Welcome to Experiment Design & Optimization

BIOE 498 PJ

BIOE 598 PJ

BIOE 598 PJO

Trebuchet Case Study Results

First Place:

Second Place:

Third Place: -- Aaron
-- Zong
-- Claire, Angelo, & Jake

Trebuchet Case Study Results

First Place:

Second Place: -- Thomas, Anna, & Bailey
-- Joshua & Rachel

Third Place: -- Aaron
-- Zong
-- Claire, Angelo, & Jake

Trebuchet Case Study Results

First Place: Lingyun & Duncan

Second Place: -- Thomas, Anna, & Bailey
-- Joshua & Rachel

Third Place: -- Aaron
-- Zong
-- Claire, Angelo, & Jake

Experimental Design

- When scientists design experiments, they must follow certain rules.
- A good experimental design has:
  - One **independent variable**
  - A **dependent variable**
  - A **control group**

https://www.slideshare.net/stubeck/the-scientific-method-and-experimental-design-9th-grade-biology

Independent variables

Experimental results are much more straightforward to interpret and analyze when there is just one independent variable (one factor changed at a time). As a general rule of thumb, especially when you are starting out in biology, you should limit yourself to one independent variable per experiment.

Once you have lots of lab experience and some background in statistics, you can consider doing experiments with two independent variables at once. For example, you might want to see how water and light levels jointly affect bean seed sprouting. A well-designed experiment with two independent variables can tell you whether the variables interact (modify each other's effects). However, experiments with more than one independent variable have to follow specific design guidelines, and the results must be analyzed using a special class of statistical tests to disentangle the effects of the two variables.

https://www.khanacademy.org/science/high-school-biology/hs-biology-foundations/hs-biology-and-the-scientific-method/a/experiments-and-observations

What was in this course?

1. DOE
2. RSM
3. Surrogate optimization
4. RL

Lesson 1: Any design is better than no design.

Scientists agree on two things:

1. You need a *proper* experiment design.
2. You need *statistics* to analyze your data.

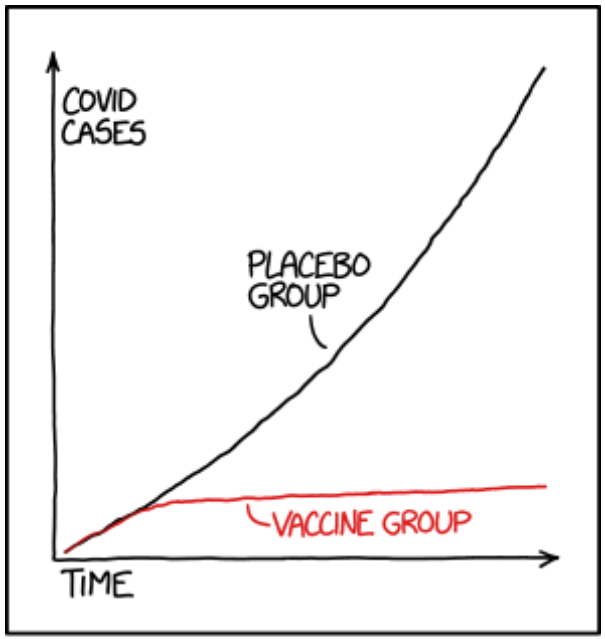
Why not use an experiment design that is optimized for statistical power?

Lesson 2: DOE requires modeling.

“All models are wrong; some are useful.” – G.P. Box

“Everyone trusts an experiment except the person who did it.
No one trusts a model except the person who built it.”

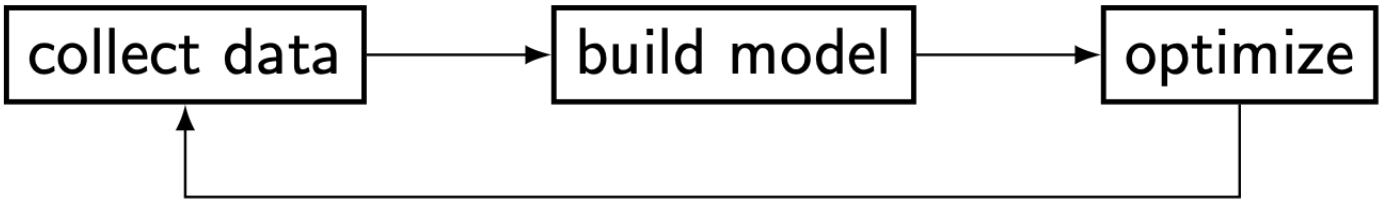
Lesson 3: Don’t rearrange deck chairs with statistics.

