A First Course in Experimental Design
Notes from Stat 263/363

Art B. Owen
Stanford University

Autumn 2020

Preface

This is a collection of scribed lecture notes about experimental design. The course covers classical experimental design methods (ANOVA and blocking and factorial settings) along with Taguchi methods for robust design, A/B testing and bandits for electronic commerce, computer experiments and supersaturated designs suitable for fitting sparse regression models.

There are lots of different ways to do an experiment. Most of the time the analysis of the resulting data is familiar from a regression class. Since this is one of the few classes about ***making data*** we proceed assuming that the students are already familiar with t-tests and regression methods for ***analyzing data***. There are a few places where the analysis of experimental data differs from that of observational data, so the lectures cover some of those.

The opening lecture sets the scene for experimental design by grounding it in causal inference. Most of causal inference is about inferring causality from purely observational data where you cannot randomize (or did not randomize). While inferring causality from purely observational data requires us to rely on untestable assumptions, in some problems it is the only choice. When one can randomize then much stronger causal claims are possible. This course is about how to use such randomization. Several of the underlying ideas from causal inference are important even when you did randomize. These include the potential outcomes framework, conveniently depicted via a device sometimes called the science table, the Stable Unit Treatment Value Assumption (SUTVA) that the outcome for one experimental unit is not affected by the treatment for any other one, and the issue of external validity that comes up generalizing from the units in a study to others. Even with randomized experiments, we should be on the lookout for SUTVA violations and issues of external validity.

The next lecture dives into how experimentation is being used now for A/B testing in electronic commerce. Then there is a lecture on bandits, especially Thompson sampling. That is an alternative to classic A/B testing that we can use when there is rapid feedback. It lets us avoid assigning a suboptimal

treatment very often.

The largest segment in the course covers classical design of experiments that is now about 100 years old. A lot of useful ideas and methods have been developed there motivated by problems from agriculture, medicine, psychology, manufacturing and more. We look at blocking and Latin squares and factorial experiments and fractional factorials. Taguchi's methods for robust design fit into this segment. Key ideas that come up are about which categorical predictors are of primary interest, which are used to get more precise comparisons among those that are of main interest and how to use confounding (ordinarily a bad thing) to our advantage. The subject matter in this portion of the course could easily fill a full semester long course. Here we consider it in about half of an academic quarter which is already shorter than a semester. As a compromise we skip over some of the most challenging issues that arise from unbalanced data sets, complicated mixed effects models, subtle nesting structures and so on, citing instead the text books one would need to read in order to handle those issues.

That long core segment is followed by a midterm and then a survey of related topics in experimental design. This includes computer experiments for investigating deterministic functions, space-filling designs and designs for sparse models.

These notes are a scribed account of lectures. The references I draw on are the ones I know about and reflect a personal view of what belongs in an experimental design course for students studying data science. I'm sure that there are more good references out there. I have a personal interest in bridging old topics in design with some newer topics. Revisiting some old ideas in the light of modern computational power and algorithms and data set sizes could uncover some useful and interesting connections.

During the lectures, given online, I got many good questions and comments. Several people contributed that way but Lihua Lei and Kenneth Tay deserve special mention for the many good points that they raised.

In these notes I cite textbooks where readers can go beyond the level of detail in a first course. In some cases I find that the most recent edition of a book is not the best one for this level of course. Also, I use as a publication date the first year in which the given edition appeared, and not the year that edition went online as a digital book which can be decades after the fact.

Art B. Owen
Stanford CA
December 2020

Contents

1

Introduction

To gain understanding from data, we must contend with noise, bias, correlation and interaction among other issues. Often there is nothing we can do about some of those things, because we are just handed data with flaws embedded. Choosing or designing the data gives us much better possibilities. By carefully designing an experiment we can gain information more efficiently, meaning lower variance for a given expenditure in time or money or subjects. More importantly, experimentation provides the most convincing empirical evidence of causality. That is, it is not just about more efficient estimation of regression coefficients and similar parameters. It is about gaining causal insight. If we think of efficiency as better handling of noise, we can think of the causal estimation as better handling of correlations among predictors as well as interactions and bias.

We all know that "correlation does not imply causation". Without a causal understanding, all we can do is predict outcomes, not confidently influence them. There are settings where prediction alone is very useful. Predicting the path of a hurricane is enough to help people get out of the way and prepare for the aftermath. Predicting stock prices is useful for an investor whose decisions are too small to move the market. However, much greater benefits are available from causal understanding. For instance, a physician who could only predict patient outcomes in the absence of treatment but not influence them would not be as effective as one who can choose a treatment that brings a causal benefit. In manufacturing, causal understanding is needed to design better products. In social science, causal understanding is needed to understand policy choices.

Our main tool will be randomizing treatment assignments. Injecting randomness into the independent variables provides the most convincing way to establish causality, though we will see that it is not perfect.

Experimental design is the science of choosing how to gather data. It has a long and continuing history spanning: agriculture, medicine and public health, education, simulation, engineering, computer experiments, A/B testing in e-commerce, philanthropy and more.

The design problem forces the statistical investigator to think carefully about the underlying domain topic, it's assumptions and costs, benefits, goals and prior history. This happens in ordinary data analysis too, but the task of choosing what data to gather amplifies the problem. More than other areas of statistics, you really need to have a some use case in mind to understand how to interpret variables models and estimators. Predictors that you can change are different from ones determined by a user or the environment set. Response variables that can be cheaply measured are different from ultimate ones. Design is often sequential, so things we learn at one stage help at the next. There are choices between safe but perhaps expensive experiments and risky but perhaps faster or cheaper experiments.

1.1 History of design

Experimental design has been systematically studied for about 100 years or more. There are many recent innovations coming from online A/B testing as well as new developments in clinical trials. We will look at new innovations but not ignore the older ones that they evolved from.

Fisher and others working on agriculture at Rothamsted starting in the 1920s. Box working on 2^{k-p} factorial experiments for industry starting in the 1960s. Computer experiments from Sacks, Welch, Wynn, Ylvisaker and others starting in the 1980s. Realization that Monte Carlo simulations are (or should be) desiged experiments starting in the 1970s. Online A/B testing in electronic commerce starting in the 2010s. (Kohavi, Tang, Xu and many others). Experiments on networks. Experiments in economics for philanthropy.

1.2 Confounding and related issues

We will begin by comparing outcomes for subjects given one of two levels. Sometimes it makes sense to call them 'treatment' versus 'control'. Control could be a default or usual or null setting. Other times the two levels are on the same footing: e.g., raspberry vs blueberry or Harvard vs Yale.

If the treatment subjects are measured in the morning and control in the afternoon, then the difference we measure is the joint effect of treatment-AM versus control-PM and any difference we measure might be partly or largely due to time of day. The treatment is ***confounded*** with time of day.

Confounding can happen easily. Maybe treatment and control were done at different labs or by different people in the same lab or on different equipment by the same person in a given lab. Treatment and control could also be confounded with some variable that we didn't measure or don't even know about.

A common source of confounding is the use of ***historical controls***. A physician might compare a survival rate in their clinic to what was the historical norm for that or some other clinic. The standard of care may have changed and then the new treatment is confounded with the time period of study. A better comparison would include concurrent controls. One ethical standard there is ***equipoise***. The physician should be genuinely uncertain about which treatment is better in order to justify doing an experiment. Cox (1958) has an example where historical controls would have supported the Lamarkian theory that learned effects are inherited by offspring. Each generation of rats performed better at a certain task. Because there was no control group, an alternative explanation is that something else about those animals or their condition was changing with time.

Confounding might not matter. Maybe there is sufficient scientific knowledge to know that the confounding variable could not affect the response. Of course that knowledge might be subject to uncertainty and debate. If we use some randomization to assign treatment versus control then we can statistically control the confounding making extreme confounding have negligible probability.

Later on in settings with k binary treatments and a budget that does not allow running all 2^k cases we will indulge in some purposeful confounding. The trick will be to control what is confounded with what.

Confounding induces a correlation of 1 (or -1) between our treatment and some other variable. When the treatment or that other variable varies continuously then we get a nonzero correlation between our treatment variable and the confounder. So perfect confounding is a special case of correlation. Randomizing the treatment will make the expected value of that correlation zero.

The confounder might be known and measured or it might be unknown (which is worse). When it is unknown, it is a missing predictor problem. A regression that is missing one or more of the predictors leads to biased estimates of the coefficient vector β. If our treatment variable is uncorrelated with that missing variable then we can put that variable into the regression error term. [Working this out might become an exercise.]

Another reason to randomly control the treatment is to put sufficient variance into its values. For a continuous variable we know that the variance of a regression coefficient is reduced by varying the corresponding predictor over a larger range.

1.3 Neyman-Rubin Causal Model

We will use the Neyman-Rubin framework to think about causality. See the book by Imbens and Rubin (2015) for full details. Begin with Lecture 1 in Wager's course notes in the class web page. For experimental unit i (e.g., a subject) we think that their response would have been Y_{i1} if treated and Y_{i0} if not treated. There is a variable $W_i \in \{0, 1\}$ with $W_i = 1$ for treated and $W_i = 0$ for control. The pair (Y_{i0}, Y_{i1}) contains the two ***potential outcomes***

for subject i. In this setting we get

$$Y_i = W_i Y_{i1} + (1 - W_i) Y_{i0}, \quad i = 1, \ldots, n.$$

We can arrange all of these outcomes into a ***science table*** $\mathcal{Y} \in \mathbb{R}^{n \times 2}$ as follows:

$$\mathcal{Y} = \begin{array}{c} \\ i{=}1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cc} W_i{=}0 & W_i{=}1 \end{array} \\ \begin{bmatrix} Y_{10} & Y_{11} \\ Y_{20} & Y_{21} \\ Y_{30} & Y_{31} \\ \vdots & \vdots \\ Y_{i0} & Y_{i1} \\ \vdots & \vdots \\ Y_{n0} & Y_{n1} \end{bmatrix} \end{array} \tag{1.1}$$

Row i shows data for subject i. There is a column each for treatment and control. If we knew $\mathcal{Y}$ then we would know every subject's own personal treatment effect $\Delta_i = Y_{i1} - Y_{i0}$.

There is an important implicit assumption involved in writing the science table this way. We are assuming that the response for subject i is $Y_i = Y_i(W_i)$ and does not depend on $W_{i'}$ for any $i' \neq i$. Imagine instead the opposite where the value Y_i depends on the whole vector $\boldsymbol{W} = (W_1, \ldots, W_n) \in \{0, 1\}^n$. Then our science table would have n rows and 2^n columns, one for each possible $\boldsymbol{W}$.

We can use the small science table in (1.1) under the Stable Unit Treatment Value Assumption (***SUTVA***). Under SUTVA, Y_i does not depend on $W_{i'}$ for any $i' \neq i$. It might not even depend on W_i, but if it depends on $\boldsymbol{W}$ at all it can only be through W_i. In applications we have to consider whether SUTVA is realistic. If patients in a clinical trial swap drugs with each other, SUTVA is violated. If we are experimenting on subjects in a network we might find that the response for one subject depends on the treatment of their neighbors. That would violate SUTVA.

1.4 Random assignment and ATE

Very often the thing we are most interested in is the average treatment effect (ATE)

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \Delta_i = \frac{1}{n}\sum_{i=1}^{n} (Y_{i1} - Y_{i0}) = \frac{1}{n}\sum_{i=1}^{n} Y_{i1} - \frac{1}{n}\sum_{i=1}^{n} Y_{i0}.$$

No matter what assignment vector $\boldsymbol{W}$ we choose, we never see any of the Δ_i. This is sometimes called the fundamental problem of causal inference. Let's write

$$\tau = \mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0))$$

where in this instance $\mathbb{E}(\cdot)$ describes a simple average of n numbers. We will use $\mathbb{E}(\cdot)$ later for other things, so lets take

$$\tau = \mu_1 - \mu_0 \quad \text{where} \quad \mu_j \equiv \frac{1}{n}\sum_{i=1}^{n} Y_{ij}, \quad j = 0, 1.$$

Given $\boldsymbol{W}$ let $n_1 = \sum_{i=1}^{n} W_i$ and $n_0 = n - n_1$ be the number of treated and control subjects, respectively. If $\min(n_0, n_1) > 0$ then we can compute

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i \quad \text{and} \quad \bar{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1 - W_i) Y_i$$

and estimate τ by

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0.$$

By choosing $\boldsymbol{W}$ randomly we can get $\mathbb{E}(\hat{\tau}) = \tau$ where this $\mathbb{E}(\cdot)$ refers to randomness in $\boldsymbol{W}$.

Suppose we have a ***simple random sample*** where all $\binom{n}{n_1}$ ways of picking n_1 subjects to treat have equal probability. We saw in class that then

$$\mathbb{E}(\bar{Y}_j) = \mu_j, \quad j = 0, 1$$

assuming that $\min(n_0, n_1) > 0$. It then follows that

$$\mathbb{E}(\hat{\tau}) = \tau.$$

What if we just tossed independent coins taking $W_i = 1$ with probability $0 < p < 1$? Then we would get $n_1 \sim \mathrm{Bin}(n, p)$. Also, conditionally on this random n_1, the units getting $W_i = 1$ would be a simple random sample. So right away

$$\mathbb{E}(\hat{\tau} \,|\, 0 < n_1 < n) = \tau.$$

Ordinarily the event $0 < n_1 < n$ has overwhelming probability under independent sampling. Also, if anybody planned an experiment and got $n_1 \in \{0, n\}$ they would almost surely just re-randomize. [Exercise: how ok is that?]

If somehow it is necessary to analyze independent assignments without assuming that $n_1 \notin \{0, n\}$ then we can regard

$$\bar{Y}_1 = \frac{\sum_i W_i Y_i}{\sum_i W_i}$$

as a ratio of two random variables and work out approximate mean and variance via the delta-method. For the gory details see Rosenman et al. (2018).

Suppose that we want $\mathrm{var}(\bar{Y}_j)$ under simple random sampling. It's kind of a pain in the neck to work that out using the theory of finite population sampling (survey sampling) from Cochran (1977) or (Rice, 2007, Chapter 7). It is even worse if we toss coins because we hit the $n_1 \in \{0, n\}$ problem with probability just enough bigger than zero to be a theoretical nuisance. Both of

those get worse if we want $\mathrm{var}(\bar{Y}_1 - \bar{Y}_0)$. Let's just avoid it! We will see later an argument by Box et al. (1978) that will let us use plain old regression methods to get inferences. That is much simpler, and there is no reason to pick the cumbersome way to do things. There are also permutation test methods to get randomization based confidence intervals by Monte Carlo sampling. Those can work well and be straightforward to use. There is also a book in the works by Tirthankar Dasgupta and Donald Rubin on using the actual randomization in more complcated settings. I'm looking forward to seeing how that goes.

After writing the above, I saw Imbens (2019) describing a conservative estimate of $\mathrm{var}(\hat{\tau})$ due to Neyman. It is

$$\frac{1}{n_1(n_1-1)}\sum_i W_i(Y_i-\bar{Y}_1)^2 + \frac{1}{n_0(n_0-1)}\sum_i (1-W_i)(Y_i-\bar{Y}_0)^2.$$

This is just the sum of the two sample variance estimators that we might have used in regular modeling (like Box et al. advise). Let's still avoid digging into why that is conservative.

1.5 Random science tables

Maybe the ij entry of $\mathcal{Y}$ should really be a distribution, like $\mathcal{N}(\mu_{ij}, \sigma^2)$. Then we get a table of random numbers. If instead we want to account for possible correlations between Y_{ij} and $Y_{i'j'}$ then we need a more general model making a random table of numbers.

Suppose $\boldsymbol{W}$ satisfies SUTVA. Then for a random science table we also want W_i to be independent of (Y_{i0}, Y_{i0}). We write this as

$$W_i \perp\!\!\!\perp (Y_{i0}, Y_{i1}).$$

Think how bad it would be otherwise. If somebody purposely set $W_i = 1$ for the largest Δ_i, they would get a very biased answer.

Let $\mathcal{Y}$ be the random science table and suppose we compute $\hat{\tau}$. Is it estimating the ATE for our given science table, or is it estimating the average ATE in the underlying process that made our science table? That is a trick question. It is actually estimating both of those quantities even though they are different from each other. Let's write $\tau(\mathcal{Y})$ for the ATE when the science table is $\mathcal{Y}$ and $\hat{\tau}(\mathcal{Y}, \boldsymbol{W})$ for our estimate making explicit that it depends on the W_i. Let's assume that $\tau_0 = \mathbb{E}(\tau(\mathcal{Y}))$ exists. It is enough for all of the Y_{ij} to have a finite mean. Assume that the randomization in $\boldsymbol{W}$ never gives $\min(n_0, n_1) = 0$. Now under the randomness in $\boldsymbol{W}$,

$$\mathbb{E}(\hat{\tau}(\mathcal{Y}, \boldsymbol{W}) \,|\, \mathcal{Y}) = \tau(\mathcal{Y}),$$

so $\hat{\tau}$ estimates the actual random ATE.

Next

$$\mathbb{E}(\tau(\mathcal{Y})) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n Y_{i1} - Y_{i0}\Big) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_{i1}) - \mathbb{E}(Y_{i0}) = \frac{1}{n}\sum_{i=1}^n \mu_{i1} - \mu_{i0}.$$

This is the average ATE over the distribution of random science tables.

We could argue whether $\tau(\mathcal{Y})$ or $\mathbb{E}(\tau(\mathcal{Y}))$ is the more important thing to estimate but in practice they may well be very close. For instance, if $Y_{ij} = \mu_{ij} + \varepsilon_{ij}$ with noise $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ then

$$\tau(\mathcal{Y}) - \mathbb{E}(\tau(\mathcal{Y})) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i1} - \varepsilon_{i0} \sim \mathcal{N}\Big(0, \frac{2\sigma^2}{n}\Big).$$

In a large experiment the two quantities are close. If the quantities are close then we may choose to study whichever one gives the most clarity to the analysis. In a theoretical study we can model reasonable science tables by specifying a distribution for $\mathcal{Y}$ that fits the applied context.

1.6 External validity

One of the difficulties of randomized controled trials is in applying lessons from a given data set to another one, such as future uses. Think of two science tables, one right now and one that we will get later:

$$\mathcal{Y}_{\text{now}} = \begin{array}{c} \\ i{=}1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cc} W_i{=}0 & W_i{=}1 \end{array} \\ \begin{bmatrix} Y_{10} & Y_{11} \\ Y_{20} & Y_{21} \\ Y_{30} & Y_{31} \\ \vdots & \vdots \\ Y_{i0} & Y_{i1} \\ \vdots & \vdots \\ Y_{n0} & Y_{n1} \end{bmatrix} \end{array} \qquad \mathcal{Y}_{\text{later}} = \begin{array}{c} \\ i{=}n{+}1 \\ n{+}2 \\ n{+}3 \\ \vdots \\ i \\ \vdots \\ n{+}m \end{array} \begin{array}{c} \begin{array}{cc} W_i{=}0 & W_i{=}1 \end{array} \\ \begin{bmatrix} Y_{n+1,0} & Y_{n+1,1} \\ Y_{n+2,0} & Y_{n+2,1} \\ Y_{n+3,0} & Y_{n+3,1} \\ \vdots & \vdots \\ Y_{i0} & Y_{i1} \\ \vdots & \vdots \\ Y_{n+m,0} & Y_{n_m,1} \end{bmatrix} \end{array} \tag{1.2}$$

These have ATEs τ_{now} and τ_{later} and our experiment will give us the estimate $\hat{\tau}_{\text{now}}$. If we think of $\hat{\tau}_{\text{now}}$ as the future ATE we incur an error

$$\hat{\tau}_{\text{now}} - \tau_{\text{later}} = (\hat{\tau}_{\text{now}} - \tau_{\text{now}}) + (\tau_{\text{now}} - \tau_{\text{later}}).$$

The first term can be studied using statistical inference on our data. We shied away from doing that once we saw the sampling theory issues it raises, but will look into it later. The second term is about whether the ATE might have changed. It is about external validity.

External validity is some sort of extrapolation. It can be extremely reasonable and well supported empirically. E.g., gravity works the same here as elsewhere. It can be unreasonable. E.g., experimental findings on undergraduates who are required to participate for a grade might not generalize to people of all ages, everywhere.

Findings on present customers might not generalize perfectly to others. Findings for mice might not generalize well to humans. Findings in the US

might not generalize to the EU. Findings in a clinical trial with given enrolment conditions might not generalize to future patients who are sicker (or are healthier) than those in the study.

You may have heard the expression 'your mileage may vary'. This refered originally to ratings of fuel effiiciency for cars. If you're considering two cars, the advertised mileages μ_0 and μ_1 might not apply to you because you drive differently or live near different kinds of roads or differ in some other way from the test conditions. It commonly holds that the difference $\mu_1 - \mu_0$ might still apply well to you. The things that make your driving different from test conditions could affect both cars nearly equally. In that case, the ATE has reasonable external validity. This is a common occurrence and gives reason for more optimism about external validity.

External validity can be judged but not on the basis of observing a subset of $\mathcal{Y}_{\text{now}}$. The exact same $\mathcal{Y}$ could appear in one problem with external validity and another without it. External validity can be based on past experiences of similar quantities generalizing where tested. That is a form of ***meta-analysis***, a study of studies. External validity can also be based on scientific understanding that may ultimately have been gleaned by looking at what generalizes and what does not organized around an underlying theory that has successfully predicted many generalizations.

1.7 More about causality

We are anchoring our causal discusion in the Neyman-Rubin framework. We emphasize 'effects of causes' not 'causes of effects'. The first involves statements like 'if I water my plant it will grow'. The second involves statements like 'my plant grew because I watered it'. We could investigate the first problem by watering some randomly chosen plants and comparing the results to similar ones that were not watered.

The second problem is a much harder thing to disentangle. The plant could have grown for some other reason. Looking backwards for a cause, there might be 10 potential causes for why a plant grew or an accident happened or a product succeeded. Maybe changing any one of those 10 things would have given a different outcome. Or maybe some of those causes could have changed the outcome on their own while others would have only changed it if paired with some other changes. Then it is quite difficult to say what the real reason was even if you know all 2^{10} potential outcomes. This is a problem of attribution, that remains hard even when all 2^{10} causal possibilities are known without error.

For coverage of methods to infer causality from purely observational data, see Imbens and Rubin (2015), Angrist and Pischke (2008) and Angrist and Pischke (2014), and the notes by Stefan Wager in the class web page. There are several excellent courses on it here at Stanford. To make a causal conclusion requires a causal assumption. The assumption may be that a treatment was applied 'as if at random' or it may be that any important causal variables have been observed. For this course we will be relying on randomizing the treatments

and more generally, randomizing treatment combinations.

There is another approach to causality using directed acyclic graphs, explained in Pearl and Mackenzie (2018). Imbens (2019) remarks that it is not widely used in substantive problems and gives reasons. Roughly, it requires inputs from the user about the true underlying causal structure that are hard to come by.

2

A/B testing

This lecture was about A/B testing primarily in electronic commerce. An A/B test is a comparison of two treatments, just like we saw for causal inference. Much of the content was cherry-picked from the text Kohavi et al. (2020) (by three illustrious authors). The toy hippopotamus on the cover illustrates the "HIghest Paid Person's Opinion". They advocate basing decisions on experimental data instead of the hippo. That book is available digitally through the Stanford library. They have further material at `https://experimentguide.com/about/`. Ronny Kohavi kindly sent me some comments that helped me improve these notes.

There are also some other materials that I've learned about through availability bias: people I know from either a Stanford or Google connection worked on them. I quite like the data science blog from Stitch Fix `https://multithreaded.stitchfix.com/blog` which includes some postings on experimentation.

The lecture began with the example of an A/B test by Candy Japan. `https://www.candyjapan.com/behind-the-scenes/results-from-box-design-ab-test` It was about whether a fancy new box to send the candy out in would reduce the number of subscription cancellations. It did not significantly do so despite costing more.

2.1 Why is this hard?

In the 1960s George Box and co-authors were working out how to study the effects of say $k = 10$ binary variables on an output, with a budget that might only allow $n = 64$ data points. We will see that they had to contend with interactions among those variables. Modern e-commerce might have $k = 1$ treatment variable to consider in an A/B test with n in the millions. So maybe

the modern problems are only 1/1024 times as hard as the old ones. Yet, they involve large teams of data scientists and engineers developing specialized experimental platforms. Why?

Online experimentation is a new use case for randomized trials and it brings with it new costs, constraints and opportunities. Here are some complicating factors about online experimentation:

- there can be thousands of experiments per year,
- many going on at the same time,
- there are numerous threats to SUTVA,
- tiny effects can have enormous value,
- effects can drift over time,
- the users may be quite different from the data scientists and engineers,
- there are adversarial elements, e.g., bots and spammers, and
- short term metrics might not lead to long term value.

There are also some key statistical advantages in the online setting. It is possible to experiment directly on the web page or other product. This is quite different from pharmaceutical or aerospace industries. People buying tylenol are not getting a random tylenol-A versus tylenol-B. When your plane pulls up to the gate it is not randomly 787-A versus 787-B. Any change to a high impact product with strong safety issues requires careful testing and perhaps independent certification about the results of those tests.

A second advantage is speed. Agriculture experiments often have to be designed to produce data in one growing season per year. If the experiment fails, a whole year is lost. Clinical trials may take years to obtain enough patients. Software or web pages can be changed much more quickly than those industries' products can. Online settings involve strong day of week effects (weekend versus work days) and hour of day effects (work hours and time zones) and with fast data arrival a clear answer might be available in one week. Or if something has gone badly wrong an answer could be available much faster.

Now, most experimental changes do almost nothing. This could be because the underlying system is near some kind of local optimum. If the changes are doing nearly nothing then it is reasonable to suppose that they don't interfere much with each other either. In the language of Chapter 1, the science table for one experiment does not have to take account of the setting levels for other concurrent experiments. So in addition to rapid feedback, this setting also allows great parallelization of experimentation. The industrial settings that Box studied often have significant interactions among the k variables of interest.

Kohavi et al. (2009, Section 4) describe reasons for using single experiments instead of varying k factors at a time. Some combinations of factors might be undesiriable or even impossible to use. Also combining factors can introduce undesirable couplings between investigations. For instance, it could be necessary to wait until all k teams are ready to start, thus delaying $k-1$ of the teams.

2.2 Selected points from the Hippo book

2.2.1 Some scale

Kohavi et al. (2020) report that some companies run thousands or tens of thousands of experiments per year. With most of them running for multiple weeks at a time, it is clear that there must be lots of simultaneous experiments happening at once. One user's experience could be influenced by many ongoing experiments, but as mentioned above there may be little interference.

They give one example of the result of an A/B test adding $100,000,000 in annual revenue to the company. Clearly, that does not happen thousands of times per year. In fact, the vast majority of experiments are for changes that bring no noticeable value or even reduce value. Despite that, there can be large year over year improvements in revenue. The book cites 15–25% revenue growth for some period in Bing. They have another example where shaving a seemingly imperceptible 4 milliseconds off a display time improves revenue by enough to pay for an engineer's salary.

2.2.2 Twyman's law

They devote a chapter to ***Twyman's law***. There does not seem to be one unique version. Two that they give are "Any figure that looks interesting or different is usually wrong" and "Any statistic that appears interesting is almost certainly a mistake". As a result the most extreme or interesting findings have to be checked carefully. The specific way to test can get very specific to implementation details.

If we are to turn the statistic or figure into a claim then we face a similar statement by Carl Sagan that "Extraordinary claims require extraordinary evidence". There is a role for Bayes or empirical Bayes here. If the most recent experiment is sharply different from the results of thousands of similar ones then we have reason to investigate further.

Suppose that we double check all of the strange looking findings but simply accept all of the ordinary ones. That poses the risk of ***confirmation bias***. In a setting where $99+\%$ of experiments are null it does not make sense to subject the apparent null findings to the same scrutiny that the outliers get. Confirmation bias seems to be the lesser evil here.

2.2.3 Randomization unit

When we write $W_i \in \{0,1\}$ and later record Y_i, that index i identifies the ***randomization unit***. In medicine that might be a subject. In agriculture it could be a plot of land and the term ***plot*** shows up in experimental design language.

They describe the usual unit as being a 'user'. A user id might be a person or a cookie or an account. To the extent possible, you want a unit in treatment A to always be in treatment A. For instance if user i returns, you want them

to get the same W_i that they got earlier. This is done by using a deterministic hash function that turns the user id into a number $b \in \{0, 1, 2, \cdots, B-1\}$ where B is a number of buckets. For instance $B = 1000$ is common. You can think of the hash function as a random number generator and the user id as a seed. Then we could give $W = 0$ whenever $b < 500$ and $W = 1$ when $b \geqslant 500$. When the user returns with the same user id they get the same treatment.

We don't want a user to be in the treatment arm (or control arm) of every A/B test that they are in. We would want independent draws instead. So we should pick the bucket b based on hash(userid + experimentid) or similar.

There is an important difference between a person and a user id. A person's user id might change if they delete a cookie, or use different hardware (e.g., their phone and a laptop) for the same service. It is also possible that multiple people share an account. When that user id returns it might be a family member. The link between people and user ids may be almost, but not quite, one to one.

They discuss other experimental units that might come up in practice. Perhaps i denotes a web page that might get changed. Or a browser session. Vaver and Koehler (2011) make the unit a geographical region. Somebody could increase their advertising in some regions, leaving others at the nominal level (or decreasing to keep spending constant) and then look for a way to measure total sales of their product in those regions.

For a cloud based document tool with shared users it makes sense to experiment on a whole cluster of users that work together. It might not even be possible to have people in both A and B arms of the trial share a document. Also there is a SUTVA issue if people in the same group have different treatments.

2.2.4 SUTVA and similar issues

They describe two sided markets, such as drivers and riders for ride hailing services or renters and hosts for Airbnb. Any treatment B that changes how some drivers work might then affect all riders in a region and that in turn can change the experience of the drivers who were in arm A. A similar example arises when A puts a large load on the servers and slows things for B. Then what we see when B is 50% of the data might not hold true when it is 100%.

2.2.5 Ramping up

Because most experiments involve changes that are unhelpful or even harmful, it is not wise to start them out at 50:50 on the experimental units. It is better to start small, perhaps with only 1% getting the new treatment. You can do that by allocating buckets 0 through 9 of 0 through 999 to the new treatment. Also, if something is going badly wrong it can be detected quickly on a small sample.

2.2.6 Guardrail metrics

We test A versus B to see if Y changes. It is helpful to also include some measure Y' that we know should not change. In biology such things are called 'negative controls'. Guardrail has connotations of safety. If your new treatment changes Y' you might not want to do it even if it improves Y. For instance Y' might be the time it takes a page to load.

One important guardrail metric is the fraction n_1/n of observations in group 1. In a 50:50 trial it should be close to 1/2. It should only fluctuate like $O_p(1/\sqrt{n})$ around 1/2 and n may be very large. So if $n_1/n = 0.99 \times 1/2$ that could be very statistically significant (judged by $n_1/n \doteq \mathcal{N}(1/2, n/4)$). They give a possible explanation: the treatment might change the fraction of downstream data that gets removed because it looks like it came from bots or spam. Also when you're looking at tiny effects losing 1% of a data stream could be as big an effect as what you were trying to detect.

The measure above is called ***sample ratio mismatch*** (SRM). It is interesting to think what the response Y_i' would be for that metric. It would be something like an indicator function about whether unit i survives into the analysis period. We ***do not*** have to make sure that the $Y_i' = 0$ cases get into the SRM analysis. Their absence is detected already by comparing n_1 to n_0.

2.2.7 Additional points

They have discussions about what makes a good metric Y to study. It needs to arrive fast enough to inform your decision. But it should ideally be closely tied to longer term goals. This is like an external validity issue, but differs a bit. A clear external validity issue would be whether the short term improvement in Y holds into the future. This issue is about whether your short term metric (e.g., immediate engagement with a web site) is a good proxy for a long term goal (e.g., some notion of long term customer/user value).

It may be possible/reasonable to do an experiment in the future reverting some of those A/B decisions to a previous setting. Then you can measure whether the effect is still present.

There is a medical experiment of similar form. It is called a randomized withdrawal. One takes a set of people habituated to some medicine and randomly revert some of them to placebos for period of time.

2.3 Questions raised in class

There were some good questions in class right after the discussion of Hippo-book ideas. They are editted for length and to fix typos. I didn't make note of all the private ones so they're recreated from memory. I don't divulge who asked them because that could be a privacy issue. Thanks to those who asked. Hopefully the answers below are at least as good as the ones I actually gave.

Q: Could you please comment on the blinding in an online A/B test setting?

A: This is very subtle. Blinding is about whether you know you're in an experiment. If A is the usual version and B is new, then the people getting A have no real way to know that they're in an experiment at all, much less what treatment they are getting. The people getting B might well recognize that it is different from usual. They could be left to guess that it is a permanent change rolled out to everybody or they might realize that they're in an experiment. If they see two different versions because they have two user ids, or a friend getting something different, then they might well guess that they're in an A/B test. Also, if an interface momentarily goes bad, then some people will speculate that they're in an A/B test.

Q: How would you test for SUTVA violations after doing the A/B test?

A: I think you have to have some kind of hunch about what kind of SUTVA violation you might be facing. The fully general science table has n rows and 2^n columns, so how could we know what the other columns are like in general? Now suppose that we suspect something about which users' treatments might affect which other users' responses. Maybe they are neighbors in a network or similar geographically.

We could do a side experiment to test the hunch. After all, a SUTVA violation is fundamentally about how causality works and randomization is the way to test it. Then again, maybe the randomization we already did is adequate to catch some kinds of SUTVA violations. That is, the needed side experiment may already be baked into our initial randomization. I think I can make a good homework question out of this. I'm thinking of a case where we have pairs (i, i') of subjects where we know or suspect that the SUTVA violation happens within those known pairs.

Q: Is an A/A test purely based on historical data? If so, wouldn't it be missing all issues of spam/bot/non-compliance/interference etc. in the A/B test? Seems that the A/A test may provide very misleading results.

A: I think you are right that there are things that an A/A test could not catch. What those are might depend on how the experimental system is configures. An A/A test might then give a false sense of security, though better than not doing it because at least it can catch some things. Maybe we should take the A subjects and split them randomly into two groups. Similarly for the B subjects. That would be non-historical. It would involve smaller sample sizes.

Q: Professor Athey once mentioned that large companies often use shared controls for experiments. Is there a correction to make for the non-random assignment across experiments?

A: It sounds like testing A versus B_1, B_2, $\cdots$, B_k by splitting the units into $k+1$ groups. Then use A for the control in all k tests. Each of the tests should be ok. But now your tests are correlated with each other. If the average in the control group A fluctuates down then all k treatment effect estimates go up and vice versa. You also get an interesting multiple comparisons problem. Suppose that you want to bound the probability of wrongly finding any of the new treatments differ significantly from the control. I believe that this gets you

to Dunnett's test. Checking just now, Dunnett's test is indeed there in Chapter 4.3.5 of Berger et al. (2018).

Q: Do people usually use A/A test to get the significance level, does it differ a lot in case you have a lot of data. I assume with a huge amount of data the test statistic might be more or less or t-distributed (generally).

A: They might. I more usually hear about it being used to spot check something that seems odd. If n is large then the t-test might be ok statistically but A/A tests can catch other things. For example if you have k highly correlated tests then the A/A sampling may capture the way false discoveries are dependent.

A t-test is unrobust to different variances. You're ok if you've done a nearly 50:50 split. But if you've done an imbalanced split then the t-test can be wrong. If the A/A test is a permutation it will be wrong there too because the usual t-test is asymptotically equivalent to a permutation test. Ouch. Maybe a clever bootstrap would do. Or Welch's t test. Or a permutation strategy based on Welch's.

If you have heavy-tailed responses, so heavy that they look like they've come from a setting with infinite variance then the t-test will not work well for you. However maybe nothing else will be very good in that case. These super heavy-tailed responses come up often when the response is dollar valued. Think of online games where a small number of players, sometimes called 'whales', spend completely unreasonable (to us at least) amounts of money. (You can use medians or take logs but those don't answer a question about expectations.)

To put A/A testing into an experimental platform you would have to find a way to let the user specify what data are the right ones to run A/A tests on for each experiment. Then you have to get that data into the system from whatever other system it was in. That would be more complicated than just using χ^2 tables or similar.

Q: Is A/A testing a form of bootstrapping?

A: It is a Monte Carlo resampling method very much like bootstrapping. It might be more accurately described as permutation testing. There's nothing to stop somebody doing a bootstrap instead. However the A/A test has very strong intuitive rationale. It describes a treatment method that we are confident cannot find any real discovery because we shift treatment labels completely at random.

2.4 Winner's curse

Suppose we adopt a treatment because we think it raises our metric 1%. Then the next one looks like 2% increase so we adopt it too followed by another at 3%. We then expect to gain about 6% overall. A bit more due to compounding. Then we do an A/B test reversing all three changes and find that the new system is only 4% better. Why would that happen?

One explanation is the ***winner's curse***. It is well known that the stock or mutual fund that did best last year is not likely to repeat this year. Also athletes that had a super year are not necessarily going to dominate the next year. These can be understood as regression to the mean `https://en.wikipedia.org/wiki/Regression_toward_the_mean`.

This section is based on Lee and Shen (2018) who cite some prior work in the area. Suppose that the B version in experiment j has true effect τ_j for $J = 1, \dots, J$. We get $\hat{\tau}_j \sim \mathcal{N}(\tau, \sigma_j^2)$. The central limit theorem may make the normal distribution an accurate approximation here. Note that this σ_j^2 is really what we would normally call σ^2/n, so it could be very small. Suppose that we adopt the B version if $\hat{\tau}_j > Z^{1-\alpha_j}\sigma_j$. For instance $Z^{1-\alpha_j} = 1.96$ corresponds to a one sided p-value below 0.025. Let $S_j = 1\{\hat{\tau}_j > Z^{1-\alpha_j}\sigma_j\}$ be the indicator of the event that the experiment was accepted. Ignoring multiplicative effects and just summing, the true gain from accepted trials is $\sum_{j=1}^J \tau_j S_j$ while the estimated gain is $\sum_{j=1}^J \hat{\tau}_j S_j$ The estimated gain is over-optimistic on average by

$$\mathbb{E}\left(\sum_{j=1}^J (\hat{\tau}_j - \tau_j) S_j\right) = \sum_{j=1}^J \int_{Z^{1-\alpha_j}\sigma_j}^\infty \frac{\hat{\tau} - \tau}{\sigma_j} \varphi\Big(\frac{\hat{\tau}_j - \tau_j}{\sigma_j}\Big)\, \mathrm{d}\hat{\tau}_j > 0,$$

where $\varphi(\cdot)$ is the $\mathcal{N}(0,1)$ probability density function. We know that the bias is positive because the unconditional expectation is $\mathbb{E}(\hat{\tau}_j - \tau_j) = 0$. Lee and Shen (2018) have a clever way to show this. If $Z^{1-\alpha_j}\sigma_j + \tau_j > 0$ then the integrand is everywhere positive. If not, then the left out integrand over $-\infty$ to $Z^{1-\alpha_j}\sigma_j$ is everywhere non-positive so the part left out has to have a negative integral giving the retained part a positive one.

They go on to plot the bias as a function of τ for different critical p-values. Smaller p-values bring less bias (also less acceptances). The bias for $\tau_j > 0$ is roughly propotional to τ_j while for $\tau_j < 0$ the bias is nearly zero. They go on to estimate the size of the bias from given data and present bootstrap confidence intervals on it to get a range of sizes.

While we are thinking of regression to the mean, we should note that it is a possible source of the ***placebo effect***. If you do a randomized trial giving people either nothing at all, or a pill with no active ingredient (placebo) you might find that the people getting the placebo do better. That could be established causally via randomization. One explanation is that they somehow expected to get better and this made them better or lead them to report being better. A real pill ought to be better than a placebo so an experiment for it could test real versus placebo to show that the resulting benefit goes beyond possible psychological effects.

If you don't randomize then the placebo effect could be from regression to the mean. Suppose people's symptoms naturally fluctuate between better and worse. If they take treatment when their symptoms are worse than average, then by regression to the mean, the expected value of symptoms some time later will be better than when they were treated. In that case, no psychological

explanation based on placebos is needed. In other words, even a placebo effect can be illusory.

By the way, nothing anywhere in any of these notes is meant to be medical advice! These examples are included because they help understand experimental design issues.

The winner's curse can also be connected to multiple hypothesis testing and selective inference. Suppose that one does N experiments and then selects out the $M \leqslant N$ significantly successful ones to implement. If tests are done at one-sided level α then the expected number of ineffectual changes that get included is αN. For large N, we can be sure of making a few mistakes. We could take all $\alpha_j \ll 1/N$ to control the probability that any null or harmful changes have been included. This may be too conservative.

2.5 Sequential testing

This section is based on Johari et al. (2017). In an A/B test, the data might come in especially fast and be displayed in a dashboard. It is then tempting to watch it until $p < .05$ or p is below a smaller and more reliable threshold and then declare significance. In some settings one could watch as each data point arises, however, when there is a strong weekly cycle it makes sense to observe for some number of weeks.

Let the p-value at time t be $p(t)$, and suppose that it is valid, meaning that $\Pr(p(t) \leqslant \alpha; H_0) = \alpha$ for all $0 < \alpha < 1$ or conservative, meaning that $\Pr(p(t) \leqslant \alpha; H_0) \leqslant \alpha$ for all $0 < \alpha < 1$.

If we declare a difference the first time that $p(t) \leqslant \alpha$ then are taking the minimum of more than one p value and that generally makes it invalid. We are essentially using $P(t) = \min_{0<s\leqslant t} p(s)$ as our p-value. We should therefore adjust the threshold and declare significance only if $P(t) \leqslant \alpha^* = \alpha^*(t)$. This stricter criterion $\alpha^* < \alpha$ must satisfy

$$\Pr(P(t) \leqslant \alpha^*(t); H_0) = \alpha.$$

We want an 'always-valid p-value'.

Computing α^* is done via sequential analysis. See the book Siegmund (1985). Sequential analysis is not a prerequisite for this course. We will see it used later for adaptive clinical trials.

We could keep $\alpha^* = \alpha$ if we set an endpoint for the study in advance and kept to it. The reason not to do that is that when the treatment difference is large, we could be very sure which treatment is better long before the experiment ended. Then continuing it is wasteful in an economic sense and unethical in some contexts.

2.6 Near impossibility of measuring returns

This section reports an observation by Lewis and Rao (2015) on how hard it can be to measure returns to advertising. The original title of their article was

"On the Near Impossibility of Measuring the Returns to Advertising".

They had an unusually rich data set connecting advertising effort to customer purchases. The amount spent by a potential customer can have an enormous coefficient of variation. For instance, if most people don't buy something and a few do buy it, then the standard deviation can be much larger than the mean. The ads might be inexpensive per person reached, so a small lift in purchase could be very valuable. Combining these facts they give a derivation in which an extremely successful advertising campaign might end up with the variable describing advertising exposure having $R^2 = 0.0000054$ in a regression. It is hard to separate such small effects from zero. It could even be hard to be sure of the sign of the effect.

One way to mitigate the small sample size in experiments is to pool multiple related experiments. Owen and Launay (2016) describe a strategy of doing B different but related experiments in G different geographical regions (Geos). An advertiser might then be able to estimate the average return of advertising for B different brands more accurately than any individual brand.

The campaign could be defined via a table X like this

	Geo 1	Geo 2	Geo 3	Geo 4	$\cdots$	Geo G
Brand 1	$+$	$+$	$-$	$-$	$\cdots$	$+$
Brand 2	$-$	$+$	$+$	$-$	$\cdots$	$-$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Brand B	$-$	$-$	$+$	$-$	$\cdots$	$+$

with + representing a brand given increased advertising in a specific Geo, perhaps paid for by reductions in the cells marked $-$. When B and G are both even then it is possible to let each column have $B/2$ settings each for $\pm$ and each row has $G/2$ for each of ± 1. That is, each row and column of the table sum to 0.

If we let $X_{ij} \in \{-1, 1\}$ for $i = 1, \dots, B$ and $j = 1, \dots, G$ we can arrange a random sampling as follows:

1) start with a checkerboard $X_{ij} = (-1)^{i+j}$, and then repeatedly

2) pick rows $i < i'$ and columns $j < j'$ and if the resulting 2×2 matrix is either $\left(\begin{smallmatrix} + & - \\ - & + \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} - & + \\ + & - \end{smallmatrix}\right)$ swap it for the other.

Step 2 runs a Markov chain from Diaconis and Gangolli (1995). In their simulations, Owen and Launay (2016) run it $50 \times G \times B$ times.

It would be unfortunate for two columns of X to be identical because then two Geos get identical treatments. Similarly we don't want any two identical rows. Neither do we want completely opposite columns $X_{ij} = -X_{ij'}$ for all i or completely opposite rows. It is only possible to avoid having one of these four problems for $\min(B, G) \geqslant 6$ (Owen and Launay, 2016).

The resulting data can be estimated by either Bayesian methods or Stein shrinkage. Both of those estimate the causal effect for advertising brand b as a weighted average of data from that brand and data from all other brands. That reduces the highest estimated returns and raises the lowest estimates, countering though not necessarily eliminating the winner's curse problem.

An interesting feature of pooling experiments here is that it becomes possible to discover from a regression model that advertising works especially well (or poorly) in one particular Geo compared to the others.

3

Bandit methods

When you say you're going to do an A/B test somebody usually suggests using bandit methods instead. And vice versa.

In the bandit framework you try to optimize as you go and ideally spend next to no time on the suboptimal choice between A and B, or other options. It can be as good as having only $O(\log(n))$ tries of any sub-optimal treatment in n trials.

We review some theory of bandit methods. The main point is to learn the goals and methods and tradeoffs with bandits. We also go more deeply into Thompson sampling proposed originally in Thompson (1933).

Puzzlers and opinions

Most texts and articles are all about facts, not opinions. Facts are better. Sometimes opinions fill in where we don't have a desired fact. I'll be putting some of those in, and I expect you will be able to tell. I have been reading the blog of Andrew Gelman `https://statmodeling.stat.columbia.edu/` for many years and have benefitted enormously. Some of the opinions he shares are things I have long believed but never saw in print. It was nice to know that I was not alone. Sometimes my opinion is different from his.

One opinion I'm planning to describe is that theorems have issues of external validity. We have to think carefully about when and how to apply them.

I'm also going to put in some 'puzzlers'. These are things that are potentially confusing or apparently contradictory about the methods and their properties. Sometimes there is a momentary puzzle while we figure things out. Other times we do not get a clean answer. One of my goals is for students to learn to find and solve their own puzzlers. When you spot and resolve a puzzler it deepens

your understanding. So, get confused and then get out of it. Spotting and resolving puzzlers is also a way to find research ideas.

Here is a quote from Paul Halmos about reading mathematics:

> Don't just read it; fight it! Ask your own questions, look for your own examples, discover your own proofs. Is the hypothesis necessary? Is the converse true? What happens in the classical special case? What about the degenerate cases? Where does the proof use the hypothesis?

It is good to poke at statistical ideas in much the same way, with a view to which problems they suit.

