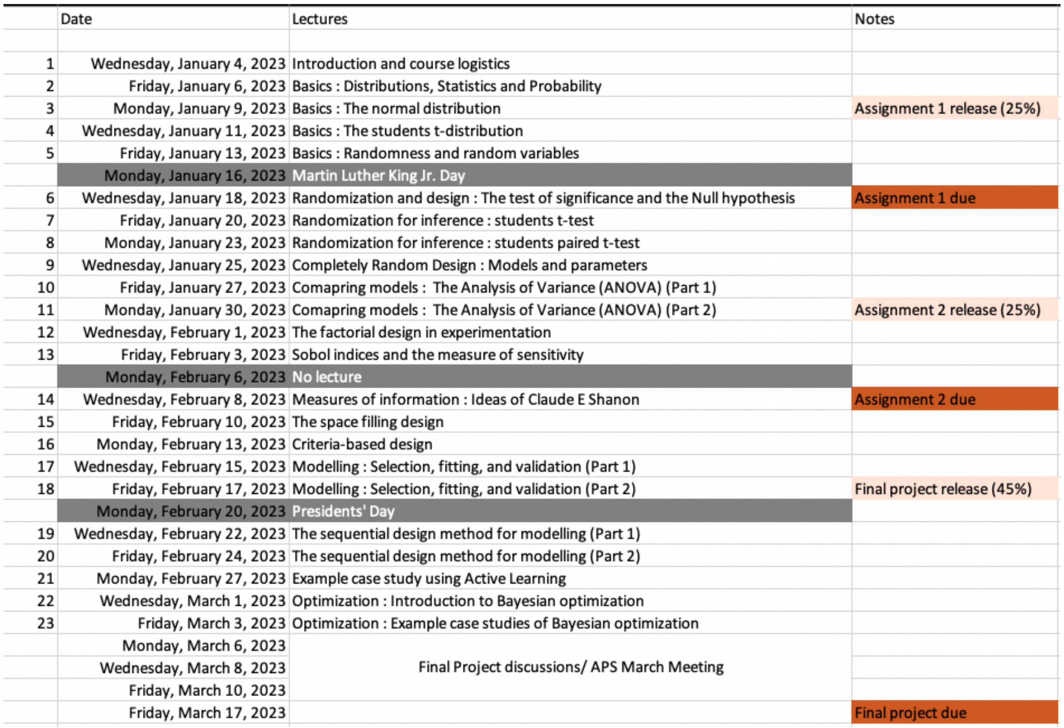
Lecture 1 : Introduction and logistics

By now all of you should've received a course description document that lists the overarching goals of this course. Today we will discuss the course logistics in a bit more detail. At the beginning of the course, we will spend a lot of time talking about what it means to design an experiment and what kind of tools one needs to have in their toolbox to do efficient design (we will also define what is efficient and how to measure it). A tentative schedule for this course is shown below:



	Date	Lectures	Notes
1	Wednesday, January 4, 2023	Introduction and course logistics	
2	Friday, January 6, 2023	Basics : Distributions, Statistics and Probability	
3	Monday, January 9, 2023	Basics : The normal distribution	Assignment 1 release (25%)
4	Wednesday, January 11, 2023	Basics : The students t-distribution	
5	Friday, January 13, 2023	Basics : Randomness and random variables	
	Monday, January 16, 2023	Martin Luther King Jr. Day	
6	Wednesday, January 18, 2023	Randomization and design : The test of significance and the Null hypothesis	Assignment 1 due
7	Friday, January 20, 2023	Randomization for inference : students t-test	
8	Monday, January 23, 2023	Randomization for inference : students paired t-test	
9	Wednesday, January 25, 2023	Completely Random Design : Models and parameters	
10	Friday, January 27, 2023	Comapring models : The Analysis of Variance (ANOVA) (Part 1)	
11	Monday, January 30, 2023	Comapring models : The Analysis of Variance (ANOVA) (Part 2)	Assignment 2 release (25%)
12	Wednesday, February 1, 2023	The factorial design in experimentation	
13	Friday, February 3, 2023	Sobol indices and the measure of sensitivity	
	Monday, February 6, 2023	No lecture	
14	Wednesday, February 8, 2023	Measures of information : Ideas of Claude E Shanon	Assignment 2 due
15	Friday, February 10, 2023	The space filling design	
16	Monday, February 13, 2023	Criteria-based design	
17	Wednesday, February 15, 2023	Modelling : Selection, fitting, and validation (Part 1)	
18	Friday, February 17, 2023	Modelling : Selection, fitting, and validation (Part 2)	Final project release (45%)
	Monday, February 20, 2023	Presidents' Day	
19	Wednesday, February 22, 2023	The sequential design method for modelling (Part 1)	
20	Friday, February 24, 2023	The sequential design method for modelling (Part 2)	
21	Monday, February 27, 2023	Example case study using Active Learning	
22	Wednesday, March 1, 2023	Optimization : Introduction to Bayesian optimization	
23	Friday, March 3, 2023	Optimization : Example case studies of Bayesian optimization	
	Monday, March 6, 2023	Final Project discussions/ APS March Meeting	
	Wednesday, March 8, 2023		
	Friday, March 10, 2023		
	Friday, March 17, 2023		Final project due