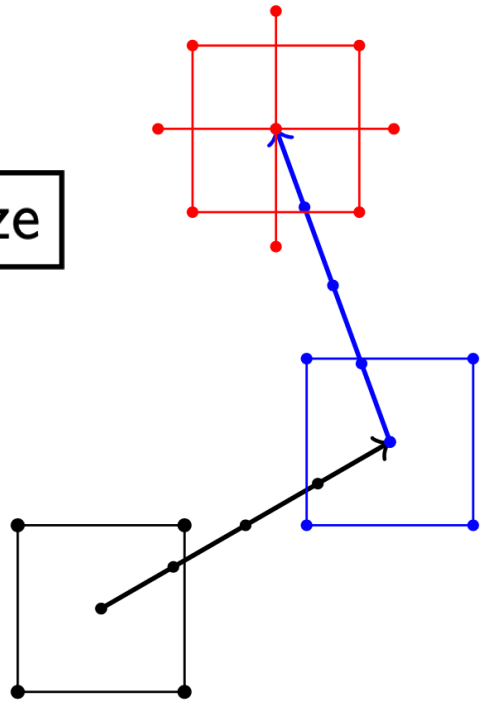


STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

https://xkcd.com/2400/

Lesson 4: DOE is an iterative process.





Lesson 5: AN-ova.

Lesson 6: Screen.

Effective screening means
starting with **more** factors
and ending with **fewer** factors
than you originally thought.

Lesson 7: Surrogate optimization is coming?

STATISTICS SPOTLIGHT

Solving quality quandaries through statistics

DATA COLLECTION AND ANALYSIS

If You Have It, Use It

Don't ignore what you already know and incorporate it into your statistical approaches

by Christine M. Anderson-Cook and Lu Lu

Most statistical data collection and analysis tools have historically started with a clean slate. That's the nature of general tools—they must be able to solve a large number of problems across different applications. That often means they want to begin with a minimum number of assumptions to allow maximum breadth of applicability.

In many practical situations, however, the choices made about what data to collect or how to proceed with an analysis can and should be based on using other available knowledge to tailor the decision to take maximum advantage of what already is known. We aren't suggesting that many of the fundamental design of experiments (DoE) and analysis tools that exist today do not serve an important and beneficial role, but rather we encourage scientists, engineers and business leaders not to restrict themselves to only those choices.

If you have preliminary data, background knowledge or previous experience that is relevant, find a way to use it to take advantage of that understanding and incorporate it into your statistical approaches. It will make better use of your resources and reduce the chance of wasted opportunities.

In some ways, this is not a new idea at all. When designing an experiment, there are key choices (such as which inputs to manipulate, what ranges to focus on for each input and what complexity of underlying model to assume) that depend on some basic understanding of the process to be studied. Indeed, it is difficult to collect relevant and useful data if there are not informed answers to the earlier questions about what to study.

In the analysis phase, the statistician brings knowledge about forms of the model that should be considered and understanding about the process to assess the appropriateness of assumptions required for a valid analysis. But we can extend this mantra of "If you have it, use it" beyond these basic principles to take maximum advantage of what our experts already know and what we have learned from previously spent resources.

Here are four examples to illustrate how we can do more with what we already know:

1. Strategic sampling

Consider sampling from a population of medical patients to judge a treatment's efficacy. To obtain accurate summaries that reflect the population's characteristics, it's important to ensure that the sampled data provide a good representation of the underlying population of interest. For example, the demographics of the patients often make important differences in the responses to treatments.

LEARN MORE

Read more about data collection and analysis tools, and related case studies and articles by visiting asq.org/quality-resources/data-collection-analysis-tools. There, you also can download free data collection tools and templates.

Therefore, if we have information on the division of the population into smaller subgroups based on the demographic characteristics such as age, gender and racial backgrounds, we can use this information to ensure a more balanced representative sample. The smaller subgroups of subjects (patients here in the medical study) often are referred to as "strata," which are formed by grouping together homogenous units from the bigger population. By using stratified proportional sampling,[3] which takes a random sample within each individual stratum of a size proportional to the stratum's size in the overall population, it ensures each sampled unit represents the same number of subjects from the population.