3.1 Exploration and exploitation

In a regular experiment we get data on n subjects estimate $\mathbb{E}(Y \mid A)$ and $\mathbb{E}(Y \mid B)$. We pick what seems to be the better of A and B from our data and retain that choice hypothetically forever. Perhaps only for some $N \gg n$ future uses before we contemplate another change.

In this setting, we use the first n subjects to ***explore*** treatment differences. Once we have the apparent best one, we ***exploit*** that knowledge for the next N subjects by using the winning treatment.

One problem with experimentation is that something like $n/2$ of the subjects will be getting the suboptimal treatment in the experiment. Maybe we can avoid much of tha cost by biasing the experiment towards the seemingly better treatment at each stage as the data come in.

The theoretically most effective way to do this is through what are called ***bandit methods***. The name comes from slot machines for gambling that are sometimes called one-armed bandits. Each time you pull that arm you win a random amount of money that has expected value less than what you paid to play. The image for a multi-armed bandit is such a machine that offers you $K \geqslant 2$ arms to pick from. Each arm has its own distribution of random payoffs. In the gambling context of pulling n arms, the goal might be to minimize your expected loss. Of course, the best move is not to play at all, so the metaphor is imperfect.

In the experimental settings we care about, the goal is to maximize your expected winnings. If we knew the best arm, we'd choose it every time. But we don't. Instead we sample from the arms to learn about payoffs on the fly while also trying to get the best payoff. We will see bandit methods that pick a suboptimal arm only $O(\log(n))$ times as the number n of subjects goes to infinity.

A very old method is called 'play the winner'. Suppose that $Y_i \in \{0, 1\}$ with 1 being the desired outcome. Then if $Y_i = 1$ we could take $W_{i+1} = W_i$, while for $Y_i = 0$ we switch to $W_{i+1} = 1 - W_i$ when the choices are $W \in \{0, 1\}$ or, for $K > 2$ choices switch to a random other arm. These methods were studied intensely in the 1960s and 1970s and the term 'play the winner' seems to be

used to describe numerous different strategies. If there is a really great strategy in the mix then play the winner can have long streaks of $Y_i = 1$. If instead the best arm has a small payoff, like $\Pr(Y_i = 1) = 0.03$, and the other arm (out of 2) has $\Pr(Y_i = 1) = 0$, then play the winner will alternate too much between the best and worst arms.

3.2 Regret

These definitions are based on Bubeck and Cesa-Bianchi (2012). Suppose that at time $i = 1, 2, 3, \ldots$ we have arms $j = 1, 2, \ldots, K$ to pick from. If at time i we pick arm j then we would get $Y_{i,j} \sim \nu_j$. Notice that the distribution ν_j here is assumed to not depend on i. We let $\mu_j = \mathbb{E}(Y_{i,j})$ be the mean of ν_j and

$$\mu_* = \max_{1 \leqslant j \leqslant K} \mu_j \equiv \mu_{j_*}.$$

So μ_* is the optimal expected payout and j_* is the optimal arm (or one that is tied for optimal).

If we knew the μ_j we would choose arm j_* every time and get expected payoff $n\mu_*$ in n tries. Instead we randomize our choice of arm, searching for the optimal one. At time i we choose a random arm $J_i \in \{1, 2, \ldots, K\}$ and get payoff Y_{i,J_i}. Because we choose just one arm, we do not get to see what would have happened for the other $K - 1$ arms. That is, we never see $Y_{i,j'}$ for any $j' \neq J_i$, so we cannot learn from those values.

There are various ways to quantify how much worse off we are than optimal play would be. The ***regret*** at time n is

$$R_n = \max_j \sum_{i=1}^n Y_{i,j} - \sum_{i=1}^n Y_{i,J_i}.$$

This is how much worse off we are compared to whatever arm would have been the best one to use continually for the first n tries. Be sure that you understand why $\Pr(R_n < 0) > 0$ with this defniition. A harsher definition is

$$\sum_{i=1}^n \max_j Y_{i,j} - \sum_{i=1}^n Y_{i,J_i}.$$

This is how much worse off we would be compared to a psychic who knew the future data. It is not a reasonable comparison so it is not the focus of our study.

The expected regret is

$$\mathbb{E}(R_n) = \mathbb{E}\left(\max_j \sum_{i=1}^n Y_{i,j} - \sum_{i=1}^n Y_{i,J_i} \right).$$

It is awkward to study because the maximizing j is inside the expectation.

Bubeck and Cesa-Bianchi (2012) define the pseudo-regret

$$\begin{aligned}\bar{R}_n &= \max_j \mathbb{E}\left(\sum_{i=1}^n Y_{i,j} - \sum_{i=1}^n Y_{i,J_i}\right)\\ &= \max_j n\mu_j - \sum_{i=1}^n \mathbb{E}(Y_{i,J_i})\\ &= n\mu_* - \sum_{i=1}^n \mathbb{E}(\mu_{J_i})\end{aligned}$$

Each time we move $\max_j$ outside of a sum or expectation, things get easier. What is random in the $\mathbb{E}(\cdot)$ of $\bar{R}_n$ is the sequence $J_1, \ldots, J_n$ of chosen arms. Other authors call $\bar{R}_n$ the expected regret.

Now let $\Delta_j = \mu_* - \mu_j \geqslant 0$ be the suboptimality of arm j and define $T_j(s) = \sum_{i=1}^s 1\{J_i = j\}$. This is the number of times that arm j was chosen in the first s tries. Then

$$\bar{R}_n = n\mu_* - \sum_{i=1}^n \mathbb{E}(\mu_{J_i}) = \sum_{j=1}^K \mathbb{E}(T_j(n))\mu_* - \sum_{j=1}^K \mathbb{E}(T_j(n))\mu_j = \sum_{j=1}^K \mathbb{E}(T_j(n))\Delta_j.$$

Our pseudo-regret comes from the expected number of each kind of suboptimal pulls time their suboptimality. To derive this notice that

$$n = \sum_{j=1}^K T_j(n) = \sum_{j=1}^K \mathbb{E}(T_j(n))$$

because exactly one arm is chosen for every i.

3.3 Upper confidence limit

Figure 3.1 depicts a hypothetical setting with three treatment arms, there denoted by A, B and C with confidence intervals for the expected value of Y in all three arms.

Based on this information, which arm should we choose? There is an argument for arm B because the point estimate at the center of its confidence interval is the highest of the three. But that does not take account of uncertainty. If we would consider two restaurants one with six ratings that were all 5 stars and another with 999 5 star ratings and one 4 star rating, we would be more confident about the second restaurant. One way to judge that and play it safe is to rank by a lower confidence limit on the expected value. By that measure, treatment C has the highest lower limit and so it seems best for the cautious user.

The answer however, spoiled by the title of this section, is to choose arm A because it has the highest upper confidence limit. Suppose that we went

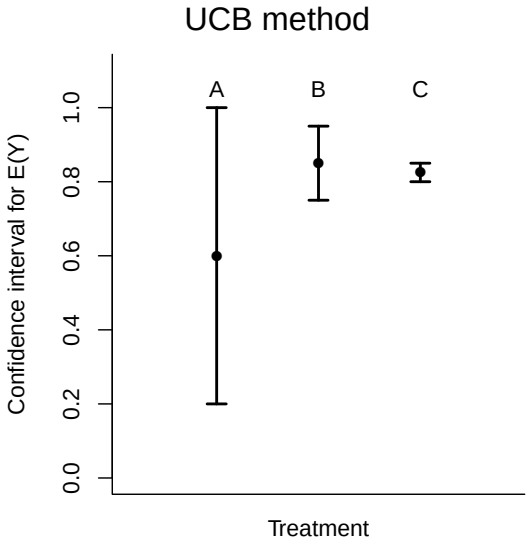


Figure 3.1: Hypothetical confidence intervals for $\mathbb{E}(Y \mid A)$, $\mathbb{E}(Y \mid B)$ and $\mathbb{E}(Y \mid C)$.

with B. Then it's confidence interval would tend to get narrower with further samples. It's center could also shift up or down as sampling goes on, tending towards the true mean which we anticipate to be somewhere inside the current confidence interval, though that won't always hold. What could happen is that the confidence interval converges on a value above the center for A but below the upper limit for A. Then if A were really as good as its upper limit, we would never sample it and find out. The same argument holds for sampling from C.

Now suppose that we sample from A. Its confidence interval will narrow and the center could move up or down. If A is really bad then sampling will move the mean down and narrow the confidence interval and it will no longer keep the top upper confidence limit. We would then stop playing it, at least for a while. If instead, A was really good, we would find that out.

Given that we want to use the upper limit of a confidence interval, what confidence level should we choose? The definitive sources on that point are Gittins (1979), Lai and Robbins (1985) and Lai (1987). We can begin with a finite horizon n, for instance the next n patients or visitors to a web page. At step i we could use the $100(1-\alpha_i)\%$ upper confidence limit.

It is easy to choose the treatment for the n'th subject. We just take the arm that we think has the highest mean. There is no $n+1$'st subject to benefit from what we learn from subject n. So we could take $\alpha_n = 1/2$. That would be the center of our confidence interval (if it is symmetric). If we are picking a fixed sequence $\alpha_1, \ldots, \alpha_n$ then it makes sense to have α_i increasing towards 0.5 because as time goes on, there is less opportuntity to take advantage of any learning. The α_i should start small, especially if n is large.

A finite horizon might not be realistic. We might not know how many subjects will be in the study. Another approach is to define the discounted regret

$$\sum_{i=1}^{\infty}(\mu_* - \mathbb{E}(\mu_{J_i}))\theta^{i-1} \qquad 0 < \theta < 1.$$

This regret is the immediate regret plus θ times a similar future quantity

$$\mu_* - \mathbb{E}(\mu_{J_1}) + \theta\sum_{i=1}^{\infty}(\mu_* - \mathbb{E}(\mu_{J_{i+1}}))\theta^{i-1}.$$

It is much more reasonable to use some constant α in the discounted setting than in the fixed n setting. Choosing θ can be complicated. The average index is

$$\sum_{i=1}^{\infty} i\theta^{i-1} \Big/ \sum_{i=1}^{\infty}\theta^{i-1} = \frac{1}{1-\theta}$$

by considering the mean of a geometric distribution. So if we pick $\theta = 0.99$ then the weighted average index is 100. Or, if we have a time horizon like $n = H$ in mind we can set $\theta = 1 - 1/H$.

Finding the critical α_i values is complicated and depends on the underlying parametric distribution one might assume for the distributions ν_j. For us, the main idea is that betting on optimism paid off by keeping pseudo-regret $O(\log(n))$.

3.4 Thompson sampling

Thompson sampling goes back to Thompson (1933). The method was forgotten and rediscovered a few times. It is comparatively recently that most articles are online and findable over the internet, so the reinvention is understandable. The idea in Thompson sampling is to choose arm i with probability $\Pr(\mu_i \text{ is the best})$. This probability is a Bayesian one, so that it can be updated based on the observations so far. Thompson's motivation was for medical problems. The current surge in interest is from internet services.

Agrawal and Goyal (2012) showed that Thompson sampling can have a pseudo-regret of $O(\log(n))$. That puts it in the same performance league as UCB methods and Thompson is easier to deploy at least in simple settings.

We will focus on the Bernoulli case. The response values are $Y_{i,j} \in \{0, 1\}$. Let $\mu_j = \Pr(Y_{i,j} = 1)$. Then the likelihood function that we will use is

$$L(\mu_1, \ldots, \mu_K) = \prod_{i=1}^{n} p(Y_{i,J_i} \,|\, J_i) = \prod_{i=1}^{n} \mu_{J_i}^{Y_{i,J_i}}(1-\mu_{J_i})^{1-Y_{i,J_i}}. \tag{3.1}$$

Each factor $\mu_{J_i}^{Y_{i,J_i}}(1-\mu_{J_i})^{1-Y_{i,J_i}}$ is a likelihood contribution for $(\mu_1, \ldots, \mu_K)$ based on the conditional distribution of $Y_i = Y_{i,J_i}$ given J_i. There's a brief discussion about using a conditional likelihood below.

Taking this conditional likelihood (3.1) as our likelihood, we then pick a conjugate prior distribution in the beta family. Taking $\mu_j \overset{\text{ind}}{\sim} \text{Beta}(a_j, b_j)$ they have joint prior distribution proportional to

$$\prod_{j=1}^{K} \mu_j^{a_j-1}(1-\mu_j)^{b_j-1} \qquad 0 \leqslant \mu_j \leqslant 1.$$

Taking $a_j = b_j = 1$ makes $\mu_j \sim \mathbb{U}[0,1]$ independently. This is a popular choice. If K is very large and we know that the μ_j are likely to be very near zero from past experience then we could work with $a_j < b_j$. The mean of $\text{Beta}(a,b)$ is $\mu = a/(a+b)$ and the variance is $\mu(1-\mu)/(a+b+1)$ and these facts might help us settle on (a_j, b_j).

Let $S_j = \sum_{i=1}^{n} Y_{i,J_i} 1\{J_i = j\}$ and $F_j = \sum_{i=1}^{n}(1-Y_{i,J_i})1\{J_i = j\}$ be the numbers of successes and failures, respectively, observed in arm j. Then we can write our conditional likelihood as

$$\prod_{j=1}^{K} \mu_j^{S_j}(1-\mu_j)^{F_j}.$$

Notice that although (S_j, F_j) are defined by summing over all $i = 1, \dots, n$, they do not depend on any unobserved Y values. Multiplying our conditional likelihood by the prior density we find a posterior density proportional to

$$\prod_{j=1}^{K} \mu_j^{a_j+S_j-1}(1-\mu_j)^{b_j+F_j-1}.$$

This means that the posterior distribution has $\mu_j \overset{\text{ind}}{\sim} \text{Beta}(a_j + S_j, b_j + F_j)$. This expression also shows that a_j and b_j can be viewed as numbers of prior pseudo-counts. We are operating as if we had already seen a_j successes and b_j failures from arm j before starting.

Figure 3.2 has pseudo-code for running the Thompson sampler for Bernoulli data and beta priors. In this problem it is easy to pick arm j with probability equal to the probability that μ_j is largest. We sample $\mu_1, \dots, \mu_K$ one time each and let J be the index of the biggest one we get.

Thompson sampling is convenient for web applications where we might not be able to update S_j and F_j as fast as the data come in. Maybe the logs can only be scanned hourly or daily to get the most recent (J_i, Y_{i,J_i}) pairs. Then we just keep sampling with the fixed posterior distribution between updates. If instead we were using UCB then we might have to sample the arm with the highest upper confidence limit for a whole day between updates. That could be very suboptimal if that arm turns out to be a poor one.

Puzzler: the UCB analysis is pretty convincing that we win by betting on optimism. How does optimism enter the Thompson sampler? We get just one draw from the posterior for arm j. That draw could be better or worse than the mean. Just taking the mean would not bake in any optimism and

Initialize:
$S_j \leftarrow a_j$, $F_j \leftarrow b_j$, $j = 1, \dots, K$ # $a_j = b_j = 1$ starts $\mu_j \sim \mathbb{U}[0,1]$

Run:
for $i \geqslant 1$
  for $j = 1, \dots, K$
    $\theta_j \sim \text{Beta}(S_j, F_j)$ # make sure $\min(S_j, F_j) > 0$
  $J \leftarrow \arg\max_j \theta_j$ # call it J_i if you plan to save them
  $S_J \leftarrow S_J + X_{i,J}$
  $F_J \leftarrow F_J + 1 - X_{i,J}$

Figure 3.2: Pseudo-code for the Thompson sampler with Bernoulli responses and beta priors. As written it runs forever.

would fail to explore. We could bake in more optimism by letting each arm take $m > 1$ draws and report its best result. I have not seen this proposal analyzed (though it might well be in the literature somewhere). It would play more towards optimism but that does not mean it will work better; optimism was just one factor. Intuitively, taking $m > 1$ should favor the arms with less data, other things being equal. Without some theory, we cannot be sure that $m > 1$ doesn't actually slow down exploration. Maybe it would get us stuck in a bad arm forever (I doubt that on intuitive grounds only). If we wanted, we could take some high quantile of the beta distributions but deciding what quantile to use would involve the complexity that we avoided by moving from UCB to Thompson. For Bernoulli responses with a low success rate, the beta distributions will initially have a positive skewness. That is a sort of optimism.

Puzzler/rabbit hole: are we leaving out information about μ_j from the distribution of J_i? I think not, because the distribution of J_i is based on the past Y_i which already contribute to the conditional likelihood terms. A bit of web searching did not turn up the answer. It is clear that if you were given $J_1, \dots, J_n$ it would be possible at the least to figure out which μ_j was μ_*. But that doesn't mean they carry extra information. The random variables are observed in this order:

$$J_1 \to Y_1 \to J_2 \to Y_2 \to \cdots \to J_i \to Y_i \to \cdots \to J_n \to Y_n.$$

Each arrow points to new information about μ. The distribution of J_1 does not depend on $\mu = (\mu_1, \dots, \mu_K)$. The likelihood is

$$p(y_1 \mid J_1; \mu) p(J_2 \mid J_1, y_1; \mu) p(y_2 \mid J_2, J_1, y_1; \mu) p(J_3 \mid y_2, J_2, J_1, y_1; \mu) \cdots$$

Now in our Bernoulli Thompson sampler our algorithm for choosing J_3 was just based on a random number generator that was making our beta random variables. That convinces me that $p(J_3 \mid y_2, J_2, J_1, y_1; \mu)$ has nothing to do with μ. So the conditional likelihood is ok. At least for the Bernoulli bandit. Phew!

3.5 Theoretical findings on Thompson sampling

Agrawal and Goyal (2012) generalize Bernoulli Thompson sampling to handle bounded non-binary inputs. They get a value $Y \in \{0,1\}$ from $Y' \in [0,1]$ by randomly taking $Y = 1$ with probabilty Y'. This is pure randomness coming from their algorrithm not their data. If the response is actually $Y'' \in [a,b]$ for known $b > a$ we can start by setting $Y' = (Y'' - a)/(b-a)$ and then sampling $Y \sim \mathrm{Bern}(Y')$.

Theorem 1. *For $K = 2$ and $Y'_{i,j} \in [0,1]$ and $Y_{i,j} \sim \mathrm{Bern}(Y'_{i,j})$ the pseudo-regret is*

$$R_n = O\Big(\frac{\log(n)}{\Delta} + \frac{1}{\Delta^3}\Big),$$

as $n \to \infty$, where Δ is the suboptimality of the second best treatment.

Proof. This is Theorem 1 of Agrawal and Goyal (2012). They refer to expected regret but it appears to be pseudo-regret in the terminology of Bubeck and Cesa-Bianchi (2012). □

The pseudo-regret grows like $O(\log(n))$ for fixed Δ. If arm 1 is the suboptimal one then this means that $\mathbb{E}(T_1(n)) = O(\log(n))$. If the cumulative number of mistakes grows logarithmically then the typical gap between mistake times has to be growing exponentially. For instance $\mathbb{E}(T_1(2n) - T_1(n)) = O(\log(2n) - \log(n)) = O(1)$ (because the constant $\log(2)$ is $O(1)$). Each doubling of n brings at most a constant expected number of suboptimal arm choices.

Now let's look into the denominator Δ. The pseudo-regret is larger when Δ is smaller. If arms return 2% and 3% respectively, the bound leads us to expect much worse results than if they are 2.99% and 3%. The reason is that a suboptimal but nearly optimal arm will get chosen much more often than one that is very suboptimal. What about $\Delta = 0$? Something discontinuous happens here. The pseudo-regret bound is ∞. The actual pseudo-regret is exactly 0. It is not a contradiction: $0 < \infty$.

I promised a discussion of ***external validity of theorems***. The big-O in our theorem means that there exist constants $C < \infty$ and $N < \infty$ such that

$$\bar{R}_n \leqslant C \times \Big(\frac{\log(n)}{\Delta} + \frac{1}{\Delta^3}\Big) \quad \text{for all } n \geqslant N.$$

Sometimes it holds for $N = 1$. In a given situation we might expect $\bar{R}_n$ to grow like $\log(n)$ but be disappointed. Maybe it happens for extremely large N (in our problem). Or maybe the value of C is so large that $C\log(n)/n$ is not very small for any n that we can afford. We need more information than the $O(\cdot)$ result in the theorem to know if things are going to be good.

It can be valuable to spot-check theorems with some simulated examples. Simulated examples on their own are unsatisfying, also for external validity reasons. The ones in the literature might have been cherry-picked. Ideally the examples are known to be similar to our use case or perhaps to cover a wide range of possibilities.

A theorem can also be misleadingly pessimistic. If an error quantity $E_n = o(g(n))$ it means that $\lim_{n\to\infty} g(n)|E_n| = 0$. Anything that is $o(g(n))$ is automatically $O(g(n))$. In this case there is a theorem from Lai (1987) showing that no method could be $o(\log(n))$, so that doesn't happen here.

In this case we can think of what is perhaps the best possible case for Thompson sampling. One arm has $\mu_j = 1$ and the other has $\mu_j = 0$.

Puzzler: In class, I wondered what would happen if instead of adding Y_{i,J_i} to S_j and $1 - Y_{iJ_i}$ to F_j we added the probabilities Y'_{I,J_i} to S_j and $1 - Y'_{iJ_i}$ to F_j. That takes some noise out of the algorithm. It leads to beta distributions with non-integral parameters, but those are ok. It would complicate the analysis behind the theorem. We know from Lai and Robbins that we would not get a better convergence rate than $O(\log(n))$. Specifically, their bound is of the form

$$\left(\sum_{j=2}^{K} \frac{\Delta_j}{\mathrm{KL}(\nu_j||\nu_*)} + o(1)\right)\log(n),$$

for Kullback-Leibler divergence

$$\mathrm{KL}(P||Q) = \int_{-\infty}^{\infty} p(x)\log\Big(\frac{p(x)}{q(x)}\Big)\,\mathrm{d}x$$

in the case of continuous distributions P and Q with a natural modification for discrete distributions. Perhaps we get a better constant in the rate from using Y' instead of Y.

Agrawal and Goyal (2012) have additional results to cover the case $K > 2$.

Theorem 2. *For $K > 2$ and optimal arm $j^* = 1$*

$$\bar{R}_n = O\left(\left(\sum_{j=2}^{K}\frac{1}{\Delta_j^2}\right)^2\log(n)\right)$$

for suboptimalities Δ_j. Also

$$\bar{R}_n = O\left(\frac{\Delta_{\max}}{\Delta_{\min}^3}\sum_{j=2}^{K}\frac{1}{\Delta_j^2}\right)\log(n)\Bigg)$$

where $\Delta_{\min} = \min_{2\leqslant j\leqslant K}\Delta_j$ and $\Delta_{\max} = \max_{2\leqslant j\leqslant K}\Delta_j$.

Proof. The first result is Theorem 2 of Agrawal and Goyal (2012) and the second is their Remark 3. □

The second bound is better for large K, while the first is better for small Δ_j.

3.6 More about bandits

Suppose that A is better than B. Then from a bandit we only get $O(\log(n))$ samples from B. Therefore we do not get a good estimate of

$$\Delta = \mathbb{E}(Y\,|\,A) - \mathbb{E}(Y\,|\,B).$$

We can be confident that we are taking the best arm but we cannot get a good estimate of the amount of improvement. For some purposes we might want to know how much better the best arm is.

Maybe $W_i = (W_{i1}, \ldots, W_{i,10}) \in \{0,1\}^{10}$ because we have 10 decisions to make for subject i. We could run a bandit with $K = 2^{10}$ arms but that is awkward. An alternative is to come up with a model, such as $\Pr(Y = 1 \,|\, W) = \Phi(W^\mathsf{T}\beta)$ for unknown β. Or maybe a logistic regression would be better. We can place a prior on β and update it as data come in. Then we need a way to sample a W with probability proportional to it being the best one. Some details for this example are given in Scott (2010). These can be hard problems but the way forward via Thompson sampling appears easier than generalizing UCB. This setting has an interesting feature. Things we learn from one of the 1024 arms provide information on β and thereby update our prior on some of the other arms.

For contextual bandits, we have a feature vector X_i that tells us something about subject i before we pick a treatment. Now our model might be $\Pr(Y = 1 \mid X, W; \beta)$ for parameters β. See Agrawal and Goyal (2013) for Thompson sampling and contextual bandits.

In restless bandits, the distributions ν_j can be drifting over time. See Whittle (1988). Clearly we have to explore more often in this case because some other arm might have suddenly become much more favorable than the one we usually choose. It also means that the very distant past observations might not be relevant, and so the upper confidence limits or parameter distributions should be based on recent data with past data downweighted or omitted.

4

Paired and blocked data, randomization inference

In this lecture we begin to look at some more traditional areas of experimental design. Much of it is based on the work by George Box and co-authors. I quite like this book: Box et al. (1978). I'm citing the first edition which I prefer to the second. Wu and Hamada (2011) cover many of the same ideas with a rich collection of examples, mostly from industrial experimentation.

These basic experimental design ideas have been used to give us about a century of exponential growth in the quality and abundance of food and medicine and industrial products. Ideas and insights from domain experts get boosted by the efficiency with which well designed experiments can speed up learning of causal relationships.

In this work we take regular regression theory as a prerequisite. Things like normal theory regression, t-tests, p-values, confidence intervals and how to analyze such data are mostly assumed. This course is mostly about making data, while most other courses are about analyzing data. One exception: we will cover the analysis of variance (ANOVA) in more detail than usual statistics courses do. The ANOVA cannot be completely understood just in terms of adding binary predictors, sometimes called a one-hot encoding. There is a bit more going on.

The class web page has notes from Stat 305A on the one way ANOVA. Read up through Chapter 1.2 and then Chapter 1.7 on random effects which we will cover later. In between there is material on statistical power, interpretation of treatment contrasts, multiple comparisons for ANOVA and false discovery rates.

4.1 The ordinary two sample t-test

Let's recall how we would do a t-test for a treatment effect. We have data Y_{ij} for treatment groups $i = 1, 2$ and observations $j = 1, \dots, n_i$. Think of i as $W + 1$, where $W \in \{0, 1\}$ is the treatment variable in causal inference from Chapter 1. The goal is to learn about $\Delta = \mathbb{E}(Y_{1i}) - \mathbb{E}(Y_{2i})$. Defining this expectation will require a model, and unlike the science tables in potential outcomes, Δ here does not depend on i. The t statistic is

$$t_{\text{obs}} = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} - \Delta}{s\sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}.$$

This t_{obs} is the observed value of a t-distributed random variable. Here $\bar{Y}_{i\bullet} = (1/n_i)\sum_{j=1}^{n_i} Y_{ij}$ and s^2 is the pooled variance estimate. This result is a miracle. We have an algebraic expression t_{obs} involving our unknown Δ and some quantities that are known after a short computation. Arithmetic that combines knowns and unknowns ordinarily returns an unknown. This result is special, because while t_{obs} is unknown it has a known distribution. It is then called a ***pivotal*** quantity.

Using the pivotal quantity we can get a 99% confidence interval for Δ as

$$\left\{\Delta \mid |t_{\text{obs}}| \leqslant t^{0.995}_{(n_1+n_2-2)}\right\}.$$

If a special value of Δ, call it Δ_0 is not in the confidence interval then we reject $H_0 : \Delta = \Delta_0$ at the 1% level. The usual Δ_0 is of course 0, corresponding to a null hypothesis of no treatment effect. We can get a p-value for $H_0 : \Delta = \Delta_0$ as

$$p = \Pr\bigl(|t_{(n_1+n_2-2)}| \geqslant |t_{\text{obs}}|\bigr).$$

We can also get these results by pooling all $n_1 + n_2$ data into a regular regression model

$$Y_j = \beta_0 + \beta_1 W_i + \varepsilon_i, \quad i = 1, \dots, N \equiv n_1 + n_2 \tag{4.1}$$

where $W_i = 1$ if observation i is from treatment 1 and $W_i = 0$ if observation i is from treatment 2. Defining

$$X = \begin{pmatrix} 1 & W_1 \\ \vdots & \vdots \\ 1 & W_{n_1} \\ 1 & W_{n_1+1} \\ \vdots & \vdots \\ 1 & W_{n_1+n_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}$$

and similarly defining $Y \in \mathbb{R}^N$ with the treatment 1 data above the treatment 2 data, we can compute

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y \quad \text{and} \quad s^2 = \frac{1}{N-2}\sum_{i=1}^{N}(Y_j - \hat{\beta}_0 - \hat{\beta}_1 W_i)^2$$

and now

$$t_{\mathrm{obs}} = \frac{\hat{\beta}_1 - \beta_1}{s\sqrt{((X^\mathsf{T}X)^{-1})_{22}}}.$$

In order to get these pivotal inferences we need to make 4 assumptions:

1) ε_i are normally distributed,

2) $\mathrm{var}(\varepsilon_i)$ does not depend on W_i,

3) ε_i are independent, and

4) there are no missing predictors.

For the last one, we need to know that $\mathbb{E}(Y_j)$ is not really $\beta_0 + \beta_1 W_i + \beta_2 U_i$ for some other variable U_i.

Assumption 1 is hard to believe, but the central limit theorem reduces the damage it causes. Assumption 2 can be serious but does little damage if $n_1 \doteq n_2$. We can also just avoid pooling the variances and use $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ in place of $s\sqrt{1/n_1 + 1/n_2}$.

Assumption 3 is critical and violations can be hard to detect. Assumption 4 is even more critical and harder to detect. We almost don't even notice we are making an assumption about U_i because U_i is missing from equation (4.1).

4.2 Randomization fixes assumptions

Box et al. (1978) consider a hypothetical neighbor with two fertilizers and 11 tomato plants. Let's go with 10 plants. We could plant them in a row like this:

A	A	A	A	A	B	B	B	B	B

That would not be a good design. Maybe there's a hidden trend variable $U_i = i$ where the plots correspond to $i = 1, \ldots, 10$ from left to right.

We could instead try:

A	B	A	B	A	B	A	B	A	B

That is better but could still be problematic. For instance there could be correlations between the yield of adjacent plants. Those would be positive if nearby locations had similar favorability. Or they could be negative if one plants roots or shade adversely affected its neighbors.

We could then try randomizing the run order perhaps getting this:

A	B	B	B	A	A	B	A	A	B

A random order cannot correlate with any trend.

Under our model, the t statistic numerator $\hat{\Delta} - \Delta = \bar{Y}_{\bullet A} - \bar{Y}_{\bullet B} - \Delta$ equals

$$\begin{aligned}
&\big(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 - \varepsilon_6 - \varepsilon_7 - \varepsilon_8 - \varepsilon_9 - \varepsilon_{10}\big)/5, && \text{A's first}\\
&\big(\varepsilon_1 - \varepsilon_2 + \varepsilon_3 - \varepsilon_4 + \varepsilon_5 - \varepsilon_6 + \varepsilon_7 - \varepsilon_8 + \varepsilon_9 - \varepsilon_{10}\big)/5, && \text{alternate}\\
&\big(\varepsilon_1 - \varepsilon_2 - \varepsilon_3 - \varepsilon_4 + \varepsilon_5 + \varepsilon_6 - \varepsilon_7 + \varepsilon_8 + \varepsilon_9 - \varepsilon_{10}\big)/5, && \text{random}
\end{aligned}$$

in our three allocations.

If there is an unknown U_i then it is within the ε_i. If $U_i = c \times (i - 5.5)$ then our model has put that U_i inside ε_i and we get a bias of

$$\mathbb{E}(\hat{\Delta}) - \Delta = \begin{cases} -5c, & \text{A's first} \\ -c, & \text{alternate} \\ 0.6c, & \text{random.} \end{cases}$$

Putting A's first gave the worst bias. The alternating plan improved a lot, but could have done very badly with some high frequency bias. The random plan came out best. The bias will be $O_p(1/\sqrt{N})$ under randomization, whether the U_i constitute a trend or an oscillation or something else.

Next, let's consider what happens if there are correlations in the ε_i. We will consider local correlations

$$\operatorname{corr}(Y_i, Y_{i'}) = \begin{cases} 1, & i = i' \\ \rho, & |i - i'| = 1 \\ 0, & \text{else.} \end{cases}$$

Now

$$\operatorname{var}(\hat{\Delta}) = \frac{1}{25} v^{\mathsf{T}} \operatorname{cov}(\varepsilon) v$$

where $v_i = 1$ for $W_i = 1$ and $v_i = -1$ for $W_i = 0$. Using σ^2 for $\operatorname{var}(\varepsilon_i)$, we get

$$\operatorname{var}(\hat{\Delta}) = \frac{2}{5}\sigma^2 + \frac{\sigma^2}{25} \times \begin{cases} 14\rho, & \text{A's first} \\ -18\rho, & \text{alternate} \\ 2\rho, & \text{random.} \end{cases}$$

The data analyst will ordinarily proceed as if $\rho = 0$ especially in small data sets where we cannot estimate ρ very well. For the plants ρ could well be positive or negative making $\operatorname{var}(\hat{\Delta})$ quite different from $2\sigma^2/5$.

Box et al. (1978) take the view that randomization makes it reasonably safe to use our usual statistical models. A forthcoming book by Tirthankar Dasgupta and Donal Rubin will, I expect, advocate for using the actual randomization that was done to drive the inferences.

4.2.1 About permutation testing

The original motivation for the t-test by Fisher was based on the asymptotic equivalence between a t-test and a permutation test. As a result we do not expect permutation tests to repair any problems that would have affected the t-test.

A t-test tests the 'small' null hypothesis $H_0 : \mathbb{E}(Y \mid A) = \mathbb{E}(Y \mid B)$ that the mean of Y is the same for $W = 0$ and $W = 1$. A permutation test addresses the 'large' null hypothesis $\mathcal{H}_0 : \mathcal{L}(Y \mid A) = \mathcal{L}(Y \mid B)$. Here $\mathcal{L}(\cdot)$ refers to the law or distribution of contents, so this hypothesis makes the strong assumption that

the distribution of Y is exactly the same for $W = 0$ and $W = 1$. It is a test of $\mathcal{H}_0$ designed to have power versus H_0.

Permutation tests have the advantage that they are very easy to explain to non-statistician users and they appear to have very clear validity.

Permutation tests can be cumbersome. In an observational setting where we get (X_i, Y_i, W_i) for a $W_i \in \{0, 1\}$ and $X_i \in \mathbb{R}$ it is tricky to use permutations to study whether $Y \perp\!\!\!\perp W$. We could permute (W, X) versus Y or we could permute W versus (X, Y). Neither gives an exact test. [This was studied by David Freedman.] Losing exactness loses a lot of the motivation behind permutations.

One of the best analyses of permutation tests is in the statistical theory book by Lehmann and Romano. They show how it comes from a group symmetry argument.

We will take the BHH view that if our experiment was randomized then we are reasonably safe to use the usual regression models.

4.3 Paired analysis

The next (hypothetical) example from BHH involves 10 kids and running shoes. There were two different materials for the soles of those shoes. Each kid gets one material on the right shoe and the other one on the left. We can diagram the situation as follows, deciding randomly whether to use left or right for material A:

$A \mid B$	$B \mid A$	$B \mid A$	$B \mid A$	$A \mid B$
$A \mid B$	$B \mid A$	$A \mid B$	$A \mid B$	$B \mid A$

BHH contemplate very big differences between the kids. Suppose that some are in the chess club while others prefer skateboarding. Figure 4.1 shows an exaggerated simulated example of how this might come out. The left panel shows that tread wear varies greatly over the 30 subjects there but just barely between the treatments. The right panel shows a consistent tendency for tread B to show more wear than tread A, though with a few exceptions.

The way to handle it is is via a paired t-test. Let $D_i = Y_{1i} - Y_{2i}$ for $i = 1, \ldots, n$ (so there are $N = 2n$ measurements). Then do a one-sample t-test for whether $\mathbb{E}(D) = \Delta$ where Δ is ordinarily 0.

The output from a paired t-test on this data is

```
t = -2.7569, df = 29, p-value = 0.009989
95 percent confidence interval: -0.59845150 -0.08868513
```

with of course more digits than we actually want. The difference is barely significant at the 1% level. An unpaired t-test on this data yields:

```
t = -0.2766, df = 57.943, p-value = 0.7831
95 percent confidence interval:  -2.829992  2.142856
```

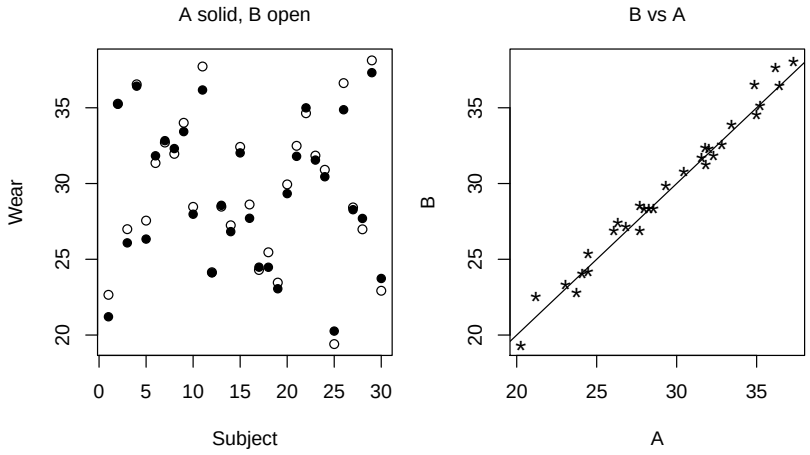


Figure 4.1: Hypothetical shoe wear numbers for 30 subject and soles A versus B.

and the difference is not statistically significant, with a much wider confidence interval.

In this setting the paired analysis is correct or at least less wrong and that is not because of the smaller p-value. It is because the unpaired analysis ignores correlations between measurements for the left and right shoe of a given kid.

In class somebody asked what would be missing from the science table for this example. We get both the A and B numbers. What we don't get is what would have happened if a kid who got $\boxed{A}\boxed{B}$ had gotten $\boxed{B}\boxed{A}$ instead. The science table would have had a row like $\boxed{LA}\boxed{LB}\boxed{RA}\boxed{RB}$ for each kid and we would only see two of those four numbers. We would never get $\boxed{LA}\boxed{LB}$ for any of the kids. It is certainly possible that there are trends where left shoes get a different wear pattern than right shows. Randomization protects against that possibility.

If we model the (Y_{1j}, Y_{2j}) pairs as random with a correlation of ρ and equal variance σ^2 then our model gives

$$\mathrm{var}(D_j) = \mathrm{var}(Y_{1j} - Y_{2j}) = 2\sigma^2(1-\rho)$$

and we see that the higher the correlation, the more variance reduction we get. Experimental design offers possibilities to reduce the variance of your data and this is perhaps the simplest such example.

The regression model for this paired data is

$$Y_{ij} = \mu + b_j + \Delta W_{ij} + \varepsilon_{ij}$$

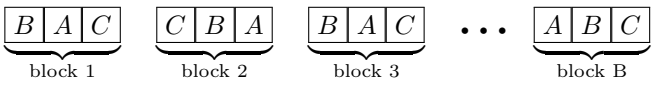
where b_j is a common effect from the j'th pair, Δ is the treatment effect and $W_{ij} \in \{0, 1\}$ is the treatment variable. This model forces the treatment difference to be the same in every pair. Then

$$D_j = Y_{1j} - Y_{2j} = (\mu + b_j + \Delta W_{1j} + \varepsilon_{1j}) - (\mu + b_j + \Delta W_{2j} + \varepsilon_{2j}) = \Delta + \varepsilon_{1j} - \varepsilon_{2j}.$$

4.4 Blocking

Pairs are blocks of size 2. We can use blocks of any size $k \geqslant 2$. They are very suitable when there are $k \geqslant 2$ treatments to compare. Perhaps the oven can hold $k = 3$ cakes at a time. Or the car has $k = 4$ wheels on it at a time.

If we have $k = 3$ treatments and block size of 3 we can arrange the treatments as follows:



with independent random assigments within each of B blocks.

Suppose that there are positive correlations for measurements within blocks but independence between blocks. Then differences of averages $\bar{Y}_{A\bullet} - \bar{Y}_{B\bullet}$, $\bar{Y}_{A\bullet} - \bar{Y}_{C\bullet}$, and $\bar{Y}_{B\bullet} - \bar{Y}_{C\bullet}$ should cancel out block effects just like we saw with paired tests and be more accurate than unblocked experiment:

A	C	B	B	C	B	A	A	$\cdots$	B

with $N = kB$ cells. This latter design would be randomized completely in one of $N!/(B!)^k$ ways.

There are lots of use cases for blocked experiments in agriculture and a few from medicine and industry. In each of the settings below we might have B blocks that each can have k experimental runs.

Treatments	Block	Response
Potato variety	Farm split into k plots	Yield
Cake recipe	Bake event, oven holds k cakes	Moisture
Diets	Litters of k animals	Weight gain
Cholesterol meds	Volunteer	Chol. levels
Sunscreen	Volunteer	Damage
Technician	Shift	Production
Ways to teach reading	School	Comprehension
Ion injection	Cassette of Si wafers	Yield or speed

A block is usually about the same size as our number of treatments. If the problem is to compare a control treatment to $k - 1$ alternatives and the block has size $k + 1$ then we might apply the control treatment twice within each block, especially if comparisons to the control of greatest importance.

4.5 Basic ANOVA

The class web page has Stat 305A notes on how to use regression to analyze this ANOVA.

The statistical model for a most basic ANOVA comparing $k \geqslant 2$ treatments is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, k \quad j = 1, \ldots, n_i. \tag{4.2}$$

This is called the one-way ANOVA because it has only one treatment factor. We will later consider multiple treatment factors. This model is not identified, because we could replace μ by $\mu-\eta$ and α_i by $\alpha_i+\eta$ for any $\eta\in\mathbb{R}$ without changing Y_{ij}. One way to handle that problem is to impose the constraint $\sum_{i=1}^k n_i\alpha_i=0$. Many regression packages would force $\alpha_1=0$. This model can be written

$$Y_{ij}=\mu_i+\varepsilon_{ij} \tag{4.3}$$

which is known as the ***cell mean model***. We can think of a grid of boxes or cells $\boxed{\mu_1}\boxed{\mu_2}\boxed{\cdots}\boxed{\mu_k}$ and we want to learn the mean response in each of them.

The null hypothesis is that the treatments all have the same mean. That can be written as

$$H_0:\mu_1=\mu_2=\cdots=\mu_k$$

or as

$$H_0:\alpha_1=\alpha_2=\cdots=\alpha_k=0.$$

The 'big null' is that $\mathcal{L}(Y_{i1})=\mathcal{L}(Y_{i2})=\cdots=\mathcal{L}(Y_{il})$ and that is what permutations test.

We can test H_0 by standard regression methods. Under H_0 the linear model is just

$$Y_{ij}=\mu+\varepsilon_{ij}. \tag{4.4}$$

We could reject H_0 by a likelihood ratio test if the 'full model' (4.3) has a much higher likelihood than the 'sub model' (4.4). When the likelihoods involve Gaussian models, log likelihoods become sums of squares and the results simplify.

Here are the results in the balanced setting where $n_i=n$ is the same for all $i=1,\dots,k$. The full model has MLE

$$\hat\mu_i=\bar Y_{i\bullet}=\frac{1}{n}\sum_{j=1}^n Y_{ij}$$

and sum of squares

$$\sum_{i=1}^k\sum_{j=1}^n(Y_{ij}-\hat\mu_i)^2=\sum_{i=1}^k\sum_{j=1}^n(Y_{ij}-\bar Y_{i\bullet})^2.$$

The sub-model from the null hypothesis has MLE

$$\hat\mu=\bar Y_{\bullet\bullet}=\frac{1}{k}\sum_{k=1}^k\bar Y_{i\bullet}=\frac{1}{nk}\sum_{i=1}^k\sum_{j=1}^n Y_{ij}.$$

Source	df	SS	MS	F
Treatments	$k-1$	SSB	$\mathrm{MSB} = \mathrm{SSB}/(k-1)$	MSB/MSW
Error	$N-k$	SSW	$\mathrm{MSW} = \mathrm{SSW}/(N-k)$	
Total	$N-1$	SST		

Table 4.1: This is the ANOVA table for a one way analysis of variance.

These sums of squared errors are connected by the ANOVA identity

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2}_{\mathrm{SST}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2}_{\mathrm{SSB}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i\bullet})^2}_{\mathrm{SSW}}.$$

The total sum of squares is equal to the sum of squares between treatment groups plus the sum of squares within treatment groups. This can be seen algebraicly by expanding $\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i\bullet}+\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2$. It is also just Pythagoras (orthogonality of the space of fits and residuals) from a first course in regression.

The F-test statistic based on the extra sum of squares principle is

$$F = \frac{\frac{1}{k-1}\left(\mathrm{SSE}_{\mathrm{null}}-\mathrm{SSE}_{\mathrm{full}}\right)}{\frac{1}{N-k}\mathrm{SSE}_{\mathrm{full}}} = \frac{\frac{1}{k-1}(\mathrm{SST}-\mathrm{SSW})}{\frac{1}{N-k}\mathrm{SSW}} = \frac{\frac{1}{k-1}\mathrm{SSB}}{\frac{1}{N-k}\mathrm{SSW}} \equiv \frac{\mathrm{MSB}}{\mathrm{MSW}}.$$

Here, $N = \sum_i n_k = nk$ is the total sample size. When we divide a sum of squares by its degrees of freedom the ratio is called a mean square. We should reject the null hypothesis if MSB is large. The question 'how large?' is answered by requiring it to be a large enough multiple of MSW. We reject H_0 if $p = \Pr(F_{k-1,N-k} \geqslant F; H_0)$ is small.

These notes assume familiarity with the simple ANOVA tables for regression and the one way analysis of variance. Table 4.1 contains the ANOVA table for this design. There are two sources of variation in this data: treatment groups and error. Because there are k treatments there are $k-1$ degrees of freedom. There are $n_i - 1$ degrees of freedom for error in each of the k treatment groups for a total of $\sum_i (n_i - 1) = N - k$. There is often another column for the p-value.

The mean square column provides information on statistical significance. The sum of squares column is about practical significance. For instance $R^2 = \mathrm{SSB}/\mathrm{SST}$ is the fraction of variation explained by the model terms.

To see why we care about mean squares consider $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_N)$. This is a vector of noise that can be projected onto a one dimensional space parallel to $(1, 1, \ldots, 1)$ where it affects $\bar{Y}_{\bullet\bullet} = \hat{\mu}$, a $k-1$ dimensional space spanned by between treatment differences $\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$ where it affects SSB and an $N-k$ dimensional space of within treatment differences $Y_{ij} - \bar{Y}_{i\bullet}$. If Y_{ij} would be just noise ε_{ij} then we would have $\hat{\mu}^2 \sim \sigma^2\chi^2_{(1)}$, $\mathrm{SSB} \sim \chi^2_{(k-1)}$ and $\mathrm{SSW} \sim \chi^2_{(N-k)}$, all independent. The χ^2 mean equals its degrees of freedom and so we normalize sums of squares into mean squares.

4.6 ANOVA for blocks

The model for a blocked analysis is

$$Y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij} \qquad i = 1, \dots, k \quad j = 1, \dots, n.$$

Note that this model does not include an interaction. The treatment differences $\alpha_i - \alpha_{i'}$ are the same in every block j. All values in block j are adjusted up or down by the same constant b_j. We denote it by b_j instead of β_j because we may not be very interested in block j per se. A block might be a litter of animals or one specific run through of our laboratory equipment. In a surfing competition it might be about one wave with three athletes on it. That wave is never coming back so we are only interested in α_i, and maybe how that wave helps us compare α_i for different i, but not b_j.