Figure 1: Course schedule with a rough set of lectures we plan to cover

The first few lectures (02-05) essentially cover probability and statistics to familiarize everyone with the language, and formalism required to develop a deeper understanding and application of DOE tools we discuss later on. We then dive into the topics of designing an experiment with an experiment where we only have to compare two entities (Lectures 06-09) before moving on to cover more practically useful cases of comparing multiple entities (Lectures 10-16). Only then would we think about analyzing the data we have collected using the design implemented for any given experiment of your choice by considering the aspects of model selection, fitting, and validation in Lectures 17,18. For the final part of the course, we would cover some of the new-ish techniques developed for designing experiments and provide a modern take on the DOE using active learning and Bayesian optimization. I expect that you'll make use of these techniques for your final project.

1 Introduction to DOE

In one sentence, one should consider the tools we will discuss in this course to simplify and accelerate the generation, testing, and development of new ideas. Think of the following iterative process of learning: we start with a possible hypothesis (a model if you will) and collect some data to deduce if our consequences matched the data. If there's some discrepancy, we will then use induction to modify our hypothesis or model. We go over this loop iteratively until our deductions match the hypothesis (up to some statistical significance of course).

Let us go over a simple example to make this concrete:

Our example consists of testing for a new catalyst whose presence in a reaction mixture would probably cause A and B to form, in high yield, a valuable product C.

Model 0: If we use Catalyst Cat in a reaction A+B -> C, the yield $> 90\%$
Deduction: Use literature to guess some initial conditions to perform this reaction. Say Temperature of 600° celsius.
Data: We observed that the resulting product is a colorless, odorless liquid with less than 1% yield.
Induction: Model and data do not agree so we have to modify them

In the above example, there are two key things we need to observe: 1) the amount of domain knowledge we used in the process of identifying four components of the experimental design; 2) The role of interpreting in case of multiple design parameters selected (we only used one so the interpretation or assignment is easy)–this is where statistics play a key role. But, domain knowledge is always going to be a key in interpreting the results, defining hypotheses to try, and identifying erroneous measurements. When we talk about experiments in general, there are three key things we need to keep in mind: 1) complexity, 2) experimental error, and 3) correlation and causation.
Complexity: We typically have more than a couple of variables that we would like to study their influence on a collection of output variables. Interaction between the input variables that you can control becomes key in accelerating the process of learning. Input variables can also have some overlapping influences on the outputs. We will learn how randomization and other design methods we discuss help us remedy these.
experimental error: We consider things that we can not explain by known influences as the basis for an experimental error. Most of you would be aware of using change in means as a way to quantify experimental error. Later in this course, we make use of a few other techniques that are designed with experiments in mind to quantify error and variance.
Correlation and Causation: There is not much to say here. We should all be aware but try to be diligent

Finally, I want to end this lecture by emphasizing the importance of experiments rather than the different statistical techniques we use or discuss in this course. We use them as a means to define ideas concretely and use computations to make a case for our observation to be the truth. Ronald A Fisher who is considered to have started the "Design of Experiments" developed this because of his interest to work with experiments. That is the only thing you need to know about this man. He was known to be involved in some pretty bad things.