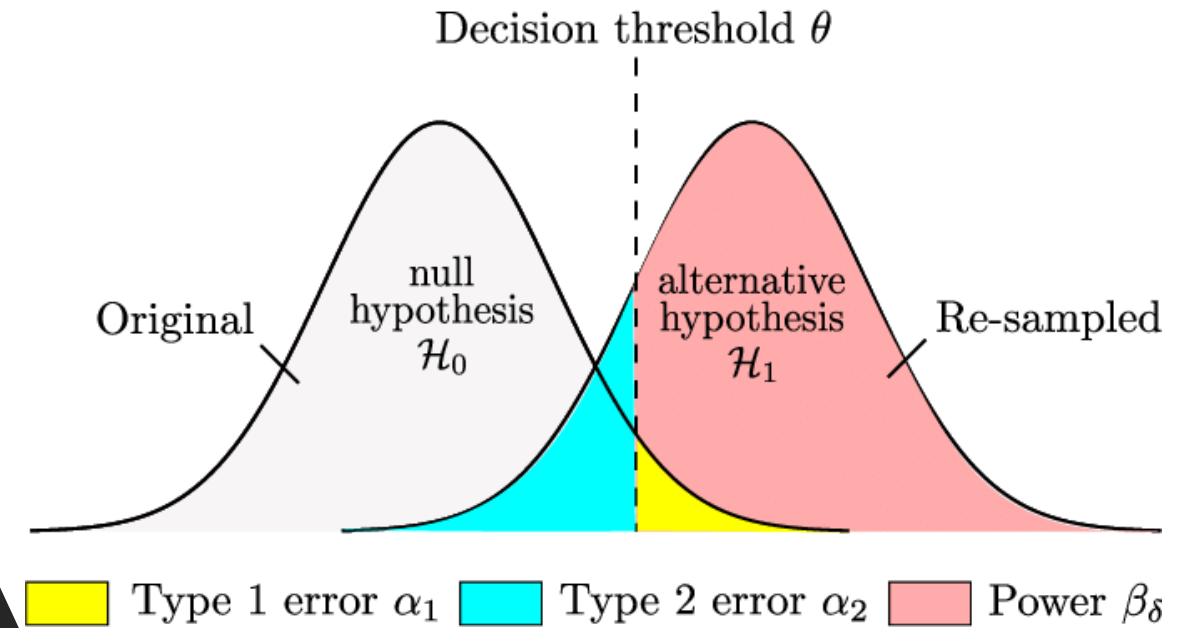


INTRODUCTION TO MACHINE LEARNING

DATA 602

Lecture 2

Hypothesis Testing: Key Concepts



What is Hypothesis Testing?

- It is about evaluating an **assumption** on a **population parameter**

Some of the Key Concepts

- The **Null Hypothesis H_0** is a general statement or default position that there is no relationship between two measured phenomena, or no association between groups
- The **Alternative Hypothesis H_1** is the hypothesis contrary to the null hypothesis
- The **level of significance** refers to the **degree of significance** in which we **accept** or **reject** the **null hypothesis**
- The **p-value** is the **probability** of **finding the observed results** when the **null hypothesis is true**
- **Type I Error** happens when we **reject the null hypothesis** when it is **true**. It is denoted by alpha (α)
- **Type II Error** occurs when we **accept the null hypothesis** when it is **false**. It is denoted by beta (β)
- This lecture focuses mainly on **causal inference**, which includes
 - Randomized control,
 - Synthetic control, and
 - Time-series forecasting as the object of another lecture

Hypothesis Testing (Cont.)