The parameter estimates here are $\hat\mu = \bar Y_{\bullet\bullet}$, $\hat\alpha_i = \bar Y_{i\bullet} - \bar Y_{\bullet\bullet}$, $\hat b_j = \bar Y_{\bullet j} - \bar Y_{\bullet\bullet}$, and

$$\hat\varepsilon_{ij} = Y_{ij} - \hat\mu - \hat\alpha_i - \hat b_j = Y_{ij} - \bar Y_{i\bullet} - \bar Y_{\bullet j} + \bar Y_{\bullet\bullet} = (Y_{ij} - \bar Y_{i\bullet}) - (\bar Y_{\bullet j} - \bar Y_{\bullet\bullet}).$$

We should get used to seeing these alternating sign and difference of differences patterns.

The ANOVA decomposition is

$$\mathrm{SST} = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSE}$$

where

$$\begin{aligned}
\mathrm{SST} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(Y_{ij} - \bar Y_{\bullet\bullet})^2,\\
\mathrm{SSA} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(\bar Y_{i\bullet} - \bar Y_{\bullet\bullet})^2,\\
\mathrm{SSB} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(\bar Y_{\bullet j} - \bar Y_{\bullet\bullet})^2, \quad \text{and}\\
\mathrm{SSE} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(Y_{ij} - \bar Y_{i\bullet} - \bar Y_{\bullet j} + \bar Y_{\bullet\bullet})^2.
\end{aligned}$$

The ANOVA table for it is in Table 4.2. You could write SSA as $\sum_i B(\bar Y_{i\bullet} - \bar Y_{\bullet\bullet})^2$ and that is definitely what you would do in a hand calculation. The way it is written is more intuitive. All the sums of squares are sums over all data points.

We test for treatment effects via

$$p = \Pr\Big(F_{k-1,(k-1)(B-1)} \geqslant \frac{\mathrm{MSA}}{\mathrm{MSE}}\Big).$$

It is sometimes argued that one ought not to test for block effects. I don't quite understand that. If it turns out that blocking is not effective, then we

Source	df	SS	MS	F
Treatments	$k-1$	SSA	$\text{MSA} = \dfrac{\text{SSA}}{k-1}$	$\dfrac{\text{MSA}}{\text{MSE}}$
Blocks	$B-1$	SSB	$\text{MSB} = \dfrac{\text{SSB}}{B-1}$	$(*)$
Error	$(k-1)(B-1)$	SSE	$\text{MSE} = \dfrac{\text{SSE}}{N-k}$	
Total	$N-1$	SST		

Table 4.2: This is the ANOVA table for a blocked design.

could just not do it in the next experiment which might then be simpler to run and have more degrees of freedom for errror. A test can be based on $\text{MSB}/\text{MSE} \sim F_{B-1,(k-1)(B-1)}$.

The very old text books going back to 1930s place a lot of emphasis on getting sufficiently many degrees of freedom for error. That concern is very relevant when the error degrees of freedom are small, say under 10. The reason can be seen by looking at quantiles of $F_{\text{num,den}}$ such as $F^{.995}_{\text{num,den}}$ and $F^{.005}_{\text{num,den}}$ when the denominator degrees of freedom den is small. Check out `qf` in R, or it's counterpart in python or matlab. It is not a concern in A/B testing with thousands or millions of observations.

4.7 Latin squares

Latin squares let you block on two sources of unwanted variation at once. Suppose that you are testing 4 battery chemistries: A, B, C, D. You have 4 different drivers and 4 different cars. The following diagram has each of A, B, C and D exactly once per row and exactly once per column.

$$\begin{array}{c} \\ \text{Car } 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} A & B & C & D \\ C & D & A & B \\ B & C & D & A \\ D & A & B & C \end{bmatrix} \end{array}$$

You could have driver 1 (column) test car 1 one with treatment A. Then driver 2 takes car 2 with B and so on through all 16 cases ending up with driver 4 taking car 4 with treatment C. Now if there are car to car differences they are balanced out with respect to treatments. Driver to driver differences are also balanced out. This design only lets one car and one driver be on the track at once.

The model for this design is

$$Y_{ijt} = \mu + \underbrace{a_i}_{\text{row}} + \underbrace{b_j}_{\text{col}} + \underbrace{\tau_k}_{\text{trt}} + \underbrace{\varepsilon_{ijk}}_{\text{err}}.$$

k:	1	2	3	4	5	6	7
#:	1	1	1	4	56	9,408	16,942,080

Table 4.3: This is integer sequence number A000315 in the online encyclopedia of integer sequences by Neil J. A. Sloane: `https://oeis.org/A000315`.

It does not allow for any interactions between cars and drivers, cars and batteries or drivers and batteries. Later when we take a closer study of interactions we will see that an interaction between cars and drivers could look like an effect of batteries. If there are no significant interactions like this then a Latin square can be an extremely efficient way to gather information. Otherwise it is risky. Sometimes a risky strategy pays off better than a cautious one. Other times not.

To use a Latin square we start with a basic Latin square, perhaps like this one

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

and then randomly permute the rows and columns. We might as well also permute the symbols in it. Even if that is not necessary, it is easy to do, and simpler to just do it than think about whether you should. The above Latin square is called a ***cyclic*** Latin square because the rows after the first are simply their predecessor shifted left one space with wraparound.

The number of distinct $k \times k$ Latin squares to start with is given in Table 4.3. Two Latin squares are distinct if you cannot change one into the other by permuting the rows and columns and symbols. The number grows quickly with k. Be sure to permute the Latin square, especially if your starting pattern is cyclic. The cyclic pattern will be very bad if there is a diagonal trend in the layout. In many of the original uses the Latin square was made up of k^2 plots of land for agriculture.

Not only are Latin squares prone to trouble with interactions, they also have only a few degrees of freedom. With k^2 data points there are $k^2 - 1$ degrees of freedom about the mean. We use up $k - 1$ of them for each of rows, columns and treatments. That leaves $k^2 - 1 - 3(k-1) = (k-1)(k-2)$ degrees of freedom for error.

Box et al. (1978, Chapter 8) provide a good description of how to analyze Latin squares. I changed their car and driver example to have electric cars. They give ANOVA tables for Latin squares and describe how to replicate them in order to get more degrees of freedom for error. In a short course like this one, we will not have time to go into those analyses.

4.8 Esoteric blocking

There are a lot of more complicated and intricate ways to design experiments in blocks. I describe a few of them below. I consider them things to "know about". If you ever find that you need them, then being able to connect the problem they solve to their name will help you search for designs and analysis strategies. They're interesting to contemplate and we can really admire them from an aesthetic point of view. We will return to one of them later when we do space filling designs for computer experiments. For the rest, we don't have time to study them carefully in a short course like this one.

In the tableaux below:

A α	B β	C γ	D δ
B δ	A γ	D β	C α
C β	D α	A δ	B γ
D γ	C δ	B α	A β

the Latin letters (A, B, C, D) form a Latin square. So do the Greek letters $(\alpha, \beta, \gamma, \delta)$. These two Latin squares are mutually orthogonal meaning that every combination of one Latin letter with one Greek letter appears the same number of times (actually once). From two mutually orthogonal Latin squares ***MOLS*** we get a Graeco-Latin square like the one shown.

We could use a Graeco-Latin square with treatments A, B, C and D blocked out against three factors: one for rows, one for columns and one for Greek letters. We are now in the setting of combinatoric existence and non-existence results. For instance, no Graeco-Latin square exists for $k = 6$. Euler thought there would be none for $k = 10$ but that was proved wrong in the 1950s.

The operational difficulties of arranging a real-world Graeco-Latin square experiment are daunting. It is easy to do in software on the computer. You can even do hyper-Graeco-Latin square experiments with three or more MOLS. For instance if k is a prime number you can have $k - 1$ MOLS and then block out k factors at k levels in addition to a treatment factor at k levels. Or you can embed k^2 points into $[0, 1]^{k+1}$ and have every pairwise scatterplot be a $k \times k$ grid. We will see this later for computer experiments and space-filling designs. Be sure to randomize!

Sometimes the number of levels in a block is less than the number of treatments we have in mind. For instance, consider a club of people are tasting 12 different wines and we don't want anybody to taste more than 6 of them. Then we would like to arrange our tastings so that each person tastes 6 wines. Those people then represent ***incomplete blocks***.

In an ideal world, each pair of wines would be tasted together by the same number of tasters. That would give us ***balanced incomplete blocks***. This makes sense because the best comparisons between wines A and B will come from people who tasted both A and B. That is, from within block comparisons. There will also be between block comparisons. For instance if many people found A better than B and many found B better than C that provides evidence (through a regression model) that A is better than C. But the within block

	10–19	20–29	30–39	40–49	Total
Treatment	7	5	4	1	17
Control	2	5	3	2	12

Table 4.4: Treatment and control sample sizes by age group in a Hodgkin's disease investigation.

evidence from having A and C compared by the same people is more informative if the block effects (people) are large.

In sporting leagues we have k teams and we compare then in games that are (ordinarily) blocks of size $B = 2$. A tournament in which each pair of teams played together the same number of times would be a balanced incomplete block design.

There are also ***partially balanced incomplete block*** designs where the number of blocks where two treatments are together is either λ or $\lambda + 1$. So, while not equal, they are close to equal.

We will not consider how to analyze incomplete block designs. If you use one in your project, the other topics from this course will prepare you to read about them and adopt them.

There are even design strategies where one blocking factor has k levels and another has fewer than k levels. So the design is incomplete in that second factor. If you find yourself facing a situation like this, look for ***Youden squares***.

4.9 Biased coins

Efron (1971) describes a small randomized controlled trial on Hodgkin's disease. The sample sizes were as in Table 4.4. Patients were assigned at random to treatment or control as the study progressed. The assignment was double blind. As show in the table, the allocations to treatment and control were not well balanced within age groups.

We could do better to adjust the sampling proportions within groups as the experiment goes on. However, it is desirable to avoid sampling probabilities that get too far from $1/2$ because then somebody with knowledge of past assignments might be able to predict future assignments in violation of the blinding protocol. The biased coin design gives treatment with some probability $q < 1/2$ if the group a subject belongs to has already had more treated than control. Efron (1971) likes $q = 1/3$. If the subject's group has had fewer treated than control, then the biased coin allocates treatment with probability $1 - q$. If those counts are equal, as must be for the first subject in a group, then the treatment probability is $1/2$. That is, a fair coin is used.

The excess of treated minus controlled subject counts follows a Markov chain and does not get far from balance. For instance, with even n, the probability of an exact split between treatment and control approaches $1/2$ when using $q = 1/3$. The balance properties are more valuable in settings with small sample

counts per group, especially when the number of groups is large.

5

Analysis of variance

This chapter goes deeper into the Analysis of Variance (ANOVA). We consider multiple factors and we also introduce the notions of fixed and random effects. Most of these notes assume familiarity with statistics and data analysis in order to study how to make data more than how to analyze it. This chapter is a bit of an exception. We will also extend the theory to cover what might be gaps in the usual way regression courses are taught. ANOVA is a bit more complicated than just running regressions with the categories coded up as indicator variables, and so we will need to add some extra theory. Where possible, the additional theory will be anchored in things that one would remember from an earlier regression course.

Suppose that we have two categorical variables, say A and B. Each has two levels. That makes four treatment combinations. We could just study it as one categorical variable with four levels. However we benefit from working with the underlying 2×2 structure. We have two choices to make (which A) and (which B) and a third thing about interaction, which we will see includes considering whether the best A depends on B and vice versa. Note that here A and B are two different treatment options. In A/B testing A and B are two different versions of one treatment option. Experimental design is complicated enough that no single notational convention can carry us through. We have to use local notation or else we would not be able to read the literature.

5.1 Potatoes and sugar

The oldest ANOVA reference I know of is Fisher and Mackenzie (1923). They were using it to study the effects of different fertilizer on potato yields. It is (historically) interesting that even in this first paper they are thinking of non-

additivity and even consider something that looks like one term of a singular value decomposition.

To illustrate how a two factor experiment works, consider the following hypothetical yields for 3 fertilizers and 4 varieties of potatoes:

$$\begin{array}{c} \text{Yield (kg)} \\ F_1 \\ F_2 \\ F_3 \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & V_3 & V_4 \end{array} \\ \begin{bmatrix} 109.0 & 110.9 & 94.2 & 125.9 \\ 104.9 & 113.4 & 110.1 & 138.0 \\ 151.8 & 160.9 & 111.9 & 145.0 \end{bmatrix}. \end{array}$$

Based on these values, we can wonder which fertilizer is best, which variety is best, and the extent to which one decision depends on the other.

Taking the yield data to be a 3×4 matrix of Y_{ij} values, the overall average yield is $\bar{Y}_{\bullet\bullet} = 123$. If we think fertilizer i raised or lowered the yield it must be about yields higher or lower than 123. So we can subtract 123 from all the Y_{ij} and then take $(1/J)\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{\bullet\bullet}) = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$ as the incremental effect of fertilizer i. We get:

F_1	F_2	F_3
-13.0	-6.4	$19.4.$

By this measure, fertilizer 1 lowers the yield by 13 while fertilizer 3 raises it by 19.4. The same idea applied to varieties yields $\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}$:

V_1	V_2	V_3	V_4
-1.1	5.4	-17.6	$13.3.$

Variety 3 underperforms quite a bit while variety 4 comes out best.

We have just computed the ***grand mean*** $\bar{Y}_{\bullet\bullet} = 123$ and the ***main effects*** for fertilizer and variety. Note that each of the main effects average to zero, because of the way that they are constructed. We can decompose the table into grand mean, main effects and a residual term, as follows:

$$\begin{bmatrix} 109.0 & 110.9 & 94.2 & 125.9 \\ 104.9 & 113.4 & 110.1 & 138.0 \\ 151.8 & 160.9 & 111.9 & 145.0 \end{bmatrix} = \begin{bmatrix} 123 & 123 & 123 & 123 \\ 123 & 123 & 123 & 123 \\ 123 & 123 & 123 & 123 \end{bmatrix}$$

$$+ \begin{bmatrix} -13.0 & -13.0 & -13.0 & -13.0 \\ -6.4 & -6.4 & -6.4 & -6.4 \\ 19.4 & 19.4 & 19.4 & 19.4 \end{bmatrix} + \begin{bmatrix} -1.1 & 5.4 & -17.6 & 13.3 \\ -1.1 & 5.4 & -17.6 & 13.3 \\ -1.1 & 5.4 & -17.6 & 13.3 \end{bmatrix}$$

$$+ \begin{bmatrix} 0.1 & -4.5 & 1.8 & 2.6 \\ -10.6 & -8.6 & 11.1 & 8.1 \\ 10.5 & 13.1 & -12.9 & -10.7 \end{bmatrix}.$$

The last term captures the extent to which the yields are not additive. It is called the ***interaction***. The grand mean, main effects and interactions we

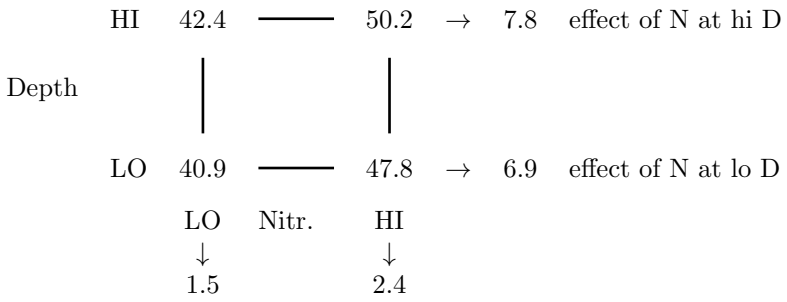


Figure 5.1: Sugar yields in a 2×2 experiment where N denotes use of nitrogen and D denotes increased depth of ploughing.

want are defined in terms of $\mathbb{E}(Y_{ij})$. The ones we get are noisy versions of those defined through Y_{ij}. Much of ANOVA is about coping with noise that makes a sample table of Y_{ij} differ from a hypothetical population table of $\mathbb{E}(Y_{ij})$. Perhaps more is about coping with interactions that complicate estimation an interpretation of data.

Here is another example motivated by agriculture. My notes say that I got it from Cochran and Cox, which has gone through many editions, but I cannot now find it in that book. It is about the yield of sugar in 100s of pounds per acre. There's an old unit called the 'hundredweight' which is about 45.4 kilograms. They considered two treatments. Treatment N involved either no nitrogen, or application of 300 lbs of nitrogen per acre. Treatment D involved either ploughing to the usual depth of 7" or going further to 11". The results, which might or might not be hypothetical are depicted in Figure 5.1.

When both variables are at their 'low' level the yield is 40.9. We see that going to the high level of N raises the yield by either 7.8 if D is at the high level (meaning greater depth) or 6.9 if D is at the low level. The overall estimate of the treatment effect is then their average, roughly 7.4. Similarly, we see two different effects for D, one at each of the high and low levels of N, that average to about 2.

There seems to be a positive interaction of about 0.9 meaning that applying both treatments gives more than we would expect from them individually. This is a kind of synnergy.

5.2 One at a time experiments

For the sugar problem, we tried all four combinations varying both factors. We could instead have done two separate experiments, one for each factor. That is called a one at a time ***OAAT*** experiment. Sometimes people even advocate for changing just one thing at a time to learn its effects. Here we show that if you have two things to investigate it is better to investigate them in a combined experiment.

Experiment A Take n observations at $(N, D) = (0, 0)$ and n more at $(N, D) = (0, 1)$ to test test D. Then take another n at $(N, D) = (0, 0)$ and n more at $(N, D) = (1, 0)$ to test n. Experiment A costs $4n$ and delivers $\hat{N} = \bar{Y}_{10} - \bar{Y}_{00}$ with

$$\mathrm{var}(\hat{N}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{2\sigma^2}{n}.$$

Similarly $\mathrm{var}(\hat{D}) = 2\sigma^2/n$.

Experiment B Take $n/2$ observations at each of $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$. Experiment B costs only $2n$. It has $\hat{N} = [(\bar{Y}_{10} - \bar{Y}_{00}) + (\bar{Y}_{11} - \bar{Y}_{01})]/2$ with

$$\mathrm{var}(\hat{N}) = \frac{1}{4}\Big(\frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2}\Big) = \frac{2\sigma^2}{n} = \mathrm{var}(\hat{D}).$$

The factorial experiment B delivers the same accuracy as the OAAT one at half the cost. By that measure, it is twice as good. It is actually better than twice as good. The factorial experiment can be used to investigate whether the factors interact while OAAT cannot.

Experiment A is not the best OAAT that we could do. It misses an opportunity to reuse the data at $(0, 0)$. We could also consider experiment C below.

Experiment C Take n observations at $(0, 0)$, and n more at $(0, 1)$ and n more at $(1, 0)$.

Experiment C costs 1.5 times as much as experiment B and has the same variance. It also makes $\mathrm{corr}(\hat{N}, \hat{D}) \neq 0$. So, while OAAT via experiment A is only half as good as a factorial experiment OAAT via experiment C is 2/3 as good. Since the observation at $(0, 0)$ is used twice, we might want to give it extra samples. Of course a better idea is not to do OAAT.

OAAT could leave us with the following information

$$\begin{array}{ccc} \text{GOOD} & & \textcircled{?} \\ \uparrow & & \\ \text{FAIR} & \longrightarrow & \text{GOOD} \end{array}$$

where two changes from our starting point are both beneficial but we don't know what happens if we make both changes. We would probably try that. We might get a result like

$$\begin{array}{ccc} \text{GOOD} & \longrightarrow & \text{BAD} \\ \uparrow & & \uparrow \\ \text{FAIR} & \longrightarrow & \text{GOOD} \end{array}$$

hopefully by trying it out first before commiting to it. In a case like this we learn that making both changes is a bad idea, but we would have learned sooner with a factorial experiment.

Another possibility is that an OAAT experiment has this result

$$\begin{array}{ccc} \text{BAD} & & (?) \\ \uparrow & & \\ \text{FAIR} & \longrightarrow & \text{BAD} \end{array}$$

where both changes are adverse. We don't know what would happen if both changes were made. Based on the sketch above, many people would not even try making both changes. It is possible that the underlying truth is like this

$$\begin{array}{ccc} \text{BAD} & \longrightarrow & \text{BEST EVER} \\ \uparrow & & \uparrow \\ \text{FAIR} & \longrightarrow & \text{BAD} \end{array}$$

where making both changes would be extremely valuable. In a factorial experiment we would learn this, while n OAAT it could very well go undiscovered.

5.3 Interactions

One severe problem with OAAT is that if there are important interactions, then we don't learn about them and might be forever stuck in a suboptimal setting.

Interactions cause severe difficulties. We can think of a failure of external validity as being an interaction between treatment choices (e.g., aspirin vs tylenol) and another variable describing the past versus the future. Or that second variable could be data in our study versus data we want to generalize to. It is bad enough that 'your mileage may vary' and worse still that mileage differences may vary. That could mean that the optimal choice changes between the data we learned from and the setting where we will make future decisions.

Interactions underly lots of accidents and disasters. An accident might only have happened because of a wet road, inattentive driver, bad street lights and poor brakes. If all of those things were needed to create the accident then it is a sort of interaction. It's good that the accident is not in the grand mean or main effects, because then we could get more of them, but having it be an interaction makes it harder to prevent.

Andrew Gelman has written that interactions take 16 times as much data to estimate as main effects do: `https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/`

It is informative to see how that 16 arises. In a 2×2 experiment the main effect estimate would have

$$\mathrm{var}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \frac{\sigma^2}{n/2} + \frac{\sigma^2}{n/2} = \frac{4\sigma^2}{n}$$

while the interaction would have variance

$$\mathrm{var}(\bar{Y}_{11} - \bar{Y}_{01} - \bar{Y}_{10} + \bar{Y}_{11})\frac{\sigma^2}{n/4} \times 4 = \frac{16\sigma^2}{n}.$$

If an interaction would have the same size as a main effect, then it would be 4 times as hard to estimate, meaning that we would need to raise n by a factor of 4 to do as well.

Now suppose that the main effect is θ_{main} and the interaction is $\theta_{\text{inter}} = \lambda\theta_{\text{main}}$. We use sample size n for the main effect and get a relative error of

$$\frac{|\theta_{\text{main}}|}{\sqrt{4\sigma^2/n}} = \frac{\sqrt{n}|\theta_{\text{main}}|}{2\sigma}.$$

If we use n' observations for the interaction our relative error would be

$$\frac{|\theta_{\text{inter}}|}{\sqrt{16\sigma^2/n'}} = \frac{\sqrt{n'}|\lambda\theta_{\text{main}}|}{4\sigma}.$$

We make these equal by solving $|\lambda|\sqrt{n'} = 2\sqrt{n}$ giving $n' = 4n/\lambda^2$. If we take $\lambda = \pm 1/2$, then we find that interactions are 16 times as hard to estimate as main effects. Relative error is important because statistical significance and power depend on the ratio of the (absolute) effect size to the standard deviation of the effect estimate. The factor of $1/2$ is a ballpark estimate. The actual size ratio between a typical interaction and typical main effect will vary with the setting.

Interactions can raise severe problems of multiple comparisons. When there are d experimental factors then there are $\binom{d}{2}$ different pairwise comparisons. If we test everything at level α then we espect up to $d\alpha$ false discoveries for main effects and up to $d(d-1)\alpha/2$ among interactions. If we think that interactions are more likely to be null (or close enough to it to be practically null) then the problem is even more severe for interactions than main effects. Or, if we use methods to control false discovery rates, then we have to be more conservative with interactions.

The problem gets worse for higher order interactions. For instance if an $A \times B$ interaction depends on the level of factor C, then we have a three factor interaction.

If in baseball a pitcher has a very good record against tall left handed batters in evening games with a light wind, then that finding is a kind of high order interaction that was probably based on slim evidence. Patterns that appear as high order interactions can sound very subtle and important but they could also be noise.

Interactions are not all bad. It is good that not everybody likes the exact same restaurants. That is an interaction between people and restaurants. More generally, personalization of services involves measuring interactions.

I believe that much of the work by George Box and his co-authors on experimental design was a response to the challenge of learning scientific facts despite the presence of many interactions. Ordinary statistical modeling is well able to handle noise. Randomization of treatments is a good way to counter missing variables and get causal estimates. Factorial designs and, later on fractional factorials, help us cope with the complexity that comes from interactions.

5.4 Multiway ANOVA

Now suppose we have 4 factors, A, B, C and D. We can write the regression model as

$$\begin{aligned}Y_{ijk\ell} &= \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{i\ell} + (\beta\gamma)_{jk} + (\beta\delta)_{j\ell} + (\gamma\delta)_{k\ell} \\ &\quad + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ik\ell} + (\alpha\gamma\delta)_{ik\ell} + (\beta\gamma\delta)_{jk\ell} \\ &\quad + \varepsilon_{ijk\ell}.\end{aligned}$$

Factor A has levels $1 \leqslant i \leqslant I$, factor B has levels $1 \leqslant j \leqslant J$, factor C has levels $1 \leqslant k \leqslant K$ and factor D has levels $1 \leqslant \ell \leqslant L$. The model above has numerous interactions. We could run out of Greek letters and so we use notation like $(\alpha\beta)$ as its own symbol to describe the parameters in an $A \times B$ interaction. Here we have an error/noise term $\varepsilon_{ijk\ell}$.

To make the model identifiable we make each main effect sum to zero and each interaction sum to zero over all values of any index in it. For instance $\sum_{j=1}^{J}(\alpha\beta\gamma)_{ijk} = 0$ for all i and k.

It is easy to work out what the estimates are when $\varepsilon_{ijk\ell} \sim \mathcal{N}(0, \sigma^2)$. We get

$$\begin{aligned}\hat\mu &= \bar Y_{\bullet\bullet\bullet\bullet} \\ \hat\alpha_i &= \frac{1}{JKL}\sum_j\sum_k\sum_\ell (Y_{ijk\ell} - \hat\mu) = \bar Y_{i\bullet\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet} \\ \widehat{(\alpha\beta)}_{ij} &= \frac{1}{KL}\sum_k\sum_\ell (Y_{ijk\ell} - \hat\mu - \hat\alpha_i - \hat\beta_j) \\ &= \bar Y_{ij\bullet\bullet} - \hat\mu - \hat\alpha_i - \hat\beta_j \\ &= \bar Y_{ij\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet} - (\bar Y_{i\bullet\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet}) - (\bar Y_{\bullet j\bullet\bullet} - \bar Y_{\bullet\bullet\bullet\bullet}) \\ &= \bar Y_{ij\bullet\bullet} - \bar Y_{i\bullet\bullet\bullet} - \bar Y_{\bullet j\bullet\bullet} + \bar Y_{\bullet\bullet\bullet\bullet}, \quad \text{and} \\ \widehat{(\alpha\beta\gamma)}_{ijk} &= \frac{1}{L}\sum_\ell \big(Y_{ijk\ell} - \hat\mu - \hat\alpha_i - \hat\beta_j - \hat\gamma_k - \widehat{(\alpha\beta)}_{ij} - \widehat{(\alpha\gamma)}_{ik} - \widehat{(\beta\gamma)}_{jk}\big) \\ &= \bar Y_{ijk\bullet} - \bar Y_{ij\bullet\bullet} - \bar Y_{i\bullet k\bullet} - \bar Y_{\bullet jk\bullet} + \bar Y_{i\bullet\bullet\bullet} + \bar Y_{\bullet j\bullet\bullet} + \bar Y_{\bullet\bullet k\bullet} - \bar Y_{\bullet\bullet\bullet\bullet}\end{aligned}$$

and the others are similar. In each case we subtract sub-effect estimates and average over variables not in the interaction of interest. The results are differences of certain data averages.

We also have an ANOVA identity. It is a bit ungainly:

$$\begin{aligned}\sum_{ijk\ell}(Y_{ijk\ell}-\hat\mu)^2 &= \sum_{ijk\ell}\hat\alpha_i^2+\sum_{ijk\ell}\hat\beta_j^2+\sum_{ijk\ell}\hat\gamma_k^2+\sum_{ijk\ell}\hat\delta_\ell^2\\ &+\sum_{ijk\ell}\widehat{\alpha\beta}_{ij}^2+\sum_{ijk\ell}\widehat{\alpha\gamma}_{ik}^2+\sum_{ijk\ell}\widehat{\alpha\delta}_{i\ell}^2+\sum_{ijk\ell}\widehat{\beta\gamma}_{jk}^2+\sum_{ijk\ell}\widehat{\beta\delta}_{j\ell}^2+\sum_{ijk\ell}\widehat{\gamma\delta}_{k\ell}^2\\ &+\sum_{ijk\ell}\widehat{\alpha\beta\gamma}_{ijk}^2+\sum_{ijk\ell}\widehat{\alpha\beta\delta}_{ij\ell}^2+\sum_{ijk\ell}\widehat{\alpha\gamma\delta}_{ik\ell}^2+\sum_{ijk\ell}\widehat{\beta\gamma\delta}_{jk\ell}^2\\ &+\sum_{ijk\ell}\widehat{\alpha\beta\gamma\delta}_{ijk\ell}^2+\sum_{ijk\ell}\hat\varepsilon_{ijk\ell}^2.\end{aligned}$$

For d factors there are 2^d-1 non-empty subsets of them that all explain some amount of variance that sums to the total variance among all N values. We ordinarily ignore $\sum_{ijk\ell}\hat\mu^2$ which when added to the above gives us $\sum_{ijk\ell}Y_{ijk\ell}^2$. The reason that μ and $\hat\mu$ are of less interest is that the grand mean has no impact on our choices of i or j or k or ℓ.

Later we will look at the ***functional ANOVA***. This was used by Hoeffding (1948) to study U-statistics, by Sobol' (1969) to study numerical integration and by Efron and Stein (1981) to study the jackknife. We will use a more abstract notation for it that simplifies some expressions. Suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ where x_j are independent random inputs. Now let $Y = f(\boldsymbol{x})$. If $\mathbb{E}(Y^2) < \infty$ then $\mathrm{var}(f(\boldsymbol{x}))$ exists and there is a way to do an ANOVA on it. We proceed by analogy. The grand mean is $\mu = \mathbb{E}(f(\boldsymbol{x}))$. Then the main effect for variable j is

$$f_{\{j\}}(\boldsymbol{x}) = \mathbb{E}(f(\boldsymbol{x})-\mu\,|\,x_j) = \mathbb{E}(f(\boldsymbol{x})\,|\,x_j)-\mu$$

and the j,k interaction is

$$\begin{aligned}f_{\{j,k\}}(\boldsymbol{x}) &= \mathbb{E}(f(\boldsymbol{x})-\mu-f_{\{j\}}(\boldsymbol{x})-f_{\{k\}}(\boldsymbol{x})\,|\,x_j,x_k)\\ &= \mathbb{E}(f(\boldsymbol{x})\,|\,x_j,x_j)-\mu-f_{\{j\}}(\boldsymbol{x})-f_{\{k\}}(\boldsymbol{x}).\end{aligned}$$

We keep subtracting sub-effects and averaging over variables not in the interaction of interest. The x_j do not have to be categorical random variables like the levels of the factors we use above. They could be $\mathbb{U}[0,1]$ random variables or vectors, sound clips, images or genomes. All that matters is that they are independent and that $f(\boldsymbol{x})$ is real-valued with finite variance.

We will look at `https://statweb.stanford.edu/~owen/mc/A-anova.pdf` later when we get to computer experiments.

5.5 Replicates

Suppose that we were interested in a four-factor interaction. We could take $R \geqslant 2$ independent measurements at each ABCD combination. Then our model

would be

$$\begin{aligned} Y_{ijk\ell r} &= \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{i\ell} + (\beta\gamma)_{jk} + (\beta\delta)_{j\ell} + (\gamma\delta)_{k\ell} \\ &\quad + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ik\ell} + (\alpha\gamma\delta)_{ik\ell} + (\beta\gamma\delta)_{jk\ell} \\ &\quad + (\alpha\beta\gamma\delta)_{ijk\ell} + \varepsilon_{ijk\ell r}. \end{aligned} \tag{5.1}$$

where $1 \leqslant r \leqslant R$.

Suppose that we are making soup. We could go to store i, buy vegetables j, use recipe k, find person ℓ and have them taste the resulting pot of soup R times for $r = 1, \dots, R$. Or, on R separate days we could go to all I stores, buy all J vegetables at each store, try all K recipes on each set of vegetables from each store, and ask all L people to try all IJK of those soups once.

These are clearly quite different things. Not all kinds of replicate are equal. Equation (5.1) is a plausible model for the setting where the tasters taste each pot of soup R times in a row. It is an entirely unsuitable model for the setting where the whole experiment is completely redone R times. For that we would at a minimum want to include a main effect η_r for replicate r. There could even be a case for making the day of the experiment be its own fifth factor E with R levels. What is going on here is that the meaning of $r = 1$ is different in the two cases. In the first case it is just the first time one person tastes a given soup. In the second case it is one entire full replication of the experiment.

Just looking at the file of $N = IJKLR$ numbers we might not be able to tell which way the experiment was done. We will think more about replicates when they come up in specific examples.

5.6 High order ANOVA tables

If factors A and B have I and J levels, respectively then their interaction has $(I-1)(J-1)$ degrees of freedom. To see why consider this table

$(\alpha\beta)_{ij}$	j=1	j=2	j=3	j=4
i=1	✓	✓	✓	?
i=2	✓	✓	✓	?
i=3	?	?	?	??

where we suppose that there are known values in all the cells marked ✓. There are $(I-1)(J-1)$ of them. For any set of choices we make, the table can be completed by first making row sums zero and them making column sums zero. Had we omitted one of those $(I-1)(J-1)$ values, there would not be a unique way to complete the table.

Table 5.1 shows a portion of the ANOVA table for a four way factorial experiment where each cell has R independent values. We see $IJKL(R-1)$ degrees of freedom for error. We could get this by subtracting all of the other degrees of freedom numbers from $N - 1 = IJKLR - 1$. Or we can view it as gathering $R - 1$ degrees of freedom for error from each of $I \times J \times K \times L$ cells.

Source	DF
A or α	$I-1$
B or β	$J-1$
$\vdots$	$\vdots$
AB or $\alpha\beta$	$(I-1)(J-1)$
$\vdots$	$\vdots$
ABC or $\alpha\beta\gamma$	$(I-1)(J-1)(K-1)$
$\vdots$	$\vdots$
ABCD or $\alpha\beta\gamma\delta$	$(I-1)(J-1)(K-1)(L-1)$
Error	$IJKL(R-1)$
Total	$IJKLR-1$

Table 5.1: Selected rows of the ANOVA table for a four way table where each cell has R independent repeats.

It is clear from the above that these full factorial experiments are big and bulky and hence probably expensive. If $I = J = K = L = 11$ then we need $N = 11^4R = 14{,}641R$ observations. They will give us 10 df in each main effect, 100 per two factor interaction, 1000 df for each three factor interaction and 10,000 df in the four factor interaction which could be the least useful of them all.

Next we will look at what happens when all factors are at 2 levels. We still need $N = 2^d$ data points for d factors or $R2^d$ if we have replicated the experiment.

A common practice is to design an experiment that learns about the main effects and low order interactions partially or completely ignoring the high order interactions. To do this seems a bit hypocritical. We earlier argued that OAAT is bad because it can miss the two factor interactions and now we are getting ready to possibly ignore some other higher order interactions.

This common practice is a gamble. In statistics, we are usually against gambling and favor a cautious approach that covers all possibilities. However the cautious approach is really expensive and could be suboptimal. Experimental design has room for both bold and cautious choices. With a bold choice we do a small experiment that could be inexpensive or fast. If it goes well, then we learn more quickly. If it goes badly, then the experiment might not be informative at all and we have to do another one.

The bold strategies we will look at are motivated by a principle of ***factor sparsity***. This holds that the important quantities are mostly lower order and perhaps even many of the main effects are unimportant. Or at least relatively unimportant. If there are $2^{10} - 1$ effects and interactions they cannot all be relatively important! There is a related ***bet on sparsity*** principle (Friedman et al., 2001) motivating the use of the lasso. If things are sparse, you win. If

they're not, maybe nothing would have worked.

5.7 Distributions of sums of squares

The ANOVA table involves a lot of algebraic manipulations. We get a table based on the values of $Y_{ijk\ell}$. We would rather have the table based on $\mathbb{E}(Y_{ijk\ell})$. So here we must study the distribution of the quantities in the table we get, using some model assumptions. The most common model assumptions are based on the Gaussian distribution. Sometimes that's robust due to the central limit theorem and sometimes it is not. Balanced experiments are more robust. The F-tests that we do for main effects and interactions are more robust than those we might do to compare variances of two sources of noise. Time permitting, we might cover this. For now, let's see how the mechanics of these tests work out.

These derivations are to the extent possible based on things that are easy to remember from an introductory regression class. To that we add some things about noncentral distributions that are not always included.

First, we recall that if $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}.$$

Also, if $Q_j \overset{\text{ind}}{\sim} \chi^2_{(n_j)}$ then

$$F = \frac{Q_1/n_1}{Q_2/n_2} \sim F_{n_1,n_2}$$

and this is the very definition of the F distribution. The F distribution has numerator and denominator degrees of freedom and we write $F_{\text{num,den}}$ for the general case.

Now we consider noncentral distributions. These appear less often in introductory courses. The main place we need them is in power calculations, such as choosing a sample size. Choosing a sample size is maybe the simplest design problem. It does not however come up if we are just looking at pre-existing data sets.

If $Z_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1)$ then

$$\sum_{i=1}^{n} Z_i^2 \sim \chi'^2_{(n)}(\lambda)$$

where $\lambda = \sum_{i=1}^{n} \mu_i^2$. This is the ***noncentral*** χ^2 distribution on n degrees of freedom with noncentrality parameter λ. There are alternative parameterizations out there so we always have to check books, articles and software documentation to see which one was used. Note particularly that the n means μ_i only affect the distribution through their sum of squares $\sum_i \mu_i^2$. That extremely convenient fact depends on the Gaussian distribution chosen for Z_i.

If $Q_1 \sim \chi'^2_{(n_1)}(\lambda)$ and $Q_2 \sim \chi^2_{(n_2)}$ then

$$F' = \frac{Q_1/n_1}{Q_2/n_2} \sim F'_{n_1,n_2}(\lambda).$$

This is the ***noncentral F*** distribution. Here is how we use it. Our F statistics will have central numerators under H_0 but noncentral ones under H_A. They will usually have central denominators under both because most of our denominators will involve sums of squares of differences of noise. If our noise model is wrong then the denominator might be noncentral. That would leave us with a ***doubly noncentral F*** distribution. That second noncentrality would make the denominator bigger and hence the F statistic smaller and that implies lower statistical power to reject H_0.

Next we look at the distributions of sums of squares for simple one way layout with $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ for $i = 1, \dots, I$ and $j = 1, \dots, n$. Recall from introductory regression that when $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ then

$$(n-1)s^2 \equiv \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \sigma^2 \chi^2_{(n-1)}.$$

Also if $Q_i \overset{\text{ind}}{\sim} \chi^2_{(n_i)}$ then

$$\sum_i Q_i \sim \chi^2_{(N)} \quad N = \sum_i n_i.$$

Now

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{I}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n} \big((\mu + \alpha_i + \varepsilon_{ij}) - (\mu + \alpha_i + \bar{\varepsilon}_{i\bullet})\big)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n} (\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet})^2 = \sum_{i=1}^{I} Q_i \qquad \text{for } Q_i \overset{\text{iid}}{\sim} \sigma^2\chi^2_{(n-1)} \\
&\sim \sigma^2 \chi^2(I(n-1)).
\end{aligned}$$

Exercise: what is $\mathcal{L}(\text{SSE})$ in the unbalanced case with n_i observations at level i of the factor A? Notice that the α_i did not affect SSE. They canceled out.

Next, under $H_0 : \alpha_i = 0$ we have

$$\begin{aligned}
\text{SSA} &= \sum_{i=1}^{I}\sum_{j=1}^{n} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = n\sum_{i=1}^{I} (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \\
&= n\sum_{i=1}^{I} ((\mu + \alpha_i + \bar{\varepsilon}_{i\bullet}) - (\mu + \bar{\alpha}_\bullet + \bar{\varepsilon}_{\bullet\bullet}))^2 \\
&= n\sum_{i=1}^{I} (\bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet})^2 \sim n\frac{\sigma^2}{n}\chi^2_{(I-1)} = \sigma^2\chi^2_{(I-1)}.
\end{aligned}$$

If H_0 does not hold then $\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet} = \alpha_i + \bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet}$. Now

$$\bar{\varepsilon}_{i\bullet} \sim \mathcal{N}\Big(\alpha_i, \frac{\sigma^2}{n}\Big) \quad \text{and so} \quad \frac{\bar{\varepsilon}_{i\bullet}}{\sigma/\sqrt{n}} \sim \mathcal{N}\Big(\frac{\alpha_i}{\sigma/\sqrt{n}}, 1\Big).$$

Therefore

$$\sum_{i=1}^{I}\Big(\frac{\bar{\varepsilon}_{i\bullet}}{\sigma/\sqrt{n}}\Big)^2 \sim \chi'^2(\lambda) \quad \text{for } \lambda = \frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2.$$

From this we find that

$$\sum_{i=1}^{I}\bar{\varepsilon}_{i\bullet}^2 \sim \frac{\sigma^2}{n}\chi'^2_{(I)}\Big(\frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2\Big) \quad \text{and so} \quad \sum_{i=1}^{I}\sum_{j=1}^{n}\bar{\varepsilon}_{i\bullet}^2 \sim \sigma^2\chi'^2_{(I)}\Big(\frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2\Big).$$

The answer we are going for is

$$\mathrm{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{n}(\bar{\varepsilon}_{i\bullet} - \bar{\varepsilon}_{\bullet\bullet})^2 \sim \sigma^2\chi'^2_{(I-1)}\Big(\frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2\Big).$$

Notice that the degrees of freedom drop by one for centered $\bar{\varepsilon}_{i\bullet}$. I have not found a nice way to get this using just things one might remember from an introductory regression class plus the noncentral distribution definitions.

Now let's look at our F statistic, under H_0:

$$\begin{aligned} F = \frac{\mathrm{MSA}}{\mathrm{MSE}} &= \frac{\frac{1}{I-1}\mathrm{SSA}}{\frac{1}{I(n-1)}\mathrm{SSE}} \\ &\sim \frac{\frac{1}{I-1}\sigma^2\chi^2_{(I-1)}}{\frac{1}{I(n-1)}\sigma^2\chi^2_{(I(n-1))}} \\ &= \frac{\frac{1}{I-1}\chi^2_{(I-1)}}{\frac{1}{I(n-1)}\chi^2_{(I(n-1))}} \\ &= F_{I-1,I(n-1)}. \end{aligned}$$

Under H_A,

$$F \sim F'_{I-1,I(n-1)}(\lambda) \qquad \lambda = \frac{n}{\sigma^2}\sum_{i=1}^{I}\alpha_i^2.$$

Under H_A, $\lambda > 0$ and the larger λ is the larger $\mathbb{E}(\mathrm{MSA})$ is. Thus H_A tends to increase F and so we should reject H_0 for unusually large values. We reject H_0 at level α when the observed value F_{obs} satisfies

$$F_{\mathrm{obs}} \geqslant F^{1-\alpha}_{I-1,I(n-1)}$$

We set a p-value of

$$p = \Pr(F_{I-1,I(n-1)} \geqslant F_{\mathrm{obs}})$$

where F_{obs} is the observed F statistic. The power of our test is

$$\Pr(F'_{I-1,I(n-1)}(\lambda) \geqslant F^{1-\alpha}_{I-1,I(n-1)}).$$

If we look at λ, then we see that it is a signal to noise ratio

$$\lambda = \frac{\sum_{i=1}^{I} \alpha_i^2}{\sigma^2/n} = I \times \frac{\frac{1}{I}\sum_{i=1}^{I} \alpha_i^2}{\sigma^2/n}.$$

Here $(1/I)\sum_{i=1}^{I} \alpha_i^2$ is the variance of α_i for a randomly chosen level i and σ^2/n is the variance of $\bar{\varepsilon}_{i\bullet}$. We get more power from larger n (get a bigger sample) or smaller σ^2 (e.g., buy better equipment) and from studying phenomena with larger effects.

5.8 Fixed and random effects

In an ANOVA we need to decide whether each of our factors represents a ***fixed effect*** or a ***random effect***. It is a fixed effect if we care about the exact set of levels in the experiment. That could be 3 headache pills, 4 gas additives or 5 machine operators. It is a random effect if we care about the population from which those levels were sampled. That could be 24 patients in a trial, 8 rolls of vinyl or those same 5 machine operators.

Whether we care about the levels in our experiment or a population that they represent depends on the uses we plan to make based on the data. While the health of 24 patients is very important, the reason for the trial is likely to be in order to learn how to treat some condition for people in general. Those 24 patients may then be viewed as a sample of a population of many millions to be treated later. It is unlikely that they really are a random sample of the population to be treated, but hopefully they are reasonably representative. Once the 8 rolls of vinyl are used up and put into tested products we have no specific interest in them per se beyond how they represent the process that made them. For machine operators we might be interested in comparing 5 specfic employee's productivity. Or we might be interested in which of two machines will work better on average for a population of operators.

In a random effects model we write

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

where $a_i \sim \mathcal{N}(0, \sigma_A^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_E^2)$ are all independent. We might be interested in learning σ_A^2/σ_E^2 and the usual null hypothesis is $H_0 : \sigma_A^2 = 0$. The ANOVA is exactly the same as before. That is SST = SSB + SSW or SST = SSA + SSE are the two different notations we have used for the one factor ANOVA. This identity is just algebraic so it holds for any numbers we would put into it. Using methods like those in Section 5.7 we find that

$$\mathrm{SSW} \sim \sigma_E^2 \chi^2_{I(n-1)} \quad \text{and} \quad \mathrm{SSB} \sim n(\sigma_A^2 + \sigma_E^2/n)\chi^2_{(I-1)}.$$

Source	DF	Expected mean square
A	$I-1$	$\sigma_E^2 + n\sigma_{AB}^2 + nJ\sigma_A^2$
B	$J-1$	$\sigma_E^2 + nI\sigma_B^2$
AB	$(I-1)(J-1)$	$\sigma_E^2 + n\sigma_{AB}^2$
Err	$IJ(n-1)$	σ_E^2

Table 5.2: Expected values of the mean squares in a mixed effects model.

Now MSB/MSW $\sim F_{I-1,I(n-1)}$ under H_0 and is larger than that under H_A, though it is not noncentrally distributed any more. So we do the same F test as before. Not much changed.

When there are two random effects we use this model

$$\begin{aligned} Y_{ijk} &= \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk} \\ a_i \sim \mathcal{N}(0,\sigma_A^2) &\qquad b_j \sim \mathcal{N}(0,\sigma_B^2) \\ (ab)_{ij} \sim \mathcal{N}(0,\sigma_{AB}^2) &\qquad \varepsilon_{ijk} \sim \mathcal{N}(0,\sigma_E^2) \end{aligned}$$

where all the Gaussian random variables are indpendent.

