DS 102 Data, Inference, and Decisions **Fall 2019**

Lecture 15: Introduction to Design of Experiments

Lecturer: Fernando Pérez

1 Motivation

What is an experiment? Notions like "finding trends," "testing on groups," and "systematic testing of hypotheses," come to mind. A fairly comprehensive definition is the following: an experiment is *an operation or procedure carried out under controlled conditions in order to discover an unknown effect or law, to test or establish a hypothesis, or to illustrate a known law.*

There's a sense of control in an experiment, in the sense that the researcher can design the systematic evaluation of a hypothesis. Experiment design as a field deals with how to systematically design procedures to discover effects, establish hypotheses, or illustrate laws.

Example 15.1. Imagine designing an experiment to test the average serve speed of professional tennis players. There are many sources of variability that we can address in designing our experiment. For example, we can purposefully sample tennis players in different locations, and observe their serve speed at the exact same time of day. Even after fixing all the reasonable sources of variability, each time a single player serves, she or he will have variation in the speed of that serve. Moreover, there is instrumental variability depending on the type of sensor we use to measure speed, and possibly different human accuracy for the researchers measuring and recording times.

Experiments can also be purely numerical or computational, as in tuning hyperparameters in a machine learning model, or trying multiple algorithms for minimize the same optimization objective to determine which algorithm runs the fastest. Another important experimental paradigm is A/B testing, where the goal of the experiment is to decide which of two possibilities (A) or (B) is better according to some pre-specified metric.

Example 15.2. Imagine assessing the *impact* of incentivizing education (by providing scholarships) on graduation rate from college. In this experiment, there's a notion of *treatment*, e.g. assignment of a scholarship, to certain individuals.

In the past two lectures we talked about causality, the relative effects *caused by* conditions. We saw methods for dealing with limited observability in observational studies. When we get to collect data to answer a question of causality, experiments are our best tool for finding causal relationships. In particular, *randomized control trials* (RCTs) randomize treatment over individuals in order to average out variation that we didn't want to consider as causal factors (for more on this, see the previous two lecture notes).

Randomization is one tool to deal with variation in the inputs to our experiments that cause variation in the outputs. In fact, the 2019 Nobel prize in economics was awarded to a team that uses

RCTs to study the effects of economic policies for alleviating poverty.[1]

In general, we want to control our inputs so we can infer what changes in inputs are causing changes in the outputs. The next section will discuss handling variation in different input variables, depending on whether we can observe and control them.

2 Experiments map inputs to outputs.

We will distinguish between three types of inputs, x, u, and v:

- **Controlled** inputs x are those that the researcher can choose (e.g. dosage of drug administered).
- Uncontrolled, but **observed inputs** u are those that the researcher can't directly control, but are recorded (e.g. exact temperature in the room).
- Uncontrolled and **unobserved** inputs v are those that the researcher both can't control, and are not recorded as part of the study (e.g. what patients had for dinner the night before).

Which uncontrolled variables to record or not record is a judgement call. Recording as much data as possible can be helpful in diagnosing unexpected results after the study is over, but recording unnecessary data can also lead to spurious trends in analyzing the data later on.

All of these inputs feed into your experiments, and feed into outputs, y. Some of these variables are thing's we'd like to study the effect of (e.g. treatment of giving loans in the example above).

The other inputs – those that affect the outputs y but are not the intent of our study – are called *nuisance* inputs. These nuisance inputs can be in x, u, or v. Nuisance inputs can affect the relationship we observed between our variables of interest and the outputs, so in general we will want to control the variability of these inputs, as well as those we intend to study.

Now how do we handle input variability? That depends on the type of variable we're dealing with:

- Controlled inputs (x):
  - **Systematic variation.** Since we can control inputs x (e.g. assignment of treatment to individuals), we do so in a systematic way that aids analysis after data collection.
- Uncontrolled, observed inputs (u):
  - **Blocking.** Here we can group experiments across reasonably constant values of u. This encodes a modeling assumption that *within blocks*, the effect of x on y should be reasonably constant. For example, in studying the effect of a drug on patients' health outcomes, we might block by hospital.