Example of Hypothesis Testing

- The t-value is used to determine a p-value (probability)
- A p-value ≤ 0.05 signifies strong evidence against the Null Hypothesis H_0 at a 95% level ($\alpha = 0.05$). Therefore, we reject the Null Hypothesis
- A p-value > 0.05 signifies weak evidence against the Null Hypothesis. As a result, we accept the Alternative Hypothesis H_1
- The process of testing a hypothesis indicates the possibility of making an error
- There are two types of errors, which are inversely related: the smaller the risk of one, the higher the risk of the other
- If we reject H_0 when it is true, then we have a Type I error
- If we accept H_0 when it is false, then we have a Type II error

Decision is:	The Null Hypothesis is	
	True	False
Accept H_0	(1- α) Confidence Level	β
Reject H_0	α	(1- β) Power of the test

		Conclusion about null hypothesis from statistical test	
		Accept Null	Reject Null
Truth about null hypothesis in population	True	Correct	Type I error Observe difference when none exists
	False	Type II error Fail to observe difference when one exists	Correct

Why Do We Need to Test Hypotheses?

- We need to determine causal inference before making a decision

How Does Machine Learning Help Make Decisions?

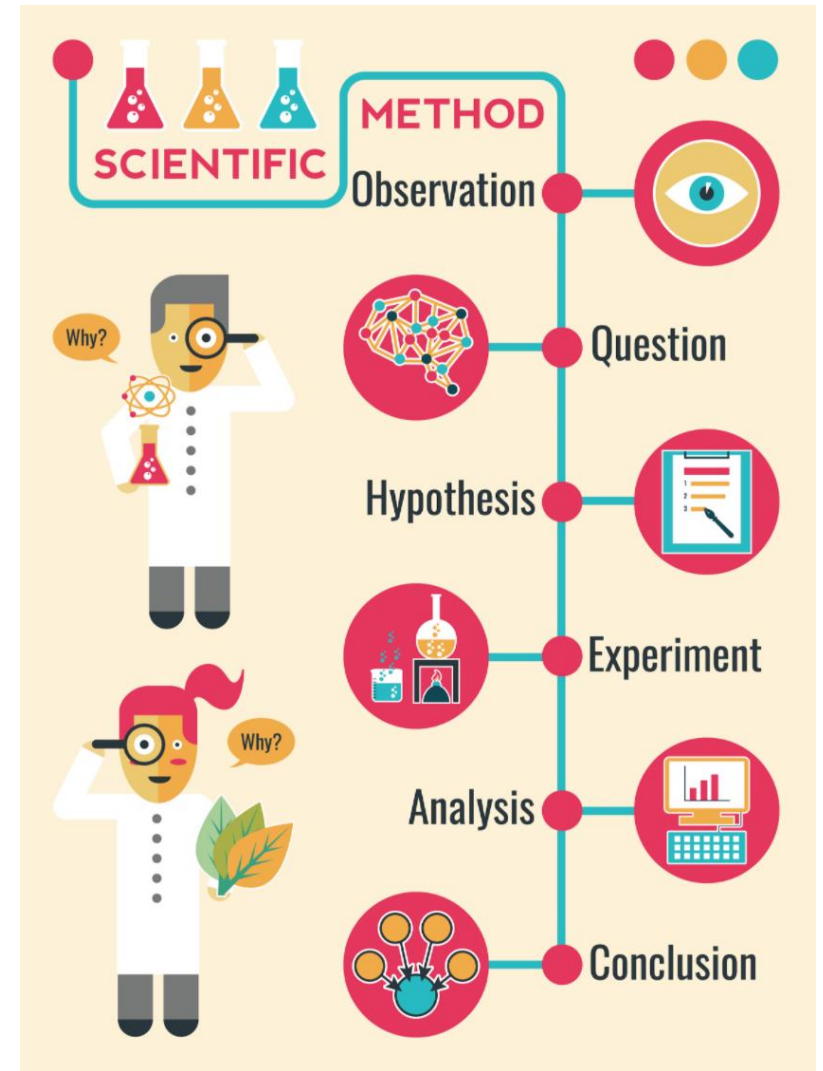
- We go through seven steps detailed below



Traditional Decision-Making Process

Source: UMASS Dartmouth

Scientific Process



What is Causal Inferences?