In the balanced case there are the same number n of observations at each ij combination. For $n \geqslant 2$ we find that

$$\begin{aligned} \mathrm{MSA} &\sim \big(\sigma_E^2 + n\sigma_{AB}^2 + nJ\sigma_A^2\big)\frac{\chi^2_{I-1}}{I-1} \\ \mathrm{MSB} &\sim \big(\sigma_E^2 + n\sigma_{AB}^2 + nI\sigma_B^2\big)\frac{\chi^2_{J-1}}{J-1} \\ \mathrm{MSAB} &\sim \big(\sigma_E^2 + n\sigma_{AB}^2\big)\frac{\chi^2_{(I-1)(J-1)}}{(I-1)(J-1)}, \quad \text{and} \\ \mathrm{MSE} &\sim \sigma_E^2\frac{\chi^2_{IJ(n-1)}}{IJ(n-1)}. \end{aligned}$$

If we use $F_A = \mathrm{MSA}/\mathrm{MSE}$, then we have a problem. It won't have the F distribution if $\sigma_A^2 = 0$ but $\sigma_{AB}^2 > 0$. What we must do instead is test it via $F_A = \mathrm{MSA}/\mathrm{MSAB}$. Similarly we test $\sigma_B^2 = 0$ using $F_B = \mathrm{MSB}/\mathrm{MSAB}$ and we test $\sigma_{AB}^2 = 0$ using $F_{AB} = \mathrm{MSAB}/\mathrm{MSE}$.

Now suppose that A is a fixed effect while B is a random effect. We are in for a bit of a surprise. This is called a ***mixed effects model***. Table 5.2 show the expected values of the mean squares for this case. What we see is that we should test the fixed effect A via the ratio MSA/MSAB and the random effect B via the ratio MSB/MSE. If the ***other effect*** is fixed use MSE in the denominator while if the ***other effect*** is random use MSAB. It is the 'other effect rule' for external validity.

To understand this result intuitively let's consider an extreme example. Suppose that the fixed effect A is about 3 headache medicines. We have 12 subjects in a random effect B. Each subject tests each medicine n times.

Let's exaggerate and suppose that $n = 10^6$. It naively looks like we have 36 million observations. But we only have 12 people in the data set. If we did let $n \to \infty$ then we would have a 3×12 table of exact means. Our sample size would actually be 36. With a small sample n our data have to be less useful than those 36 values would have been. For purposes of learning the effect of 3 medicines on a population we have something less informative than 36 observations and nothing like 36,000,000 observations worth of information.

Exercise: work out the limit as $n \to \infty$ of the test for A.

6

Two level factorials

Suppose that we have k factors to explore. If we go with 2 levels each that yields 2^k different treatment combinations which is the smallest possible product of k integers larger than 1. A variable could originally be binary, such as a choice between two polymers, or the choice between doing or not doing a step in a process. Or the variable could have more level or even take on a continuum of levels such as an oven temperature. For such a variable we could select two levels such as $1020°C$ and $1050°C$.

When we choose two levels for a continuous variable, some judgment must be exercised. If the effect of that variable on the response is monotone or even linear then we will have the greatest statistical power using widely separated values. The flip side is that we should avoid closely spaced values because then there will be low power and we might not get good relative accuracy on the effect. Then if we suspect a variable is not important we might use more widely separated values just to be able to check on that. If we know a good operating range then spanning that range makes sense for 'external validity' reasons. If we space things too far apart then our experimental analysis might give that variable too much importance compared to the others. The possibilities that we want to study might not be perfectly orthogonal. For instance if the experiment has an oven at two temperatures for two different time periods, then either the (LO,LO) combination or the (HI,HI) combination might be unsuitable. In that case one can use ***sliding levels*** such as 30 vs 50 minutes when the temperature is high and 45 vs 75 minutes when it is low. Wu and Hamada (2011) discuss these tradeoffs.

If we suspect curvature then two values will not be enough. We will look at 'response surface' designs later that let us estimate curved responses.

If we see each treatment combination$n$ times then we need to gather $n2^k$

Source	df
TRTs	$2^k - 1$
DAYS	$n - 1$
ERR	$(n-1)(2^k-1)$
TOT	$n2^k - 1$

Source	df
TRTs	$2^k - 1$
ERR	$2^k(n-1)$
TOT	$n2^k - 1$

Table 6.1: Left: Randomized blocks ANOVA table. Right: completely randomized design ANOVA table.

data points.

6.1 Replicates and more

Suppose that we get n data points from each of our 2^k treatment combinations. There are different ways that we could have done this and the way we did it can and should affect the analysis.

Suppose that we are making soup. We have two different stores, we use two different vegetables (carrots vs cabbage), cook at two different temperatures, and add two amounts of salt. There is more to our soup than just these factors but those other aspects are not varying. We now have $2^4 = 16$ precedures. The value Y is a measure of how good the soup tastes.

We are going to get $16n$ data points for some $n > 1$. One way to do this is to make all 16 kinds of soup on n different days. On each of those days we go to both stores, buy both kinds of vegetable cook at both temperature and use both salt levels. If we do it this way then we have 16 experimental treatments in n blocks, a ***randomized blocks*** design. Within each block (i.e., each day) we make the 16 soups in a random order.

We could also simply spend one day making soup 16 ways and taste each pot n times. This is really very different. It is sometimes called a ***repeated measures*** design. It is much faster, easier and cheaper than the blocked analysis but also much less informative. The order in which those tastes were made could be impactful, due to correlations between consecutive measurements. Maybe the soup is cooling off as we keep on tasting. We might just average those n values into one presumably better value.

A third way to do this is to make soup once on each of $16n$ different days. Each of the 16 combinations appears n times in a randomized order. This is a ***completely randomized design*** and it does not impose the balance that the randomized blocks version would. Table 6.1 shows ANOVA tables for the randomized blocks and completely randomized 2^k factorial experiment. If we did just average the repeated measures then we would not have any degrees of freedom for error and so Table 6.1 does not have a corresponding table. Later on we will see ways to analyze such an unreplicated experiment.

A B C D	Y
− − −−	(1)
+ − −−	a
− + −−	b
+ + −−	ab
− − +−	c
+ − +−	ac
− + +−	bc
+ + +−	abc
− − −+	d
+ − −+	ad
− + −+	bd
+ + −+	abd
− − ++	cd
+ − ++	acd
− + ++	bcd
+ + ++	abcd

Table 6.2: Standard notation for 16 runs in a 2^4 experiment.

6.2 Notation for 2^k experiments

Let's consider factors A, B, C and so forth at two levels each. Because these factors have 2 levels, they have $2-1=1$ degrees of freedom. Furthermore the interaction between a factor with I levels and one with J levels has $(I-1)(J-1)$ levels and so in this setting the two factor interactions also have 1 degree of freedom. Indeed, any interaction among any number of two level factors has just one degree of freedom.

There is something special about an effect having just one degree of freedom. We can attach a sign to it, and consider factors that increase or decrease a response. That is, given levels μ_1 and μ_2 we can compare them via $\mu_2-\mu_1$. We cannot easily and uniquely attach a sign to $(\mu_1-\mu_2, \mu_1-\mu_3, \mu_2-\mu_3) \in \mathbb{R}^3$. We could reduce them to $\mu_1-(\mu_2+\mu_3)/2$ and $\mu_3-\mu_2$ but that is just one of many potential comparisons to describe 2 degrees of freedom among three means.

To exploit this signing possibility we use a special notation for the observations in a 2^k experiment. For each factor we decide that one of its levels will be called the ***high level*** and the other will be the ***low level***. For a numerical value we would ordinarily take the higher value to be the high level. For an attribute like presence or absences of something, presence makes sense as a high level. In other settings the high level could be the new treatement, or the more expensive one, or the one that we anticipate is more likely to increase $\mathbb{E}(Y)$.

Table 6.2 shows all 16 possible combinations in a 2^4 experiment using + for high levels and − for low levels of factors A, B, C and D. The name of an observation is the string of letters at which it takes the high level. For instance "abd" is the observation at high levels of A and B and D with the low level

Y		A		B		AB	
(1)	b	−	−	−	+	+	−
a	ab	+	+	−	+	−	+

Table 6.3: The first square diagrams Y for a 2^2 experiment. The next ones show which observations are at high and low levels of A, B and AB.

of C. When everything is at the low level we use the symbol '(1)' instead of a blank or null string. We will be using some multiplicative formulas in which (1) makes a natural multiplicative identity.

The treatment effect for A is defined to be the average of $\mathbb{E}(Y)$ over 2^{k-1} observations at the high level of A minus the average of $\mathbb{E}(Y)$ over 2^{k-1} observations at the low level of A. It is called α_A and its estimate $\hat{\alpha}_A$ is the average of observed Y at the high level of A minus the average of observed Y at the low level of A. For instance with $k = 3$ and no replicates

$$\alpha_A = \frac{1}{4}\big(\mathbb{E}(Y_a) + \mathbb{E}(Y_{ab}) + \mathbb{E}(Y_{ac}) + \mathbb{E}(Y_{abc})\big) - \frac{1}{4}\big(\mathbb{E}(Y_{(1)}) + \mathbb{E}(Y_b) + \mathbb{E}(Y_c) + \mathbb{E}(Y_{bc})\big) \quad \text{and}$$
$$\hat{\alpha}_A = \frac{1}{4}\big(Y_a + Y_{ab} + Y_{ac} + Y_{abc}\big) - \frac{1}{4}\big(Y_{(1)} + Y_b + Y_c + Y_{bc}\big).$$

If there are replicates, we could call the obvservations $Y_{a,j}$ and $Y_{b,j}$ for $j = 1, \dots, n$ et cetera.

This is different from what we had before. With $k = 1$ we used to have $\mathbb{E}(Y_{1j}) = \mu + \alpha_1$ and $\mathbb{E}(Y_{1j}) = \mu + \alpha_2$ with $\alpha_1 + \alpha_2 = 0$. With that notation $\mathbb{E}(Y_{2j}) - \mathbb{E}(Y_{1j}) = \alpha_2 - \alpha_1 = 2\alpha_2$. Now we have $\mathbb{E}(Y_{a,j}) - \mathbb{E}(Y_{(1),j}) = \alpha_A$. That is $\alpha_a = 2\alpha_2$. In a 2^1 experiment we write

$$Y_{a,j} = \mu + \frac{1}{2}\alpha_a + \varepsilon_{a,j} \quad \text{and} \quad Y_{(1),j} = \mu - \frac{1}{2}\alpha_a + \varepsilon_{(1),j}.$$

Because interactions have one degree of freedom, we can give them a sign too. The AB interaction is the effect of A at the high level of B minus the effect of A at the low level of B. We can write it as

$$(ab - b) - (a - (1)) = ab - a - b + (1).$$

If instead we look at effect of B at the high level of A minus the effect of B at the low level of A we get

$$(ab - a) - (b - (1)) = ab - a - b + (1)$$

again That, is the AB interaction is the same as the BA interaction. Observations ab and (1) are at the high level of the AB interaction while a and b are at the low level. Table 6.3 depicts the AB interaction as a diagonal comparison in a 2×2 square of data values.

For a 2^2 experiment we may write the expected observatoins values as follows:

$$\begin{aligned}
\mathbb{E}(y_{ab}) &= \mu + \frac{1}{2}\big[\ \ \alpha_a + \alpha_b + \alpha_{ab}\big]\\
\mathbb{E}(y_{a}) &= \mu + \frac{1}{2}\big[\ \ \alpha_a - \alpha_b - \alpha_{ab}\big]\\
\mathbb{E}(y_{b}) &= \mu + \frac{1}{2}\big[-\alpha_a + \alpha_b - \alpha_{ab}\big] \quad \text{and}\\
\mathbb{E}(y_{(1)}) &= \mu + \frac{1}{2}\big[-\alpha_a - \alpha_b + \alpha_{ab}\big].
\end{aligned}$$

The estimated AB interaction parameter is $\hat{\alpha}_{ab}$ equal to the average of 2^{k-1} data values at the high level of the AB interaction minus the average of 2^{k-1} data values at the low level of the AB interaction. Now we find that

$$\begin{aligned}
&\mathrm{var}(\hat{\alpha}_a) = \mathrm{var}(\hat{\alpha}_b) = \mathrm{var}(\hat{\alpha}_{ab}) = \mathrm{var}(\hat{\alpha}_a) = \mathrm{var}(\hat{\alpha}_{abc}) = \cdots = \mathrm{var}(\hat{\alpha}_{abc\cdots z})\\
&\quad = \frac{\sigma^2}{N/2} + \frac{\sigma^2}{N/2} = \frac{4\sigma^2}{N} = \frac{4\sigma^2}{n2^k} = \frac{\sigma^2}{n2^{k-2}}.
\end{aligned}$$

Every main effect and every interaction of whatever order are all estimated with the same precision.

We can define interactions of all orders using this approach. The ABC interaction is the BC interaction at the high level of A minus the BC interaction at the low level of A:

$$\begin{aligned}
&(abc - ab - ac + a) - (bc - b - c + (1))\\
&= \underbrace{(abc + a + b + c)}_{\text{HI level of ABC}} - \underbrace{(ab + ac + bc + (1))}_{\text{LO level of ABC}}.
\end{aligned}$$

Just like other main effects and interactions, half the data are at the high level of ABC and half are at the low level.

Because each effect has one degree of freedom, we can use a t test for it instead of an F test. The t statistic for $H_0 : \alpha_a = 0$ is

$$\frac{\hat{\alpha}_a}{s/\sqrt{2^{k-2}n}},$$

where s is the standard error. The degrees of freedom are $(n-1)(2^k-1)$ if the experiment was run in n blocks of 2^k runs (and the model should then include a block effect) and the degrees of freedom are $n2^k - 1$ if it is a completely randomized allocation.

6.3 Why $n = 1$ is popular

If we take $n = 1$ then there are no replicates and there are then 0 degrees of freedom for error. Were we to compute s^2 we would get 0/0 (not a number).

To see why $n = 1$ is popular suppose that we have a choice between a 2^4 experiment to study factors ABCD with 2 replicates, and a 2^5 experiment to study factors ABCDE without replication. Given the chance to explore one more factor without a variance estimate many people would do that. Experimental design has room for both cautious approaches and bold ones.

In class we considered a famous example about an experiment to improve ball bearings. The engineers were planning $n = 4$ plant runs for two levels of one factor. A statistician persuaded them to use those same 8 plant runs to investigate 3 factors in a 2^3 layout. The additional factors proved to be much more effective than the originally contemplated one. The story is at this link `https://en.wikipedia.org/wiki/Factorial_experiment` after scrolling down a bit. What I like about this story is that it has the statistician playing the part of the bold person. As always, there is selection bias in the stories that get included in a course or text book or article.

There is a notion of ***factor sparsity*** that underlies this decision. The idea is that most of the α's are close to 0. Then the estimates that we get are mostly $\mathcal{N}(0, 4\sigma^2/N)$ plus a few real effects providing outliers. Then using QQ-plots of the estimated effects or their absolute values we can hope to spot the outliers. This approach was developed by Cuthbert Daniel (1959). Even if the α's are not exactly zero, it could happen that most of them are relatively small. They cannot all be relatively important. Suppose that the small ones can be modeled by effects that are $\mathcal{N}(\mu_*, \sigma_*^2)$. Then their estimates will look like $\mathcal{N}(\mu_*, \sigma_*^2 + 4\sigma^2/2^k)$. If the large effects have α values that are much greater than $|\mu_*| + \sqrt{\sigma_*^2 + 4\sigma^2/N}$ then they will still look like outliers.

When this factor sparsity is in play, then it would be wasteful to take $n > 1$ replicate instead of inspecting additional factors. If you replicate and leave out the most important factor, that seemingly safe choice can be very costly.

It is common to see that the important interactions in a QQ plot involve the important main effects. We might see $\hat{\alpha}_A$, $\hat{\alpha}_B$, $\hat{\alpha}_C$ and $\hat{\alpha}_{BC}$ as the outliers. In class we saw such a QQ plot in a the fractional factorial class.

Daniels' article closes with an interesting comment about why we just analyze one response at a time:

> The third set of questions plaguing the conscience of one applied statistician concerns the repeated use of univariate statistical methods instead of a single multivariate system. I know of no industrial product that has but one important property. And yet, mainly because of convenience, partly because of ignorance, and perhaps partly because of lack of full development of methods applicable to industrial research problems, I find myself using one-at-a-time methods on responses, even at the same time that I derogate a one-at-a-time approach to the factors influencing a response. To what extent does this simplification invalidate the judgments I make concerning the effects of all factors on all responses?

Factor sparsity supports a concept called ***design projection***. Suppose that in a 2^3 experiment that factor A is completely null: $\alpha_a = \alpha_{ab} = \alpha_{ac} = \alpha_{abc} = 0$.

Then our experiment is just like a 2^2 experiment in B and C with two replicates, depicted as follows:

$$\begin{bmatrix} A & B & C \\ -1 & -1 & -1 \\ -1 & -1 & +1 \\ -1 & +1 & -1 \\ -1 & +1 & +1 \\ +1 & -1 & -1 \\ +1 & -1 & +1 \\ +1 & +1 & -1 \\ +1 & +1 & +1 \end{bmatrix} \longrightarrow \begin{bmatrix} B & C \\ -1 & -1 \\ -1 & +1 \\ +1 & -1 \\ +1 & +1 \\ -1 & -1 \\ -1 & +1 \\ +1 & -1 \\ +1 & +1 \end{bmatrix} = \begin{bmatrix} B & C \\ -1 & -1 \\ -1 & -1 \\ -1 & +1 \\ -1 & +1 \\ +1 & -1 \\ +1 & -1 \\ +1 & +1 \\ +1 & +1 \end{bmatrix}.$$

In the rightmost array the runs are reordered to show the replication. The same thing happens if either B or C is null. In the multi-response setting that Daniels' was concerned with, there might be a different null factor for each of the responses being measured.

In practice it is unlikely that A is perfectly null but it might be relatively null with some other $|\alpha|$'s being orders of magnitude larger than the ones involving A. It could be tricky to decide whether A is null and would involve the issues raised in the next section on factorial experiments with no replicates.

6.4 Factorial with no replicates

If we had and $I \times J$ experiment run just $n = 1$ time, then our $N = I \times J \times 1$ data points could be used to get SSA, SSB and SSE. If the factors interact, then SSE will be inflated by that interaction. It will have a noncentral χ^2 distribution (under our Gaussian assumption). Then $F = \text{MSA}/\text{MSE}$ will be doubly noncentral. Noncentrality in SSE will make MSE larger, making F smaller, making for less signficance (larger p values). So our test would be conservative having lower power. If the resulting p-value is below 0.01 or other threshold then it likely would have been even further below it in the event of no interaction.

In the 2^k experiment we could designate a priori some high order interactions to use in forming an error estimate. We could take a mean square of the corresponding $\hat{\alpha}$'s as our estimate of $4\sigma^2/N$. This would have the same conservative property as the doubly noncentral F described above.

We could possibly just take the 10 smallest $\hat{\alpha}^2$'s but then we would have to adjust them for this selection. For instance, a Monte Carlo computation can find for us the distribution of $\sum_{j=1}^{10}(\hat{\alpha}^2)_{(i)}/(4\sigma^2/N)$.

There are lots of methods to get approximate inferences without replication. Hamada and Balakrishnan (1998) describe 24 different methods. One prominent one is due to Lenth (1989). In recent years there has been much interest in inference after model selection and perhaps progress in that area could lead to some new methods here.

y	M	A	B	C	AB	AC	BC	ABC
(1)	+	−	−	−	+	+	+	−
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
ab	+	+	+	−	+	−	−	−
c	+	−	−	+	+	−	−	+
ac	+	+	−	+	−	+	−	−
bc	+	−	+	+	−	−	+	−
abc	+	+	+	+	+	+	+	+

Table 6.4: Each row is an observation in a $2^k = 2^3$ experiment. Each column shows an effect of interest with M representing the grand mean. The estimate of that effect is the inner product of the corresponding column of ± 1's divided by 2^{k-1}, except that for the grand mean (column M) we divide by 2^k.

6.5 Generalized interactions

For $n = 1$, the effects that we are interested in are dot products of a vector of Y values with some coefficients ± 1 and then divided by 2^k for $\hat{\mu}$ or 2^{k-1} for the αs. Table 6.4 shows those coefficents for $k = 3$.

Using these signs we define some ***generalized interactions***. The generalized interaction of AB and BC is what we get if we mutiply the column of signs for AB times that for BC. We get $AB \times BC = AB^2C = AC$ because B^2 is a column of +1s. Similarly $AB \times ABC = C$.

We can compare our 2^k effects to a plain linear regression model

$$Y = X\beta + \varepsilon \in \mathbb{R}^N$$

with parameter estimates $\hat{\beta} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ and $\mathrm{var}(\hat{\beta}) = (X^\mathsf{T}X)^{-1}\sigma^2$ We could take X to be the matrix in Table 6.4 or a generalization of it to some other k. Now $X^\mathsf{T}X = 2^k \times I_{2^k}$ so $\hat{\beta}$ would have $\hat{\mu}$ in it and $2\hat{\alpha}$ for the other effects.

What we see here is that all of the $\hat{\beta}_j$ (and hence all of our effects) are uncorrelated because $\mathrm{var}(\hat{\beta}) = \sigma^2 I_{2^k}/2^k$. Ordinarily a regression with N observations and p parameters has a computational cost proportional to Np^2. Here $N = p = 2^k$ and the cost would be $O(N^3)$. In this instance, the orthogonality of X reduces the cost to just that of p one at a time regressions actually dot products, and the cost is $O(N^2)$.

Yates's algorithm (Yates, 1937) reduces the cost to $O(N\log(N))$. When $N = 16$ we are not much interested in the cost difference between N^2 and $N\log(N)$. If however $N = 2^{30}$ as it might in an experiment carried out solely in software, then Yates' algorithm is very useful. His algorithm is a fast Fourier transform used for hand calculations decades before the fast Fourier transform became prominent in signal processing.

We can get an algebraic expression to show which observations are at the high and low levels of an interaction. Consider the BCD interaction. We can

b	ab	(1)	a
c	ac	bc	abc
d	ad	cd	acd
bcd	abcd	bd	abd

Table 6.5: The 8 observations on the left are at the high level of BCD. The 8 observations on the right are at the low level of BCD.

write it as

$$\begin{aligned} BCD &= \frac{1}{2^{k-1}}(a+1)(b-1)(c-1)(d-1) \\ &= \frac{1}{2^{k-1}}(ab+b-1-b)(cd-c-d+1) \\ &= \cdots \\ &= \frac{1}{2^{k-1}}(b+c+d+bcd+ab+ac+ad+abcd) \\ &\quad - \frac{1}{2^{k-1}}((1)+bc+cd+bd+a+abc+acd+abd). \end{aligned}$$

The high and low levels of BCD are depicted in Table 6.5. What we see are that bcd is at the high level. Anything that is missing an odd number of those letters is at the low level. Anything missing an even number is at the high level (just like zero which is even). The presence of another factor like A does not change it.

6.6 Blocking

We might want to (or have to) run our 2^k experiment in blocks. One common choice has blocks of size 2^{k-1}. So our experiment has 2 blocks of size 2^{k-1}. Blocking could be done for greater accuracy (balancing out some source of noise) or we might be forced into it: the equipment holds exactly 8 pieces, no more.

The best plan for 2 blocks is run one block at the high level of the k-factor interaction and one at the low level. For $k = 3$ it would be

$$\boxed{a \quad b \quad c \quad abc} \quad \text{vs} \quad \boxed{1 \quad ab \quad ac \quad bc}$$

with run order randomized within the blocks. With this choice the ABC interaction is confounded with the blocks. The effects for A, B, C, AB, BC and AC are orthogonal to blocks (because they're orthogonal to ABC).

If we only have these 3 factors we might replicate the block structure as follows:

$$\boxed{\begin{array}{|c|c|}\hline abc & ab \\ a & ac \\ b & bc \\ c & 1 \\ \hline\end{array}} \quad \boxed{\begin{array}{|c|c|}\hline abc & ab \\ a & ac \\ b & bc \\ c & 1 \\ \hline\end{array}} \quad \boxed{\begin{array}{|c|c|}\hline abc & ab \\ a & ac \\ b & bc \\ c & 1 \\ \hline\end{array}}$$

Source	df
Replicates	2
Blocks = ABC	1
Blocks × replicates	2
A, B, C, AB, AC, BC	1 each
Error	12
Total	23

Table 6.6: This is the ANOVA table for a blocked 2^3 experiment with $n = 3$ replicates each having 2 blocks.

Of course if there were additional factors to investigate we might introduce those instead of replicating our blocked 2^3 experiment.

The ANOVA table for this setup is shown in Table 6.6. With 3 replicates, there are 2 degrees of freedom for replicates. One degree of freedom goes to comparing the block at the high level of ABC to the one at the low level. There are also 2 degrees of freedom for blocks by replicates and 1 degree of freedom for each of our main effects and two factor interactions. This table will let us study the main effects and two factor interactions.

To study the 3 factor interaction we can test block level averages. We form the 2×3 matrix

$$\begin{array}{c} \\ \text{ABC}{=}+ \\ \text{ABC}{=}- \end{array} \begin{array}{c} \begin{array}{ccc} r{=}1 & r{=}2 & r{=}3 \end{array} \\ \begin{bmatrix} \bar{Y}_{+1\bullet} & \bar{Y}_{+2\bullet} & \bar{Y}_{+3\bullet} \\ \bar{Y}_{-1\bullet} & \bar{Y}_{-2\bullet} & \bar{Y}_{-3\bullet} \end{bmatrix} \end{array}$$

and test the row effect. Wu and Hamada (2011) warn that there will be trouble if the blocks interact with treatments. That will also mess up other blocked designs.

With many replicates, there is also the possibility of confounding ABC with blocks in some replicates and other interactions such as BC in different replicates. The analysis of such combinations could get pretty complicated. In class I mentioned that it might be wise to simply use a probabilistic programming language (e.g., Stan or others mentioned here `https://en.wikipedia.org/wiki/Probabilistic_programming#Probabilistic_programming_languages`) to analyse such a thing instead of wrangling mean squares and ANOVA tables. That would handle the analysis. To actually choose a design one could simulate data from it many times and pipe the simulated output through the probabilistic programming language.

7

Fractional factorials

Even at just two levels, studying k factors brings in 2^k levels. It is a curse of dimension. We cannot necessarily afford that. And maybe we don't have to. Fractional factorials look at k factors of 2 levels each using fewer than 2^k runs. We will study k factors using only $N = 2^{k-p}$ runs for some integer $1 \leqslant p < k$. These are called ***fractional factorials***. The name stems from the fact that they are fractional replicates of 2^k. It is as if we are doing n replicates for $n = 1/2$.

There are many good references for this topic, such as Wu and Hamada (2011) and Box et al. (1978) and Montgomery (1997). The last two examples are not the most recent editions of those books. I prefer the cited editions. For instance later versions of the book by Montgomery seem less suitable for a graduate level class, though they are better suited for their target audience than the cited edition.

We will study k binary inputs in 2^{k-r} runs. This necessarily involves aliasing/confounding relationships among the 2^k estimands of interest. When the phenomenon is dominated by main effects and low order interactions and the high order interactions are small, then aliasing the high order interactions with each other makes sense. We still get to learn about the important quantities at low cost.

One problem with a 2^k factorial experiment is that as k grows larger, most of the degrees of freedom are used up in interactions of order about $k/2$. We have much more interest in learning about main effects and low order interactions. There's a sense that they're more likely to be large and they are also going to be simpler to interpret if we can estimate them. It can be puzzling to wonder why degrees of freedom are so important. The explanation is that each extra degree of freedom costs us one more data point. See Table 7.1 which illustrates the point with $k = 7$ and runs costing \$1000 each. We end up spending \$8000

Interaction order:	0	1	2	3	4	5	6	7
Degrees of freedom:	$\binom{7}{0}$	$\binom{7}{1}$	$\binom{7}{2}$	$\binom{7}{3}$	$\binom{7}{4}$	$\binom{7}{5}$	$\binom{7}{6}$	$\binom{7}{7}$
Degrees of freedom:	1	7	21	35	35	21	7	1
Cost	\$1k	\$7k	\$21k	\$35k	\$35k	\$21k	\$7k	\$1k

Table 7.1: For a hypothetical 2^7 experiment where each run costs \$1000, the table shows how much of the budget goes to interactions of each size.

on the grand mean and main effects along with \$59,000 on interactions of order three and higher.

7.1 Half replicates

Let's do 2^{k-1} runs, i.e., half of the full 2^k experiment. We will get confounding. To make the confounding minimally damaging, we will confound "done" versus "not done" with the k-fold interaction.

For $k = 3$ we could do the 4 runs at the high level of ABC:

$$\begin{array}{c} \\ a \\ b \\ c \\ abc \end{array} \begin{array}{c} \begin{array}{cccccccc} I & A & B & AB & C & AC & BC & ABC \end{array} \\ \left[\begin{array}{cccccccc} + & + & - & - & - & - & + & + \\ + & - & + & - & - & + & - & + \\ + & - & - & + & + & - & - & + \\ + & + & + & + & + & + & + & + \end{array}\right] \end{array}$$

We see right away that the intercept column I is identical to the one for ABC. We say that the grand mean is ***aliased*** with ABC. It is also confounded with ABC. We then get

$$\mathbb{E}(\hat{\mu}) = \mu + \alpha_{ABC}.$$

In our half replicate, $I = ABC$ holds. If we multiply both sides by A, we get $A \times I = A \times ABC$. Of course $A \times I = A$ and $A \times ABC = A^2BC = BC$. It follows that $A = BC$ so the main effect for A is aliased with the BC interaction. Therefore

$$\mathbb{E}(\hat{\alpha}_A) = \alpha_A + \alpha_{BC}.$$

By the same argument $B = AC$ and $C = AB$. If we do a 2^{k-1} experiment running everything at the high level of the k-fold interaction then every effect will be aliased with one other, the one that complements it's set of variables.

The set of all eight runs looks like this:

	I	A	B	AB	C	AC	BC	ABC
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
c	+	−	−	+	+	−	−	+
abc	+	+	+	+	+	+	+	+
bc	+	−	+	−	+	−	+	−
ac	+	+	−	−	+	+	−	−
ab	+	+	+	+	−	−	−	−
(1)	+	−	−	+	−	+	+	−

We could do either the top half, with $ABC = I$ or the bottom half with $ABC = -I$. That is $ABC = \pm I$ give a 2^{3-1} experiment. If we use the bottom half we get

$$\mathbb{E}(\hat{\alpha}_A) = \alpha_A - \alpha_{BC}.$$

If any of the k factors in our 2^{k-1} factorial experiment are "null" then we get a full 2^k factorial experiment in the remaining ones. Practially, null means "relatively null" in comparison to some larger effects.

Geometrically the sampled points of a 2^{3-1} experiment lie on 4 corners of the cube $\{-1, 1\}^3$ (or if you prefer $[0, 1]^3$). If we draw in the edges on each face of the cube, no pair of sampled points share an edge.

A 2^{k-1} design looks like one block of a blocked 2^k where the blocks are defined by high and low levels of the k-factor interaction. If we have just done the $ABC\cdots Z = I$ block we could analyze it as a 2^{k-1} design and perhaps decide that we have learned what we need and don't have to do the other block.

7.2 Catapult example

The lightning calculator Cat-100 catapult is described here `http://www.qualitytng.com/cat-100-catapult/`. In class we looked at data from a 2^{5-1} experiment in it. The data are shown in Table 7.2. Distance is in centimeters. It appears that most of the energy from the surgical tubing goes into moving the wooden arm, so the projectiles remain safe for classroom use (and are not suitable as siege weapon).

Table 7.3 shows those same data in standard (Yates) order. That is (1), a, b, ab, et cetera for four of the five variables. The reason to do this is that the data were analyzed by Yates' algorithm which is not really necessary on a problem this small.

Yates' algorithm pair off the data into $n/2$ consecutive pairs. It takes sums within those $n/2$ pairs and places them above differences. If we do that k times in a full factorial or $k-1$ times in a fractional factorial we get a column with main effects and interaction estimates in the standard order (scaled up by n for

	Front	Back	Fixed	Moving	Bucket	Dist
1	1	-1	-1	1	1	210.3
2	1	1	1	1	1	343.0
3	1	-1	-1	-1	-1	50.0
4	-1	-1	1	1	1	263.5
5	-1	-1	-1	1	-1	134.5
6	-1	-1	-1	-1	1	94.5
7	-1	1	-1	1	1	310.8
8	1	1	-1	-1	1	94.8
9	-1	1	-1	-1	-1	91.5
10	1	1	-1	1	-1	168.5
11	-1	1	1	-1	1	277.4
12	1	1	1	-1	-1	145.5
13	1	-1	1	-1	1	157.5
14	-1	1	1	1	-1	266.5
15	-1	-1	1	-1	-1	120.5
16	1	-1	1	1	-1	166.5

Table 7.2: Experimental output from a catapult experiment.

	Front	Back	Fixed	Moving	Bucket	Dist
6	-1	-1	-1	-1	1	94.5
3	1	-1	-1	-1	-1	50.0
9	-1	1	-1	-1	-1	91.5
8	1	1	-1	-1	1	94.8
15	-1	-1	1	-1	-1	120.5
13	1	-1	1	-1	1	157.5
11	-1	1	1	-1	1	277.4
12	1	1	1	-1	-1	145.5
5	-1	-1	-1	1	-1	134.5
1	1	-1	-1	1	1	210.3
7	-1	1	-1	1	1	310.8
10	1	1	-1	1	-1	168.5
4	-1	-1	1	1	1	263.5
16	1	-1	1	1	-1	166.5
14	-1	1	1	1	-1	266.5
2	1	1	1	1	1	343.0

Table 7.3: Catapult data in Yates' order.

the main effect and $n/2$ for the others). For $k = 2$ we get

$$\begin{array}{llll} y_{(1)} & y_a + y_{(1)} & y_{ab} + y_b + y_a + y_{(1)} & \to \quad 4\hat{\mu} \\ y_a & y_{ab} + y_b & y_{ab} - y_b + y_a - y_{(1)} & \to \quad 2\hat{\alpha}_A \\ y_b & y_a - y_{(1)} & y_{ab} + y_b - y_a - y_{(1)} & \to \quad 2\hat{\alpha}_B \\ y_{ab} & y_{ab} - y_b & y_{ab} - y_b - y_a + y_{(1)} & \to \quad 2\hat{\alpha}_{AB} \end{array}$$

and for $k = 2^{30}$ it would only take 30 of these operations to compute 2^{30} effects of interest (most likely in a computer experiment). In a fractional factorial we need to account for the aliasing.

Here it is (or at least part of it) for $k = 3$:

$$\begin{array}{llll} y_{(1)} & y_a + y_{(1)} & y_{ab} + y_b + y_a + y_{(1)} & y_{ab} + y_b + y_a + y_{(1)} y_{ab} + y_b + y_a + y_{(1)} \\ y_a & y_{ab} + y_b & y_{abc} + y_{bc} + y_{ac} + y_c & y_{abc} - y_{bc} + y_{ac} - y_c + y_{ab} - y_b + y_a - y_{(1)} \\ y_b & y_{ac} + y_c & y_{ab} - y_b + y_a - y_{(1)} & \\ y_{ab} & y_{abc} + y_{bc} & y_{abc} - y_{bc} + y_{ac} - y_c & \text{et cetera} \\ y_c & y_a - y_{(1)} & y_{ab} + y_b - y_a - y_{(1)} & \\ y_{ac} & y_{ab} - y_b & y_{abc} + y_{bc} - y_{ac} - y_c & \\ y_{bc} & y_{ac} - y_c & y_{ab} - y_b - y_a + y_{(1)} & \\ y_{abc} & y_{abc} - y_{bc} & y_{abc} - y_{bc} - y_{ac} + y_c & \end{array}$$

The digression on Yates' algorithm demystifies the axis label in Figure 7.1. There we see that main effects for 'back stop', 'fixed arm', 'moving arm' and 'bucket' all have positive values clearly separated from the noise level. The effect for 'front stop' appears to be negative but is not clearly separated from the others. The reference line is based on a least squares fit to the 11 smallest effects (F and the interactions). Effects are named after their lowest alias.

In this instance there was a clear break between large effects and most small effects and some reasonable doubt as to whether F belongs with the large or the small ones. In many examples in the literature some main effects end up in the bulk of small effects and a handful of two factor interactions show large effects. Usually one or both of the interacting factors appears also as a large main effect.

One reason to be interested in statistical insignificance is that when it happens we clearly do not know the sign of the effect, even if we're certain that an exact zero cannot be true. Statistical significance makes it more reasonable that you can confidently state the sign of the effect. There can however be doubt about the sign if the confidence interval for the effect has an edge too close to zero. This could happen in a low power setting. See Owen (2017).

Here is a naive regression analysis of the data.

```
Coefficients:
             Value Std. Error t value
(Intercept)  180.96      7.441  24.318
      Front  -13.94      7.441  -1.874
       Back   31.29      7.441   4.205
      Fixed   36.59      7.441   4.918
```

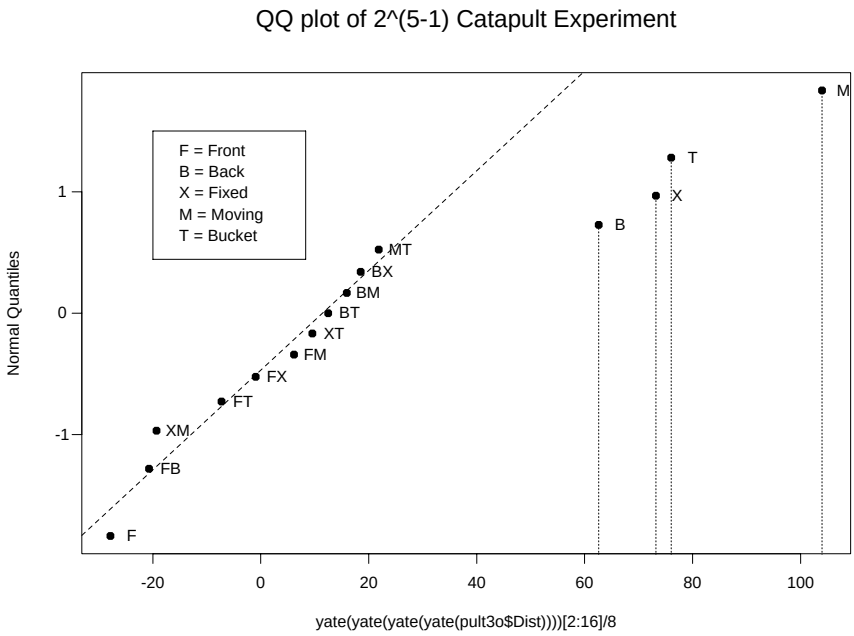


Figure 7.1: QQ plot of catapult experiment.

```
Moving   51.99      7.441   6.987
Bucket   38.02      7.441   5.109

Residual standard error: 29.77 on 10 degrees of freedom
Multiple R-Squared: 0.9233
```

It would be interesting to see how modern post-selective inference methods might work with factorial and fractional factorial models.

7.3 Quarter fractions

For a 2^{k-2} experiment we can set two combinations of effects equal to I. Or we could set one or both of them equal to $-I$. For $k = 6$, if we have set $ABCDEF = I$ we might first consider (bad idea) also setting $ABCDE = I$. This is bad because then the product of $ABCDEF$ and $ABCDE$ becomes $I^2 = I$ and this product is

$$ABCDEF \times ABCDE = F.$$

We certainly don't want to alias one of our main effects to the grand mean.

A better choice is to set $ABCD = I$ and $CDEF = I$. We then also get $ABCD \times CDEF = ABEF = I$. Then we find that

$$A = A \times (ABCD, CDEF, ABEF) = BCD = ACDEF = BEF$$

and so we get

$$\begin{aligned}
\mathbb{E}(\hat{\alpha}_A) &= \alpha_A + \alpha_{BCD} + \alpha_{ACDEF} + \alpha_{BEF}, \quad \text{and similarly}\\
\mathbb{E}(\hat{\alpha}_{AB}) &= \alpha_{AB} + \alpha_{CD} + \alpha_{ABCDEF} + \alpha_{EF}, \quad \text{and}\\
\mathbb{E}(\hat{\alpha}_{ABC}) &= \alpha_{ABC} + \alpha_D + \alpha_{ABDEF} + \alpha_{CEF}.
\end{aligned}$$

In a quarter replicate, any effect that we compute estimates a combination of four effects. One is the desired effect. The other three are aliases. The aliases above all have a positive sign. If we had set $ABCD = -I$ then $\mathbb{E}(\hat{\alpha}_A)$ would include some aliased effects with negative signs.

Here are some of the aliasing patterns in this design:

$$\begin{aligned}
I &= ABCD = CDEF = ABEF\\
AB &= CD = ABCDEF = EF\\
AC &= BD = ADEF = CBEF\\
AD &= BC = AEF = BDEF\\
AE &= BCDE = ACDF = BF\\
AF &= BCDF = ACDE = BE
\end{aligned}$$

The cited books have tables of design choices for fractional factorial experiments. Those tables can run to several pages. Probably nobody actually memorizes or even uses them all but it is good to have those choices. One of the key quantities in those designs is the "resolution" that we discuss next.

7.4 Resolution

The ***resolution*** of a fractional factorial experiment is the number of letters in the shortest word in the defining relation. A defining relation is an expression like $ABCD = \pm I$. In this case the ***word*** is ABCD (not I!) and it has length 4. The quarter fraction example from the previous section had defining relations $ABCD = I$, $CDEF = I$ and $ABEF = I$, so the shortest one is 4. In a 2^{k-1} fraction with defining relation setting the k-factor interaction equal to $\pm I$ the shortest work length is k.

When there are k factors of two levels each and we have 2^{k-p} runs in a fractional factorial of resolution R, then the design is labeled 2^{k-p}_R. Resolution is conventionally given in Roman numerals. The three most important ones are III, IV and V.

For resolution $R = III$, no main effects are aliased/confounded with each other. Some main effects are confounded with two factor interactions.

For resolution $R = IV$, no main effects are confounded with each other, no main effects are confounded with any two factor interactions, and some two factor interactions are confounded with each other.

For esolution $R = V$, no main effects are confounded with each other, no main effects are confounded with two factor interactions and no two factor interactions are confounded with each other. However, some main effects are confounded with four factor interactions, and some two factor interactions confounded with three factor interactions.

Table 7.4 has an informal summary of what these resolutions require. Resolution V has the least confounding but requires the most expense. Resolution III has the least expense but could be misleading if we do not have enough factor sparsity. Resolution IV is a compromise but it requires careful thought. Some but not all of the two factor interactions may be aliased with others. We can check the tables or defining relations to see which they are. If we have good knowledge or guesses ahead of time we can keep the interactions most likely to be important unconfounded with each other. Similarly, after the experiment a better understanding of the underlying science would help us guess which interaction in a confounded pair might actually be most important.

For a specific pattern 2^{k-p}_R there can be multiple designs and they are not all equally good. For instance with $R = IV$ we would prefer a design with the smallest number of aliased $2FI$'s (***two factor interactions***). If there's tie we would break it in favor of the design with the fewest aliased $3FI$s. Carrying on until the tie is broken we reach a ***minimum aberration*** design. An investigator would of course turn to tables of minimum aberration designs constructed by researchers who specialize in this.

For resolution R	you need to be	because you need
III	lucky	few significant effects, all of low order
IV	smart	to untangle confounded two factor inter.s
V	rich	the largest sample size

Table 7.4: Informal synopsis of what we you need for the three most common resolutions.

For the 2^{k-1} experiment with the k-fold interaction aliased to $\pm I$ we find that $R = k$. So $k = 3$ gives 2^{3-1}_{III}, $k = 4$ gives 2^{4-1}_{IV} and $k = 5$ gives 2^{5-1}_{V}.

There is a projection property for resolution R. If we select any $R-1$ factors then we get all 2^{R-1} possible combinations the same number of times. That has to be $2^{k-p}/2^{R-1}$ times, that is $2^{k-p-R+1}$ times.

7.5 Overwriting notation

We need to figure out how to actually conduct our 2^{k-p} experiment. For a 2^{4-1} we can get a table of runs in factors A, B, and C. Then we replace/overwrite the ABC column by D, like this

	I	A	B	AB	C	AC	BC	$D{=}ABC$
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
c	+	−	−	+	+	−	−	+
abc	+	+	+	+	+	+	+	+
bc	+	−	+	−	+	−	+	−
ac	+	+	−	−	+	+	−	−
ab	+	+	+	+	−	−	−	−
(1)	+	−	−	+	−	+	+	−

Now $AD = BC$ as before (check that for yourself).

The more recent edition of Box Hunter and Hunter (Box et al., 2005) describes a ***nodal design***. It has $n = 16$ runs. You could analyze 15 effects, A, B, C, D, and interactions up to ABCD.

For 2^{5-1}_{V} they alias a fifth effect to ABCD. For 2^{8-4}_{IV} they alias four more effects as follows $L = ABC$, $M = ABD$, $N = ACD$ and $O = BCD$. For 2^{15-11}_{III} they alias effects as follows $E = AB$, $F = AC$, $G = AD$, $H = BC$, $J = BD$ and $K = CD$. They skip the letter I because it can mean 'intercept'.

7.6 Saturated designs

Using 2^{15-11}_{III} we can estimate a grand mean and 15 main effect, but no interactions and we get no degrees of freedom for error. Or if we like we could study only $15 - r$ effects and get r degrees of freedom for error.

There is a special setting where we can dispense with any concern over interactions. The classic example is when we are weighing objects. The weight of a pair of objects is simply the sum of their weights, with interactions being zero. Experimental designs tuned to a setting where interactions are known to be impossible are called ***weighing designs***.

A special kind of weighing design is known as ***Plackett-Burman*** designs. They exist when $n = 4m$ for (most) values of m, and so n does not have to be a power of 2. We will see them later as Hadamard designs.

7.7 Followup fractions

Suppose we have done a 1/4 fraction. Then we can follow up with a second 1/4 in more than one different way. We would ordinarily want to treat that second fraction as a block.

Suppose that factor A looks really important in the first 1/4 that we do. Maybe the aliasing pattern is

$$\mathbb{E}(\hat{\alpha}_A) = \alpha_A + \alpha_{BC} \pm \cdots .$$

In the second part of our experiment, we could flip the signs of the A assignments but leave everything else unchanged. In that second half we will have

$$\mathbb{E}(\hat{\alpha}_A) = \alpha_A - \alpha_{BC} \mp \cdots .$$

If we pool the two parts of our experiment, then

$$\begin{aligned}\hat{\alpha}_A &= \frac{1}{2}\big(\text{first expt } \hat{\alpha}_A + \text{second expt } \hat{\alpha}_A\big)\\ \mathbb{E}(\hat{\alpha}_A) &= \frac{1}{2}\big((\alpha_A + \alpha_{BC} + \cdots) + (\alpha_A - \alpha_{BC} - \cdots)\big) = \alpha_A.\end{aligned}$$

We get rid of any aliasing for A. Of course if B had looked important after the first 1/4 we could have flipped B. We would not have to know which one is important before doing the first 1/4 of the factorial.