[1] `https://www.nature.com/articles/d41586-019-03125-y`

  - **(Statistically) controlling for u.** If we believe u does have an effect on y, we can model the impact of u and remove its effect from the model $y = f(x) - g(u)$.

- Uncontrolled, unobserved inputs (v):
  - **Randomization.** By assigning units randomly to control inputs, we hope to average out the remaining hidden sources of variability. Note: this will require big enough sample sizes.

Summarized in single piece of advice: "control what you can, randomize the rest." Control as we've been discussing it has two meanings: an experimental meaning and a statistical meaning. In the experimental sense, the researcher can manually control variables and parameters as part of the design (e.g. controlling treatment assignment or designing the wording of survey questions). In the statistical sense, controlling for a variable means understanding and modelling what variation this variable produces toward your data (as in the bullet point above).

Inputs x are experimentally controlled. Inputs u are partly experimentally and partly statistically controlled. Units v are statistically controlled.

A very important aspect of experiments is that they are repeated and replicated. Colloquially, these two words may mean the same thing, but in the context of experiment design, they have precise definitions:

Replicates are repeated experimental runs, where the whole experiment is *fully repeated* from start to finish. That is, each replicate is independently subject to the full variability of the experiment (say a complete block). **Repeats**, on the other hand, duplicate the experiment on some data within one run.

Replication is expensive, since you're essentially repeating the entire experiment to test whether there is hidden variation within your blocks. For example, a replication for a drug-effectiveness study could re-conduct the entire experimental methodology at a new set of hospitals. Repeats typically don't regenerate all sources of variations, they might instead run the same experiment a month later at the original set of hospitals, thereby adding little bit of variability to things that otherwise should be identical. In general repeats quicker than replication, but provide less exhaustive testing.

3 Design of Experiments (DoE)

The design of an experiment is just one stage of the entire experimental pipeline, which is followed by running the experiment (gathering data), and running analysis on these data to reach conclusions. Experiment design includes decisions that effect these later steps: *(i) what measurement to make* (the response), *(ii) what conditions to compare* (the treatment), and *(iii) what material to apply the treatment to* (the units)?

Design of Experiments (DoE) is used in many contexts. In exploratory work, like comparison between alternatives, or screening which factors affect a response, a good design will allow us to

reach high level conclusions, possibly enabling more targeted data collection in a later step. DoE is also used to optimize components of a process, where a good design could mean any of the following: (a) obtaining and maintaining a response with minimum variability (process control), (b) maximizing or minimizing a response output, or (c) reducing overall variability of response (process robustness). A third context in which DoE is valuable is in modeling and regression. In DS100, you've seen multiple methods for modeling relationships between data. A DoE problem asks how to optimally collect data to enable precise and accurate fitting of these models.

Addressing any of these goals in any of these contexts requires controlling variability in our outputs due to variability in our inputs. We'll discuss two powerful tools for achieving this control: blocking, and randomization. Remember that there are two meanings of control. At a high level, blocking aims to achieve experimental control, whereas randomization aims to achieve statistical control.

Optimal Design: algorithmically search design space and optimize a specific statistical metric – which metric to optimize??

3.1 Blocking

X, Y: statistics that result from treatment A, B. Define the difference in statistics across treatments as $Z = X - Y$.

$$var(Z) = var(X) + var(Y) - 2cov(X, Y) \tag{15.1}$$

If we could somehow set the blocks to increase the covariance between X and Y within blocks, that would reduce the variance in our estimate of the differences. The following example makes this a little more concrete using linear models.