- **Causal inference** refers to an intellectual discipline that considers the **assumptions**, **study designs**, and **estimation strategies** that allow researchers to draw **causal conclusions** based on data
 - Understanding the relationship between causes and effects
 - Identifying the causes of observations
 - Using the outcomes of carefully designed experiments to determine the cause of a decision
 - Using naturally available data to extract experiments
- In science, we constantly need to accept or reject hypotheses to reach conclusions
- This explains why **causal inference** is a necessity if we want to reach **valid conclusions**

Scientific Process

- Make an observation
- Ask a question
- Form a hypothesis, or testable explanation
- Make a prediction based on the hypothesis
- Test the prediction
- Iterate: use the results to make new hypotheses or predictions

- When we consider **scientific theory** as **hypothesis**
 - Scientific theory explains the phenomena
 - The phenomena is supported by experiments

- **Falsifiability** is the cornerstone of scientific method:
 - All experiments prove the theory → the theory is valid
 - One experiment disproves the theory → the theory is invalid

- **Randomization**

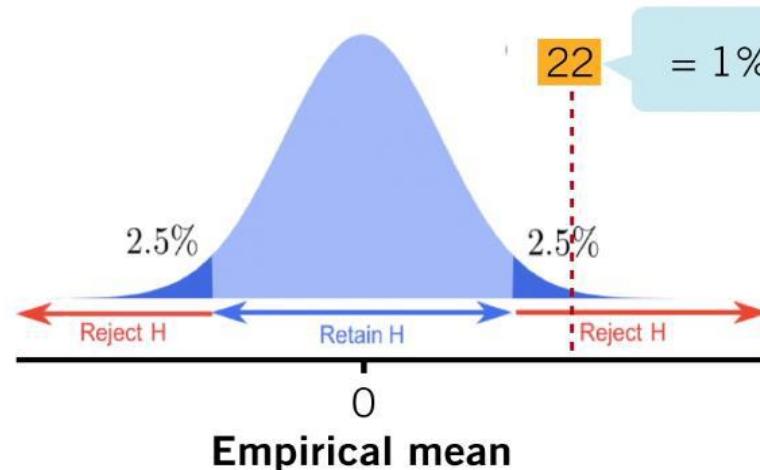
- It takes care of **the confounding factors** and stabilizes the experiment
 - Let's say we determine that all bald men are beer drinkers. The confounding factor is **gender bias**
 - In a drug test, the confounding factor can be **patient selection**. We will use examples later to illustrate applications of ML to experiments
- In a drug test experiments, for instance,
 - We select a set of patients
 - We create two sub-sets of patients for respective administration of the two drugs
 - We observe patient performance, and
 - We make a decision


- **Synthetic Control**

- When it is difficult to control an experiment, **synthetic control** allows to evaluate the effect of an intervention in comparative case studies
- This method can account for the effects of confounders changing over time, by weighting the control group to better match the treatment group before the intervention
- It allows researchers to systematically select comparison groups, especially in the case of political science, health policy, and criminology, among others

Example of Drug Testing Experiment and Hypothesis Validation

- If in a drug experiment, we have an existing drug provided to Group 1 and a new drug provided to Group 2
 - We compute health measurement for both groups
 - We compute average health measurement of all patients (say A1 for Group 1 and A2 for Group 2),
 - Compute $A1 - A2$
 - Compute standard deviation (SD) across all patients, and
 - Compute $(A1 - A2)/SD$
- The basic hypothesis is the two drugs used in Group 1 and Group 2 have the **same** performance → the goal is to reject the basic hypothesis
 - Group 1 mean – Group 2 mean = 0 → drugs are the same. Both drugs are not different.
 - Group 1 mean – Group 2 mean \neq 0 → drugs are different, and one drug is better than the other



- If $\frac{G1 \text{ mean} - G2 \text{ mean}}{SD} = 22$
- Area below the curve at 22 = 1%
- p -value = 0.01
- p -value significance: Reject the hypothesis
-  Basic (null) hypothesis
Drug 1 performance = Drug 2 performance
- Smaller the p -value, larger the significance

Source: MIT

- When testing your hypothesis, you may have to decide if your statistical test should be a one-tailed test or a two-tailed test
- A one-tailed test is appropriate if you only want to determine if there is a difference between groups in a specific direction
- So, if you are only interested in determining if Group A scored higher than Group B, and you are completely uninterested in possibility of Group A scoring lower than Group B, then you may want to use a one-tailed test