A second kind of followup is the ***foldover***. We flip the signs of all the factors. (Note that we cannot and don't attempt to flip the intercept.) Foldovers deconfound an main effect from any 2FI that it was confounded with. To see this, suppose that $A = BC$ in the first experiment. Flipping signs of everything makes $(-A) = (-B)(-C)$. That means $-A = BC$ or $A = -BC$. Then

$$\mathbb{E}(\hat{\alpha}_A) = \frac{1}{2}\Big[(\alpha_A + \alpha_{BC} + \cdots) + (\alpha_A - \alpha_{BC} + \cdots)\Big] = \alpha_A + \cdots .$$

If however $A = BCD$ then flipping the signs makes $(-A) = (-B)(-C)(-D) = -BCD$ so A still equals BCD.

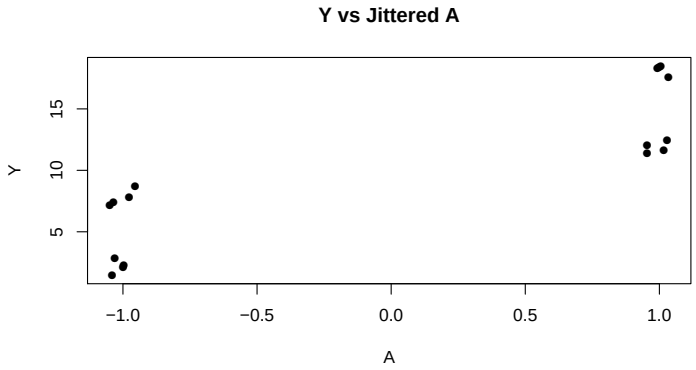


Figure 7.2: Exaggerated plot when A is a strong main effect and there is a second strong main effect.

7.8 More data analysis

Figure 7.2 shows what we might see in a plot of Y versus an important factor A. The bimodal pattern is what we would see if there were a second important factor. If A were a relatively very weak (almost null) factor then we would see almost the same point pattern at each side of the plot. Figure 7.3 shows what we might see of A has a strong interaction with one or more other factors. The variance of Y depends strongly on A. A graphical way to look for such patterns is to compute $F_A = \log(\mathrm{var}(y \mid A = 1)/\mathrm{var}(y \mid A = -1))$ and similar things for other main effects and interactions and produce a QQ plot of them. The outliers could be factors that are involved in many interactions. This could be fooled: it would be possible to have AB place a lot of variance at the low level of A while AC places a lot at the high level of A making F_A close to zero.

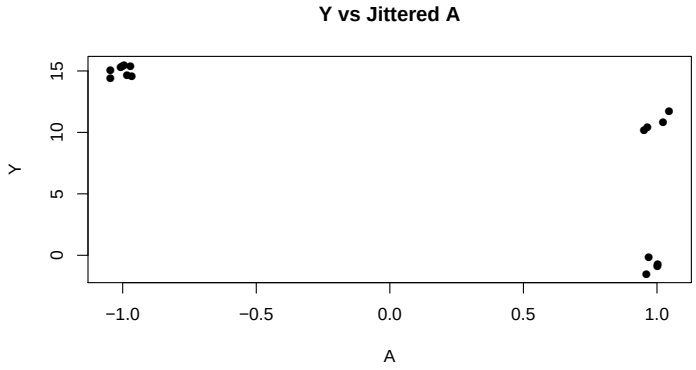


Figure 7.3: Exaggerated plot when A is a strong main effect and it has a strong interaction with one or more other factors.

8

Analysis of covariance and crossover designs

What if we have a categorical treatment variable and some continuous ones? It is easy to fold them all into one regresson. The combination is known as the analysis of covariance ***ANCOVA***. The name derives from some specialized shortcut algorithms to fit this model. Those were very useful before computers became routinely used. Now tt would take a staggeringly large data set for those algorithms to bring a meaningful savings now. Cutting 10^{-6} seconds in half is not interesting. What is more interesting and useful is the statistical thinking involved in using this method, especially in regard to measurements taken both before and after at treatment is applied.

In this chapter we also consider a related way compare before and after measurements. It is known as the cross-over design and it applies two or more treatments to the same subject, one treatment at a time.

8.1 Combining an ANOVA with continuous predictors

Suppose that we have Y_{ij} for treatments $i = 1, \dots, k$ and replicates $j = 1, \dots, n$. Now suppose that we additionally have continuously distributed covariates $\boldsymbol{x}_{ij} \in \mathbb{R}^p$. We can combine the ANOVA model with the continuous predictors by writing

$$Y_{ij} = \mu + \alpha_i + \boldsymbol{x}_{ij}^{\mathsf{T}}\beta + \varepsilon_{ij}.$$

It seems better to center the $\boldsymbol{x}_{ij}$ and fit

$$Y_{ij} = \mu + \alpha_i + (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\bullet\bullet})^{\mathsf{T}}\beta + \varepsilon_{ij}$$

where

$$\bar{\boldsymbol{x}}_{\bullet\bullet} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}\boldsymbol{x}_{ij}.$$

This way $\mu = \mathbb{E}(\bar{y}_{\bullet\bullet})$. We also still impose $\sum_i \alpha_i = 0$.

We can similarly add the $(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\bullet\bullet})^{\mathsf{T}}\beta$ term to any of our other models: blocked designs, Latin squares, factorials and fractional factorials. For instance, in a randomized block experiment we would fit

$$Y_{ij} = \mu + \alpha_i + b_j + (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{\bullet\bullet})^{\mathsf{T}}\beta + \varepsilon_{ij}$$

where b_j are block effects.

8.2 Before and after

An extremely important special case arises when for $x_{ij} \in \mathbb{R}$ is the pre-treatment version of the same quantity we measure as Y_{ij} after the treatment. Even with a random assignment of treatments to experimental units there can still be a meaningfully large amount of variance in x_{ij} between treatment groups $i = 1, \dots, k$. The randomization still means that we get valid/reliable variance estimates, confidence intervals for the treatment effect and p-values for H_0 : $\mathbb{E}(Y \mid A) = \mathbb{E}(Y \mid B)$ for treatments A and B. We can often still improve our analysis by getting reliable inferences with smaller variance by taking account of the x_{ij}.

Consider treatments to help people lose weight. The subjects might vary considerably in weight prior to the treatment. A useful weight reduction could be much smaller than the person to person differences. Those would then make it hard to properly compare the treatments. For two treatments, if we did not take account of x_{ij} and used a model like

$$Y_{ij} = \alpha + \mu_i + \varepsilon_{ij}$$

then $\mathrm{var}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})$ under random assignment would be very large.

In a setting like this we might opt to analyze differences $D_{ij} = Y_{ij} - x_{ij}$ and then test $H_0 : \mathbb{E}(D \mid A) = \mathbb{E}(D \mid B)$. This can be a great improvement over comparing unadjusted post-treatment means.

In class we saw a dental example from Fleiss (1986). There x_{ij} and Y_{ij} were measurements of gingivitis (gum inflammation) prior to and following upon one of two treatments. Fleiss analyzes $D_{ij} = x_{ij} - Y_{ij}$, the opposite of the difference described above. With this choice positive values indicate better outcomes, since gingivitis is undesirable.

The post treatment data are as follows:

Trt	n	Mean	s.dev.
1	74	0.5514	0.3054
2	64	0.3927	0.1988

Fleiss gives four significant figures for most numerical values in this example. These data lead to an estimated benefit for treatment 2 of 0.1587. A t-test (pooling the variance estimates) gives $t = 3.56$ which is significant enough. [Exercise: figure out the degrees of freedom and the p-value.]

Prior to the treatment, however, the subjects started off as follows:

Trt	n	Mean	s.dev.
1	74	0.6065	0.2541
2	64	0.5578	0.2293

So group 2 started out with an advantage of 0.0487. That is roughly one third of the apparent post-treatment gain.

We could reasonably want to get a sharper estimate of the treatment benefit. In some problems it might be enough to simply be confident about which of two treatments is best. For other purposes it is important to estimate how much better that treatment is.

The table of treatment differences $x_{ij} - Y_{ij}$ is

Trt	n	Mean	s.dev.
1	74	0.0551	0.2192
2	64	0.1651	0.2235

The estimated benefit from treatment 2 is now 0.1100 with $t = 2.91$ (still significant).

Fleiss articulates two goals. One is to properly account for pre-treatment differences. Another is to reduce the variance of the treatment effect estimate.

Differences do not alway achieve the second goal. To see why, let $\rho = \mathrm{corr}(x_{ij}, Y_{ij})$ and $\sigma_y^2 = \mathrm{var}(Y_{ij})$ and $\sigma_x^2 = \mathrm{var}(x_{ij})$. Then $\mathrm{var}(x_{ij} - Y_{ij}) = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$. Now $\mathrm{var}(x_{ij} - Y_{ij}) < \mathrm{var}(Y_{ij})$ if and only if

$$\rho > \frac{1}{2}\frac{\sigma_x}{\sigma_y}.$$

If $\sigma_x \approx \sigma_y$, then we need $\rho \gtrapprox 1/2$ for differencing to improve accuracy. In the hypothetical weight loss example it seems quite plausible that ρ would be large enough to make differences more accurate than using post-treatment weight.

Now let's look at a regression model

$$Y_{ij} = \mu_i + \beta(x_{ij} - \bar{x}_{\bullet\bullet}) + \varepsilon_{ij}, \quad i = 1, 2 \quad j = 1, \ldots, n_i.$$

If we take $\beta = 0$, we get the post-treatment analysis. If we take $\beta = 1$, we get an analysis of differences (in this case future minus past). There is a value of β that would minimize the variance of $Y_{ij} - \beta(x_{ij} - \bar{x}_{\bullet,\bullet})$. If we knew that β we could test $H_0 : \mathbb{E}(Y - \beta x \mid A) = \mathbb{E}(Y - \beta x \mid B)$ which is the same as $H_0 : \mathbb{E}(Y - \beta(x - \bar{x}) \,|\, A) = \mathbb{E}(Y - \beta(x - \bar{x}) \,|\, B)$ so centering does not affect the null hypothesis we are testing. It is also the same as

$$H_0 : \mathbb{E}(Y - \mu - \beta(x - \bar{x}) \,|\, A) = \mathbb{E}(Y - \mu - \beta(x - \bar{x}) \,|\, B).$$

In this version we can think of $\mu + \beta(x - \bar{x})$ as an estimate of where the subject would be post-treatment and then the experiment is comparing the average amount that Y exceeds this baseline, between levels A and B of the treatment. We don't know β, but we can estimate it by least squares, and that's what ANCOVA does.

For the gingivitis data, the estimated treatment difference from ANCOVA was 0.1263 and the$t$ statistic was 3.57 (vs 3.56 for the original data and 2.91 for differences). The standard error of the estimate was 0.0354 vs 0.0446 (for Y) or 0.0374 (for $x - Y$). In this instance the precision was improved by regression on the prior measurement.

8.3 More general regression

The regression does not have to be on a prior version of Y. Montgomery (1997) considers a setting where Y_{ij} is the strength of a cable and x_{ij} is the diameter of the cables prior to treatment. Given multiple variables $\boldsymbol{x}_{ij} \in \mathbb{R}^p$, measured prior to treatment, we could put them all into the ANCOVA. If p becomes large then it is possible that ANCOVA would increase the variance of the adjusted responses $Y_{ij} - (\boldsymbol{x}_{ij} - \boldsymbol{x}_{\bullet\bullet})^{\mathsf{T}}\hat{\beta}$ beyond the variance of Y_{ij} and give a less precise comparison. In some cases, even the reduced degrees of freedom could widen the confidence interval on the treatment difference. Deciding the number and identity of predictors to include is a common tradeoff in statistics.

We can think of ANCOVA as adjusting for pre-treatment differences in much the same way that blocking does. The difference is that blocking is used for categorical prior variables while ANCOVA can handle continuous ones.

Suppose that

$$Y_{ij} = \mu_i + g(\boldsymbol{x}_{ij}) + \varepsilon_{ij}$$

where $g(\cdot)$ is not a linear function. In his causal inference class notes, Wager shows that regressing on prior variables will be asymptotically better than not doing it even if the linear model used is not right. We would like to compare $\mathbb{E}(Y_{ij} - g(\boldsymbol{x}_{ij}) \,|\, A)$ and $\mathbb{E}(Y_{ij} - g(\boldsymbol{x}_{ij}) \,|\, B)$. We get to compare $\mathbb{E}(Y_{ij} - \tilde{g}(\boldsymbol{x}_{ij}) \,|\, A)$ and $\mathbb{E}(Y_{ij} - \tilde{g}(\boldsymbol{x}_{ij}) \,|\, B)$ for some imperfect function $\tilde{g}(\cdot)$, such as a linear model approximation to $g(\cdot)$.

One way to see this is to consider just doing a plain ANOVA without the regression on x_{ij}. That will have validity from the randomization while it also corresponds to taking $\beta = 0$, or $\tilde{g}(\cdot) = 0$, which will ordinarily not be the best regression model to have used.

8.4 Post treatment variables

Suppose that we measure x_{ij} prior to treatment, then apply the treatment and then measure a response Y_{ij} and additional post-treatment variables z_{ij}. We might contemplate the model

$$Y_{ij} = \mu_i + (x_{ij} - \bar{x}_{\bullet\bullet})^{\mathsf{T}}\beta + (z_{ij} - \bar{z}_{\bullet\bullet})^{\mathsf{T}}\beta + \varepsilon_{ij} \qquad \text{(bad idea!)}.$$

This model is a very bad choice if our goal is to test whether the treatment has a causal impact on Y. Suppose that the treatment has a causal impact on Z which then has a causal impact on Y. We might then find that Z explains Y so well that the treatment variable appears to be insignificant. For instance, the causal implications in an agriculture setting might be that a pesticide treatment has a causal impact on the quantity X of insects in an orchard, and that in turn has a causal impact on the size Y of the apple harvest. Including Z in the regression could change the magnitude of the treatment differences and lead us to conclude that the treatment does not causally affect the size of the apple harvest. This could be exactly wrong if the treatment really does increase the apple harvest ***because*** of its impact on the pests.

With data like this we could do two ANCOVAs. One relates Y to the treatment and any pre-treatment predictors. Another relates z to the treatment and any pre-treatment variables. We could then learn whether the treatment affects z and that might be a plausible mechanism for how it affects Y.

If there is a strong predictive value in $\boldsymbol{x}_{ij}^{\mathsf{T}}\beta$ then we get a more precise comparison of treatment groups from the ANCOVA than we would from an ANOVA that ignored $\boldsymbol{x}_{ij}$.

8.5 Crossover trials

Consider two painkillers A and B. In a regular trial, n people get A and n people get B by a random assignment. Then later we compare the two groups of people.

In a cross-over trial, each of the people would get both A and B and we would then compare their experiences with the two medications. The potential advantage is that now the different people are analogous to blocks. If there are strong person to person variations in the response we can balance them out. It is like the running shoe example from before, except that while kids can wear two kinds of running shoes at once, we cannot have people take two medicines at once because we would not know how to separate the effects that they had on the people. For the shoes, the response was the effect that the person had on two separate shoes and that's different.

In the cross-over trial n people get A for a period (say one week) then they wait some times (perhaps also a week) for the effects to wash out and then they get B for the same length of time that they got A. Another n people go through the trial in the opposite order taking B first then waiting for it to wash out and then taking A. We can sketch the layout like this:

	Period 1	Period 2
Subject Group 1	A	B
Subject Group 2	B	A

The complication in cross-over designs is that the treatment in period one might affect the measurement in period two. If we are confident that the washout

period is long enough then we might bet on a cross-over design. Clearly if the first period treatment can bring a permanent cure (or irreversible damage) then the second period is affected and a cross-over is problematic. Cross-over designs are well suited for chronic issues that require continual medication.

Here is a sketch from Brown (1980) of how cross-over data looks.

		Group 1		Group 2
Period	Trt	Subjects $S_{11}, \cdots, S_{1n_1}$	Trt	Subjects $S_{21}, \ldots, S_{2n_2}$
1	A	$Y_{111}, \cdots, Y_{1n_11}$	B	$Y_{211}, \cdots, Y_{2n_21}$
2	B	$Y_{112}, \cdots, Y_{1n_12}$	A	$Y_{221}, \cdots, Y_{2n_22}$

There are two groups of subjects, two treatments and two periods. For instance, in period 1, group 1 gets treatment A and yields observations $Y_{111}, \cdots, Y_{1n_11}$.

Here is his regression model after the carry-over term has been removed.

$$\begin{aligned} Y_{ijk} &= \mu + \pi_k + \tau_{u(i,k)} + \eta_{ij} + \varepsilon_{ijk} \\ i &= \text{group 1 or 2} \qquad j = \text{subject 1 to } n_i \qquad k = \text{period 1 or 2} \\ u &= u(i,k) = \text{treatment A or B} \qquad \text{A iff } i = k \\ \tau_u &= \text{treatment effect} \\ \eta_{ij} &= \text{subject effect.} \end{aligned}$$

Notice the subscript on the treatment effect τ. The treatment level is $u(i,k)$ which is A if it is group $i = 1$ in period $k = 1$ or group $i = 2$ in period $k = 2$. If $i \neq k$ then the treatment is B.

As in ANCOVA, we will consider some differences. Let D_j be the period 2 value minus the period 1 value for subject j. Then

$$\mathbb{E}(D_j) = \begin{cases} \underbrace{\pi_1 - \pi_2}_{\text{period effect}} + \underbrace{\tau_1 - \tau_2}_{\text{treatment effect}} & \text{Group } i = 1 \\ \underbrace{\pi_1 - \pi_2}_{\text{period effect}} + \underbrace{\boldsymbol{\tau_2 - \tau_1}}_{\text{treatment effect}} & \text{Group } i = 2 \end{cases}$$

and so

$$\mathbb{E}(D_j \,|\, i = 1) - \mathbb{E}(D_j \,|\, i = 2) = 2(\tau_1 - \tau_2).$$

That is, the group differences of D_j inform us about treatment difference $\tau_1 - \tau_2$. We can use this to test for treatment differences via a t-statistic

$$t = \frac{\bar{D}_{gp1} - \bar{D}_{gp2}}{s_D} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad s_D^2 = \frac{(n_1 - 1)s_{Dgp1}^2 + (n_2 - 1)s_{Dgp2}^2}{n_1 + n_2 - 2}.$$

Or we could do an unpooled t-test.

Similarly

$$\mathbb{E}(D_j \,|\, i = 1) + \mathbb{E}(D_j \,|\, i = 2) = 2(\pi_1 - \pi_2).$$

Therefore we can test for period differences by summing group differences of D_j.

From the Table that Brown gives we can see that the A/V treatment effect is exactly the same as an interaction between periods and groups. A crossover would be bad choice if we thought such an interaction likely.

Now suppose that we want to test for a carryover effect. We could add a predictor variable to the regression model. We would construct a predictor variable that takes the value 0 for all period one data, takes the value +1 in period 2 if the period 1 treatment was A and itakes the value −1 in period 2 if the period 1 treatment was B. This predictor times its coefficient does nothing in period 1, as appropriate, because there was not prior effect that could carry over to period 1. In period 2 it models twice the difference at period 2 from having period A versus B at period 1.

Some authors advise against crossovers if carryover is at all possible. Some revert to period 1 after testing for carryover and finding that it is present. There is some controversy over this method. The test for carryover may not be reliable enough to use. See Brown (1980). There is a numerical example, also from dentistry, in Brown's paper at `https://www.jstor.org/stable/2530496`.

8.6 More general crossover

Suppose that we have 3 treatments. We could arrange them in a 3 period cross-over as follows.

$$\begin{matrix} Per1 & Per2 & Per3 \end{matrix}$$
$$\begin{bmatrix} A & B & C \\ A & C & B \\ B & A & C \\ B & C & A \\ C & B & A \\ C & A & B \end{bmatrix}$$

This uses 6 groups. Cox (1958) give an example using 3 groups

$$\begin{matrix} Per1 & Per2 & Per3 \end{matrix}$$
$$\begin{bmatrix} B & A & C \\ C & B & A \\ A & C & B \end{bmatrix}$$

This is a Latin square. Each treatment appears once in each of the three periods. It misses one kind of balance that we might want. If we thought that carryover were possible then we might want every treatment to have a balanced set of immediate predecessors. For instance each of AB, AC, BC, BA, CA, CB should appear the same number of times in consecutive time periods. The Latin square above has BA twice but never has BC. The top design has all 6 possible permutations once and so by symmetry it is balanced. For instance A is first twice, follow B twice and follows C twice. Letters B and C are similarly balanced.

If we were interested in carryover effects and wanted to measure them well, we could add more groups. With two treatments, we could have four groups: AB, BA, AA and BB. Or we could have groups like: ABB, ABA, BAB, and BAA.

These and similar ideas are taken up at length in the book Jones and Kenward (2014).

9

Split-plot and nested designs

In this chapter we look at ***split-plot*** designs. The terminology comes from agriculture. One treatment factor might be applied to large plots of land. A second treatment treatment factor is then applied to smaller areas, nested within the plots. The original plots are then split up for the second treatment. The more general term is ***split-unit*** design because the experimental unit doesn't have to be a plot of land. For instance in steel production one factor might be applied to 350 tons of steel while it is being produced and a second factor might be applied to ingots weight 10 kilograms from one production run (called a "heat").

There is a clear economic advantage to split-plot experiments. For instance, it would be extremely expensive to increase the number of heats of steel being made and quite inexpensive to look at a large number of ingots from each heat. Cox (1958, Chapter 7.4) mentions another issue. We might have a limit on the natural size of blocks that stops us from putting all AB combinations into one block. Split-plot designs can fit into the block size more easily.

We will also look at two closely related topics. These are nested ANOVAs and cluster randomized trials.

9.1 Split-plot experiments

In an agriculural setting, one might drive a tractor for one kilometer while applying factor A (e.g., fertilizer) to plots of land. Then factor B (e.g., seed type) could be applied to smaller units within that kilometer, called sub-plots.

We do this when factor A is expensive to change in time or money, while factor B is inexpensive to change. The plots for factor A serve almost exactly like blocks for factor B. They're somewhat different from the usual blocks because

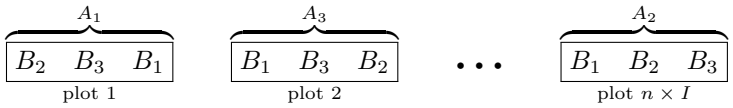
we want have purposely made them differ and want to study them in their own right.

When we do computer experiments or Monte Carlo simulations, it often makes sense to analyze them as designed experiments. If one factor A is set at the beginning of an hour long computation and then a second factor B can be set when there are just two minutes left in the computation, then a split-plot design makes sense. We choose A, compute for 58 minutes, save the internal state of the computation, and then vary factor B several times.

A split-plot design will ordinarily give us better comparisons for the inner factor B because it is blocked by the outer factor and the outer factor is not blocked. This is perhaps not surprising. If it is so much cheaper to vary factor B then it is expected that we can study it with more precision.

To begin with, we will suppose that both A and B are fixed effects. We consider random effects and mixed effects later.

Let's vary factor A at $I \geqslant 2$ levels. We will have n plots at each of those levels for a total of $n \times I$ plots. We depict them as follows



Here, factor A varies at the level of whole plots. Each level i appears n times. Factor B varies at the level of sub-plots: $j = 1, \ldots, J$, with $J = 3$ in the diagram. Each level j appears nI times

The n appearances of each level of A could be from n replicates or from a completely randomized design where I treatments were applied n times each in random order to $n \times I$ plots. When replicates are used it is usual to include an additive shift for them in the model.

We can compare levels of B using 'within-plot' differences such as $Y_{ijk} - Y_{ij'k}$ for levels $j \neq j'$ of factor B. We can compare levels of A using 'between-plot' differences such as $\bar{Y}_{i\bullet k} - \bar{Y}_{i'\bullet k}$. We expect between-plot differences to be less informative when the plots vary a lot.

The AB interaction is estimated with 'within-plot' differences. More precisely, it uses between plot differences of within plot differences that are also within plot differences of between plot differences:

$$\begin{aligned}
\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet} &= \bigl(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet}\bigr) - \bigl(\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}\bigr)\\
\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} &= \frac{1}{J}\sum_{j'=1}^{J} \bar{Y}_{ij\bullet} - \bar{Y}_{ij'\bullet} \qquad \text{within}\\
\bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet} &= \frac{1}{I}\sum_{i=1}^{I} \bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} \qquad \text{also within.}
\end{aligned}$$

We see from the last line that the interaction is an average of within-plot differences and so it gets within-plot accuracy.

The ANOVA for Factor A is based on $\bar{Y}_{i\bullet k}$ for $i = 1, \ldots, I$ and $k = 1, \ldots, n$. If we the plots are in a randomized block design then our analysis uses those $I \times n$ numbers in a table like the following:

Source	df
Replications	$n-1$
A	$I-1$
Whole plot error	$(I-1)(n-1)$
Total	$In-1$

If the plots are not in n blocks of I units then we use

Source	df
A	$I-1$
Whole plot error	$I(n-1)$
Total	$In-1$

Let's prefer the blocked analysis. Then the sub-plot ANOVA table is

Source	df	
B	$J-1$	
AB	$(I-1)(J-1)$	
"Sub-plot error"	$I(J-1)(n-1)$	by subtraction
Total	$I(J-1)n$	by subtraction

The subtraction to get the subplot degrees of freedom is

$$(IJn-1) - \underbrace{In-1}_{\text{whole plots}} - (J-1) - (I-1)(J-1) = I(J-1)(n-1).$$

It is the same df as for I replicates of a $J \times n$ experiment. The subtraction to get the total degrees of freedom is

$$(IJn-1) - (In-1) = I(J-1)n.$$

We still have to figure out the sums of squares that go in these tables. We could do a full $I \times J \times n$ ANOVA with replicates $k = 1, \ldots, n$ treated as a third factor C crossed with factors A and B. Then the subplot error sum of squares is $\mathrm{SS}_{ABC} + \mathrm{SS}_{BC}$. The whole plot error sum of squares is SS_{AC}. The replicates sum of squares in the whole plot analysis is SS_C. The sums of squares for A, B and AB are, unsurprisingly, SS_A, SS_B and SS_{AB}.

There is a different analysis in Montgomery (1997, Chapter 12-4). In our notation, his model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + c_k + (\beta c)_{jk} + (\alpha\beta c)_{ijk} + \varepsilon_{ijk},$$

where γ_k is the variable for block $k = 1, \ldots, n$. That analysis treats blocks as random effects and also allows them to have interactions with the fixed effects. Most other authors choose models in which blocks simply do not interact with

other effects, for better or worse. There would be no way to estimate $\mathbb{E}(\varepsilon_{ijk}^2)$ separately from SS_{ABC} in that model, without another level of replication. We will stay with the analysis based on the two tables described above. However, if you are in a setting where you suspect that there could be meaningful interactions between blocks and treatments, then Montgomery's approach provides a way forward.

Yandell (1997, Chapter 23) is a whole chapter on split-plot designs mostly motivated by agriculture.

9.2 More about split-plots

Jones and Nachtsheim (2009) focus on split-plot designs in industry. They consider unsuspected split-plots and quote Cuthbert Daniel as saying that most or all industrial experiments are split-plots. They raise an issue of unrecognized split-plot experiments. Suppose that an experiment varies oven temperature A, position in the oven B and recipe C. We think of an $A \times B \times C$ experiment.

Suppose that temperature is at three levels and is done on a random order schedule

350 400 375 375 350 400 $\cdots$

The question that arises is whether after run three they turned the oven off and back on again or just kept it going at 375 degrees. If the other temperature changes involved resetting the oven temperature but this one didn't, then the experiment may really be partly of split plot type, with the two consecutive runs at 375 degrees being a double-size whole plot. It is also possible that the operators might undo the randomization into:

350 350 400 400 375 375 $\cdots$

to save time and expense which would be a genuine split-plot. Jones and Nachtsheim (2009) advocate taking care to make sure that the analysis matches how the experiment was done.

In a ***split-split-plot*** experiment, the sub-plots are split further into sub-sub-plots for a third treatment factor C. The analysis involves three tables, one at the plot level, one at the sub-plot level and a new third one at the sub-sub-plot level.

In a ***strip-plot*** experiment, there are two whole plot factors that when crossed define the whole plots. The name comes from agriculture. A tractor might drive North to South placing 8 different kinds of fertilizer on the ground. Later that year a crop-dusting plain might fly East to West trajectories over the farm land, spraying 12 different pesticide treatments. That crossed structure generates 96 different whole plots. Each of those whole plots can then be divided into 4 subplots for four kinds of brocolli. Yandell (1997, Chapter 24.2) discusses strip-plot experiments.

9.3 Nested ANOVAs

Sometimes the levels of B are only defined and make sense with respect to a specific setting of factor A. Kirk (2013) describes an experiment where rats were exposed to ionized air. There were four animals per cage. There were eight cages. Each cage and the animals in it got exposed to either positive or negative ionization. They then studied a measure of the animals' activity level. We can sketch the setup as follows:

Cage	1	2	3	4	5	6	7	8
+ve:	••••	••••	••••	••••				
−ve:					••••	••••	••••	••••

We could also depict it this way:

Cage	1	2	3	4
+ve:	••••	••••	••••	••••
−ve:	••••	••••	••••	••••

However, in this diagram there is no meaningful connection between cage 1 at positive ionization and cage 1 at negative ionization. If we treat 'cage' as a factor then its meaning is dependent on the ionization level. It would not make sense to make comparisons between cage numbers 1, 2, 3 and 4 in general.

In the setting above we say that the factor 'cage' is nested within the factor 'ionization'. If we were studying schools we might label first grade classrooms with numbers 1 through 4 but classroom would ordinarily be nested within school. Schools can be nested within school boards and those can be nested within counties within states. Nesting is a hierarchical relationship often drawn using branching tree diagrams instead of a grid of boxes formed by horizontal and vertical lines as we have for crossed effects.

When B is nested within A, we write $B(A)$. This is a little odd because we just put A inside the parentheses. But since we read left to right, "B nested within A" gets the label $B(A)$.

Whether something is nested or crossed can become subtle. Ingots might ordinarily be nested within heats. For instance, we might randomly select three ingots from each heat to study. Then they are clearly nested. Then again somebody might always take the first and last and middle ones from a subsequent step in the production line. In that case the first one out of one heat does have a meaningful connection with the first one out of another heat and we have a crossed structure.

Suppose we get Y_{ijk} for animal k in cage j that gets treatment i. Then for $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J$ and $1 \leqslant k \leqslant n$ our model has

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \qquad \text{or } \varepsilon_{k(ij)}.$$

For identifiability we set $\sum_{i=1}^{I} \alpha_i = 0$ and $\sum_{j=1}^{J} \beta_{j(i)} = 0$ for each $i = 1, \ldots, I$. We do not require $\beta_{j(i)}$ to sum to zero over i for any j because that would be a

sum of values that had no meaningful connection. The ANOVA decomposition for this setting is

$$\begin{aligned}
\mathrm{SS}_{E(A,B)} &= \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\bullet})^2 \\
\mathrm{SS}_{B(A)} &= \sum_{ijk}(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet})^2 = n\sum_{ij}(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet})^2 \\
\mathrm{SS}_{A} &= \sum_{ijk}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2 = Jn\sum_{i}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2, \quad \text{and} \\
\mathrm{SST} &= \sum_{ijk}(Y_{ijk} - \bar{Y}_{\bullet\bullet\bullet})^2 = \mathrm{SS}_{E(A,B)} + \mathrm{SS}_{B(A)} + \mathrm{SS}_{A}.
\end{aligned}$$

The new quantity is

$$\mathrm{SS}_{B(A)} = n\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet})^2.$$

For each level i, it has $J-1$ degrees of freedom and so it has $I(J-1)$ degrees of freedom in total.

Recall that the AB interaction has $(I-1)(J-1)$ df. This $B(A)$ sum of squares gets $I(J-1)-(I-1)(J-1) = J-1$ more df. They are the df for the B main effect. The B main effect is meaningless when $j=1$ has no persistent meaning as i varies. As a result we lump the B main effect in with the prior AB interaction to get $\mathrm{SS}_{B(A)} = \mathrm{SS}_B + \mathrm{SS}_{AB}$.

9.4 Expected mean squares and random effects

Now let's consider a model with a random effect B nested within another random effect A. The model has

$$Y_{ijk} = \mu + a_i + b_{j(i)} + \varepsilon_{ijk}$$

where $a_i \overset{\text{iid}}{\sim} (0, \sigma_A^2)$, independently of $b_{j(i)} \overset{\text{iid}}{\sim} (0, \sigma_B^2)$, and $\varepsilon_{ijk} \overset{\text{iid}}{\sim} (0, \sigma^2)$. When we say something has distribution (μ, σ^2) it is like $\mathcal{N}(\mu, \sigma^2)$ but without assuming normality. Derivations of F distributions require a normal distribution but expected mean squares do not. The expected mean squares in this setting are

$$\begin{aligned}
\mathbb{E}(\mathrm{MS}_A) &= \sigma^2 + n\sigma^2_{B(A)} + nJ\sigma^2_A \\
\mathbb{E}(\mathrm{MS}_{B(A)}) &= \sigma^2 + n\sigma^2_{B(A)}, \quad \text{and} \\
\mathbb{E}(\mathrm{MS}_E) &= \sigma^2.
\end{aligned}$$

Let's derive the first one. We start with $\mathrm{SS}_A = nJ\sum_{i=1}^{I}(\bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2$. So it is just nJ times a sample variance among

$$\bar{Y}_{i\bullet\bullet} = \mu + a_i + \bar{b}_{\bullet(i)} + \bar{\varepsilon}_{i\bullet\bullet} \sim \Big(\mu, \sigma_A^2 + \frac{\sigma^2_{B(A)}}{J} + \frac{\sigma^2}{nJ}\Big).$$

We know from formulas for a sample variance that

$$\mathbb{E}(\mathrm{SS}_A) = nJ(I-1)\Big(\sigma_A^2 + \frac{\sigma^2_{B(A)}}{J} + \frac{\sigma^2}{nJ}\Big)$$

and so

$$\mathbb{E}(\mathrm{MS}_A) = nJ\Big(\sigma_A^2 + \frac{\sigma^2_{B(A)}}{J} + \frac{\sigma^2}{nJ}\Big) = \sigma^2 + n\sigma^2_{B(A)} + nJ\sigma_A^2.$$

The others are similar.

If B is a random effect nested in a fixed effect A, then

$$\begin{aligned}
\mathbb{E}(\mathrm{MS}_A) &= \sigma^2 + n\sigma^2_{B(A)} + \frac{nJ\sum_{i=1}^I \alpha_i^2}{I-1}\\
\mathbb{E}(\mathrm{MS}_{B(A)}) &= \sigma^2 + n\sigma^2_{B(A)}\\
\mathbb{E}(\mathrm{MS}_E) &= \sigma^2
\end{aligned}$$

and we see that σ_A^2 is replaced by a sample variance among the α_i. If both A and B are fixed, then

$$\begin{aligned}
\mathbb{E}(\mathrm{MS}_A) &= \sigma^2 + \frac{nJ\sum_{i=1}^I \alpha_i^2}{I-1}\\
\mathbb{E}(\mathrm{MS}_{B(A)}) &= \sigma^2 + n\frac{\sum_{i=1}^I\sum_{j=1}^J \beta^2_{j(i)}}{I(J-1)}, \quad \text{and}\\
\mathbb{E}(\mathrm{MS}_E) &= \sigma^2.
\end{aligned}$$

9.5 Additional models

There are more nesting patterns. We might have A, $B(A)$ and C nested within B within A, i.e., $C(B(A))$. Or we could have B and C and $B \times C$ nested within A. Or we could have A crossed with B while C is nested within A. For instance hospital A crossed with drug B and ward C nested within A. These designs can be completely randomized or arranged in randomized block structures. There is a comprehensive treatment of these situations in Kirk (2013, Chapter 11). The online version has 71 pages with worked examples and formulas for expected mean squares.

It is striking how complicated an analysis can become based on combinations of a small number of choices based on how things are nested or crossed and the way the blocks are structured. We are faced with a small but combinatorial explosion of cases. It would be very useful to have a tool such as a probabiliistic programming language that lets a user describe how the data were organized and then sets up the analysis.

9.6 Cluster randomized trials

In cluster randomized trials, we might apply a treatment at random to a whole village or a school or a sports team or a county or a marketing region, like the Bay Area versus Pheonix or Chicago. The experimental unit is a cluster of one of those types. We might also be able to get data on individuals within the cluster. Perhaps people or, in a marketing context, stores.

Let the data be Y_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, n_i$. Suppose that cluster i got treatment $\mathrm{trt}(i) \in 1, \dots, I$ where ordinarily I is much less than n. We can model the individual data via

$$Y_{ij} = \mu + \alpha_{\mathrm{trt}(i)} + \varepsilon_{ij},$$

and we can model the cluster data via

$$\bar{Y}_{i\bullet} = \mu + \alpha_{\mathrm{trt}(i)} + \bar{\varepsilon}_{i\bullet}, \quad i = 1, \dots, n.$$

The most straitforward analysis is to model the cluster level data using either a randomization (permutation) approach or a one way ANOVA. It seems like a shame to greatly reduce the sample size from $N = \sum_{i=1}^{n} n_i$ individuals to just $n \ll N$ clusters. It is then tempting to analyze the data on the individual level. It would however be quite wrong to analyze the individual data as if they were N independent measurements. There are instead inter-cluster correlations. An individual level analysis can be quite unreliable if it considers the individuals to be independent when they are in fact correlated. Suppose for instance that ε_{ij} have variance σ^2, correlations ρ within a cluster and are independent between clusters. Then

$$\begin{aligned}
\mathrm{var}(\bar{Y}_{i\bullet}) = \mathrm{var}(\bar{\varepsilon}_{i\bullet}) &= \mathrm{var}\Big(\frac{1}{n_i}\sum_{j=1}^{n_i}\varepsilon_{ij}\Big) \\
&= \frac{1}{n_i^2}\Big(n_i\sigma^2 + n_i(n_i-1)\rho\sigma^2\Big) \\
&= \frac{\sigma^2}{n_i}\big(1 + (n_i-1)\rho\big).
\end{aligned}$$

It is larger by a factor $1 + (n_i - 1)\rho$ than it would be under independence. This factor is called the ***design effect***.

If we knew ρ then we could consider an individual level analysis. For two treatments we could work out

$$\mathrm{var}\big(\mathrm{avg}(Y_{ij} \,|\, A) - \mathrm{avg}(Y_{ij} \,|\, B)\big)$$

in terms of ρ and σ^2 and all the n_i. Murray (1998) is an entire book on cluster randomized trials also known as ***group randomized trials***.

10

Taguchi methods

Genichi Taguchi developed a methodology also called robust design. It is used in manufacturing to drive variance out of a product. Product quality can be much more sensitive to variance than the mean. A table whose legs are all 3mm short is better than one with just one leg 3mm short, because that one will wobble.

If a product has hundreds or thousands of measured attributes, then variance in their levels can cause unpredictable problems. A door that is slightly too big combined while the frame it has to fit into is slightly too small, becomes a severe problem. Quality is not well described by a rectangular region in the space of attributes.

Our prior chapters were mostly about "getting more". By modeling $\mathbb{E}(Y \mid x)$ we could learn how to get more potatoes or survival or speed or learning by choosing x well. In robust design, we seek x to get a lower value of $\mathrm{var}(Y \mid x)$.

The reason to reduce variance is that we want to get closer to a target. Methods that reduce variance could move the mean response away from the target. In the robust design world there is usually some way to adjust the mean onto target after the variance has been reduced. They call it a ***signal factor***. This is not something that we would anticipate just thinking in terms of distributions and regression models and expected mean squares. It may however be quite obvious to people working on a product. For instance, if the quantity of interest is the thickness of paint on a car, then that amount will be more or less directly prportional to the amount of time that the paint is being sprayed and so the mean thickness can be controlled that way after designing a process to reduce variance. That only really works if adjusting the signal factor does not more than undo the variance reduction we have obtained.

10.1 Context and philosophy

The topic of robust design and Taguchi methods is not purely mathematical or statistical. There are elements of philosophy and it takes advantages of empirically observed phenomena common to many physical systems.

The history of the problem is also an important component. In the 1980s there was significant concern in the US about a loss of quality in manufactured products, especially compared to high quality output from Japan. Methods developed by Taguchi in Japan were introduced to the US, especially at AT&T. This was to some extent returning a favor of Deming who had earlier worked to help Japanese manufacturers improve their quality, some decades earlier.

One of the ideas in robust design is to consider a quadratic loss function instead of specification limits. If the ideal value of a variable Y is a target T then instead of keeping score by the fraction of times that $|Y - T| \leqslant \delta$ or more generally that $L \leqslant T \leqslant U$, we could study instead $k(Y - T)^2$ for some value $k > 0$.

This idea is often accompanied by the experience of Sony making television sets in both the US and Japan. It is described in Phadke (1989) who references the newspaper 'The Asahi' from April 17, 1979. The color density of the US made sets were all within the specification limits but were widely scattered in the interval from L to U. For the Japanese sets, 0.3% were outside the interval $[L, U]$ but the distribution was more sharply peaked around the target level T. Phadke (1989) reports that US consumers thought the Japanese made sets had higher quality. (Somehow I remember the story being lower return rates for Japanese-made sets. That would be harder data than just a survey of perceptions, but I'm unable to find that aspect of the story in print.)

A quadratic loss function has several advantages. It allows one to keep track of continued progress beyond the point where 100% of products are meeting a given set of specs, without having to keep 'moving the goalposts'.

Another reason to avoid using a binary spec to track quality is that those specs may well have been set by some arbitrary tradeoffs between the interests of producers and consumers of a product along with its costs. A spec like $T \pm \delta$ might not really be from a law of nature, but some negotiation based on what is possible or reasonable. Then getting zero defects does not imply perfect quality. A rectangular spec on many features, such as $|Y_j - T_j| \leqslant \delta_j$ for all $j = 1, \ldots, J$ is even less likely to be physically appropriate.

Some sort of binary rule is necessary if one has to accept some product and reject others. However continuous data are also more informative than binary data when it comes to understanding and improving a system. Even with a spec such as 0.013 ± 0.001 getting values $(0.01301, 0.01303, 0.01298, 0.01302)$ inspires more confidence than just getting 'OK' four times out of four tries.

The quadratic loss is sometimes motivated as the total cost to society of a product missing its target value. It is difficult to judge the whole cost to society of a product, but perhaps reasonable that a quadratic formula is more useful than a binary one.

For some quantities Y that cannot be negative this smaller they are the

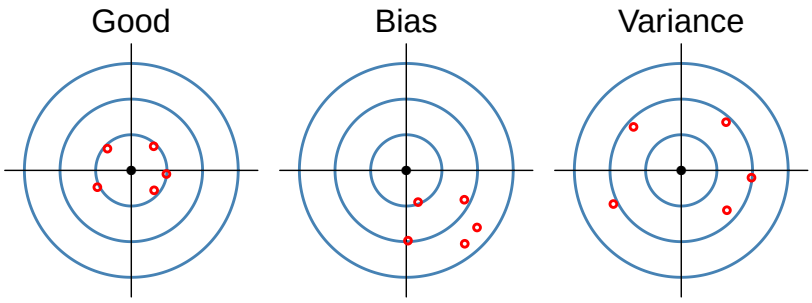


Figure 10.1: An archery example to illustrate bias and variance. The first archer is quite good, or at least best. The second has higher bias. The third has higher variance.

better they are. These then have a target of $T = 0$ and are scored by kY^2. For other quantities, the larger they are, the better, and those are scored via kY^{-2}. That is smaller values of $1/Y$ are better.

10.2 Bias and variance

A great convenience of a quadratic loss function is that it splits into variance and bias squared. Given n observations we have

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - T)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + (\bar{Y} - T)^2.$$

The first, variance term, is generally harder to control. A common analogy is made to a coach in archery as shown in Figure 10.1. The coach can might be able to help the second archer with one or two suggestions about where to aim. The third archer might have issues that involve many variables, just like the noise in a model is often the result of many uncontrolled variables.

Whether the archery model is appropriate to manufacturing is not something we can prove philosophically or mathematically. It might be known to hold empirically in some problems. Then the strategy in robust design is to first find a way to reduce the variance. Then later we fix any bias created by the first step assumed to be the hard one (without causing the variance to go back up).

One model to explain robust design has

$$Y = f(x_1, \ldots, x_k,\ z_1, \ldots, z_n) + \varepsilon = f(\boldsymbol{x}, \boldsymbol{z}) + \varepsilon$$

where $\boldsymbol{x} = (x_1, \ldots, x_k)$ is a vector of variables that we can control and $\boldsymbol{z} = (z_1, \ldots, z_r)$ is a vector of ***noise variables*** that we cannot control. At first this sounds like it will be tough to experiment on variables we cannot control. What is happening is that those are variables we can control in an experiment but

not afterwards. In an example mentioned by Shin Taguchi in a round table discussion (Nair et al., 1992) a company making a paper feeder has noise variables including "paper type, paper size, paper warp, paper surface, paper alignment, stack height, roller wear, and humidity." They can vary all of those things in their experiment to understand how they affect paper jams, but they cannot control them in their customers' uses once the product has been sold. Variables $\boldsymbol{x}$ are different. They might describe how the paper feeder was constructed. In an automotive application, $\boldsymbol{z}$ might include the weather that a car will be driven in.

In robust design, we get to choose $\boldsymbol{x}$ but our product has to work very generally for lots of $\boldsymbol{z}$. We might formulate the problem as

$$\min_{\boldsymbol{x}} \operatorname{var}(Y \mid \boldsymbol{x})$$

where the randomness in Y is from an assumed distribution on $\boldsymbol{z}$ and from ε. Then change some ***signal variable*** to adjust $\mathbb{E}(Y \mid \boldsymbol{x}) = T$. That variable cannot be one of the z_j. It must be one of the x_j or some other variable either implicitly or explicitly in $f(\cdot)$. A good example is picking $\boldsymbol{x}$ to minimize variance of the paint thickness on a car. Then if the mean is off target, adjust the spray time.

There is a very famous example about the Ina Tile company. See Phadke (1989). Their tiles came out of the kiln in unequal sizes and quality. One of the noise factors was the position of a tile in the kiln. The tiles were packed within the kiln. No matter how you do that some tiles will be central and some at the edge. So that position z could be a noise factor that you don't want to affect product quality. After doing a robust design experiment they found that they were able to reduce the variability of their tiles. The solution even involved using less of an expensive ingredient and more of an inexpensive one and so the result was greater quality at lower cost.

10.3 Taylor expansion

Suppose that we set targets $\boldsymbol{x}_0$ for $\boldsymbol{x}$ and $\boldsymbol{z}_0$ for $\boldsymbol{z}$. Under random assignment $\mathbb{E}(\boldsymbol{z}) = \mu$ which is not necessarily $\boldsymbol{z}_0$ and $\operatorname{var}(\boldsymbol{z}) = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2)$. Here we are making a great simplification that components z_j are uncorrelated. We could weaken that. With this assumption a Taylor expansion gives

$$f(\boldsymbol{x}_0, \boldsymbol{z}) \doteq f(\boldsymbol{x}_0, \boldsymbol{z}_0) + \sum_{j=1}^{n} \frac{\partial}{\partial z_j} f(\boldsymbol{x}_0, \boldsymbol{z}_0)(z_j - z_{0j})$$

and then

$$\operatorname{var}(f(\boldsymbol{x}_0, \boldsymbol{z})) \approx \sum_{j=1}^{n} \Big(\frac{\partial}{\partial z_j} f \Big)^2 \sigma_j^2.$$

Since $\boldsymbol{z}$ and hence σ_j^2 is out of our control, our best chance to reduce this variance is to reduce $(\partial f/\partial z_j)^2$ through our choice of $\boldsymbol{x}_0$. That is, we want to reduce the sensitivity of our output to fluctuations in the input.