This approach provides an unbiased sample that represents a miniature version of the population and also a more precise estimate of the population characteristic than using a simple random sample. When some strata exist that are extremely small in size, however, using the proportional sampling could result in a sample insufficient to produce an accurate estimation of the subgroup. In this case, it may be possible to use stratified disproportional sampling,[4] which intentionally oversamples the small strata, and then uses post-sampling adjustments to combine the results to restore the original population proportions and ensure unbiased estimates with adequate precision of results.

2. Space-filling designs

Space-filling designs can be advantageous for computer experiments or when little is known about the underlying model form for the relationship between inputs and responses.[5] The majority of the available space-filling designs assume that the goal is to place experimental runs evenly or uniformly throughout the input region specified.

But if we have some knowledge of what to expect in the region, having the option to fill the space nonuniformly could allow exploration of the input region while still allowing some greater emphasis on the parts of the input space where there is more interest.

Figure 1 (p. 54) shows two designs:

1. The first (left) is a standard space-filling design,[6] a maximin distance design, which maximizes the minimum pairwise distance between designs points to achieve optimum uniform spread across the input space.
2. The second (right) is from a new type of nonuniform space-filling (NUSF) design[7,8] based on a weighted maximin distance criterion in which larger weights (shown by the contours) are assigned toward the bottom right of the input space.

We see how the NUSF design differs from the standard uniform space-filling (USF) design with increased density of design points in higher weight region. With these NUSF designs, the experimenter can specify when or she wants to emphasize and the degree of nonuniform desired (as specified by a design parameter, the maximum weight ratio).

Here are some reasons why you might want to place a greater concentration of design points in one region over another:

- **Goal: exploration**—If there is knowledge of "interesting features," such as rapid changes or higher variability in the function, in some portion of the input space.
- **Goal: model refinement**—If an existing model suggests that some regions of the input space have poorer prediction precision, placing more data in these areas can improve overall performance.
- **Goal: model calibration**—If there are discrepancies between the model and previously observed data in some input regions, emphasizing these regions can provide better understanding of these differences.
- **Goal: optimization**—If the goal is to identify the best performance in the input space, previous knowledge might suggest where this is likely to exist. Emphasizing the anticipated optimum, while still allowing some exploration to explore the surrounding region, would be desirable.

In each case, having more understanding about the input space and what is likely to occur there can suggest where to emphasize that region. The default of a standard USF design might end up wasting resources with too much placed in less interesting input regions. Imagine if the right corner of Figure 1(b) is the anticipated optimum and how many resources would be wasted in the top left region of least interest if a USF design is used.

3. Sequential DoEs

For many applications, collecting all of the data in an experiment at a single time is not a requirement. In these cases it can be advantageous to collect data in stages[9] and learn from the results of each stage to inform future stages.[10]

Suppose at the beginning stage of an experiment that involves two input factors, four pilot runs (shown as blue points in Figure 2, p. 55) were collected to provide an initial exploration of the input region and validate the measure-

FIGURE 2

Comparison of USF designs that (a) ignore or (b) account for previous data

(a) USF design

(b) Augmented USF design

USF - uniform space filling

Note: Blue indicates previously collected data and red indicates new design.

knowledge about the system structure and the expected performance of the individual components and subsystems, which jointly determine the appropriate functionality and reliability of the entire system.

This prior knowledge can be included in a Bayesian analysis through the appropriate formulation of prior distributions and models that capture the system's structure. This enables combining different sources of data and information to improve the accuracy and precision of the estimated system reliability.[11] Bayesian analyses provide powerful methods and tools to leverage a variety of forms of relevant information for improved estimation and prediction.

While adapting the data collection and analysis strategies to incorporate the additional knowledge that exists can require some more specialized tools and be a bit more complicated, the benefits of not ignoring what you know already generally outweigh the additional effort required. So if you have knowledge or previous data, find a way to use them. QP

EDITOR'S NOTE

References listed in this column can be found on the column's webpage at qualityprogress.com.

Christine M. Anderson-Cook is a research scientist in the Statistical Sciences Group at Los Alamos National Laboratory in New Mexico. She is a fellow of ASQ and the American Statistical Association.

Lu Lu is an associate professor in the department of mathematics and statistics at the University of South Florida in Tampa. She is a member of ASQ and the American Statistical Association.

Lesson 8: Neural networks require deep understanding.

Lesson 9: The race for AI is a race for data.

“Advancing AI by collecting huge personal profiles is laziness, not efficiency” – Tim Cook, CEO of Apple

That’s it.

Thanks for a fun semester of (online) learning.

Your first DOE consultation is on the house.