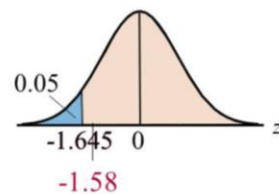
Solution: Testing with Rejection Regions

- $H_0: \mu = \$45,000$

- $H_a: \mu < \$45,000$

- $\alpha = 0.05$

- Rejection Region:



- Test Statistic

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{43,500 - 45,000}{5200 / \sqrt{30}} = -1.58$$

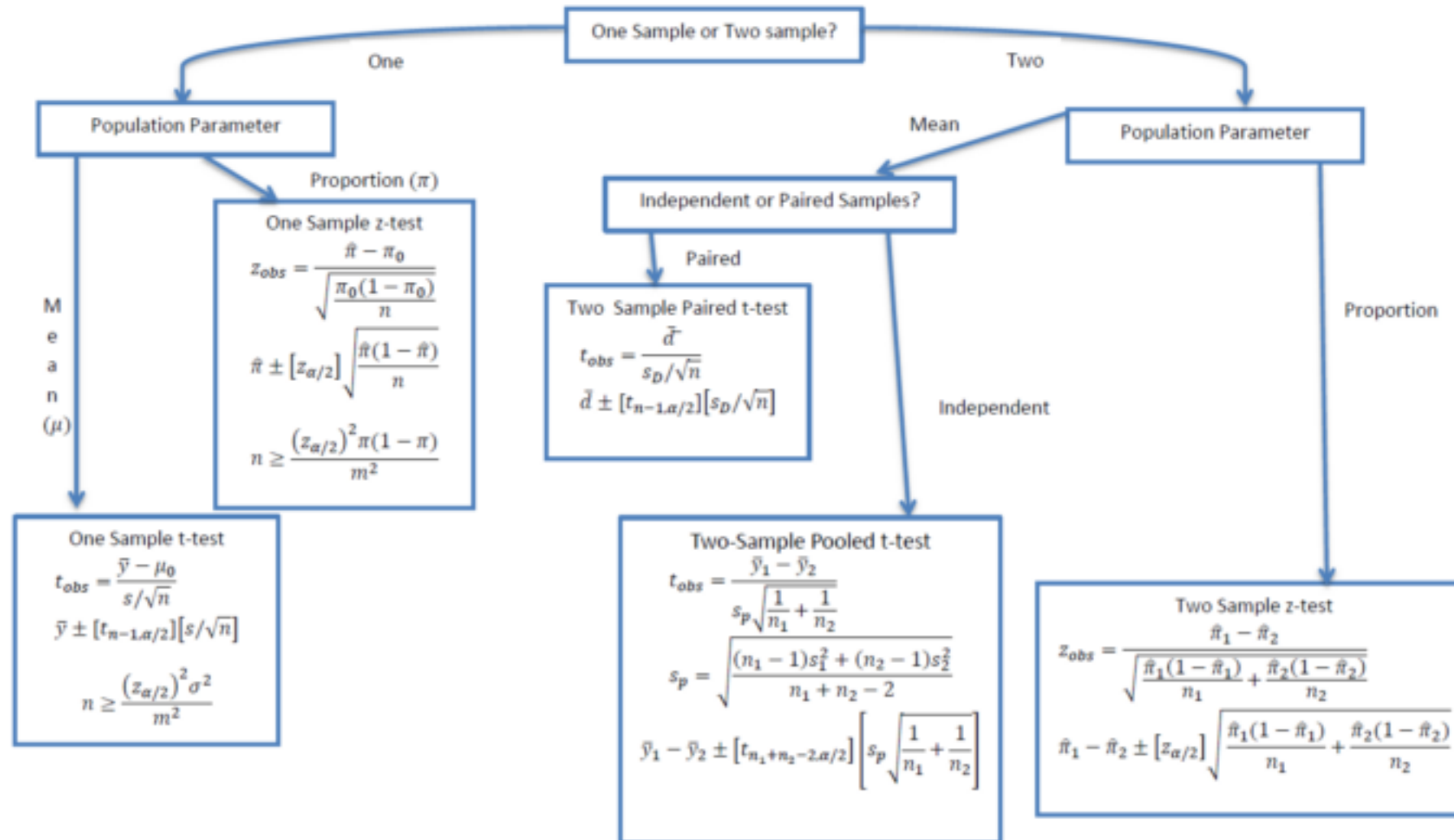
- Decision: **Fail to reject H_0**

Since the p-value falls outside the rejection region, we fail to reject the null hypothesis and can say that there is not sufficient evidence to support the employees' claim that the mean salary is less than \$45,000.

