependent Variables
>2	
Continuous Normal	

Real Life EX For:

One-Way ANOVA

Let's say you are a researcher studying the impact of different teaching methods on student performance in a particular subject. You have three different teaching methods: Method A, Method B, and Method C. You randomly assign students to these three groups and provide instruction using the respective methods over a semester.
At the end of the semester, you administer the same exam to all students and collect their scores. Now, you want to analyze whether there are any significant differences in the mean scores among the three teaching methods using one-way ANOVA.
Here, the independent variable is the teaching method (with three levels: Method A, Method B, and Method C), and the dependent variable is the exam score. By conducting a one-way ANOVA, you can determine if there are any significant differences in the mean scores between the three teaching methods.
The null hypothesis for the one-way ANOVA would state that there are no significant differences in the mean exam scores among the three teaching methods, while the alternative hypothesis would suggest that at least one teaching method differs from the others in terms of mean exam scores.
By performing the one-way ANOVA analysis and examining the resulting F-value and p-value, you can assess whether there is enough evidence to reject the null hypothesis and conclude that there are significant differences in student performance based on the teaching methods employed.

Repeated Measure ANOVA

Imagine you are a psychologist studying the effects of a new therapy on anxiety levels in individuals. You have a group of participants who suffer from anxiety, and you want to investigate how their anxiety levels change over time during different phases of the therapy.
To conduct this study, you measure the anxiety levels of each participant at three different time points: before the therapy (baseline), halfway through the therapy, and at the end of the therapy. You administer a standardized anxiety questionnaire to each participant at each time point and record their anxiety scores.
In this scenario, repeated measures ANOVA would be suitable for analyzing the data. The independent variable is the time points (with three levels: baseline, halfway, and end), and the dependent variable is the anxiety score.
By performing repeated measures ANOVA, you can examine if there are significant differences in anxiety scores across the three time points. The advantage of using a repeated measures design is that it allows you to control for individual differences by measuring the same participants at multiple time points, reducing the impact of inter-individual variability.
The null hypothesis for the repeated measures ANOVA would state that there are no significant differences in the mean anxiety scores across the three time points, while the alternative hypothesis would suggest that at least one time point differs from the others in terms of mean anxiety scores.
By analyzing the data with repeated measures ANOVA and assessing the resulting F-value and p-value, you can determine if there is enough evidence to reject the null hypothesis and conclude that there are significant changes in anxiety levels over time during the therapy.

Task6

Gradient Descent vs Stochastic Gradient Descent

SGD :Stochastic Gradient Descent
BGD: Batch Gradient Descent
MBGD: Mini Batch Gradient Descent

Certainly! Here's a table comparing stochastic gradient descent (SGD), batch gradient descent (BGD), and mini-batch gradient descent (MBGD) based on various factors commonly considered in machine learning:

Factor	SGD	BGD	MBGD
Update Frequency	After each training example	After processing entire dataset	After processing a mini-batch of examples
Computational Efficiency	Fastest iteration time per update	Slowest iteration time per update	Faster than BGD, slower than SGD
Variance in Parameter Updates	High	None	Moderate
Convergence Speed	Fast convergence due to frequent updates	Slower convergence	Faster than BGD, slower than SGD
Memory Requirement	Low	High	Moderate
Utilization of Hardware Resources	Low	High	Moderate
Suitable for Large Datasets	Yes	No	Yes
Suitable for Small Datasets	No	Yes	Yes
Noise in Updates	Yes	No	Moderate

The main difference between GD and SGD is

NonStatistical ML Models

In machine learning, non-statistical models refer to approaches that do not rely heavily on statistical assumptions or techniques. These models are often based on alternative methodologies or concepts. Here are a few examples of non-statistical models in machine learning:

1. Rule-based models: These models make decisions based on a set of predefined rules. They use logical statements and conditions to determine the outcome. Rule-based models are transparent and interpretable, making them useful in domains where explainability is important.

2. Decision trees: A decision tree is a flowchart-like structure where internal nodes represent features, branches represent decisions, and leaf nodes represent the outcomes. These models make predictions by traversing the tree based on the values of input features. Decision trees can be easily visualized and are widely used in classification and regression tasks.

3. Random forests: Random forests are an ensemble learning method that combines multiple decision trees. Each tree in the forest is trained on a random subset of the training data and features. The final prediction is made by aggregating the predictions of individual trees. Random forests are known for their high accuracy and robustness.

4. Neural networks: Neural networks, particularly deep learning models, have gained significant popularity in recent years. They are composed of interconnected layers of artificial neurons that can learn complex patterns and relationships from data. Neural networks are capable of automatically extracting features and are used in various tasks, including image and speech recognition.

5. Support Vector Machines (SVM): SVM is a non-probabilistic, non-statistical machine learning model that classifies data by finding the best hyperplane that separates the classes. SVM aims to maximize the margin between the classes, allowing for better generalization to unseen data. It has been widely used in both classification and regression tasks.

6. Genetic algorithms: Genetic algorithms are a type of evolutionary computation that mimics the process of natural selection. They use principles from genetics to search for optimal solutions in a large search space. Genetic algorithms are commonly employed in optimization problems and feature selection tasks.

7. Reinforcement learning models: Reinforcement learning is a branch of machine learning where an agent learns to make decisions by interacting with an environment. These models use a reward-based approach, where the agent learns through trial and error to maximize cumulative rewards. Reinforcement learning has been successfully applied in areas like game playing and robotics.

These are just a few examples of non-statistical models in machine learning. There are various other models and techniques that fall into this category, each with its own strengths and weaknesses. The choice of model depends on the problem at hand, the available data, and the desired outcomes.