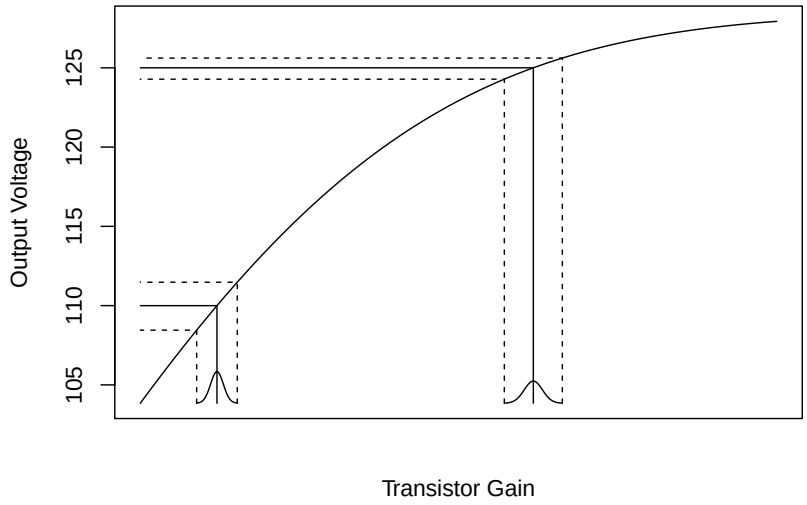


Figure 10.2: Hypothetical curve modeled after the account in Phadke (1989) of power supply design.

Phadke (1989) has an example where z is gain of a transistor Y is output voltage of a power supply. Figure 10.2 illustrates his setting. The horizontal axis depicts a transistor gain quantity z (without units). There is a curve that translates this gain into output voltage $Y = f(z)$. There is a value z_{110} that gives $Y = f(z_{110}) = 110$, the desired value, but in that region z has some noise that gets amplified by the steep slope $f'(z_{110})$ to give a very noisy voltage. Moving to a higher value of z, such as z_{125} with $f(z_{125}) = 125$ leads to less noisy Y. This happens despite increased variance in z there because $f'(z_{125})$ is much smaller than $f'(z_{110})$. In order to bring the system on target he uses a resistor that reduces the voltage from a value centered around 125 to one centered around the target $T = 110$. The impact of this signal factor does not change variance due to linearity of the voltage versus resistance relationship.

10.4 Inner and outer arrays

Taguchi's strategy is to choose an experiment varying $\boldsymbol{x}$ in n_1 runs. This is called the ***inner experiment***. For each of those n_1 runs he has a second experiment in the noise factors $\boldsymbol{z}$ using n_2 runs. That is the ***outer experiment***. The total experiment then has $n_1 \times n_2$ runs, yielding y_{ij} for $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$.

At each level of the inner experiment, the method computes

$$\eta_i = -10 \log_{10}\Big(\frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_{ij} - \bar{y}_{i\bullet})^2\Big).$$

This is the sample variance over the outer experiment recorded in decibels. Recall that $10\log_{10}(\cdot)$ converts a quantity to decibels. A logarithmic scale is convenient because any model we get for $\mathbb{E}(\eta)$ when exponentiated to give a variance (or inverse of a variance) will never give a negative value. It is also more plausible that the factors in $\boldsymbol{x}$ might have mutiplicative effects on this variance than an additive effect.

In its most basic form Taguchi's method finds settings that maximize the signal to noise ratio η. It is often an additive model in the components of $\boldsymbol{x}$. In the "bigger the better" setting, the analysis is of

$$\eta_i = -10\log_{10}\Big(\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{1}{y_{ij}}\Big).$$

while in the "bigger the better" setting, the analysis is of

$$\eta_i = -10\log_{10}\Big(\frac{1}{n_2}\sum_{j=1}^{n_2} y_{ij}\Big).$$

The experimental designs are usually orthgonal arrays like the following one at 3 levels:

$$\begin{array}{cccc}
0 & 0 & 0 & 0\\
0 & 1 & 1 & 2\\
0 & 2 & 2 & 1\\
1 & 0 & 1 & 1\\
1 & 1 & 2 & 0\\
1 & 2 & 0 & 2\\
2 & 0 & 2 & 2\\
2 & 1 & 0 & 1\\
2 & 2 & 1 & 0
\end{array}$$

This design is called an orthogonal array because in any pair of columns each of the 9 possible combinations appears the same number of times, i.e., once. We will see more orthogonal arrays when we look at computer experiments. We recognize this as a 3^{4-2} design because it handles 4 variables at 3 levels each using only 9 runs not 81.

It is usual for the outer experiment to also be an orthogonal array. Then the design is made up of an ***inner array*** and an ***outer array***.

There is a good worked example of robust design in Byrne and Taguchi (1987) (which seems not to be available online). The value Y was the force needed to pull a nylon tube off of a connector. This pull-off force was studied with an inner array varying 4 quantities at three levels each (interference, wall thickness, insertion depth and % adhesive). The outer array was a 2^3 experiment in the time, temperature and relative humidity during conditioning. There were thus 72 runs in all.

10.5 Controversy

There were some quite heated discussions about which aspects of Taguchi's robust design methods were new and which were good. The round table discussion in Nair et al. (1992) includes many points of view and dozens of references.

One issue is whether it is a good idea to use that combination of inner and outer arrays. There every value of $\boldsymbol{x}$ is paired with every value of $\boldsymbol{z}$. It might be less expensive to run a joint factorial experiment on $(\boldsymbol{x}, \boldsymbol{z}) \in \mathbb{R}^{k+n}$ instead. As usual, costs come into it. If the cost is dominated by the number of unique $\boldsymbol{x}$ runs made, then a split plot structure like Taguchi uses is quite efficient. If changing $\boldsymbol{z}$ and changing $\boldsymbol{x}$ cost the same then a joint factorial experiment and the corresponding analysis will be more efficient.

Another issue is whether an analysis of signal to noise ratios is the best way to solve the problem. In the roundtable discussion James Lucas says "The designs that Taguchi recommends have the two most important characteristics of experimental designs: (1) They have factorial structure, and (2) they get run." That is, an analysis that is not optimal may be better because more people can understand it and use it.

11

Some data analysis

Most of these notes are about designing how to experimentally gather data with the assumption that they can be largely analyzed with methods familiar from linear regression. Here we look at some ways of analyzing data that are especially suited to designed experiments.

The course begin by grounding experimentation in causal inference. The notions of potential outcomes, randomization, the SUTVA assumption and external validity help us think about experimentation. Then A/B testing and bandit methods bridge us to problems of great current interest in industry. Then we began with more classical experimental design.

The chapters so far have included a number of categorical quantities. There are experimental units which may be plots or subjects and there are sub-units. There are experimental factors. A combination of factors comprises a treatment which may or may not involve important interactions. Those factors can be fixed or random, nested or crossed. We also saw blocks and replicates and repeated measures.

11.1 Contrasts

Beyond just rejecting H_0 or not rejecting it, we have an interest in the different expected values of Y. For a one way fixed effects model with

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i \equiv \mu_i$$

the comparisons of interest involve certain differences among the μ_i or α_i. We might want to compare two expected outcomes through $\mu_2 - \mu_7$. If we are compariing effectiveness of five soaps where the first three contain phosphates

and the other two do not then we might be interested in

$$(\mu_1 + \mu_2 + \mu_3)/3 - (\mu_4 + \mu_5)/2.$$

If we are comparing a new product to three old ones we might study

$$\mu_1 - (\mu_2 + \mu_3 + \mu_4)/3.$$

These are all examples of ***contrasts***. Contrasts take the form $\sum_{i=1}^I \lambda_i \mu_i$ where $\sum_i \lambda_i = 0$ and, to remove an uninteresting case, $\sum_i \lambda_i^2 > 0$. A contrast also satisfies $\sum_{i=1}^I \lambda_i \alpha_i$. The reason why we have so much less interest in μ than α_i is that μ does not affect any comparisons of the levels of this factor and so does not affect many of our choices. Perhaps if μ is bad enough we might not want to use any of the levels of our factor, but when as usual we have to choose, μ plays no role in $\mathbb{E}(Y)$.

In the one way layout we can test a contrast with a t test, via

$$t = \frac{\sum_i \lambda_i \bar{Y}_{i\bullet}}{s\sqrt{\sum_i \lambda_i^2/n}} \sim t_{(N-k)} \qquad s = \sqrt{\mathrm{MSE}} \qquad N = \sum_{i=1}^I n_i.$$

We saw earlier that the presence of a random effect can complicate the inference on a fixed effect with which it is crossed. If A is fixed and B is random we can use

$$t = \frac{\sum_i \lambda_i \bar{Y}_{i\bullet\bullet}}{s\sqrt{\sum_i \lambda_i^2/(nB)}} \sim t_{((I-1)(J-1))} \qquad s = \sqrt{\mathrm{MSAB}}.$$

This formula is for a balanced setting. When MSAB is the appropriate denominator for our F test it provides the appropriate value of s for our t-test. The degrees of freedom to use are the number underlying the estimate s.

Suppose that we have a factor that represents I equispaced levels of a continuous variable. For instance 20kg, 40kg, 60kg and 80kg of fertilizer. It is then interesting to test the extent of a linear trend in the average responses Y. Centering these levels produces a contrast $\lambda = (-30, -10, 10, 30)$. A test of $\sum_i \lambda_i \alpha_i = 0$ is equivalent to one with $\lambda = (-3, -1, 1, 3)$. This is a ***linear contrast***. When there are an odd number of levels then the central element in the contrast has $\lambda_i = 0$. A test for curvature can be based on a quadratic contrast. If the levels are linearly related to i then we can take

$$\lambda_i = (i - \bar{i})^2 - \frac{1}{I}\sum_i (i - \bar{i})^2$$

where $\bar{i} = (I+1)/2$.

Two contrasts λ and λ' are orthogonal if $\sum_i \lambda_i \lambda_i' = 0$. Then $\sum_i \lambda_i \bar{Y}_{i\bullet}$ and $\sum_i \lambda_i' \bar{Y}_{i\bullet}$ are uncorrelated.

11.2 Normality assumption

We have used reference distributions like the t and F distributions derived from an assumption that the errors are normally distributed.

By the central limit theorem, $\bar{Y}_{i\bullet}$ is more nearly normally distributed than the Y_{ij} are. Similarly $\sum_i \lambda_i \bar{Y}_{i\bullet}$ involves more averaging than the individual $\bar{Y}_{i\bullet}$ do and so we anticipate that they more nearly follow a normal distribution. There is a special aspect of contrasts that helps. The main departure of an average from the normal distribution is typically from it's skewness $\mathbb{E}((Y-\mu)^3)/\sigma^3$ being nonzero. If $\bar{Y}_{1\bullet}$ and $\bar{Y}_{2\bullet}$ have similar skewness then much of it cancels in a difference like $\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$. Since λ_i sum to zero, contrasts must also introduce some degree of cancellation in skewnesses.

Our t-test for a contrast is based on the approximation

$$\sum_i \lambda_i \bar{Y}_{i\bullet} \approx \mathcal{N}\Big(\sum_i \lambda_i \alpha_i, \sum_i \lambda_i^2 \frac{\sigma^2}{n}\Big)$$

or, for unbalanced samples,

$$\sum_i \lambda_i \bar{Y}_{i\bullet} \approx \mathcal{N}\Big(\sum_i \lambda_i \alpha_i, \sum_i \lambda_i^2 \frac{\sigma^2}{n_i}\Big).$$

The F test for H_0 is similarly robust to small departures from normality by the CLT because

$$\begin{pmatrix} \bar{Y}_{1\bullet} \\ \bar{Y}_{2\bullet} \\ \vdots \\ \bar{Y}_{I\bullet} \end{pmatrix} \doteq \mathcal{N}\left(\begin{pmatrix} \mu+\alpha_1 \\ \mu+\alpha_2 \\ \vdots \\ \mu+\alpha_I \end{pmatrix}, \operatorname{diag}\Big(\frac{\sigma^2}{n}, \frac{\sigma^2}{n}, \cdots, \frac{\sigma^2}{n}\Big) \right)$$

This is all we need for our usual derivation. The central limit theorem yields approximately Gaussian $\bar{Y}_{i\bullet}$ values and then sums of squares among them are approximately χ^2. The denominator in the F test uses a mean square such as MSE or MSAB as an estimate of σ^2. It commonly has many more degrees of freedom than the numerator. We do not need a central limit theorem for the denominator just that it yields a good approximation to σ^2.

Tests for a variance or ratio of variances are not robust to non-Gaussianity. A typical MSE is like

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Under Gaussianity $s^2 \sim \sigma^2 \chi^2_{(n-1)}/(n-1)$. More generally

$$\operatorname{var}(s^2) = \Big(\frac{2}{n-1} + \frac{\kappa}{n}\Big)\sigma^4$$

(Miller, 1997, Chapter 7) where

$$\kappa = \frac{\mathbb{E}((Y-\mu)^4)}{\sigma^4} - 3$$

is the kurtosis of Y. The kurtosis is zero for Gaussian random variables but not necessarily for other variables. When Y has 'heavier tails' than the Gaussian distribution has, then $\kappa > 0$ and s^2 has higher variance than under a Gaussian assumption (and is not χ^2). When $\hat\sigma_1^2$ and $\hat\sigma_2^2$ are not scaled χ^2 random variables then we cannot expect their ratio $\hat\sigma_1^2/\hat\sigma_2^2$ to be approximately F distributed.

The situation is a better for

$$\frac{1}{n-1}\sum_{i=1}^{n}(\bar Y_{i\bullet}-\bar Y_{\bullet\bullet})^2$$

because the CLT is making each $\bar Y_{i\bullet}$ more nearly normally distributed than individual Y_{ij} are.

11.3 Variance components

Our model for a one way layout with random effects is

$$Y_{ij}=\mu+a_i+\varepsilon_{ij},\quad i=1,\dots,I\quad j=1,\dots,n$$

where $a_i\sim\mathcal{N}(0,\sigma_A^2)$ and $\varepsilon_{ij}\sim\mathcal{N}(0,\sigma_E^2)$ are all independent. We have renamed σ^2 to be σ_E^2 here. The variances σ_A^2 and σ_E^2 are called ***variance components***. For a thorough treatment of variance components see Searle et al. (1992).

In this simple variance components model we have

$$\mathbb{E}(\mathrm{MSA})=n\sigma_A^2+\sigma_E^2\quad\text{and}\quad\mathbb{E}(\mathrm{MSE})=\sigma_E^2.$$

It is quite common in more complicated variance components settings to have σ_E^2 in every expected mean square. The reason is that the errors ε_{ij} contribute variance to every observation and there is no way to cancel them out.

We are often most interested in estimating σ_A^2 and σ_E^2 and related ratios such as $\sigma_A^2/(\sigma_A^2+\sigma_E^2)$. We can get unbiased estimates of them by taking

$$\hat\sigma_E^2=\mathrm{MSE}\quad\text{and}\quad\hat\sigma_A^2=\frac{\mathrm{MSA}-\mathrm{MSE}}{n}.$$

The estimate of σ_A^2 is potentially awkward because it can be negative. It is then common to take

$$\hat\sigma_A^2=\max\Big(\frac{\mathrm{MSA}-\mathrm{MSE}}{n},0\Big).$$

This estimate is no longer unbiased. It satisfies $\mathbb{E}(\hat\sigma_A^2)>\sigma_A^2$ because we sometimes increase an unbiased estimate to zero, but never decrease it. If we are averaging estimates like this over many data sets we might prefer to use any negative values we get so as not to get a biased average.

We can also look at this setting through the correlation patterns in the data. If $j\neq j'$ then

$$\mathrm{cov}(Y_{ij},Y_{ij'})=\mathrm{cov}(a_i+\varepsilon_{ij},a_i+\varepsilon_{ij'})=\mathrm{cov}(a_i,a_i)=\sigma_A^2$$

and so

$$\rho \equiv \mathrm{corr}(Y_{ij}, Y_{ij'}) = \frac{\sigma_A^2}{\mathrm{var}(Y_{ij})} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}.$$

We can interpret $\hat{\sigma}_A^2$ as an indication that $\rho < 0$. Negative correlations are impossible under our random effects model but distributions with those negative correlations do exist. For instance the correlation matrix for Y_{ij} for $j = 1, \ldots, n$ could be

$$\begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

for $-1/(n-1) \leqslant \rho \leqslant 1$. The lower limit on ρ is there to keep the correlation matrix positive semi-definite.

In this correlation model

$$\mathrm{var}\left(\sum_{j=1}^{n} Y_{ij} \right) = n\sigma^2 + n(n-1)\rho\sigma^2$$

or equivalently

$$\mathrm{var}(\bar{Y}_{i\bullet}) = \frac{\sigma^2}{n}(1 + (n-1)\rho).$$

Here $1 + (n-1)\rho$ is the design effect we saw in cluster randomized trials. What we see with $\rho < 0$ is that the variance of $\bar{Y}_{i\bullet}$ or of $\sum_j Y_{ij}$ is less than what it would be for independent observations. If there is some mechanism keeping the total more constant than under independence that could explain negative correlations. Cox (1958) considers animals that share a pen into which some constant amount of food is placed. That could introduce negative correlations in their weights. In a ride hailing setting with a fixed number of passengers we might see negative correlations among the number of rides per driver. In both of those settings we could get positive correlations too. The quantity of food or of passengers could fluctuate up and down generating positive correlations.

If negative correlations are statistically convincing then we can move away from the ANOVA and model the covariance matrix of the data instead.

11.4 Unbalanced settings

In most of these notes we look at balanced data settings. In a few cases the unbalanced sample sizes cause no difficulty. For instance this is true for the one way layout with fixed effects. In other settings unbalanced sample sizes cause severe complications that we do not delve into in a first course on experimental design.

For an illustration consider the one way random effects model

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

and suppose that we want to estimate the grand mean μ. Two natural estimates are

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} \quad \text{and} \quad \hat{\mu} = \frac{1}{I}\sum_{i=1}^{I}\bar{Y}_{i\bullet}.$$

That is we can average all of the data or average all of the group means. If we actually knew σ_A^2 and σ_E^2 then we could compute the minimum variance unbiased linear combination of $\bar{Y}_{i\bullet}$ as

$$\begin{aligned}\hat{\mu} &= \sum_{i=1}^{I} \bar{Y}_{i\bullet}/\mathrm{var}(\bar{Y}_{i\bullet}) \Big/ \sum_{i=1}^{I} 1/\mathrm{var}(\bar{Y}_{i\bullet}) \\ &= \sum_{i=1}^{I} \bar{Y}_{i\bullet}/(\sigma_A^2 + \sigma_E^2/n_i) \Big/ \sum_{i=1}^{I} 1/(\sigma_A^2 + \sigma_E^2/n_i).\end{aligned}$$

Now if $\sigma_A^2 \gg \sigma_E^2/n_i$ for all i then averaging the $\bar{Y}_{i\bullet}$ would be nearly optimal. If instead, $\sigma_A^2 \ll \sigma_E^2/n_i$ for all i then averaging all the data would be nearly optimal. In practice we don't ordinarily know these variance components but this analysis would let us make a reasonable choice between the two natural estimates above given a guess or assumption on the variance components.

For much more about variance components and unbalanced data, see Searle et al. (1992).

11.5 Estimating or predicting the a_i

There are settings where we actually want to know something about a_i for a specific experimental unit i, even though a_i are thought to be sampled from some distribution.

Searle et al. (1992) give an example from dairy science. Suppose that i represents a bull and j represents a cow that is a daughter of bull i. The setting is nested because cow $j = 1$ for bull i' has nothing to do with cow $j = 1$ for bull i. Now let

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

be some measure of milk yield or quality from cow $j \in \{1, 2, \ldots, \ldots, n_i\}$ of bull i. We might want to estimate a_i in order to judge whether to keep using bull i. Sometimes this problem is described as ***predicting*** a_i because a_i is random. The term "predicting" seems unnatural here because a random effect is not necessarily a quantity defined through the future.

To see how this works we once again make a simplifying assumption that we know μ and σ_A^2 and σ_E^2. If we want to estimate $\mu + a_i$ we can do better than just using $\bar{Y}_{i\bullet}$. Following Searle et al. (1992, Chapter 3.4) suppose that

$$\begin{pmatrix} a_i \\ \bar{Y}_{i\bullet} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 + \sigma_E^2/n_i \end{pmatrix} \right).$$

Our best estimate of a_i is $\mathbb{E}(a_i \,|\, \bar{Y}_{i\bullet})$ (after arguing that observations from $i' \neq i$ don't help and that only the sufficient statistic $\bar{Y}_{i\bullet}$ is useful). Using properties of the bivariate Gaussian distribution

$$\begin{aligned}\mathbb{E}(a_i \,|\, \bar{Y}_{i\bullet}) &= \mathbb{E}(a_i) + \mathrm{cov}(a_i, \bar{Y}_{i\bullet})\mathrm{var}(\bar{Y}_{i\bullet})^{-1}(\bar{Y}_{i\bullet} - \mu) \\ &= \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/n_i}\big(\bar{Y}_{i\bullet} - \mu\big).\end{aligned}$$

Under very strong assumptions of normality and knowing μ, σ_A^2 and σ_E^2 we would estimate (predict) a_i by

$$\frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/n_i}\big(\bar{Y}_{i\bullet} - \mu\big).$$

We ***shrink*** $\bar{Y}_{i\bullet} - \mu$ towards zero, shirking it a lot if $\sigma_A^2 \ll \sigma_E^2/n_i$. So we estimate $\mu + a_i$ by

$$\frac{\sigma_E^2/n_i}{\sigma_A^2 + \sigma_E^2/n_i}\mu + \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/n_i}\bar{Y}_{i\bullet}.$$

This is a linear combination of the population mean μ and the average for unit i. As n_i increases we trust $\bar{Y}_{i\bullet}$ more. This estimate is the ***BLUP***, for best linear unbiased predictor. It minimizes variance among linear combinations of data. With our simplifying assumptions here, the data is just $\bar{Y}_{i\bullet}$. The approach generalizes but becomes complex to depict.

11.6 Missing data

Suppose we have randomized blocks

$$Y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$$

viewing the block as a random effect, and the observation $Y_{i'j'}$ is missing. We could replace it with whatever minimizes

$$\sum_i \sum_j (Y_{ij}^* - \mu - \alpha_i - b_j)^2$$

where Y_{ij}^* is Y_{ij} if we have it and a parameter if we don't.

This amounts to running a regression on the row and column indicators with a special variable X with $X_{ij} = 1$ if $i = i'$ and $j = j'$ and $X_{ij} = 0$ otherwise. Because we have fit one more parameter we subtract one from the error df getting $(I-1)(J-1)-1$. We can adjust for a small number of missing responses this way. For more details see Montgomery (1997).

11.7 Choice of response

Suppose that $\mathbb{E}(Y_{ijk}) \doteq e^{\mu+\alpha_i+\beta_j+\gamma_k}$. We can still write

$$\mathbb{E}(Y_{ijk}) = \tilde{\mu} + \tilde{\alpha}_i + \tilde{\beta}_j + \widetilde{\alpha\beta}_{ij} + \tilde{\gamma}_k + \widetilde{\alpha\gamma}_{ik} + \widetilde{\beta\gamma}_{jk} + \widetilde{\alpha\beta\gamma}_{ijk}$$

for some new parameters. But we may have made the problem much harder by introducing high order interactions.

In a setting like this, $\log(Y)$ may have a more nearly additive model than Y does. If $\log(Y)$ is nearly additive then Y may not be. The expression above has $\log(\mathbb{E}(Y_{ijk}))$ additive which is not the same has having $\mathbb{E}(\log(Y_{ijk}))$ additive. Conversely, sometimes Y is more nearly additive than $\log(Y)$.

There is a strong simplification from modeling on a nearly additive scale because interactions bring in so many more parameters. Also many of our models and methods use aliasing of the interactions and that is less harmful when they are much smaller. It may then require some after thought to translate a model for transformed Y to get conclusions for $\mathbb{E}(Y)$. A very difficult situation arises when Y is measured in dollars and the model works with $\log(Y)$¿

As a second example, suppose that

$$\mathbb{E}(Y_{ijk}) \doteq \mu + \alpha_i + \beta_j + \gamma_k,$$

and let

$$\tilde{Y} = \begin{cases} 0, & |Y-\tau| > \delta \\ 1, & |Y-\tau| \leqslant \delta. \end{cases}$$

I.e. $\tilde{Y}$ = "Y is ok". Even if we ultimately care about $\tilde{Y}$ it can be much simpler to model Y because $\tilde{Y}$ can have lots of interactions. For instance, suppose that larger i implies larger α_i. Then $\tilde{Y}$ increases with i when $\beta_j + \gamma_k$ is small but decreases with i when $\beta_j + \gamma_k$ is large. That translates into greater impact from interactions. It is better to model Y statistically and then derive consequences for $\tilde{Y}$ from the model for Y.

12

Response surfaces

Very often we want to model $\mathbb{E}(Y \mid \boldsymbol{x})$ for continuosly distributed $\boldsymbol{x}$, not just binary variables as we could handle with 2^k factorials. Those can be used to study continuous variables by choosing two levels for them. However, once we know which variables are the most important and have perhaps a rough idea of the range in which to explore them we may then want to map out $\mathbb{E}(Y \mid \boldsymbol{x})$ more precisely for the subset of most important variables. We would like to estimate an arbitrary ***response surface*** $\mathbb{E}(Y \mid \boldsymbol{x})$ in those key variables. The literature on response surface models is mostly about estimating first order (linear) and second order (quadratic) polynomial models in $\boldsymbol{x}$, so most of the practical methods do not have the full generality that the term 'response surface' suggests.

If we are operating near an optimum value of $\mathbb{E}(Y \mid \boldsymbol{x})$ then a quadratic model might capture the most important aspects of $\mathbb{E}(Y)$. If that optimum is on the boundary of a constraint region then a local linear model might be very suitable. Local linear models are also very suitable in screening out the most important variables.

The material for this lecture is based largely on these texts Box and Draper (1987), Myers et al. (2016) and Wu and Hamada (2011) and these survey articles Myers et al. (1989) and Khuri and Mukhopadhyay (2010). Additional material on optimal design comes from Atkinson et al. (2007).

12.1 Center points

Designs at just two levels for each component of $\boldsymbol{x}$ let us fit the first order model

$$\mathbb{E}(Y\,|\,\boldsymbol{x}) \doteq \beta_0 + \sum_{j=1}^{p}\beta_j x_j. \tag{12.1}$$

We can also fit a "ruled surface" model like

$$\mathbb{E}(Y\,|\,\boldsymbol{x}) \doteq \beta_0 + \sum_{j=1}^{p}\beta_j x_j + \sum_{1\leqslant j<k\leqslant p}\beta_{jk}x_j x_k \tag{12.2}$$

though some of the cross terms might be subject to aliasing with higher order interactions, with each other (resolution IV) or with main effects (resolution III). Equation (12.2) leaves out term for $\beta_{jj}x_j^2$.

If x_{ij} takes only two levels, then the most we can do with it is fit a two parameter model such as a linear one. To fit a third parameter, such as curvature, we need a third level. For that we can take some center points. When we have been sampling $x_{ij}\in\{-1,1\}$ we might then take some additional runs with $x_{ij}=0$.

The simplest strategy is to add one or more center points with $\boldsymbol{x}_i = 0$. Put in a center point (maybe several). E.g. for $p=2$ we could use

$$\begin{matrix} x_1 & x_2 \end{matrix}\\ \begin{bmatrix} -1 & -1\\ -1 & 1\\ 1 & -1\\ 1 & 1\\ 0 & 0\\ \vdots & \vdots\\ 0 & 0 \end{bmatrix}.$$

From the repeated center points, we can estimate σ^2 or at least $\mathrm{var}(Y\,|\,\boldsymbol{x}=0)$. We can also use that data to estimate this model

$$\beta_0 + \sum_{j=1}^{p}\beta_j x_j + \sum_{1\leqslant j<k\leqslant p}\beta_{jk}x_j x_k + \gamma\sum_{j=1}^{p}x_j^2.$$

Notice that there is only one coefficient γ for all of the squared terms. This is $\gamma = \sum_{j=1}^{p}\beta_{jj}$ in our usual notation. The reason is that in a design with two levels plus a center point we have $x_{ij} = \pm x_{ij'}$ for all $i=1,\dots,n$ and all $1\leqslant j<j'\leqslant d$. This then implies that $x_{ij}^2 = x_{ij'}^2$ and so all of the quadratic terms are perfectly confounded.

We might run this centerpoint design in a case where we expect little curvature but just want to be able to make a check on it. If there is convincing evidence that γ differs from zero by an important amount, then we know there

is curvature and a first order model is problematic. If $\hat{\gamma}$ is not significantly different from zero then this is consistent with there being no curvature but does not prove it. The true β_{jj} might sum to nearly zero. In other words we have a ***one way diagnostic***. When it provides evidence of curvature we can be confident that it is there, but when it does not provide such evidence we cannot be confident that there isn't any curvature. Readers might be familiar with one way diagnostics in Markov chain Monte Carlo methods: they can reliably detect slow mixing but ordinarily cannot establish good mixing.

Another use for centerpoints is that, as remarked above, we might want an estimate of $\mathrm{var}(Y \mid \boldsymbol{x} = 0)$. The variance of Y at the centerpoint might be a reasonable variance to use for planning even if the true variance depends on $\boldsymbol{x}$.

Yet another use for them is to serve as 'control runs' that this page from NIST `https://www.itl.nist.gov/div898/handbook/pri/section3/pri337.htm` describes as "To provide a measure of process stability and inherent variability". Their advice is ***not*** to include the centerpoints in the randomized experimental order. Instead they recommend spacing those points out evenly through the run. For instance with 16 points in a 2 level design, they might add a centerpoint at the beginning, middle and end of experimentation, bringing the total to $n = 19$ runs. The other 16 points would be placed in a random order in the remaining 16 experimental positions.

12.2 Three level factorials

There is a theory of 3^k factorial designs and 3^{k-p} fractional factorial designs that parallels the case for two level designs. Not surprisingly, the expense grows more quickly with k than we get for 2 level designs.

In a three level design we take $x_{ij} \in \{-1, 0, 1\}$ after rescaling. For variables that take widely different values these levels might be what we get after a logarithmic transformation. For instance we might use $x_{ij} \in \{-1, 0, 1\}$ to encode a quantity that originally took values 100, 200 and 400.

When $k = 1$ our experiment at 3 level has two degrees of freedom for treatments. These are usually expressed through two contrasts: a linear contrast $\bar{Y}_1 - \bar{Y}_{-1}$, and a quadratic contrast $2\bar{Y}_0 - \bar{Y}_1 - \bar{Y}_{-1}$. These are orthogonal contrasts.

With k effects A, B, C, $\cdots$ we find 2 degrees of freedom for A, 4 degrees of freedom for a two factor interaction like AB, 8 degrees of freedom for a three factor interaction like ABC, and so on. So things get expensive. The ANOVA table for a three level design can be partitioned as in Table 12.1.

The ANOVA table for a three level design has terms for the product of k quadratic effects such as $A_Q \times B_Q \times C_Q$. We might well use some of those interactions in an error term, just as we did for two factor designs to mitigate the high cost of three level experiments. We could also plot estimated effects in a QQ plot to identify important variables.

Source	df
A	2
A_L	1
A_Q	1
B	2
B_L	1
B_Q	1
AB	4
$A_L \times B_L$	1
$A_L \times B_Q$	1
$A_Q \times B_L$	1
$A_Q \times B_Q$	1

Table 12.1: An ANOVA table for a three level factorial.

x	y	x+y mod 3	x+2y mod 3
0	0	0	0
0	1	1	2
0	2	2	1
1	0	1	1
1	1	2	0
1	2	0	2
2	0	2	2
2	1	0	1
2	2	1	0

Table 12.2: This is a 3^{4-2} fractional factorial. It is also an orthogonal array in that every pair of columns has all 9 possible rows the same number of time (i.e., once).

Given data from a 3^k factorial experiment we can fit the two level model

$$\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \sum_{j=1}^{k} \beta_{jj} x_{ij}^2 + \sum_{1 \leqslant j < j' \leqslant k} \beta_{jj} x_{ij} x_{ij'}$$

by least squares. We might prefer to center the pure quadratics

$$\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \sum_{j=1}^{k} \beta_{jj} (x_{ij}^2 - 2/3) + \sum_{1 \leqslant j < j' \leqslant k} \beta_{jj} x_{ij} x_{ij'}$$

to make them orthogonal to the intercept term.

Table 12.2 shows a 3^{4-2} fractional factorial experiment. It is known as an 'orthogonal array' because each pair of columns has all 9 possible combinations of variables the same number of times. We will see much more about orthogonal

arrays later. The top of Table 12.2 shows a construction in modular arithmetic. We will see more of that construction too.

When using an orthogonal array, be sure to randomize the levels. That is, there are 6 possible ways to map the levels 0, 1 and 2 of the array onto the experimental levels -1, 0 and 1 and one of those should be chosen at random. An independent randomization should be made for each column. It would be a very poor practice to just subtract 1 from each entry in the array. The run order should also be randomized.

These 3^{k-p} designs can also be run in blocks whose size is a power of 3.

There is an extensive selection of three level designs here: `http://neilsloane.com/oadir/#3_2`. For a comprehensive account of orthogonal arrays, see Hedayat et al. (1999).

12.3 Central composite designs

In the ***central composite design*** of Box and Wilson (1951) we begin with a two level design with values $\{-1,1\}^k$, ordinarily a 2^{k-p} fractional factorial, then add some number n_0 of center points $(0,0,0,\dots,0)$ and then $2k$ "star points" varying one factor at a time. These take the form $(\pm\alpha,0,0,\dots,0)$, $(0,\pm\alpha,0,\dots,0)$, $(0,0,\pm\alpha,\dots,0)$, and so on, up to $(0,0,0,\dots,\pm\alpha)$, for some $\alpha>0$. The points are then used in random order, sometimes within blocks.

In setting up a central composite design we have to choose our three components: the two level design to use, the number of center points, and the value α for the star points.

The analysis is usually a quadratic linear regression. Given a point $\boldsymbol{x}\in\mathbb{R}^d$ we form a vector of features

$$f(\boldsymbol{x}) = (1, x_1,\dots,x_d, x_1x_2,\dots,x_{d-1}x_d, x_1^2,\dots,x_d^2)^{\mathsf{T}} \in \mathbb{R}^p$$

for $p = 1+d+d(d-1)/2+d = 1+d(d+3)/2$ and fit by least squares. That is

$$\hat\beta = (F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}Y$$

where $F\in\mathbb{R}^p$ has i'th row $\boldsymbol{f}_i = f(\boldsymbol{x}_i)$ and $Y\in\mathbb{R}^n$ with i'th element Y_i.

In choosing the two-level experiment, we will want all quadratic and cross terms to be estimable. That is, $F^{\mathsf{T}}F$ should be invertible, which we can easily check before commencing to experiment. Naively that would be solved by using a resolution V experiment that keeps the cross terms uncounded with each other. By the time the center points and star points are included, the true condition becomes much more subtle. See Wu and Hamada (2011, Chapter 9.7) for a very careful exposition. For instance, one can use what they call resolution III* designs. Those have resolution III with no words of length exactly four. That is one cannot have $ABCD=\pm I$. They also point out that one can use Plackett-Burman (i.e., Hadamard) points for the two level design. In dimension $k=2$, even a 2^{2-1} experiment plus center points and axial points can make the model estimable.

$2k$ points OAAT $\pm\alpha \times e_j$
n_0 center points
https://www.jstor.org/stable/2983966
https://www.google.com/search?q=central+composite+design

It is also common to code the pure quadratic features as $x_{ij}^2 - (1/n)\sum_{i'=1}^n x_{i'j}^2$ to make them orthogonal to the intercept. This also makes them orthogonal to the linear terms because, breaking the design into its three parts we find that

$$\begin{aligned}\sum_i (x_{ij}^2 - \overline{x_j^2}) x_{ij'} &= \sum_i x_{ij}^2 x_{ij'} \qquad \text{from } \sum_i x_{ij'} = 0\\ &= \sum_{i \in \text{Factorial}} x_{ij'} + (\alpha * 0) + (-\alpha * 0)\\ &= 0.\end{aligned}$$

Exercise: show that $x_{ij}^2 - \overline{x_j^2}$ is orthogonal to $x_{ij}x_{ij'}$ for $j' \neq j$ and to $x_{ij'}x_{ij''}$ when no two of j, j' and j'' are equal.

One very tricky issue is choosing the value of α. Taking $\alpha = 1$ is convenient because it keeps all factors at three levels. We will have a subset of a 3^k factorial experiment. Another choice is to take $\alpha = \sqrt{k}$ because this keeps $\|\boldsymbol{x}_i\|^2 = k$ on the star points just like it is for the factorial points. This is called a "spherical design" because then both the star and factorial points are embedded within a sphere of radius $\sqrt{k}$. When we choose a spherical design we need some zero points or else $\sum_{j=1}^k x_{ij}^2 = k$ for all i and we will have a singular matrix F. Exercise: is this exactly the same problem that we saw with a centerpoint design and the parameter γ or is it different?

If we choose $\alpha = \sqrt{k}$ then we might find that values $x_{ij} \in \pm\sqrt{k}$ are too far from the region of interest even though they are exactly the same distance from the center as the factorial points are. The issue stems from factor sparsity. If x_1 is a very important variable, much more so than the others, then taking $x_{i1} = \pm\sqrt{k}$ represents a much more consequential change than just taking everything in $\{-1, 1\}$. Something is too far from the region of interest if the quadratic model that serves over $[-1, 1]^k$ does not extrapolate well there. Perhaps changing a geometric parameter for a transistor by that much turns it into a diode. It is even possible that operating at $x_{ij} = \pm\sqrt{k}$ is unsafe if x_j represents temperature or pressure. This sort of non-statistical issue can be much more important than designing to reduce $\text{var}(\hat{\beta})$ and it requires the input of a domain expert.

One possible way to choose α is to obtain orthogonality, that is a diagonal matrix $F^{\mathsf{T}}F \in \mathbb{R}^{p\times p}$. If the pure quadratic terms are centered then it remains possible that they are not orthogonal to each other. There is one value of α that makes them orthogonal. After some algebra, this is

$$\alpha = \Big(\frac{QF}{4}\Big)^{1/4}$$

for $Q = [(F+T)^{1/2} - F^{1/2}]^2$ where F is number of factorial observations and $T = 2k + n_0$. This then makes $\text{corr}(\hat{\beta}_{jj}, \hat{\beta}_{j'j'}) = 0$.

Another way to choose α is to obtain ***rotatability***:

$$\mathrm{var}(\hat{\mathbb{E}}(Y\,|\,\boldsymbol{x})) = f(\boldsymbol{x})^{\mathsf{T}}(F^{\mathsf{T}}F)^{-1}f(\boldsymbol{x})\sigma^2 = g(\|\boldsymbol{x}\|)$$

for some function $g(\cdot)$. Now the statistical information when predicting $\mathbb{E}(Y\,|\,\boldsymbol{x})$ depends only on $\|\boldsymbol{x}\|$ and not on the angle between $\boldsymbol{x}$ and any of the coordinate axes. There is not a strong motivation for choosing rotatability. Rather it is a tie-breaker condition when choices are otherwise equal. Also, when factor sparsity is present then sparse vectors $\boldsymbol{x}$ such as $(1,0,0,\ldots,0)$ and $(0,1,0,\ldots,0)$ should be more important than others of the form $(\cos(\theta),\sin(\theta),0,\ldots,0)$ for arbitrary $0<\theta<\pi/2$.

Box and Draper (1987) include some blocking schemes for central composite designs. In a blocked analysis we would use indictor variables taking the value 1 in a given block and zero outside of that block. It is then important to have those indicator variables be orthogonal to the linear, quadratic and mixed terms in the second order regression model. NIST shows some examples of central composite designs in blocks at `https://www.itl.nist.gov/div898/handbook/pri/section3/pri3364.htm`. If that link doesn't work well, look for section 5.3.3.6.4 entitled "Blocking a response surface design" in their online engineering statistics handbook.

12.4 Box-Behnken designs

The second major class of response surface designs are the Box-Behnken designs from Box and Behnken (1960). Those designs have three levels 0 and ±1. A small example for factors A, B and C looks like this:

A	B	C
±1	±1	0
0	±1	±1
±1	0	±1
0	0	0

The first row denotes a 2^2 factorial in A and B with factor C held at zero. There follow two similar rows with A and then B held at zero. Finally there is a row representing n_0 runs at $\boldsymbol{x}=0$. If, for instance $n_0=3$ then this experiment would have 15 runs. According to Box and Behnken (1960), "The exact number of center points is not critical". Geometrically this design has 12 points on the surface of the unit cube $[-1,1]^3$, one at the midpoint of each of the 12 edges connecting the 8 vertices within six faces. There are also center points.

We can recognize the strategy in this design. There is a balanced incomplete block structure designating some number $r<k$ of the factors that take values ±1 while the remaining $k-r$ factors are held at zero. Then some number of center points are added.

Table 12.3 shows another Box-Behnken design, this time for 4 factors. It is arranged in three blocks each of which has its own center point.

A	B	C	D
±1	±1	0	0
0	0	±1	±1
0	0	0	0
±1	0	0	±1
0	±1	±1	0
0	0	0	0
±1	0	±1	0
0	±1	0	±1
0	0	0	0

Table 12.3: A schematic for a Box-Behnken design in four factors with three blocks and one center point per block.

Exercise: are the block indicator variables for the Box-Behnken design in 12.3 orthogonal to the regression model? If we change it to $n_0 = 4$ are the orthogonal? Since there are three block variables you can drop the intercept column.

Like 3^{k-p} designs Box-Behnken designs can easily handle categorical variables at 3 levels. The tabled designs in the literature and textbooks involve only modest dimensions k. For large k, Box-Behnken designs use many more runs than parameters. While that may be useful in some settings, in others there is a premium on minimizing n by taking it just slightly larger than the number of parameters.

12.5 Uses of response surface methodology

One of the main uses is to fit a quadratic model, and estimate the direction of steepest ascent. Then, assuming that larger $\mathbb{E}(Y \mid \boldsymbol{x})$ is better move the region of interest in the direction of apparent improvement and run another experiment. The approach features a 'human in the loop' deciding which variables to explore and how at each iteration of the experiment. This is called ***evolutionary operation*** by Box (1957).

There is a large related field of stochastic optimization that takes possibly noisy data and uses it to decide where next to sample with the goal of finding an optimum. See Kushner and Yin (2003) and Spall (2003). Those methods often emphasize iterations taking one or two new data points at each round.

12.6 Optimal designs

This section is based primarily on Atkinson et al. (2007). We will pick $\boldsymbol{x}_i \in \mathbb{R}^k$ and then compute features $f(\boldsymbol{x}_i) \in \mathbb{R}^p$ (such as for a quadratic regression). Then our model is $Y_i = f(\boldsymbol{x}_i)^\mathsf{T}\beta + \varepsilon_i$ we estimate β by $\hat\beta = (F^\mathsf{T}F)^{-1}F^\mathsf{T}Y$

and our accuracy is described by $\mathrm{var}(\hat\beta) = (F^\mathsf{T}F)^{-1}\sigma^2 \in \mathbb{R}^{p\times p}$. When we predict Y for a given $\boldsymbol{x}$ then we may use $\widehat{Y}(\boldsymbol{x}) = f(\boldsymbol{x})^\mathsf{T}\hat\beta$ with $\mathrm{var}(\widehat{Y}(\boldsymbol{x})) = f(\boldsymbol{x})^\mathsf{T}(F^\mathsf{T}F)^{-1}f(\boldsymbol{x})\sigma^2$.

We will want to choose $\boldsymbol{x}_i$ in some way that $\mathrm{var}(\hat\beta)$ is small and there are many ways that a matrix can be considered small. Before that however it is worth noting that in this setting $\mathrm{var}(\hat\beta)$ does not depend on the true β! The fact that β does not appear in the formula $(F^\mathsf{T}F)^{-1}\sigma^2$ is an enormous simplification, that we usually take for granted. Otherwise the best design for learning β would depend on the unknown β and then the design problem would be intrinsically circular. This actually happens for logistic regression which we consider briefly below. Once again, finding that a symbol ***is not*** in a formula is quite special.

Before getting started we should rule out three approaches. First, we don't want to solve the problem by letting $n \to \infty$. That's expensive and may be wasteful. We want to be as efficient as we can with the n that we can afford. Second, we don't want to solve it by letting $\sigma^2 \to 0$. We want to be as efficient as we can with a given quality of measurement. Finally, while small $\mathrm{var}(\hat\beta)$ corresponds to large $F^\mathsf{T}F$, we don't want to solve the problem by letting $\|\boldsymbol{f}_i\| \to \infty$ where $\boldsymbol{f}_i = f(\boldsymbol{x}_i)$. The linear model is only approximate and so we need to keep $\boldsymbol{x}_i$ in or near the region of interest. Also, the extreme $\|\boldsymbol{x}_i\|$ that we would ordinarily need for extreme $\|\boldsymbol{f}_i\|$ may be expensive or unsafe.

We formulate the design problem by fixing n and requiring $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$. Depending on the problem we might require $\boldsymbol{x}_i \in \mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| \leqslant 1\}$ or $\boldsymbol{x}_i \in \mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \max_j |x_j| \leqslant 1\}$.

There are numerous notions of optimality, including A-optimality, D-optimality, E-optimality, G-optimality and I-optimality. Using

$$\mathrm{var}(\hat\beta) = (F^\mathsf{T}F)^{-1}\sigma^2 = \frac{\sigma^2}{n}M^{-1} \quad \text{where} \quad M \equiv \frac{F^\mathsf{T}F}{n} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{f}_i\boldsymbol{f}_i^\mathsf{T}$$

we can describe the optimality criteria via the matrix M.

Perhaps the most famous and widely used notion is ***D-optimality***. A D-optimal design minimizes $\det(F^\mathsf{T}F)^{-1}\sigma^2$ (sometimes called the generalized variance of $\hat\beta$) or equivalently it maximizes $\det(M)$. This is the product of the eigenvalues of M. The 'D' stands for determinant.

A-optimality with 'A' for average refers to minimizing $\sum_j \mathrm{var}(\hat\beta_j)$ or maximizing the sum or average of the eigenvalues of M^{-1}. E-optimality with 'E' for extreme refers to minimizing $\max_j 1/\lambda_j$ where λ_j are the eigenvalues of M.

G-optimality refers to minimizing $\max_{\boldsymbol{x}\in\mathcal{X}} \mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta)$ where $\mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta) \propto f(\boldsymbol{x})^\mathsf{T}M^{-1}f(\boldsymbol{x})$. This notion goes back to Smith (1918) for polynomial regression, which is the first optimal design paper.

