# STA461: Introduction

Lynn Lin

August 20, 2018

# Principles of experimental design

Experimental design allows *data-driven* approach to *decision making*

# A simple example

- A farmer has been fertilizing her large corn fields with brand X fertilizer, but hears from a friend that brand Y fertilizer works better
- Should she trust her friend and switch to brand Y exclusively?

# A simple example

- Let's run a small experiment!
- She takes ten plots within a field and applies fertilizer X to 5 of them and fertilizer Y to the other 5. She then waits to the end of the year and measures the corn yield from each of the 10 plots.
- The results are as follows

| Plot | Fertilizer | Yield |
|---|---|---|
| 1 | X | 14 |
| 2 | X | 12 |
| 3 | X | 11 |
| 4 | X | 18 |
| 5 | X | 20 |
| 6 | Y | 21 |
| 7 | Y | 20 |
| 8 | Y | 17 |
| 9 | Y | 23 |
| 10 | Y | 25 |

# Data-driven decision making

- The farmer compares the average yield of the fields using fertilizer X (avg = 15) with the average yield of the fields using fertilizer Y (avg = 21.2)
- She decides that fertilizer Y improves corn yield
- The next year, she applies fertilizer Y to her entire corn crop

# Data-driven decision making

- Data-driven decision making (as illustrated in the previous example) can be very powerful, but there are many possible pitfalls
- It is worth considering the assumptions implicit in the decision made by the farmer
- What the farmer really wants to know is which fertilizer **WILL** give her the most yield next year, which is impossible to know in almost all cases
- Instead, the farmer assumes the following

1. The results of the 10-plot experiment are *generalizable* to all of her corn crop
2. The *cause* of the observed difference in average yield between plots is the fertilizer

# Generalizability

- Generalizability is a measure of how well an experimental result from a sample can be extended to the population as a whole
- The mean (average) of the experimental yields is a good (close) approximation to the average yield of any plot in the entire corn crop

# Causation

- There are **MANY** factors that influence corn yield, such as daily rain, soil nutrients, daily sunlight,. . .
- We can **NEVER** control every factor influencing a response
- Uncontrolled factors are called nuisance factors or nuisance variables or confounders or confounding factors

# Nuisance factors

We can address nuisance factors in multiple ways:

1. Statistically model unmeasured factors

$$y_i = \mu_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

- $y_i$ is the yield of the i-th plot
- $\mu_i$ is the mean (or average, or expected) yield of the i-th plot
- We model the mean as a function of the factors we control

$$\mu_i = \mu_X \text{ if the i-th plot gets fertilizer X}$$

$$\mu_i = \mu_Y \text{ if the i-th plot gets fertilizer Y}$$

# Nuisance factors

- In this model we allow the choice of fertilizer to influence the mean (average) plot yield
- $\epsilon_i$ captures all other factors that influence $y_i$. We don't measure these other factors, but assume that they can be represented by a normally-distributed random variable with mean 0 and a shared variance
- The assumption that $\epsilon_i \sim N(0, \sigma^2)$ is a good representation of all other factors needs to be checked!

# Nuisance factors

2. We try to control for confounders through *replication* and *randomization*

- What if fertilizer X and fertilizer Y are exactly the same? (e.g., They are different brands but have the identical active chemicals)
- Statistically, $\mu_X = \mu_Y$
- BUT... the 10 plots in the experiment were set up as follows, with the last 5 plots bordering a stream
- Then the difference in mean yields between plots 1-5 and plots 6-10 may be caused by the extra water/nutrients available to the plots that border the stream (NOT by the fertilizer)

# Randomization

- In this case assigning fertilizer X to the first 5 plots and fertilizer Y to the last 5 plots (bordering a stream) allowed the effect of the fertilizer to be confounded with the effect of the stream on yield
- This was NOT a randomized trial. Rather, the treatments (different fertilizers) were assignment systematically
- Instead, a randomized approach to setting up this experiment would get the 10 plots and then randomly assign 5 of the plots to be treated with fertilizer X and the rest to be treated with fertilizer Y
- The purpose of randomization is to prevent systematic and personal biases from being introduced into the experiment by the experimenter

# Randomization

- One way to do this: put the numbers 1-10 on papers in a bag, shuffle them, and randomly draw out 5 of them. Treat these 5 fields with X, and the other 5 with Y
- Then the results of the experiment might be:

| Plot | Fertilizer | Yield |
|---|---|---|
| 1 | Y | 14 |
| 2 | X | 12 |
| 3 | X | 11 |
| 4 | Y | 18 |
| 5 | Y | 20 |
| 6 | X | 21 |
| 7 | Y | 20 |
| 8 | Y | 17 |
| 9 | X | 23 |
| 10 | X | 25 |

# Randomization

- Under this randomization: $\mu_X = 18.4$, $\mu_Y = 17.8$
- The observed experimental average yields for each fertilizer treatment are now much closer to each other (recall - fertilizer doesn't matter, so this is good!)
- Randomizing the assignment of treatments made the effect of the unmeasured confounder (proximity to stream) much less than when we assigned treatments systematically
- There is still a difference in the mean yields. Is it an important difference? (more on this later - statistical significance)

# Randomization

Randomization is a key idea in experimental design

- It allows us to study the difference between different treatments with little worry about unmeasured confounding factors
- Randomization let us conclude causation: the gold standard of science
- If we do not assign treatments randomly, we run the risk of any result being caused by an unmeasured confounder (like the stream) instead of the treatments under study (like the fertilizer)