		Conclusion about null hypothesis from statistical test	
		Accept Null	Reject Null
Truth about null hypothesis in population	True	Correct	Type I error Observe difference when none exists
	False	Type II error Fail to observe difference when one exists	Correct

- $CI = \text{Point estimate} \pm \text{Margin of error}$
- $CI = \text{Sample mean} \pm z \text{ value} \times \text{Standard error of the mean (SEM)}$
- $CI = \text{Sample mean} \pm z \text{ value} \times (\text{Standard deviation} / \sqrt{n})$
- If we are calculating the 95% CI of the mean, the z value to be used would be 1.96.
- A 95% CI does not mean that 95% of the sample data lie within that interval
- A CI is not a range of plausible values for the sample. Rather, it is an interval estimate of plausible values for the population parameter

```
import numpy as np
from scipy import stats
N = 10000
a = np.random.normal(0, 1, N)
mean, sigma = a.mean(), a.std(ddof=1)
conf_int_a = stats.norm.interval(0.68, loc=mean, scale=sigma)
print('{:0.2%} of the single draws are in conf_int_a'
      .format(((a >= conf_int_a[0]) & (a < conf_int_a[1])).sum() / float(N)))
M = 1000
b = np.random.normal(0, 1, (N, M)).mean(axis=1)
conf_int_b = stats.norm.interval(0.68, loc=0, scale=1 / np.sqrt(M))
print('{:0.2%} of the means are in conf_int_b'
      .format(((b >= conf_int_b[0]) & (b < conf_int_b[1])).sum() / float(N)))
```



Experimental Designs



Experimental Design

- Experimental design refers to **how participants are allocated to the different conditions** (or Independent Variable levels) in an experiment
- The most common way to design an experiment is to divide the participants into two groups:
 - The **experimental group**, and
 - The **control group**
 - Then, we introduce a change to the experimental group and not the control group
- There are three types of experimental designs:
 - **1. Independent Measures:** This type of design is also known as **between groups**
 - Different participants are used in each condition of the independent variable
 - This means that each condition of the experiment includes a different group of participants
 - This should be done by random allocation, which ensures that each participant has an equal chance of being assigned to one group or the other
 - Independent measures involve using two separate groups of participants; one in each condition

Experimental Design

- **2. Repeated Measures:** This type of design is also known as **within groups**
 - The same participants take part in each condition of the independent variable
 - This means that each condition of the experiment includes the same group of participants
- **3. Matched Pairs:** Each condition uses different but similar participants
 - An effort is made to **match the participants** in each condition in terms of any important characteristic which might affect performance, e.g., gender, age, intelligence, etc.
 - One member of each matched pair must be randomly assigned to the experimental group and the other to the control group

Suggested Reading on Experimental Designs

<https://www.statisticshowto.datasciencecentral.com/experimental-design/>

- **Ecological Validity:** The degree to which an investigation represents real-life experiences
- **Experimenter Effects:** The ways an experimenter can accidentally influence the participant through their appearance or behavior
- **Demand Characteristics:** The clues in an experiment that lead the participants to think they know what the researcher is looking for (e.g. experimenter's body language)
- **Independent Variable (IV):** Variable the experimenter manipulates (i.e. changes) – assumed to have a direct effect on the dependent variable
- **Dependent Variable (DV):** Variable the experimenter measures. This is the outcome (i.e. result) of a study
- **Extraneous Variables (EV):** All variables, which are not the independent variable, but could affect the results (DV) of the experiment. EVs should be controlled where possible
- **Confounding Variables (CV):** Variable(s) that have affected the results (DV), apart from the IV. A confounding variable could be an extraneous variable that has not been controlled