Example 15.3. Suppose a simplified setting, where the experimental population you'd like to study is comprised of two groups defined by attributes, such that each individual is either in group $a = 1$ or in group $a = 0$. We'd like to study the effect of treatment t on the output variable y. For each individual we either see the result of treatment

$$y_i = f(1) + g(a_i) + \epsilon_{yi}$$

or of control

$$x_i = f(0) + h(a_i) + \epsilon_{xi}$$

Where $f(t)$ models the relationship from treatment to outcome (as a fixed function), $g(a)$ and $h(a)$ model group-specific trends for treatment or non-treatment (thse can be random functions), and the ϵ's model additional noise (as usual, we assume these are zero-mean and independent of all other variables in the model). Under this model, we can expand the variance of our difference in treatment groups, Z:

$$\begin{aligned} var(Z) &= var(X) + var(Y) - 2cov(X, Y) \\ &= var(g(a)) + var(\epsilon_x) + var(h(a)) + var(\epsilon_y) - 2cov(g(a), h(a)) \ . \end{aligned}$$

Note that there is no variance due to $f(\cdot)$ because that is a fixed function. The terms $var(\epsilon_x)$ and $var(\epsilon_y)$ are somewhat unavoidable given the model above, but we see something interesting in the variance and covariance associated with the group-specific random variables $g(a)$. As a special case, if $g(\cdot) = h(\cdot)$, then $var(g(a)) + var(h(a)) = 2cov(g(a), h(a))$, and these contributions to the overall variance cancel. In general, positive correlation between $g(\cdot)$ and $h(\cdot)$ *within each block a* will lower the overall variance of Z computed separately on each block. In terms of the language we've seen so far: we'd like to design blocks such that within a given block, the variation in the outcome due to factors other than treatment is highly correlated between treated and non-treated individuals in the group.

3.2 Randomization

When dealing with controllable data, it is generally advisable to block that data whenever possible. However, there are many cases in which we can't control, or even observe, certain nuisance inputs. In the previous lectures on causal inference, we saw that randomization is a powerful tool in this setting. The high level idea is that by randomly assigning individuals to treatments, the distribution over nuisance inputs would be similar between the control group and the treatment group, in which case variation due to nuisance inputs would average out for large enough sample sizes.

3.3 Primary vs. alternate models.

Example 15.4. Consider a linear model that maps x to y, where x lies in the range $[0, 2]$:

$$y = mx + b\epsilon$$

And the goal of the experiment is to determine m and b with the highest precision possible. In collecting a data point, you get to choose x, and then some process generates y so that you get an (x, y) pair. Suppose we have enough funding to collect 50 data points. How would you choose these 50 x values? One option would be to evenly space x along the domain of interest. However, if data truly come from a linear model, placing half of the datapoints at each extreme of the domain results in lower variance estimates of the parameter m.

The example above is one in which your "primary model" is a linear model with parameters m and b. If we are confident that our primary model holds, then a good experiment design will "anchor the model," by sampling points at the extremes of the domain for which that model holds. In Example 15.4 that corresponds to placing half of the data points on either end of the $[0, 2]$ domain. For linear models in higher dimensions, we would want to sample toward the extremes of the higher dimensional domains where we want to apply the linear models.

In some cases, we might not be certain that a linear model is the right choice. For exploratory work, we may or may not have a clear idea of our model, and in this case we'd like our data collection to guide a choice of what model to use. In many cases, we may have a clear primary model that we think it most applicable, as well as some potential alternative models. These scenarios span a spectrum of trade offs, where we want to fit your primary model as well as possible, but also *hedge*

against the primary model being wrong. Depending on the level of confidence in your model, good experiments will allow you to precisely and accurately fit your primary model, while allowing you to assess whether that primary model is preferable over the alternative models.

4 Conclusion

Experiment design is a fundamental component of a data science pipeline. In particular, the way in which data collection is conducted will affect all downstream analysis and conclusion steps of the pipeline. Accounting for the different types of input variables (observed/unobserved, experimentally controllable/uncontrollable) allows us to systematically account for how their variability influences variation in outcomes. In order to control the variability due to nuisance inputs (those whose effect we do not wish to study), we studied blocking and randomization.

As a final, for our purposes more philosophical, point, in practice an experimental pipeline a single component in a closed feedback loop. The inferences and decisions made at the and of a single experiment will feed into the next questions that future experiments will test, as well as set precedent for the data acquisition, modelling, and analysis strategies. While the unit of an individual experiment is natural and extremely useful, it is in reality situated within the context of past data, decisions, and experiments, and contributes to the context of future experiments.