I-optimal design also called V-optimal refers minimizing $\int_{\mathcal{X}} \mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta)g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where $g(\cdot) \geqslant 0$ measures interest level. It could be a distribution but does not have to be. One could also minimize $\int_{\mathcal{X}'} \mathrm{var}(f(\boldsymbol{x})^\mathsf{T}\hat\beta)g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where the set $\mathcal{X}'$ includes extrapolations to points that are not in $\mathcal{X}$. D_A optimality refers to minimizing $\det(\mathrm{var}(A^\mathsf{T}\hat\beta))$ for some matrix A.

These definitions raise the question "which optimality is best?" To address that we need to consider ***design measures***. In a design measure, we generalize from $M(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^\mathsf{T}$ to

$$M(\mu) = \int_{\mathcal{X}} f(\boldsymbol{x})f(\boldsymbol{x})^\mathsf{T}\mu(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \mathbb{E}(f(\boldsymbol{x})f(\boldsymbol{x})^\mathsf{T}), \qquad \boldsymbol{x} \sim \mu$$

for a distribution μ. Then instead of finding the best list of n points in $\mathcal{X}$ we relax to problem to just seeking the optimal distribution μ. While $M(\mu)$ is written above as an integral as if μ were continuous, μ can also be discrete as all we need is the expectation. In fact, most of the solutions we see will involve discrete distributions μ. Given an optimal or near optimal design measure μ we can then pick $\boldsymbol{x}_i$ to approximate μ.

Atkinson et al. (2007) consider a problem of quadratic regression through origin. The model is $\mathbb{E}(Y \mid x) = \beta_1 x + \beta_2 x^2$ with $0 \leqslant x \leqslant 1$. That is, $\mathcal{X} = [0, 1]$. This is not a model we would often want to fit, but it is an excellent small illustration of how optimal design works. They find that to maximize the D-optimality condition $\det(\mathbb{E}(f(\boldsymbol{x})f(\boldsymbol{x})^\mathsf{T}))$ under $\boldsymbol{x} \sim \mu$ one should choose $\mu(\sqrt{2}-1) = 1/\sqrt{2}$ and $\mu(1) = 1 - 1/\sqrt{2}$. This optimal design measure only uses two different points x. Because the fraction of data at $\sqrt{2}-1$ is not a rational number we cannot actually find a finite set of points with empirical distribution μ. Instead, we would approximate it taking roughly $n/\sqrt{2}$ observations at $\boldsymbol{x} = 1/\sqrt{2}$ and the rest at $\boldsymbol{x} = 1$.

For a design measure D-optimality maximizes $-\log\det(M(\mu))$, If $\mu = \sum_{i=1}^n \omega_i 1_{\boldsymbol{x}_i}$ then for fixed $\boldsymbol{x}_i$, we get a convex function $\sum_{i=1}^n \omega_i \boldsymbol{f}_i\boldsymbol{f}_i^\mathsf{T}$ of $(\omega_1, \ldots, \omega_n)$ because $\log(\det(\cdot))$ is a convex operation on matrices. Some algorithms use a large n and then get most ω_i equal to zero or close to it. It is typical for the optimal design measure to have p points $\boldsymbol{x}_i$ with positive probabililty where p is the number of parameters in the regression model. This leaves us with no way to estimate σ^2 or test whether a model with more than p parameters would be suitable. Perhaps that is not surprising. Criteria like D-optimality do not include either variance estimation or testing goodness of fit.

The ***general equivalence theorem*** is an important result of Kiefer and Wolfowitz (1960). It is that D-optimality and G-optimality are equivalent for design measures that continuously weight the same set of $\boldsymbol{x}_i$.

A special property of D-optimality is ***equivariance***. If we change units from meters to centimeters, the D-optimal designs scale accordingly and we would run exactly the same set of physical experiments either way. Generally replacing $\boldsymbol{f}_i$ by $A \times \boldsymbol{f}_i + b$ the new optimal points have $\tilde{f}_i = Af_i^* + b$ and the same ω_i.

Optimal designs are not always good enough to use because they focus only on variance and might require awkward input combinations instead of for instance using just three levels of a continuous factor which could be convenient for implementation. In other words, we might not have been able to put all of the important criteria into the objective function or encoded all of the desired constraints. We can however compare a design such as a central composite or Box-Behnken to the optimal design and see how close to 100% efficiency we get.

Optimal design can also help us when we have a more complicated constraint region $\mathcal{X}$ to work with than the usual cubes and balls. It is not always possible to do the global optimization that we would want in optimal design.

Box and Draper (1987) are skeptical about the optimal design approaches sometimes called "alphabetic optimality". The work some examples that optimize mean squared error taking account of bias and variance. The bias comes from the possibility that the model is a polynomial of higher order than the one fit by the response surface model. They find that the optimal designs for mean squared error are similar to those we would get optimizing for bias squared and they can be quite different from what we would find optimizing just for variance like the alphabetic optimality designs do. They 'match moments' over the region, making $(1/n)\sum_{i=1}^n \boldsymbol{f}_i\boldsymbol{f}_i^{\mathsf{T}} = \mathbb{E}(\boldsymbol{f}\boldsymbol{f}^{\mathsf{T}})$.

Now we turn briefly to logistic regression. It is not a pre-requisite for this course but many readers will have encountered it. Logistic regression is for binary responses $Y \in \{0, 1\}$. The model relating Y to $x \in \mathbb{R}$ is

$$\Pr(Y = 1 \,|\, X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \equiv p(x; \beta)$$

and in general it uses $\Pr(Y = 1 \,|\, \boldsymbol{x}) = \exp(\beta^{\mathsf{T}} f(\boldsymbol{x}))/(1 + \exp(f(\boldsymbol{x})^{\mathsf{T}}\beta))$. The likelihood function is

$$L(\beta) = \prod_{i=1}^n p(x_i; \beta)^{Y_i}(1 - p(x_i; \beta))^{1-Y_i}$$

and $\hat{\beta}$ is optained by maximizing $\log(L(\beta))$ numerically.

The design problem is about where to take x_i. If we were to set $n/2$ of the $x_i = \infty$ (or as large as possible) and $n/2$ of the $x_i = -\infty$ (or as small as possible) we would not ordinarily get the best design. We might well get all $Y_i = 1$ at one extreme and all $Y_i = 0$ at the other, with no idea of the shape of the $\Pr(Y = 1 \,|\, \boldsymbol{x})$ curve (and a degenerate log likelihood as well). It turns out that the optimal design for estimating β takes $n/2$ observations at a point x with $p(x; \beta) \approx 0.15$ and $n/2$ observations and x with $p(x; \beta) \approx 0.85$. This design has the same number of distinct design points as parameters. Those design points depend on the true β. A starting point in this literature is Chaloner and Larntz (1989).

12.7 Mixture designs

Suppose that $x_{ij} \geqslant 0$ is proportion of input j in used in observation i, with $\sum_{j=1}^J x_{ij} = 1$. In some settings Y_i depends mostly on the proportions and not the absolute levels of the inputs. For instance, when mixing paints the ratios will matter more than the absolute amounts to the color of the final produce, assuming that the mixing has been done well. In many recipes, proportions matter much more than absolute amounts.

For $k = 3$ we then have an experimental region defined by an equilateral triangle with corners $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. In general, our input space is a simplex $\{\boldsymbol{x} \in [0, 1]^k \mid \sum_{j=1}^k x_j = 1\}$ with intrinsic dimension $k - 1$ embedded in the k dimensional unit cube. This different shape motivates different experimental designs. See the book by Cornell (0002).

The changed space also has consequences for polynomial models. The first order model is

$$\mathbb{E}(Y \mid \boldsymbol{x}) = a_0 + \sum_{j=1}^k a_j x_j = \sum_{j=1}^k (a_0 + a_j) x_j \equiv \sum_{j=1}^k b_j x_j,$$

where $b_j = a_0 + a_j$. We don't need an intercept term. The Second order model is

$$\mathbb{E}(Y \mid \boldsymbol{x}) = \sum_{j=1}^k b_j x_j + \sum_{j=1}^{k-1} \sum_{j'=j+1}^k b_{jj'} x_j x_{j'}.$$

We don't need pure quadratic terms because $x_1^2 = x_1(1 - x_2 - x_3 - \ldots - x_k)$, which can be written as the linear term and some cross terms, and of course, the same holds for all x_j.

13

Super-saturated designs

Sometimes the number p of regression variables is larger than the number n of observations we can take. Just as ***saturated*** models have one parameter per obervation, ***supersaturated*** models have $p > n$ or even $p \gg n$.

In this chapter we look at how do design for such cases. This leads us to consider Hadamard matrices, some history of 'random balance designs', compressed sensing, and the Johnson-Lindenstrauss lemma. There is a survey of supersaturated designs in Georgiou (2014). Krahmer and Ward (2011) discusses experimental design for compressed sensing.

We have already seen supersaturated designs defined via fractional factorials where there are fewer observations than parameters. Here we focus on problems where there are fewer observations than main effects (plus intercept). A plain least squares fit will then interpolate the data, both signal and noise, assuming as is reasonable that no two predictors $\boldsymbol{x}_i$ are equal. There is no obvious way to estimate the error variance and now obvious way to check for lack of fit.

The designs we consider are also called ***screening designs*** because their goal is to identify the perhaps small number of relatively important predictors. They are well suited to settings where the regression model is thought to be sparse with the majority of regression coefficients either zero or at least negligible.

13.1 Hadamard matrices

We begin with Hadamard matrices that are suitable for saturated settings. The matrix $H \in \{-1, 1\}^{n\times n}$ is a ***Hadamard matrix*** if $H^\mathsf{T}H = HH^\mathsf{T} = nI$. For

instance, with $n = 4$,

$$\begin{pmatrix} + & + & + & + \\ + & + & - & - \\ + & - & + & - \\ + & - & - & + \end{pmatrix}$$

is a Hadamard matrix. We could use it as a saturated design to fit an intercept and 3 binary variables. We have seen it before as a 2^{3-1} factorial. There is a good account of Hadamard matrices in (Hedayat et al., 1999, Chapter 7). It is definitive apart from a few recent contributions that one can find online either at Wikipedia or Neil Sloane's web site.

The ***Sylvester construction***, which actually pre-dates Hadamard's interest in these matrices is as follows:

$$H_1 = \begin{pmatrix} 1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} H_1 & H_1 \\ H_1 & -H_1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} H_2 & H_2 \\ H_2 & -H_2 \end{pmatrix}$$

and so on with

$$H_{2^{k+1}} = \begin{pmatrix} H_{2^k} & H_{2^k} \\ H_{2^k} & -H_{2^k} \end{pmatrix}$$

for $k \geqslant 1$ in general.

Sylvester's construction is a special case of a ***Kronecker construction*** that works as follows. If $A \in \{-1, 1\}^{n\times n}$ and $B \in \{-1, 1\}^{m\times m}$ then their Kronecker product is

$$A \otimes B = \begin{pmatrix} A_{11}B & A_{12}B & \cdots & A_{1n}B \\ A_{21}B & A_{22}B & \cdots & A_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1}B & A_{n2}B & \cdots & A_{nn}B \end{pmatrix} \in \{-1, 1\}^{nm\times nm}$$

If A is a H_n and B is an H_m then $A \otimes B$ is an $H_{n\times m}$

Now if A is an H_n and B is an H_m then $A \otimes B$ is an $H_{n\times m}$. The proof is simple and it illustrates some basic rules for manipulating Kronecker products:

$$\begin{aligned} (A \otimes B)^{\mathsf{T}}(A \otimes B) &= (A^{\mathsf{T}} \otimes B^{\mathsf{T}})(A \otimes B) \\ &= (A^{\mathsf{T}}A) \otimes (B^{\mathsf{T}}B) \\ &= (nI_n) \otimes (mI_m) \\ &= (nm)I_n \otimes I_m \\ &= nmI_{nm}. \end{aligned}$$

Every step follows directly from the definition of the Hadamard product, so readers seeing Kronecker products for the first time should take a moment to check each step.

It is known that if a matrix H_n exists then $n = 1$ or 2 or $4m$ for some integer $m \geqslant 1$. The ***Hadamard conjecture*** is that H_n exists whenever $n = 4m$ for $m \geqslant 1$. There is no known counter example, but matrices for $n \in \{668, 716, 892\}$

n	# distinct H_n
1,2,4,8,12	1
16	5
20	3
24	60
28	487
32	13,710,027

Table 13.1: Number of distinct Hadamard matrices of sizes up to 32. From `https://oeis.org/A007299` as of October 2020.

have not (as of October 2020) been found and there are 10 more missing cases for $n \leqslant 2000$. These missing values are not a problem for experimental design. If we want H_{668} but have to use H_{772} instead, it would not be a costly increase in sample size. The most plausible uses for such large matrices are in software and four additional function evaluations are unlikely to be problematic.

Two Hadamard matrices are ***equivalent*** if one can be turned into the other by permuting its rows, or by permuting its columns, or by flipping the signs of an entire row or by flipping the signs of an entire column. The number of distinct (non-equivalent) Hadamard matrices that exist for some small values of n are in Table 13.1.

Given a Hadamard matrix we can always find an equivalent one whose first row and column are all $+1$. Hadamard matrices are often depicted in this form. In an experiment we would then use the first column to represent the intercept and the next $n-1$ columns for $n-1$ predictor variables. What we get is a Resolution III design (main effects clear) in $n-1$ binary variables.

In addition to the Sylvester construction there is a construction of Paley (1933) that is worth noting. If $n = 4m$ and $s = n-1 = p^r$ for a prime number p and exponent $r \geqslant 1$, then Paley's first construction provides H_n. The construction is available whenever $p^r \equiv 3 \bmod 4$. Figure 13.1 shows one of these matrices for $n = 252$ and prime number $p = 251 \equiv 3 \bmod 4$. Apart from a border of $+1$ at the left and top, each row of this matrix is a cyclic shift of the row above it. That means we can construct the matrix 'on the fly' and need not store it. That feature would be very useful for $n > 10^6$. There is a construction in (Hedayat et al., 1999, Theorem 7.1) that is quite simple to use for a prime number $p \bmod 4$. For $n-1 = p^r$ it would require finite field arithmetic that when $r = 1$ reduces to arithmetic modulo p. Note that the theorem gives a first row of H_n equal to $\begin{pmatrix}1 & -1 & -1 & \cdots & -1\end{pmatrix}$ instead of all 1s. However, we would randomize all the columns once constructed. (Be sure to pick one randomization for each of the $n-1$ columns and use it for all n rows.) Paley (1933) has a second construction, but it does not have quite the same simply implemented striping pattern.

We can use foldovers of Hadamard matrices. First split the intercept column

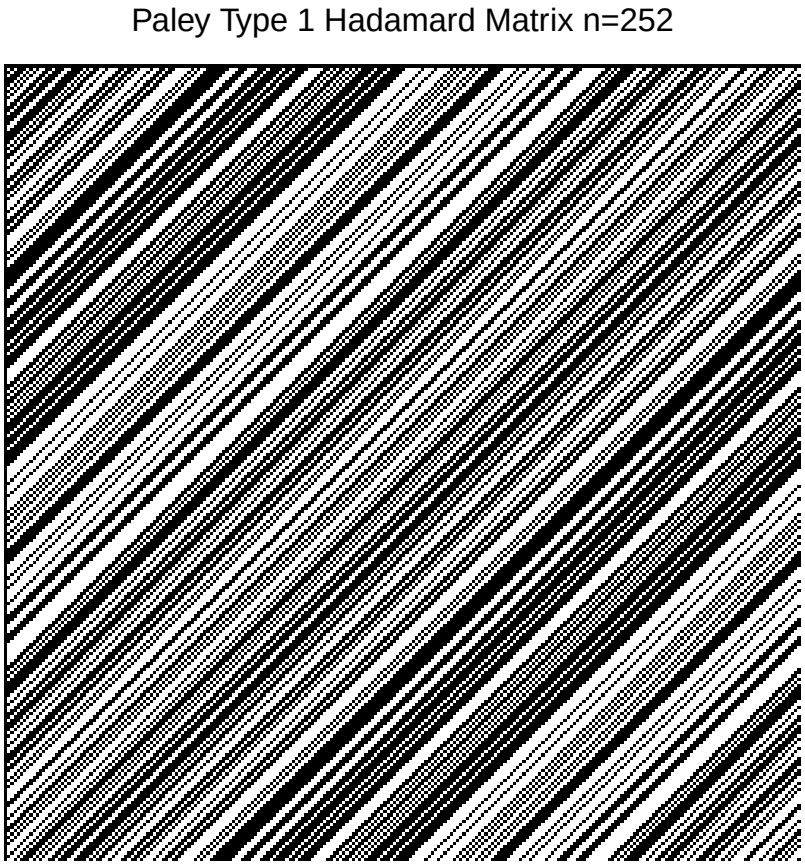


Figure 13.1: Image of a Hadamard matrix constructed using Paley's first construction with prime $p = 251$. Raw data from `http://neilsloane.com/hadamard/had.252.pal.txt` downloaded October 2020.

off producing $\tilde{H}_{4m} \in \{-1,1\}^{4m\times 4m-1}$ as follows

$$H_{4m} = \begin{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tilde{H}_{4m} \end{pmatrix} \in \{-1,1\}^{4m\times 4m}.$$

Then flip all the non-intercept columns yielding

$$\begin{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} & \begin{pmatrix} \tilde{H}_{4m} \\ \\ -\tilde{H}_{4m} \end{pmatrix} \end{pmatrix} \in \{-1,1\}^{8m\times 4m}$$

Any three columns of this matrix have all eight of $\{-1,1\}^3$ m times each.

13.2 Group testing and puzzles

Suppose that

$$Y = X\beta + \varepsilon \in \mathbb{R}^n \qquad X \in \mathbb{R}^{n\times p} \quad p > n.$$

Finding β is doable if β is sparse, having only a few nonzero entries.

As a familiar example of recent interest, suppose that one is doing group testing of blood samples. For convenience we might be testing $p = 1000$ people and we give them labels $000, 001, \ldots, 999$. Now suppose that we know that with high probability either none of them have covid or just one of them has it. Suppose further that if we pool samples from 30 people that we can still get a valid indication of whether at least one of those 30 has covid.

We could then use group testing. We take three samples from each person. We pool one sample from everybody whose first digit was 0. If that comes back positive then we have narrowed the set of candidates to 100 people. If it does not we can do nine more tests for those with first digits 1 through 9. Next we test groups of people based on their second digits and third digits. If all 30 tests come back negative then we have learned that none of those 1000 people have it. If somebody does have it, then three of the tests will come back positive and we will have identified the person.

As an exercise, formulate this group testing into the linear regreassion model above. Figure out what X_{ij} would be and what β is and even what is assumed about ε.

Group testing is faster and less expensive when the phenomenon is sparse.

There are some closely related ideas in old puzzles about finding which coin in a set has the wrong weight in a small number of weighings. When the person who has to figure this out has a equal arm balance then putting a coin on one sides corresponds to $X_{ij} = 1$, the other side corresponds to $X_{ij} = -1$ and leaving the coin off the balance corresponds to $X_{ij} = 0$. Those puzzles are usually sequential where the outcome of one test informs the following ones.

13.3 Random balance

Random balance is an idea proposed by Satterthwaite (1959) and supported by industry experience of Budne (1959) with a discussion by Youden et al. (1959). The idea is to simply take $x_{ij} \overset{\text{iid}}{\sim} F_j$ for $i = 1, \ldots, n$. There are versions of random balance where all variables are sampled this way and in other settings perhaps only some of them are sampled randomly while others are balanced carefully in Latin squares, factorial experiments or other such designs. In a two level design we might take all $x_{ij} \overset{\text{iid}}{\sim} \mathbb{U}\{-1, 1\}$ after rescaling the predictors. The author Satterthwaite is now best known for using the method of moments to approximate a sum of weighted χ^2 random variables by a random variable having a weighted χ^2 distribution (Satterthwaite, 1946).

Random balance was a provocative proposal and it evoked a strong response. For instance, Box wrote in his discussion that "The only thing wrong with

random balance is random balance". Tukey was more supportive.

Satterthwaite defined ***exact balance*** between two variables as what we now call orthogonality: under the empirical distribution on that pair of variables, they are independent. Then random balance is just that they are sampled from a distribution where they're independent with observed values that could violate orthogonality. A variable has random balance with respect to a set of other variables if it is independently sampled conditionally on them. When all variables are sampled randomly and independently the design is one of "pure random balance". If a bad randomly drawn experiment is discarded in favor of trying again, the design is "restricted pure random balance". This process is of renewed interest in the causal inference literature, where it is known as ***rerandomization***. See Morgan and Rubin (2012) and Li and Ding (2020) for more.

Much of the controversy around random balance is about its efficiency or lack of same. Satterthwaite mentions several settings that favor random balance. Sometimes the data can be collected very cheaply. Sometimes random balance simplifies the administration of an experiment. A related point is that people with very little statistical training might be more able to run random balance experiments than others.

The reason to include random balance in this section is that one of the use cases was for high dimensional input spaces and the random balance proposal was instrumental in raising this issue. Satterthwaite claims that they're nearly efficient as exactly balanced designs, becoming more so as the number of data sets increases (page 121). However the reasoning behind his evaluations is not given.

Budne (1959) writes in favor of random balance for ***screening experiments***. Those settings have a large number of variables but only a small number of them affect the mean response, or, only a small number affect the variance of the response. He then shows some example experiments with a graphical analyiss that identifies the few important variables.

It is interesting to see the issues brought up in the paper and discussion. The discussion reveals things that the experimenters knew about but might not have emphasized in many of their other writings. Multiple comparisons were well known (Tukey had worked on them a few years earlier). Pooling bias refers to selecting the small effects to create an error estimate (without adequate adjusments). Budne refers to screening experiments for both the mean and the variance. Youden reports running experiments with "dummy treatments", where the same treatment is given two labels, like the A/A tests now used in industry. Youden notes that an experiment to identify the large effects without balancing the smaller ones adds those smaller ones to the noise variance, reducing power. It will then be difficult to find the moderately sized effects. Kempthorne makes a similar comment with more detail about power. Kempthorne also notes that random balance is being proposed without regard to what we now call selective inference issues: the multiple comparisons underlying screening many variables and the bias in forming a variance estimate from the ones found to be smaller. Box makes some analyses of the efficiency of ran-

dom balance and in particular a stepwise approach of Satterthwaite's. He finds random balance mostly inefficient but one exception arises when there is only one nonzero coefficient (extreme sparsity) and the stepwise approach is used.

Tukey was more supportive than the other authors. He mentioned that broadening the base of inference (which is a form of external validity) is a worthy tradeoff for lower efficiency. He also remarked that he learned nearly as much from Budne's scatterplots as from the complete analysis of variance. There was also some discussion about random balance being so easy to do that it gets used more than classical designs that are hard to understand.

In a stepwise analysis, we might begin by estimating a slope for each x_{ij} individually, via

$$\tilde{\beta}_j \equiv \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{\bullet j})Y_{ij} \Big/ \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{\bullet j})^2.$$

For a binary x_{ij} the most efficient allocation would have half of the observations at each of the two levels. With random balance they would be somewhat unequally split. The second inefficiency in random balance is that the terms $x_{ij'}\beta_{j'}$ would raise the variance of Y_{ij} in a regression model to $\sigma^2+\sum_{j'\neq j}\beta_{j'}^2\mathrm{var}(x_{ij'})$. On the other hand, in a full regression model the matrix $X^{\mathsf{T}}X/n$ is far from identity and then $\det((X^{\mathsf{T}}X)^{-1})$ is infinite for $p>n$ no matter what design is used.

13.4 Quasi-regression

Suppose that we know the distribution of the feature vector $f(\boldsymbol{x}_i)\in\mathbb{R}^p$. Then the regression parameter that minimizes squared error in predicting a finite variance value Y_i with a linear combination of these features is

$$\beta = (\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}}))^{-1}\mathbb{E}(f(\boldsymbol{x}_i)Y_i). \tag{13.1}$$

This β minimizes $\mathbb{E}((Y-f(\boldsymbol{x})^{\mathsf{T}}\beta)^2)$ whether or not $\mathbb{E}(Y\,|\,\boldsymbol{x})=f(\boldsymbol{x})^{\mathsf{T}}\beta$. In linear regression with $n>p$ we estimate β by

$$\hat{\beta} = \Big(\frac{1}{n}F^{\mathsf{T}}F\Big)^{-1}\Big(\frac{1}{n}F^{\mathsf{T}}Y\Big) \tag{13.2}$$

where $F\in\mathbb{R}^{n\times p}$ has i'th row $f(\boldsymbol{x}_i)$ and $Y\in\mathbb{R}^n$ has i'th component Y_i. The expectations in (13.1) have been replaced by corresponding sample averages in (13.2) to get $\hat{\beta}$.

A very popular choice is to take $f(\boldsymbol{x})$ to be products of orthogonal polynomials. The resulting expansion approximating $\mathbb{E}(Y\,|\,\boldsymbol{x})$ is known as ***polynomial chaos***.

Now suppose that $\boldsymbol{x}_i$ have been sampled at random. When we choose the sampling distribution we may well know $\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}})$ because we also chose the features. For instance, if the features are all polynomials and $\boldsymbol{x}_i$ have a

convenient distribution such as uniform or Gaussian we would have easily computable moments of f. In this case, we can estimate β by

$$\tilde{\beta} = (\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}}))^{-1}\Big(\frac{1}{n}F^{\mathsf{T}}Y\Big), \tag{13.3}$$

replacing the estimate $(F^{\mathsf{T}}F)/n$ by its expected value.

The estimate (13.3) is known as ***quasi-regression***. See An and Owen (2001) and Jiang and Owen (2003) who used quasi-regression to get interpretable approximations to black box functions. Maybe $\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}})$ is diagonal or block structured. If $\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}}) = nI$ then

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)Y_i$$

which can be computed at $O(np)$ cost instead of $O(np^2)$ that least squares costs. Quasi-regression can be used when $p > n$ and it avoids the $O(p^2)$ space required for linear regression.

When $p > n$ then shrinkage estimators as in Jiang and Owen (2003) are advised. Ordinarily $\mathrm{var}(\tilde{\beta}) \geqslant \mathrm{var}(\hat{\beta})$ when both are possible. This is a counterexample to the usual rule in Monte Carlo sampling where plugging in a known expectation in place of a sampled quantity ordinarily helps. Blatman and Sudret (2010) report that sparse regression methods outperform methods based on numerically estimating $\mathbb{E}(ff^{\mathsf{T}})$ and $\mathbb{E}(fY)$.

13.5 Supersaturated criteria and designs

We can see in some of the discussions of Satterthwaite (1959) the beginnings of criteria to improve upon random balance for super saturated settings. We clearly cannot use D-optimality because we are sure to have the regression matrix X satisfy $\mathrm{rank}(X^{\mathsf{T}}X) \leqslant n < p$. Then $X^{\mathsf{T}}X$ is singular with $\det(X^{\mathsf{T}}X) = 0$ and then effectively "$\det((X^{\mathsf{T}}X)^{-1}) = \infty$". Georgiou (2014) is a survey of criteria and algorithms for supersaturated designs.

Another thing that changes in the supersaturated setting is that we will need adaptive methods that decide which β_j to estimate and which to leave at a default value like 0. These adaptive methods will have to use the observed Y_i values. In the end the estimate of β is not ordinarily a linear combination of Y_i like it is in least squares. The variance of the estimated β becomes more complicated and will depend on what the true β is. For instance if the true β is all zeros except the intercept then methods based on sparsity could be very accurate and even get most of the coefficients exactly right. If instead β has all nonzero elements of roughly equal size, perhaps described by $\beta_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ then no algorithm can find them. In other words, the accuracy with which β can be estimated now depends on the true value of β, because even though the model relating $\mathbb{E}(Y)$ to X is linear in X, the estimator of β is not linear in Y_i.

Booth and Cox (1962) citing Box's discussion of random balance introduce some criteria. The consider the matrix $X \in \{-1, 1\}^{n\times p}$ with $p > n-1$ and each column of X containing $n/2$ values for each of ± 1. Then they consider the matrix $S = X^{\mathsf{T}}X$ which is proportional to sample correlations among the predictors. Their criterion is $\max_{1\leqslant j<j'\leqslant p} |S_{jj'}|$. They break ties by preferring designs where a smaller number of column pairs attain the maximum of $|S_{jj'}|$. They give some small examples with n and p of a few dozens. The examples were found by computer search. When compared to random balance the designs they obtain have much better values of S. They also report the variance of $|S_{jj'}|$ values.

Georgiou (2014) considers the criterion

$$\mathbb{E}(S^2) \equiv \frac{1}{\binom{p}{2}} \sum_{1\leqslant j<j'\leqslant p} S_{jj'}^2$$

and remarks that algorithms that try to optimize it may possibly yield identical columns. Incidentally, Georgiou (2014) gives this as the first criterion that Booth and Cox (1962) consider but in preparing these notes I do not find that in their article.

Georgiou (2014) reports some lower bounds on $\mathbb{E}(s^2)$. The precise bounds depend on things like the value of n modulo 4. In all cases

$$\mathbb{E}(S^2) \geqslant \frac{n^2(p-n+1)}{(n-1)(p-1)}$$

where his definition of $\mathbb{E}(S^2)$ includes the intercept column of all ones. If we normalize each $S_{jj'}$ to $S_{jj'}/n$ then

$$\mathbb{E}((S/n)^2) \geqslant \frac{p-n+1}{(n-1)(p-1)}.$$

Let's suppose that $n \gg 1$ and $p - n \gg 1$. Then ignoring the ± 1s above, we get

$$\mathbb{E}((S/n)^2) \geqslant \frac{p-n+1}{(n-1)(p-1)} \approx \frac{p-n}{np} = \frac{1-n/p}{n}.$$

Exercise: what would we get for $X_{ij} \overset{\text{iid}}{\sim} \mathbb{U}\{-1, 1\}$? What would we get if half of column j were randomly chosen to be each of ± 1 and column j' were chosen that way too but independently of column j?

Georgiou (2014) presents numerous design strategies for the supersaturated setting. Of special interest is the proposal of Lin (1993). This design works with a Hadamard matrix H_n. It picks one of the columns and keeps only the $n/2$ runs with $+1$ in that column. This is appealing because Hadamard matrices are so abundant and both the Sylvester and first Paley constructions are easy to use. Lin's approach provides $n/2$ experimental runs in $p = n-2$ variables. One of the n colunns is lost to the intercept term and one more is lost because of the column chosen to define the selected runs. The specific column chosen

makes makes little difference. Exercise: compare the $\mathbb{E}(S^2)$ criterion that Lin gets to what he would get choosing runs with -1 in the given column.

Lin (1993) runs forward stepwise regression on data from his design. Phoa et al. (2009) use L_1 regularization based on the Dantzig selector of Candes and Tao (2007).

Lin (1995) looks for ways to maximize the number p of binary variables in a model subject to a constraint on $\max_{j\neq j'}|S_{jj'}|$. He breaks ties based on the number of pairs at the same maximum level.

Practical investigations: how would it go to choose 1/4 or 1/8 et cetera of a Hadamard design? Would Paley or Sylvester constructions be about equally good or would one be better? This last point requires a way to make a fair comparison between Hadamard designs with different values of n.

13.6 Designs for compressed sensing

Supersaturated designs are well suited to settings where we have $p > n$ but can reasonably expect β to be sparse or nearly so. Donoho (2006) describes compressed sensing and Tibshirani (1996) introduces the ***lasso***, both of which can be used when $p > n$. Krahmer and Ward (2011) discusses experimental design for this setting. This section only surveys the issue, because a detailed discussion goes beyond prerequisites for this course.

In this context it is desireable for the matrix $X \in \mathbb{R}^{n\times p}$ (excluding intercept) to satisfy the ***restricted isometry property*** (RIP), defined in terms of a level $\delta \in (0,1)$ and an integer order $k > 0$. The vector $v \in \mathbb{R}^p$ is said to be ***k-sparse*** if $\sum_{j=1}^p 1_{v_j\neq 0} \leqslant k$. Then X satisfies the (k,δ)–RIP if

$$(1-\delta)\|v\|^2 \leqslant \|Xv\|^2 \leqslant (1+\delta)\|v\|^2$$

holds for all k-sparse v. Krahmer and Ward (2011) give RIP conditions (and references) where minimizing $\|X\beta\|_1$ subject to $X\beta = Y$ exactly recovers a sparse β. If β is sparse and there is no noise it can be recovered by L_1 penalized regression. That is, a very tractable convex optimization problem can be used to find β where searching for a sparse solution would otherwise be quite costly. There are generalizations to handle additive $Y = X\beta + \varepsilon$ (i.e., measured with noise) but sparsity remains a critical ingredient.

One of the most basic constructions of a matrix with an RIP property is to take a random subset of rows of a Hadamard matrix. The resulting design will satisfy an RIP property with very high probability. Specifically, for p predictors we can ignore the intercept column and take n randomly selected rows (without replacement) from H_{4m} where $4m \geqslant p+1$. Compared to the method of Lin (1993) this approach allows n other than $4m/2 = 2m$. When we want half of the rows it makes more sense to use Lin's design because then every column will be equally split between ± 1 values. The designs described for compressed sensing require only very small values of n, just over $\delta^{-2}k\log(p)^4$.

A related approach is to choose a random subset from a discrete Fourier matrix. That matrix has complex entries $X_{jk} = \omega^{jk}/\sqrt{n}$ for $0 \leqslant j,k < n$ (note

the zero based indexing) where $\omega = \exp(2\pi\sqrt{-1}/n)$. They split out the real and imaginary parts of X_{ij} doubling the number of columns obtained. (Those $2n$ real valued vectors in $\mathbb{R}^n$ cannot of course be mutually orthogonal.)

Another approach they describe is to take $X \in \mathbb{R}^{n\times p}$ with IID $\mathcal{N}(0,1)$ or IID $\mathbb{U}\{-1,1\}$ entries multiplied by $\sqrt{p/n}$. The amazing thing is that this, after many years, provides some justification for Satterthwaite's random balance.

Part of the argument in Krahmer and Ward (2011) is based on the ***Johnson-Lindenstrauss*** lemma (Johnson and Lindenstrauss, 1984). Think of a saturated design as $p-1$ column vectors in $\mathbb{R}^p$, orthogonal to each other and to the vector of ones. Now we project those columns $v_1,\dots,v_{p-1}$ into a lower dimensional space by multiplying them by a matrix Φ. That is we get $u_i = \Phi v_i \in \mathbb{R}^n$ for $i = 1,\dots,n$ for a matrix $\Phi \in \mathbb{R}^{n\times p}$. The Johnson-Lindenstrauss lemma shows that there is a mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$ that preserves interpoint distances to within ϵ. When that mapping is the linear one described above this means that

$$(1-\epsilon)\|v_j - v_{j'}\|^2 \leqslant \|u_j - u_{j'}\|^2 \leqslant (1+\epsilon)\|v_j - v_{j'}\|^2.$$

The original Johnson-Lindenstrauss Lemma allowed a nonlinear but Lipschitz continuous function $u_j = g(v_j)$ instead of the $u_j = \Phi v_j$ that we use here. If the interpoint distances are nearly preserved in this mapping then so are angles. Think of a triangle defined by v_j, $v_{j'}$ and $v_{j''}$. If all three interpoint distances are nearly preserved then so are all three angles defined by the points. Now when v_j are orthogonal (right angles) then the u_j are nearly orthogonal.

In the Johnson-Lindenstrauss lemma the number n does not have to be very large compared to p. It only needs to be $O(\epsilon^{-2}\log(p))$.

14

Computer experiments

In a computer experiment we investigate a deterministic function $f(\boldsymbol{x})$ on what may be a high dimensional domain and where f might be very expensive to evaluate. We then have to carefully pick out which input values to use and how to interpolate/extrapolate to the uncomputed values. That extension can then be a cheap surrogate function which we can explore intensively.

The odd thing is that we are used to applying statistical methods in a setting with $Y = f(\boldsymbol{x}) + \varepsilon$. Here we are ordinarily missing the $+\varepsilon$ term which is the usual entry point for statistical methods into a problem setting. What we will do is sample $Y_i = f(\boldsymbol{x}_i)$ for strategically chosen $\boldsymbol{x}_i$, essentially turning numerical problems into statistical ones. The designs we use are ordinarily descibed informally as ***space filling***. When $f(\cdot)$ has no measurement error then we normally turn to interpolation methods to estimate $f(\boldsymbol{x})$ at points where we have not sampled. The most common choices are based on ***kriging*** which stems from modeling $f(\cdot)$ is if it were a randomly chosen continuous function.

For background on computer experiments, see the texts Santner et al. (2018) and Fang et al. (2006) as well as the articles Sacks et al. (1989), Koehler and Owen (1996) and Roustant et al. (2012).

This chapter is based on two lectures. It might have been good to have one lecture for design and another for analysis. That however is not a good fit. The designs and analyses are more naturally developed jointly. This chapter is also a survey of computer experiment ideas; readers will need to follow up in the references for more details.

14.1 Motivating problems

In class we first looked at animations of some computer experiment output. One was a finite element analysis of a car crashing into a post and another showed fuel sloshing in the back of a transport truck that was braking. Similar videos are readily found online though any specific one might disappear. The users of that code can extract quantities of interest like the amount of energy absorbed by the front of the car (and hence not delivered to the passenger compartment) in a collision or the amount of force on the sides and ends of the fuel container. The designer can then vary baffles in the truck or the configuration of the car to get the best tradeoff among numerous quantities of interest.

Codes of this type are used in the design of aircraft, semiconductors and factories. They are also used in modeling of climate and oil reservoirs. For instance we might have a model that predicts how much the arctic will warm at night in 10 years time given some assumptions about CO_2 emissions and reflectivity of clouds among other things. In settings like this we have

$$Y = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d, \quad Y \in \mathbb{R}^q.$$

The function $f(\cdot)$ is usually deterministic and expensive (e.g. 24 hours to run). The number d of input variables may be large and the output dimension q can be large too. The output might even be a curve such as a time trajectory of some scalar quantity.

The authors of these codes put knowledge from physics and chemistry into them in order to describe as faithfully as possible what Y will be like given $\boldsymbol{x}$. The users often have the opposite goal. They may want the $\boldsymbol{x}$ values that give some specific value of Y or that optimize some function $g(Y)$. It is as if they want "$f(\cdot)^{-1}$", the inverse of the physics or chemistry that was used to write f.

In computer experiments we make a strategic choice of input points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and compute $f(\boldsymbol{x}_i)$. We then use those function values to answer questions about $f(\cdot)$. Those questions usually depend on values like $f(\boldsymbol{x}_0)$ for new points $\boldsymbol{x}_0$ that are not among those n points. We do this by finding an ***emulator*** function $\tilde{f}(\cdot)$. Ordinarily $\tilde{f}(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$. That is, the emulator usually interpolates the known values exactly. We might be able to evaluate $\tilde{f}$ millions of times if it is fast enough.

There are compelling advantages to computer experiments. They are usually cheaper than physical experiments. Safer too. Design iterations can happen more quickly in computation than physically. For problems involving the future or objects deep in space, computer experiments may be the only option.

An interesting feature of computer experiments for statisticians is that we are very used to $Y = f(\boldsymbol{x}) + \varepsilon$. Now, all we have is $Y = f(\boldsymbol{x})$ without the noise term that is usually the invitation to think statistically. Experimental design has more to offer than just reducing the impact of noise. It also helps one cope with the complexity arising from interactions. In computer experiments, there may be no issue at all about causality. If we change $\boldsymbol{x}$ then $f(\boldsymbol{x})$ changes in a perfectly predictable way. There can still be issues of external validity arising from the assumptions baked into f.

As an example function, consider the ***wing weight function*** from Surjanovic and Bingham (2013). This is available online at `https://www.sfu.ca/~ssurjano` with some code. The function is

$$0.036 S_{\mathrm{w}}^{0.758} W_{\mathrm{fw}}^{0.0035} \Big(\frac{A}{\cos^2(\Lambda)}\Big)^{0.6} q^{0.006} \lambda^{0.04} \Big(\frac{100 t_{\mathrm{c}}}{\cos(\Lambda)}\Big)^{-0.3} (N_{\mathrm{x}} W_{\mathrm{dg}})^{0.49} + S_{\mathrm{w}} W_{\mathrm{p}},$$

It represents the weight of the wing of an aircraft depending on variables defined at Surjanovic and Bingham (2013).

Diaconis (1988) asks whether seeing the functional form is enough to understand a function. We might ask: which variables are most important? how do we get wing weight below some target?, which variables affect the function monotonically and what sizeable interactions are there? Even with a closed form expression like the above it is difficult to address these questions. Usually in computer experiments, we don't even have a formula, just code.

The questions above are actually trick questions too. Their answers depends on set $\mathcal{X}$ of interesting values $\boldsymbol{x}$. We may also need to introduce a distribution on $\boldsymbol{x}$ to pose the questions well. Even with a known distribution for $\boldsymbol{x}$ on $\mathcal{X}$, the point that Diaconis raises remains. It is not obvious how to interpret even a moderately complicated function from its formula. Variable importance is mentioned briefly near the end of this chapter.

14.2 Latin hypercube sampling

We seldom want to take a grid of $N = n^d$ points because the cost grows too quickly with d. We could just take $\boldsymbol{x}_i \sim \mathbb{U}[-1,1]^d$ or $\mathbb{U}[0,1]^d$ or $\mathbb{U}(\mathcal{X})$ for a set $\mathcal{X}$ of interest. We can however do better than this.

The first space filling design to consider is the ***Latin hypercube sample*** of McKay et al. (1979). We spread out $x_{1j}, \dots, x_{nj}$ to fill the range 0 to 1 for each $j = 1, \dots, d$. That is we take $\mathcal{X} = [0,1]^d$. For a Latin hypercube sample of n points in d dimensions we take

$$x_{ij} = \frac{\pi_j(i-1) + u_{ij}}{n} \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant d$$

where π_j are uniform random permutations of $0, 1, \dots, n-1$ and $u_{ij} \sim \mathbb{U}[0,1)$. All d permutations and nd uniform random variables are mutually independent. Many computing environments include methods to make a uniform random permutation. Figure 14.1 shows a small Latin hypercube sample. We see that the n values of x_{ij} for $i = 1, \dots, n$ are equispaced (stratified) and this is true simultaneously for all $j = 1, \dots, d$. It balances nd prespecified rectangles in $[0,1]^d$ using just n points. If any one of those inputs is extremely important, we have sampled it evenly without even knowing which one it was.

A notable strength of a Latin hypercube sample is that it allows $d > n$ or even $d \gg n$. It is also easy to show that each $\boldsymbol{x}_i \sim \mathbb{U}[0,1]^d$. The name of this

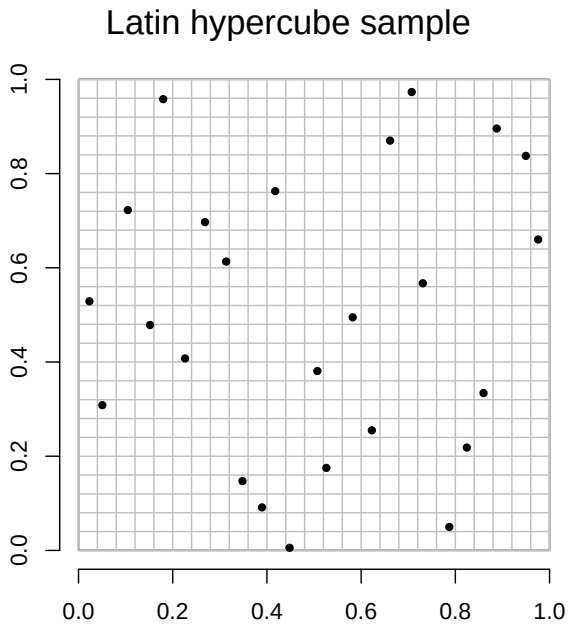


Figure 14.1: This shows a Latin hypercube sample of $n = 25$ points in $d = 2$ dimensions. From Owen (2020).

design is connected to Latin squares. For Figure 14.1, imagine placing letters A, B, C, $\cdots$, Y in the 25×25 grid with each letter appearing once per row and once per column. That would be a Latin square. In an LHS we sample within the cells occupied by just one of those 25 letters, perhaps 'A'. This design is also used in computer graphics by Shirley (1991) who gives it the name ***n rooks*** because if the sampled points were rooks on an $n \times n$ chessboard, no rook could take any other.

Figure 14.1 showed points uniformly and randomly distributed inside tiny squares. If we prefer, we could evaluate f at the centers of those squares by taking

$$x_{ij} = \frac{\pi_j(i-1) + 1/2}{n} \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant d.$$

This centered version of Latin hypercube sampling was described by Patterson (1954) for an agricultural setting (crediting Yates).

14.3 Orthogonal arrays

Orthogonal array designs are a generalization of Latin hypercube samples that allow us to balance more than one input variable at a time in our sampling. Randomized orthogonal arrays, described below, are suitable designs for ex-

ploring deterministic functions. The definitive reference for orthogonal arrays is Hedayat et al. (1999). There is also Neil Sloane's web site `http://neilsloane.com/oadir/` which has results from after the publication of that book as well as some files containing orthogonal arrays.

For an integer base $b \geqslant 2$, a dimension $d \geqslant t \geqslant 1$, and an integer $t \geqslant 1$, the discrete matrix

$$A \in \{0, 1, \dots, b-1\}^{n \times d}$$

is an ***orthogonal array of strength t*** if each $n \times t$ submatrix of A has all b^t possible distinct rows the same number λ of times. The nomenclature for it is $\mathrm{OA}(n, d, b, t)$. It is clear that $n = \lambda b^t$. Often $\lambda = 1$. Geometrically, if we were to pick any t columns of A make a t-dimensional scatter plot, we would get a full b^t grid, with λ copies per point.

In computer experiments, odd things can happen in the "corners". Those are points where two variables are both at extremes be they maxima or minima or a mix of the two. For instance, our codes might crash there. Orthogonal arrays get points into the corners that Latin hypercube samples could miss.

There are many constructions of orthogonal arrays. We will look at one from Bose (1938). It is very useful and the construction can be understood using just one fact about prime numbers. A prime number is an integer $p \geqslant 2$ whose only divisors are 1 and itself. That is not the fact about primes; that's the definition. The fact is that if positive integers a and b are such that ab is a multiple of p, then at least one of a or b is a multiple of p. This follows from the prime factorization theorem.

The Bose construction will give us $\mathrm{OA}(p^2, p+1, p, 2)$. Therefore we can handle $p+1$ variables and every $p \times 2$ submatrix has all p^2 rows once. We will be able to use that design to get into all 4 corners of all $p(p-1)/2$ pairwise scatterplots of our data. We will be sampling in more corners than there are data points. Any two columns of the orthogonal array could be reordered to look like columns labeled x and y in Table 14.1. Exercise: explain why this gives us $p-1$ mutually orthogonal Latin squares.

It is especially convenient that $p = 11$ is prime. Using $p = 11$ we get $n = 121$ points in $[0,1]^{12}$ after rescaling. All bivariate plots are 11×11 grids. We could use round number values 0.0, 0.1, 0.2, $\cdots$, 1.0. Similarly, for $p = 101$ we get $n = 10{,}201$ points in $[0,1]^{102}$. All bivariate plots are 101×101 grids. We could use round number values 0.00, 0.01, 0.02, $\cdots$, 1.00.