- **Random Allocation:** Randomly allocating participants to independent variable conditions means that all participants should have an **equal chance** of taking part in each condition. The principle of random allocation is to **avoid bias** in the way the experiment is carried out and to **limit the effects** of participant variables
- **Order Effects:** Changes in participants' performance due to their repeating the same or similar test more than once. They include:
 - **Practice Effect:** an improvement in performance on a task due to repetition, for example, because of familiarity with the task
 - **Fatigue Effect:** a decrease in performance of a task due to repetition, for example, because of boredom or tiredness

ANOVA Test

- An ANOVA test is a way to find out if survey or experiment results are significant
- It helps determine whether we need to reject the null hypothesis or accept the alternate hypothesis
- We are testing whether there is a difference between groups
- **One-way or two-way** refers to the number of independent variables (IVs) in ANOVA test
 - **One-way** has **one independent variable** (with **2 levels**), and
 - **Two-way** has **two independent variables** (and it can have **multiple levels**)
 - A one-way Analysis of Variance could have one IV (brand of cereal) and a two-way Analysis of Variance has two IVs (brand of cereal and calories)
- **Groups or levels** are different groups in the same independent variable. In the above example, your levels for “brand of cereal” might be Lucky Charms, Raisin Bran, Cornflakes — a total of three levels. Your levels for “calories” might be sweetened, unsweetened — a total of two levels

ANOVA Test (Cont.)

- **Nested ANOVA** implies a hierarchy within groups
- **Replication** means whether we are replicating the test(s) with multiple groups
- A **Student's t-test** will tell us if there is a **significant variation between groups**
- A **t-test compares means**, while the ANOVA compares **variances between populations**
- As groups grow in number, paired comparisons become unmanageable
- ANOVA will provide a single number (the **F-statistic**) and **one p-value** to help you support or reject the null hypothesis

Types of ANOVA and Hypothesis Testing

There are two main types: one-way and two-way. Two-way tests can be with or without replication.

- **One-way ANOVA between groups:** used when you want to test **two groups** to see if there is a difference between them
- **Two-way ANOVA without replication:** used when you have **one group** and you are **double-testing** that same group. For example, you are testing one set of individuals before and after they take a medication to see if it works or not
- **Two-way ANOVA with replication:** there are **two groups** and the members of those groups are **doing more than one thing**. For example, two groups of patients from different hospitals trying two different therapies

Types of ANOVA and Hypothesis Testing (Cont.)

- **Two null hypotheses** are tested if we are placing one observation in each cell. For this example, those hypotheses would be:
 - H_{01} : All the income groups have equal mean stress
 - H_{02} : All the gender groups have equal mean stress
- **For multiple observations** in cells, we would also be testing a third hypothesis:
 - H_{03} : The factors are independent, *or* the interaction effect does not exist
- An **F-statistic** is computed for each hypothesis you are testing

MANOVA

- Suppose you wanted to find out whether a difference in textbooks affected students' scores in math *and* science. Improvements in math *and* science means that there are **two** dependent variables. Therefore, a **MANOVA** is appropriate
- An **ANOVA** will give us a single (“univariate”) F-value, while a MANOVA will provide a multivariate F-value
- **MANOVA** tests the **multiple dependent variables** by creating new, artificial, dependent variables that maximize group differences
- These new dependent variables are linear combinations of the measured dependent variables
- **Interpreting the MANOVA results**
 - If the multivariate F-value indicates the test is statistically significant, this means math scores may have improved, or science scores, or both
 - Once we have a significant result, we would need to examine each individual component (the univariate F tests) to see which dependent variable(s) contributed to the statistically significant result

- **Time-Series Forecasting** is appropriate when
 - Causality is closely-related to time as an important factor in prediction and forecasting
 - Assuming the history of time-series is related to the future, we are interested in determining the amount of change in value in the next time step
 - Is the change in the next time period going to be larger than X but smaller than y ? → That is a classification problem
 - We will focus on time series analysis in the next lecture