The construction goes as follows. We start with the two columns in Table 14.1, and then we adjoin columns $x + y$, $x + 2y$, $x + 3y$, $\dots$, $x + (p-1)y$ all in arithmetic mod p. Note that this construction does not work if p isn't prime. There is a generalization for prime powers $b = p^r$ but when $r > 1$ the generalization does not use arithmetic modulo b.

To see why it works, let's ignore the y column for now (leaving an exercise for later). Then any column can be written $x + cy \bmod p$ for some integer c. It is enough to show that any pair of columns has all p^2 possible vectors $(a_1, a_2) \in \{0, 1, \dots, p-1\}^2$. Since there are only $n = p^2$ rows, each of those vectors must then be present exactly once.

x	y
0	0
0	1
⋮	⋮
0	$p-1$
1	0
⋮	⋮
1	$p-1$
⋮	⋮
$p-1$	0
⋮	⋮
$p-1$	$p-1$

Table 14.1: Here are the first 2 columns of a Bose OA$(p^2, p+1, p, 2)$ orthogonal array. We have labeled them x and y for future use.

Let's pick any two of these columns indexed by c_1, c_2 with $c_1 < c_2$. The (a_1, a_2) that we need satisfies

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 & c_1 \\ 1 & c_2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \bmod p$$

where $(x, y)^{\mathsf{T}} \in \{0, 1, \ldots, p-1\}^2$ indexes the row of Table 14.1 that we want.

We can solve this formally by taking

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & c_1 \\ 1 & c_2 \end{pmatrix}^{-1} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \frac{1}{c_2 - c_1} \begin{pmatrix} c_2 & -c_1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}.$$

The matrix multiplication above is well defined in arithmetic modulo p but dividing by the determinant $d = c_2 - c_1$ has to be checked.

First, because $c_1 \neq c_2$ (for the two distinct columns) we have $d = c_2 - c_1 \neq 0 \bmod p$, so the determinant is not zero. Next we show that the value "$1/d$" is an integer $d^{-1} \in \{0, 1, \ldots, p-1\}$ such that $dd^{-1} = 1 \bmod p$, and that there is only one such integer. There are $p-1$ candidates $\{1, 2, \ldots, p-1\}$ for d^{-1}. Suppose that we multiply all of them by d getting $\{d, 2d, \ldots, (p-1)d\} \bmod p$. None of these can be zero because $0 < j, d < p$ with $jd = \bmod p$ would make jd a multiple of p contradicting the prime number fact above. This means that one of them has to be equal to $1 \bmod p$. As for uniqueness, suppose that there are two values $0 < j_1 < j_2 < p$ with $j_1 d = j_2 d = 1 \bmod p$. Then $(j_2 - j_1)d = 0 \bmod p$ which we have already ruled out.

It follows that we get all p^2 possible rows from any two columns of the Bose OA apart from cases where one of those columns is labeled 'y' above. Exercise: use a similar argument for the case of columns y and $x + cy \bmod p$ for $c \neq 0$.

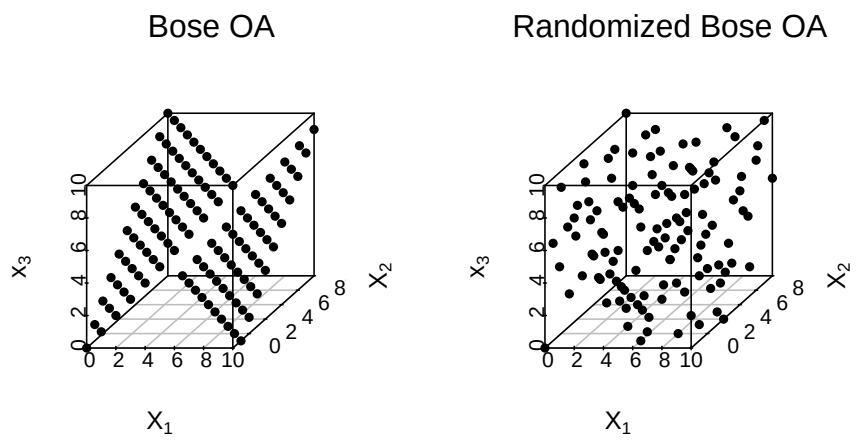


Figure 14.2: The left panel shows three columns of OA$(121, 12, 11, 2)$ unscrambled. The right panel shows a scramble of them. From Owen (2020).

The Bose OA "balances" $\binom{p}{2}p^2 = p^3(p+1)/2$ hyperrectangular subsets of $[0,1]^{p+1}$ with just p^2 points. For $p = 101$ we have balanced more than 5×10^7 strata with only about 10^5 points.

To use an orthogonal array as a space filling design it is important to randomize the levels. In a ***randomized orthogonal array*** (Owen, 1992) we begin with an OA(n, d, b, t) matrix A and take

$$x_{ij} = \frac{\pi_j(a_{ij}) + u_{ij}}{b} \quad \text{or} \quad x_{ij} = \frac{\pi_j(a_{ij}) + 1/2}{b}$$

for independent uniform random permutations π_j of $\{0, 1, \dots, b-1\}$ and $u_{ij} \sim \mathbb{U}[0,1)^d$. Random offsets u_{ij} produce $\boldsymbol{x}_i \sim \mathbb{U}[0,1)^d$ (which is the same distribution as $\mathbb{U}[0,1]^d$). These points are dependent because by construction they must avoid each other in any t-dimensional coordinate projection. The centered versions might be better for plotting contours of f in plane given by x_j and $x_{j'}$ for two variables $1 \leqslant j, j' \leqslant d$. Exercise: is a Latin hypercube sample a randomized orthogonal array?

It is important to apply the permutations π_j. Without permutation, the points will lie in or near two planes and then not fill the space well. See Figure 14.2. Figure 14.3 shows several pairwise scatterplots from a randomized orthogonal array.

Tang (1993) perturbs the points of an orthogonal array slightly, to make them also a Latin hypercube sample. Compared to a randomized orthogonal array the result is more uniform univariate marginal distributions.

There are many more orthogonal arrays to chose from. See Hedayat et al. (1999). As mentioned above, the Bose construction can be generalized to

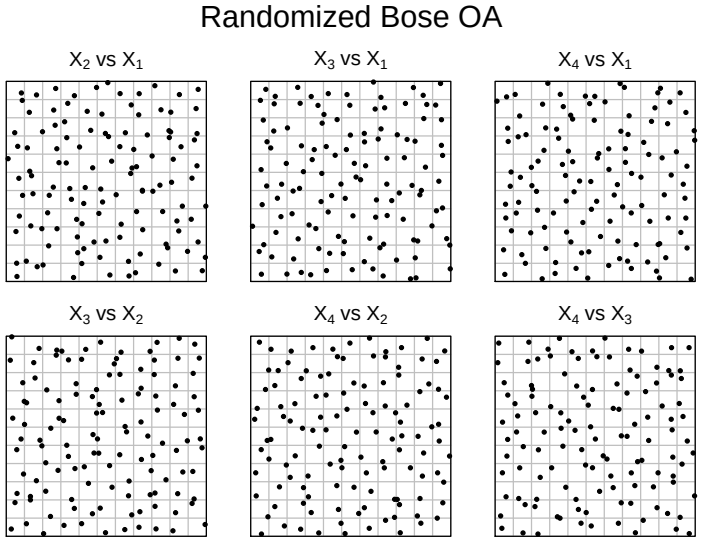


Figure 14.3: Some coordinate projections of randomized orthogonal arrays based on OA$(121, 12, 11, 2)$. From Owen (2020).

OA$(b^2, b+1, b, 2)$ for prime powers $b = p^r$. There are constructions of the form OA$(2b^2, 2b+1, b, 2)$ for $b = p^r$. These constructions from Bose & Bush or Addleman & Kempthorne nearly double the number of variables that can be explored by doubling the number n of function evaluations. Higher strength constructions due to Bush in 1952 produce OA$(b^t, b+1, b, t)$ for prime powers $b = p^r$ with $b \geqslant t-1 \geqslant 0$. There are also arrays with mixed levels. That is the number of distinct values can be different from one column to another. Some can be downloaded from `http://neilsloane.com/oadir/`.

14.4 Exploratory analysis

Given a space filling sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we can compute $Y_i = f(\boldsymbol{x}_i)$ yielding data pairs $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$. We can explore these using statistical and graphical methods. For instance we could select the "good Y values" and see what $\boldsymbol{x}_i$ they have. If f is expensive to compute thane we could also do that for the values $\tilde{f}(\boldsymbol{x}_i)$ for $i = 1, \ldots, N$ with $N \gg n$ once we have worked out how to extend f to an emulator $\tilde{f}$.

Sobol' sequences are a kind of quasi-Monte Carlo (QMC) sampling described below. Like orthogonal arrays there are advantages to scrambling QMC points. Using 1024 scrambled Sobol' points in $[0,1]^{10}$ as inputs to the wing weight function, we obtain Figure 14.4. When taking this many points is infeasible for our original function f, we might do it for an emulator.

It is surprising to see that the range of interesting wing weights has results

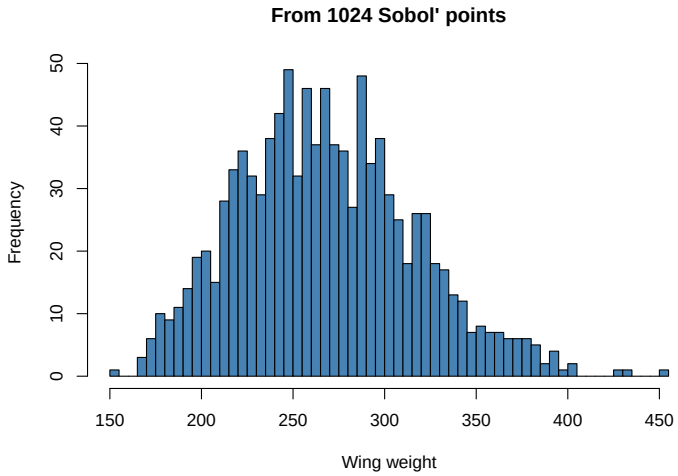


Figure 14.4: Histogram of the wing weight function evaluated at 1024 scrambled Sobol' points.

that vary by a factor of about 3–fold. Perhaps the input ranges are quite wide or there are strong effects in the corners. If we were interested in the lightest wings, then we could select out the points with $f(\boldsymbol{x}_i) < 200$ and plot their input values as in Figure 14.5. In many of the scatterplots we see very empty corners and densely sampled corners. For instance variables 3 and 8 together seem to have a strong impact on whether the wing weight is below 200. Figures like this are exploratory in nature; we might see something we did not anticipate, or we might not see anything that we can interpret.

Figure 14.6 show linear regression output for the wing weight function on the randomized Sobol' inputs. This function is nearly linear with $R^2 \doteq 0.9833$ (adjusted R^2 virtually identical at 0.9831). Because the sampling model is not linear plus noise the usual interpretation of regression output does not apply. Notwithstanding that, this function is surprisingly close to linear, even though the formula did not look linear. By Taylor's theorem, a smooth and very nonlinear function could look locally linear especially in a small region not centered at a point where the gradient vanishes. Reading the variables' described ranges online does not make them appear to be very local. Also, a small region of interest in design space would seem like it ought to restrict the wing's weight to a much narrower range than 3–fold. It appears that this nonlinear looking function is actually close to linear over a wide range of inputs. A plain regression as a quadratic polynomial with cross terms and pure quadratic terms scores an adjusted R^2 of 0.9997. The function is simpler than it looks.

This is by no means what always happens with computer experiments. It is however also quite common that potentially quite complicated functions that come in real applications are not maximally complex. There are additional

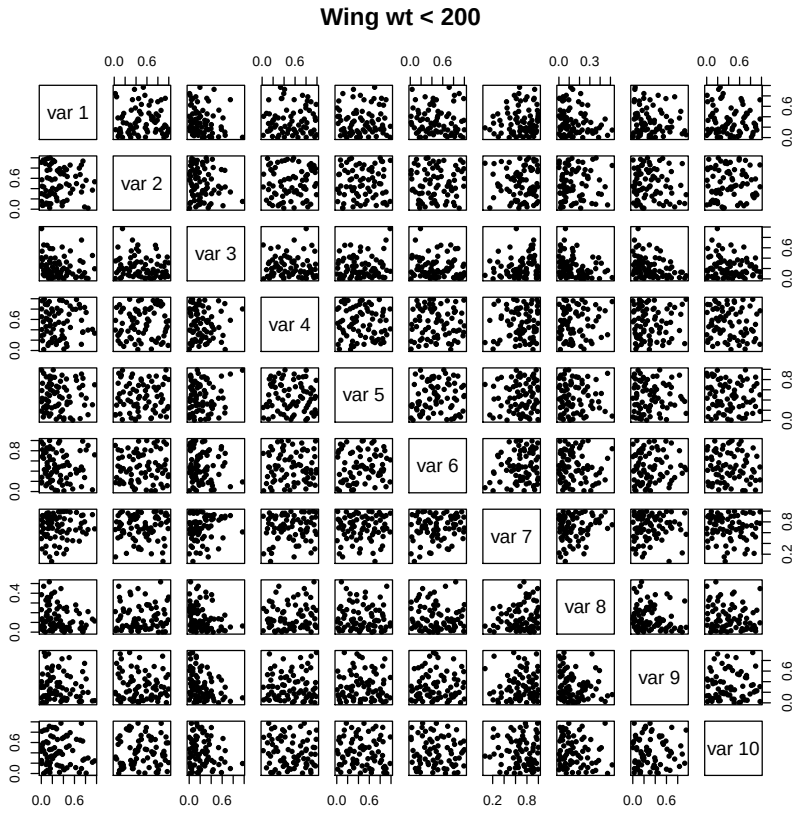


Figure 14.5: Scatterplot matrix of sample points for which the wing weight was below 200.

examples in Constantine (2015). See also Caflisch et al. (1997) who find that a complicated 360-dimensional function arising in financial valuation is almost perfectly additive.

14.5 Kriging

The usual way to predict $f(\boldsymbol{x}_0)$ at a point $\boldsymbol{x}_0 \notin \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ where we have evaluated f is based on a method called kriging from geostatistics. Kriging originated in the work of Krige (1951). Kriging is based on Gaussian process models of the unknown function f. Stein (2012) gives the theoretical background. Sacks et al. (1989) applies it to computer experiments drawing on a body of work developed by J. Sacks and D. Ylvisaker. Much of the description below is based on Roustant et al. (2012) who present software for kriging.

```
> summary(lm( wingwt1024pts ~ rsobo1024pts ))
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    143.64941    1.08843 131.978   <2e-16 ***
rsobo1024pts1   58.76020    0.67690  86.807   <2e-16 ***
rsobo1024pts2    0.16170    0.67693   0.239   0.8112
rsobo1024pts3   78.22236    0.67691 115.557   <2e-16 ***
rsobo1024pts4   -0.00758    0.67697  -0.011   0.9911
rsobo1024pts5    1.50798    0.67699   2.227   0.0261 *
rsobo1024pts6    7.45687    0.67693  11.016   <2e-16 ***
rsobo1024pts7  -62.02511    0.67697 -91.622   <2e-16 ***
rsobo1024pts8  106.70114    0.67698 157.614   <2e-16 ***
rsobo1024pts9   48.60273    0.67693  71.798   <2e-16 ***
rsobo1024pts10   9.46825    0.67694  13.987   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.253 on 1013 degrees of freedom
Multiple R-squared:  0.9833,    Adjusted R-squared:  0.9831
F-statistic:  5956 on 10 and 1013 DF,  p-value: < 2.2e-16
```

Figure 14.6: Some regression output for the wing weight function.

For an intuitive understanding of kriging consider Figure 14.7 and suppose we would like to know how much gold there might be per cubic meter at the location marked '?' given the other measured values in that figure. We would like to estimate it by a weighted average of those values. Because the point labeled 1.9 is closest we should give it more more weight than the point labeled 1.7. The three values somewhat larger than 2 should get more weight than the value 1.1 because there are three of them, and the distances from the target point are similar between them and the point labeled 1.7. They should not get in triple the weight in aggregate because, by being so close together we anticipate that the three readings might be somewhat redundant.

To turn these intuitive notions into a formula for computation we introduce a Gaussian model for the gold values with a covariance function to describe how similar nearby points are to each other.

First we review properties of the ***Multivariate Gaussian*** distribution. We assume that the standard normal distribution $\mathcal{N}(0,1)$ is familiar. If $\boldsymbol{z} \in \mathbb{R}^d$ has independent $\mathcal{N}(0,1)$ components then we write this as $\boldsymbol{z} \sim \mathcal{N}(0, I)$ where here 0 is a vector of d zeroes and I is the d-dimensional identity matrix. The vector $\boldsymbol{x} = \mu + C\boldsymbol{z}$ then has the $\mathcal{N}(\mu, CC^{\mathsf{T}})$ distribution. The general case is written $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = \mathbb{E}(\boldsymbol{x})$ and $\Sigma = \mathrm{cov}(\boldsymbol{x}) = \mathbb{E}((\boldsymbol{x}-\mu)(\boldsymbol{x}-\mu)^{\mathsf{T}})$. If $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$, then $\boldsymbol{x} = \mu + C\boldsymbol{z}$ for some matrix C satisfying $CC^{\mathsf{T}} = \Sigma$. The matrix square root C is not uniquely determined. Linear combinations of

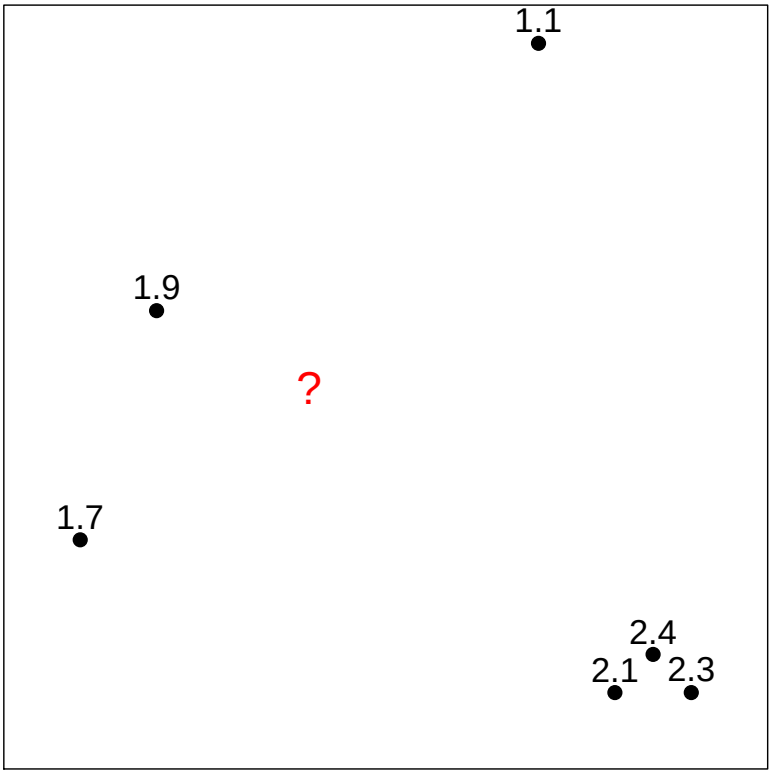


Figure 14.7: Hypothetical setting to illustrate kriging.

Gaussian vectors are also Gaussian: $A\boldsymbol{x}+b \sim \mathcal{N}(A\mu+b, A\Sigma A^{\mathsf{T}})$.

If $\boldsymbol{x}_1 = (x_1, \ldots, x_r)^{\mathsf{T}}$ and $\boldsymbol{x}_2 = (x_{r+1}, \ldots, x_d)^{\mathsf{T}}$ and

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

then

$$\boldsymbol{x}_1 \perp\!\!\!\perp \boldsymbol{x}_2 \iff \Sigma_{12} = 0.$$

For our present purposes, the most useful property of the multivariate Gaussian distribution is that, if Σ_{22} invertible, then

$$\mathcal{L}(\boldsymbol{x}_1 \,|\, \boldsymbol{x}_2) = \mathcal{N}\Big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big).$$

This condition is not very restrictive. If Σ_{22} is singular then some component of $\boldsymbol{x}_2$ is linearly dependent on the others. We could just condition on those others if they have a nonsingular covariance. More generally, we need only condition on a subvector of $\boldsymbol{x}_2$ that has an invertible covariance.

In a computer experiment context, we have $Y_i = f(\boldsymbol{x}_i)$, $i = 1, \ldots, n$ and we want to know about $f(\boldsymbol{x}_0)$. We adopt a multivariate Gaussian model

$$\boldsymbol{Y} = \begin{pmatrix} f(\boldsymbol{x}_0) \\ f(\boldsymbol{x}_1) \\ \vdots \\ f(\boldsymbol{x}_n) \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma),$$

with μ and Σ chosen as described below. Then, to predict $f(\boldsymbol{x}_0)$ we may use

$$\tilde{f}(\boldsymbol{x}_0) = \mathbb{E}(f(\boldsymbol{x}_0) \,|\, f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)).$$

We also get $\mathrm{var}(f(\boldsymbol{x}_0) \,|\, f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n))$ for uncertainty quantification, from the multivariate Gaussian distribution.

Because $\boldsymbol{x}_0$ could be anywhere in the set $\mathcal{X}$ of interest, we need a model for $f(\boldsymbol{x})$ at all $\boldsymbol{x} \in \mathcal{X}$. For this, we select functions $\mu(\boldsymbol{x}) = \mathbb{E}(f(\boldsymbol{x}))$ defined on $\mathcal{X}$ and $\Sigma(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{cov}(f(\boldsymbol{x}), f(\boldsymbol{x}'))$ defined on $\mathcal{X} \times \mathcal{X}$. The covariance function $\Sigma(\cdot, \cdot)$ must satisfy

$$\mathrm{var}\left(\sum_{i=1}^{n} a_i f(\boldsymbol{x}_i)\right) = \sum_{i=1}^{n}\sum_{i'=1}^{n} a_i a_{i'} \Sigma(x_i, x_{i'}) \geqslant 0 \tag{14.1}$$

for all $\boldsymbol{a} \in \mathbb{R}^n$ and all $\boldsymbol{x}_i \in \mathcal{X}$ for all $n \geqslant 1$, or else it yields negative variances that are invalid. Interestingly, the condition (14.1) is sufficient for us to get a well defined Gaussian process.

There is an interesting question about in what precise sense is $f(\cdot)$ random? We simply treat it as if the function f were drawn at random from a set of possible functions that we could have been studying. The function is usually chosen to fit a scientific purpose, though perhaps a random function model describes our state of knowledge about $f(\cdot)$. Perhaps not. However, the kriging method is widely used because it often gives very accurate emulator functions $\tilde{f}(\cdot)$. They interpolate because for $1 \leqslant i \leqslant n$,

$$\tilde{f}(\boldsymbol{x}_i) = \mathbb{E}(f(\boldsymbol{x}_i) \,|\, f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)) = f(\boldsymbol{x}_i).$$

Knowledge about $f(\cdot)$ can be used to guide the choice of $\mu(\cdot)$ and $\Sigma(\cdot, \cdot)$.

After choosing $\Sigma(\cdot, \cdot)$, we may pick a model like

$$Y = f(\boldsymbol{x}) = \mu(x) + Z(x)$$

where $\mu(\cdot)$ is a known function and Z is a mean zero Gaussian process with covariance Σ. This method is known as ***simple kriging***. For $\mu(\cdot)$, we might choose an older/cheaper/simpler version of the function f.

A second choice, known as ***universal kriging*** takes

$$Y = f(\boldsymbol{x}) = \sum_j \beta_j f_j(\boldsymbol{x}) + Z(x)$$

where once again $Z(\cdot)$ is a mean zero Gaussian process. Here β_j are unknown coefficients, while $f_j(\cdot)$ are known predictor functions. They could for instance be polynomials or sinusoids, or as above, some prior versions of f. Sometimes β_j are given a Gaussian prior, and other times they are treated as unknown constants. That makes $\sum_j \beta_j f_j$ a fixed effect term which in combination with a random effect $Z(\cdot)$ makes this model one of mixed effects. It has greater complexity than simple kriging. Roustant et al. (2012) describe how to analyze this situation.

The third model we consider is ***ordinary kriging***. Here

$$Y = f(\boldsymbol{x}) = \mu + Z(x)$$

for an unknown $\mu \in \mathbb{R}$ and a Gaussian process $Z(\cdot)$. We can view this as Z absorbing all of the $f_j(\cdot)$ from universal kriging. Ordinary kriging is the most commonly used method for computer experiments. It has a simpler theory and estimation strategy than universal kriging.

14.6 Covariance functions for kriging

The usual choice of covariance has $\Sigma(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 R(\boldsymbol{x}, \boldsymbol{x}')$ for a correlation function

$$R(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{corr}(f(\boldsymbol{x}), f(\boldsymbol{x}')).$$

It is also common to choose a ***stationary*** covariance, meaning one with

$$R(\boldsymbol{x}, \boldsymbol{x}') \equiv R(\boldsymbol{x} - \boldsymbol{x}').$$

For $\boldsymbol{x} \in \mathbb{R}^d$ we now have R depending just on a difference vector in $\mathbb{R}^d$ instead of depending on two vectors in $\mathbb{R}^d$. A ***radial*** or ***isotropic*** corelation function takes the form

$$R(\boldsymbol{x}, \boldsymbol{x}') = \rho(\|\boldsymbol{x} - \boldsymbol{x}'\|; \theta)$$

for a parameter θ. While these are commonly used in earth sciences (see Stein (2012)), computer experiments more commonly use a ***tensor product*** model

$$R(\boldsymbol{x}, \boldsymbol{x}') = \prod_{j=1}^{d} \rho_j(x_j - x_j') = \prod_{j=1}^{d} \rho(x_j - x_j'; \theta_j)$$

This model fits with factor sparsity. The value of θ_j may make some x_j very important and others unimportant. It reduces our problem to finding covariances for the $d = 1$ case. If all d correlations $\rho_j(\cdot)$ are valid then so is their tensor product.

One common choice is the ***squared exponential covariance***

$$\rho(s - t; \theta) = \exp(-\theta(s - t)^2)$$

for a parameter $\theta > 0$. This covariance resembles a Gaussian probability density function and so it is sometimes called a Gaussian correlation though that term

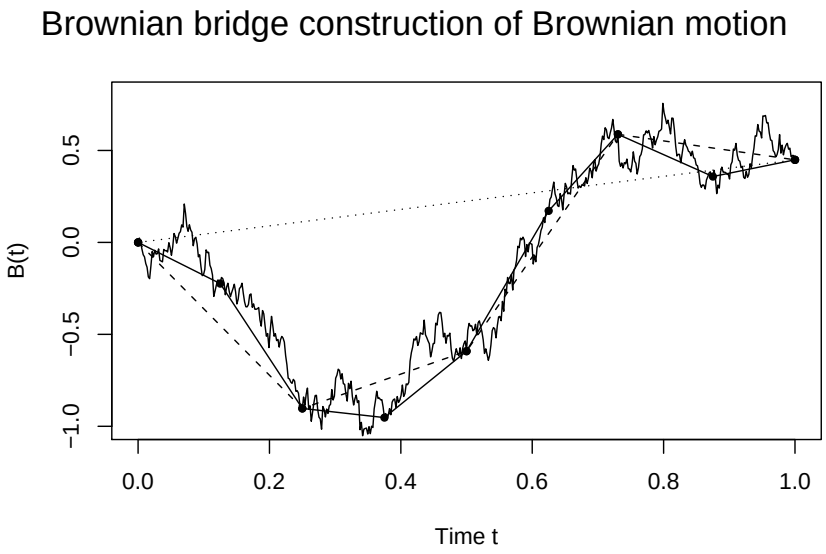


Figure 14.8: A Brownian motion path and some 'skeletons of it'. From Owen (2020).

is potentially confusing since the adjective 'Gaussian' is already being used in another sense to describe the distribution of $f(\boldsymbol{x})$. For $s - t = \epsilon$, we get $\exp(-\theta\epsilon^2) \approx 1 - \theta\epsilon^2$. Nearby points are ***very strongly*** correlated and as a consequence $f(\cdot)$ must be very smooth. Stein (1989) points out that the realizations of f are infinitely differentiable, which is unrealistically smooth for many applications.

The ***exponential covariance*** has $\rho(s-t;\theta) = \exp(-\theta|s-t|)$, for a parameter $\theta > 0$. For $s - t = \epsilon$, we get $\exp(-\theta\epsilon) \approx 1 - \theta\epsilon$. The correlation drops off much faster than for the squared exponential covariance. The realizations of $f(\cdot)$ are much less smooth. They are not even differentiable once, resembling instead Brownian motion. This of course is not smooth enough for many applications. Figure 14.8 shows a sample path of Brownian motion which has a covariance with a decay similar to that of the exponential correlation function. The piecewise linear 'skeletons' there give $\tilde{f}(\boldsymbol{x})$ based on the $(t, f(t))$ values that they connect.

The effect of θ also has implications on the smoothness of the realizations. A large θ implies that the correlation between $f(t)$ and $f(t+\delta)$ drops rapidly as $\delta > 0$ is increased. If we think of t as a time, then the process rapidly forgets $f(t)$ when θ is large. This can induce more rapid oscillations whether the sample paths of f are smooth or rough.

A compromise between exponential and square exponential correlations can be attained via the ***Matern kernels***. They are defined in terms of Bessel functions. When the paramter ν is set to $m + 1/2$, then the realizations have m derivatives and the correlation function has a closed form. Here are the first

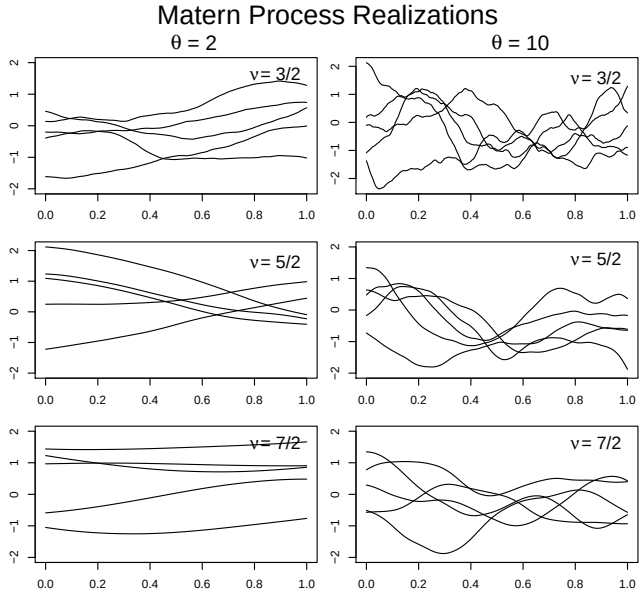


Figure 14.9: Some realizations of Gaussian processes with one dimensional Matern covariances. Larger values of the parameter ν increase their number of derivatives while larger θ value increase the speed at which they oscillate by forgetting their past. From Owen (2020).

four of them:

$$\begin{aligned}
\rho(s,t;1/2) &= \exp(-\theta|s-t|),\\
\rho(s,t;3/2) &= \exp(-\theta|s-t|)(1+\theta|s-t|),\\
\rho(s,t;5/2) &= \exp(-\theta|s-t|)\Big(1+\theta|s-t|+\frac{1}{3}\theta^2|s-t|^2\Big), \quad \text{and}\\
\rho(s,t;7/2) &= \exp(-\theta|s-t|)\Big(1+\theta|s-t|+\frac{2}{5}\theta^2|s-t|^2)+\frac{1}{15}\theta^3|s-t|^3\Big).
\end{aligned}$$

Notice that $\nu = 0$ gives the exponential covariance. Letting $\nu \to \infty$ produces the squared exponential covariance. Figure 14.9 shows some sample paths.

To understand the connection between process smoothness and the correlation function, we proceed informally. We begin with the correlation between f at one point and a divided difference of f some place else:

$$\operatorname{cov}\Big(\frac{f(x+h)-f(x)}{h}, f(\tilde{x})\Big) = \frac{1}{h}\operatorname{cov}\Big(f(x+h)-f(x), f(\tilde{x})\Big).$$

Next, if we let $h \to 0$ on both sides, and are not rigorous about limits, we find that

$$\operatorname{cov}\Big(\frac{\mathrm{d}}{\mathrm{d}x}f(x), f(\tilde{x})\Big) = \frac{\mathrm{d}}{\mathrm{d}x}\rho(x,\tilde{x})$$

A similarly informal argument gives

$$\operatorname{cov}\Big(\frac{\mathrm{d}}{\mathrm{d}x}f(x), \frac{\mathrm{d}}{\mathrm{d}\tilde{x}}f(\tilde{x})\Big) = \frac{\mathrm{d}^2}{\mathrm{d}x\mathrm{d}\tilde{x}}\rho(x,\tilde{x}).$$

For stationary correlations this becomes

$$\operatorname{cov}\Big(\frac{\mathrm{d}}{\mathrm{d}x}f(x), \frac{\mathrm{d}}{\mathrm{d}\tilde{x}}f(\tilde{x})\Big) = \frac{\mathrm{d}^2}{\mathrm{d}x\mathrm{d}\tilde{x}}\rho(x-\tilde{x}) = -\rho''(x-\tilde{x})$$

and then

$$\operatorname{var}\Big(\frac{\mathrm{d}}{\mathrm{d}x}f(x)\Big) = -\rho''(0).$$

The exponential covariance $\rho(\delta) = \exp(-\theta|\delta|)$ is not twice differentiable at the origin and that resulting process does not have a derivative. For a properly rigorous account, see Stein (2012).

14.7 Interpolation, noise and nuggets

Suppose that we obtain $f(\boldsymbol{x}_1) \cdots f(\boldsymbol{x}_n)$ and some partial derivatives like $\partial f(\boldsymbol{x}_1)/\partial x_{1j}$ and even some integrals $\int_A f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. When f is a Gaussian process for which the derivative exist, then all of these quantities are jointly Gaussian along with any unmeasured function values, derivatives and integrals.

This is a very powerful property. It means that using the recipe

$$\mathbb{E}(\text{any unknowns}\,|\,\text{all the knowns})$$

We get a function $\tilde{f}$ that matches all known values and derivatives and integrals while also being consistently extendable to the unknown values and derivatives and integrals. It is quite common for automatic differentiation codes to deliver gradients along with function values. Kriging can make use of known gradients to better predict function values.

If there is some Monte Carlo sampling embedded in $f(\boldsymbol{x})$, then we may need to model f as having an unknown true value somewhat different from the value we computed. Let's suppose that $Y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ for measurement noise ε_i that is independent of the Gaussian process distribution of f. Then

$$\begin{aligned}\operatorname{cov}(Y_i, Y_{i'}) &= \operatorname{cov}(f(\boldsymbol{x}_i) + \varepsilon_i, f(\boldsymbol{x}_{i'}) + \varepsilon_{i'})\\ &= \operatorname{cov}(f(\boldsymbol{x}_i), f(\boldsymbol{x}_{i'}) + \operatorname{cov}(\varepsilon_i, \varepsilon_{i'})\\ &= \sigma_1^2 R(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) + \sigma_0^2 1_{i=i'}\end{aligned}$$

if we assume that ε_i are IID with mean zero and variance σ_0^2. Here $R(\cdot,\cdot)$ may be any one of our prior correlation functions.

Now suppose that there is sporadic roughness in $f(\cdot)$ such as small step discontinuities that we described as numerical noise above. We can model that using what is called a ***nugget effect***. The covariance is

$$\operatorname{cov}(f(\boldsymbol{x}_i), f(\boldsymbol{x}_{i'})) = \sigma_1^2 R(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) + \sigma_0^2 1_{\boldsymbol{x}_i = \boldsymbol{x}_{i'}}.$$

This looks a lot like the way we handled measurement noise above. What changed is that $1_{i=i'}$ has become $1_{\boldsymbol{x}_i=\boldsymbol{x}_{i'}}$. These would be different if we had two observations $i \neq i'$ with $\boldsymbol{x}_i = \boldsymbol{x}_{i'}$, that is replicates. A nugget effect is a kind of "reproducible noise".

In class we looked at and discussed Figures 1, 2 and 3 from Roustant et al. (2012). Figure 1 shows $\tilde{f}(x)$ and 95% confidence bands. It is simple kriging with a quadratic curve $11x + 2x^2$. The covariance is Matern with $\nu = 5/2$, $\theta = 0.4$ and $\sigma = 5$. Figure 2 shows three different θ. The leftmost panel of Figure 3 shows a mean reversion issue. As we move $\boldsymbol{x}_0$ away from the region sampled by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ then $\tilde{f}(\boldsymbol{x}_0)$ reverts towards (the esimated value of) μ. The dissertation Lee (2017) provides an alternative form of kriging that reverts toward the nearest neighbor.

The interpolating prediction for simple kriging is

$$\tilde{f}(\boldsymbol{x}_0) = \mu(\boldsymbol{x}_0) + c(\boldsymbol{x}_0)^{\mathsf{T}} C^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$$

where

$$\begin{aligned} \boldsymbol{Y} &= (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)^{\mathsf{T}}, \\ c(\boldsymbol{x}_0) &= (\Sigma(\boldsymbol{x}_1, \boldsymbol{x}_0), \ldots, \Sigma(\boldsymbol{x}_n, \boldsymbol{x}_0))^{\mathsf{T}}, \quad \text{and} \\ C &= \begin{pmatrix} \Sigma(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & \Sigma(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ \Sigma(\boldsymbol{x}_n, \boldsymbol{x}_1) & \ldots & \Sigma(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix}. \end{aligned}$$

This follows from the multivariate Gaussian model. For $\Sigma(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sigma^2 R(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ we can replace $\Sigma(\cdot, \cdot)$ by $R(\cdot, \cdot)$.

One difficulty with kriging is that the cost of the linear algebra ordinarily grows proportionally to n^3. This may be ok if $f(\cdot)$ is so expensive that only a tiny value of n is possible. Otherwise we might turn to polynomial chaos or quasi-regression. A second difficulty with kriging is that C is very often nearly singular. Indeed that is perhaps very good. For instance if $f(\boldsymbol{x}_n)$ is almost identical to the linear combination of $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_{n-1})$ that we would have used to get $\tilde{f}(\boldsymbol{x}_n)$ from the first $n-1$ points then C is nearly singular.

Ritter (1995) shows that kriging can attain prediction errors $\tilde{f}(\boldsymbol{x}_0) - f(\boldsymbol{x}_0)$ that are $O(n^{-r-1/2+\epsilon})$ for n evaluations of a function with r derivatives. Here $\epsilon > 0$ hides powers of $\log(n)$. The interpolations in kriging have some connections to classical numerical analysis methods that may explain why they work so well (Diaconis, 1988). When the process is Brownian motion $f(t) = B(t)$, then the predictions are linear splines (in one dimension). For a process that is once integrated Brownian motion $f(t) = \int_0^t B(x)\,\mathrm{d}x$, we find that $\tilde{f}$ is a cubic spline interpolator.

14.8 Optimization

One of the main uses of kriging is to approximately find the optimum of the function $f(\boldsymbol{x})$ on $\mathcal{X}$. That is, we might seek $\boldsymbol{x}_* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$.

Given $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$, we could estimate $\boldsymbol{x}_*$ by $\tilde{\boldsymbol{x}}_* = \arg\min_{\boldsymbol{x}} \tilde{f}(\boldsymbol{x})$ where $\tilde{f}(\boldsymbol{x}) = \mathbb{E}(f(\boldsymbol{x}) \mid f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))$. However, if we are still searching for $\boldsymbol{x}_*$ then $\tilde{\boldsymbol{x}}_*$ is not necessarily the best choice for $\boldsymbol{x}_{n+1}$. Suppose for instance that a confidence interval yields $f(\tilde{\boldsymbol{x}}_*) = 10 \pm 1$ while at a different location $\boldsymbol{x}'$ we have $f(\boldsymbol{x}') = 11 \pm 5$. Then $\boldsymbol{x}'$ could well be a better choice for the next function evaluation.

The DiceOptim package of Roustant et al. (2012) chooses $\boldsymbol{x}_{n+1}$ to be the point with the greatest ***expected improvement*** as described next. First let $f^* \equiv \min_i f(\boldsymbol{x}_i)$. Then, if $f(\boldsymbol{x}) < f^*$ we improve by $f^* - f(\boldsymbol{x})$. Otherwise, we improve by 0. The expected improvement at $\boldsymbol{x}$ is

$$\mathrm{EI}(\boldsymbol{x}) \equiv \mathbb{E}((f^* - f(\boldsymbol{x}))_+ \,|\, f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)).$$

There is a closed form expression for EI in terms of $\varphi(\cdot)$ and $\Phi(\cdot)$. We could then take

$$\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x}} \mathrm{EI}(\boldsymbol{x}).$$

Like bandit methods, this approach involves some betting on optimism. The optimization of EI may be difficult because that function could be very multimodal. However, if each evaluation of $f(\cdot)$ takes many hours, then there is plenty of time available to search for the optimum of EI, since it will ordinarily be inexpensive to evaluate. Figure 22 of Roustant et al. (2012) illustrates this process. That paper also describes how to choose multiple candidates for the improvement of EI in case they can be computed in parallel. Balandat et al. (2019) use randomized QMC in their search for the best expected improvement.

14.9 Further designs for computer experiments

Here we consider some optimal designs for computer experiments. There are figures illustrating some of those designs in the dissertation of Koehler and Owen (1996) as well Koehler and Owen (1996). One criterion is to maximize entropy, choosing $\boldsymbol{x}_i$ to

$$\min_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n} \mathbb{E}(-\log(p(\boldsymbol{Y}))).$$

Figures 8(abc) of Koehler and Owen (1996) show some of these for $d = 2$. Another criterion is

$$\min_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n} \int_{[0,1]^d} \mathbb{E}((\widetilde{f}(\boldsymbol{x}) - f(\boldsymbol{x}))^2)\, \mathrm{d}\boldsymbol{x}$$

to minimize the integrated mean square error (MISE). Figures 9(ab) show some examples for $d = 2$. The maximum entropy designs place points on the boundary of $[0, 1]^2$ while the MISE designs are internal.

Some other design approaches presented in Johnson et al. (1990) are called ***minimax*** and ***maximin*** designs. They are related to ***packing*** and ***covering*** as described in Conway and Sloane (2013).

The minimax design chooses $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n} \max_{\boldsymbol{x}_0 \in [0,1]^d} \|\boldsymbol{x}_0 - \boldsymbol{x}_i\|.$$

If you were placing coffee shops at points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ you might want to minimize the maximum distance that a customer (who might be at $\boldsymbol{x}_0$) has to go to get to one of your shops. We can also think of "covering" the cube $[0,1]^d$ with n disks of small radius and centers $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

The maximin design chooses $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to

$$\max_{1 \leqslant i \leqslant n} \min_{i' \neq i} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|.$$

In coffee terms, we would not want any two coffee shops to be very close to each other. We can think of this as successfully "packing" n disks into $[0,1]^d$ without any of them overlapping.

Johnson et al. (1990) show that minimax designs are G-optimal in the $\theta \to \infty$ limit. This is the limit in which values of $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$ most quickly become independent as $\boldsymbol{x}$ and $\boldsymbol{x}'$ move away from each other.

Park (1994) looks a ways to numerically optimize criteria such as the above among Latin hypercube sample designs.

14.10 Quasi-Monte Carlo

Quasi-Monte Carlo methods are designed for numerical integration. They can attain much better convergence rates than plain Monte Carlo does. Niederreiter (1992) and Dick and Pillichshammer (2010) provide comprehensive references.

In quasi-Monte Carlo methods we choose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to minimize a "discrepancy". There are many different notions of discrepancy. We can think of them as distances $\big\|\mathbb{U}[0,1]^d - \mathbb{U}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}\big\|$ between the discrete uniform distribution on our n points and the continuous uniform distribution on the unit cube. As such they have a "space filling" interpretation. See Figure 14.10.

14.11 Variable importance

Variable importance is a harder problem than causal inference. Even when there is absolutely no doubt about the causal effect of $\boldsymbol{x}$ on f we can still debate about ways to measure importance of variables and rank them. Variable importance involves the "causes of effects" that we avoided in the causal inference chapter in favor of studying "effects of causes".

Variable importance is about the extent to which $f(\boldsymbol{x})$ changes when one component x_j is changed. There are local versions called ***sensitivity analysis*** about small changes to components of $\boldsymbol{x}$. There are also global sensitivity analysis methods where components of $\boldsymbol{x}$ are change completely at random.

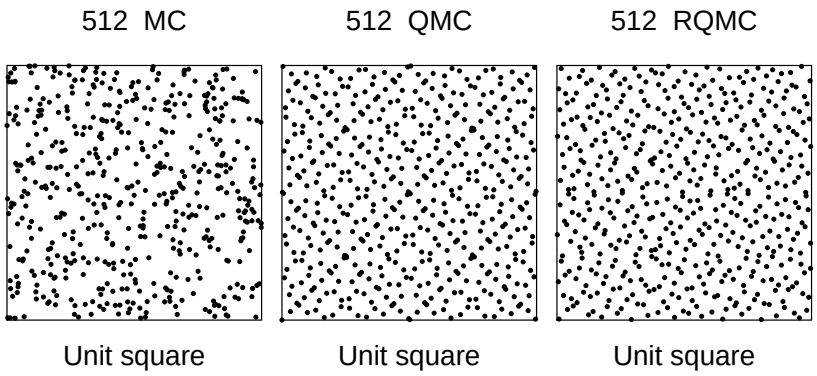


Figure 14.10: The left panel shows 1024 random points. The middle panel shows a Sobol' sequence. The right panel shows some scrambled Sobol' points.

For an introduction to global sensitivity analysis, see the book by Saltelli et al. (2008). Some of the key quantities are Sobol' indices defined in Sobol' (1993).

There is currently a lot of interest in explaining black box functions, such as those in artificial intelligence and machine learning. See Molnar (2018), Ribeiro et al. (2016) and Lundberg and Lee (2017) for an entry to that literature. For some methods that take special care about dependent input variables, see Owen and Prieur (2017) and Mase et al. (2019).

15

Guest lectures and hybrids of experimental and observational data

We had two guest lectures by people using experimental design and developing new methods to handle the new problems. We also had a lecture on methods to mix some randomization in with what would otherwise be an observational causal inference.

15.1 Guest lecture by Min Liu

We had a lecture by Min Liu from LinkedIn. Min Liu has an M.S. in Statistics from Stanford where she took this course. Her talk was entitled "Online Experimentation at LinkedIn". They face great challenges in measuring the causal impact of changes to their product.

People with LinkedIn accounts (members) are connected to each other. Changes to the experience of one member might affect behavior of others. That may then produce al SUTVA violation.

Very small and hard to detect effect sizes can be economically meaningul because of the scale (675 million members at the time of that presentation).

There are over 3000 different metrics to track when deciding whether to launch a new feature or not. It is not reasonable to expect a change that improves some metrics to not be detrimental to some others.

They like where possible to get an experiment to completion within two weeks.

They need to go beyond mean responses and they find that quantiles are very useful. For instance a variable like page load time is important to the user experience. Raising all page load times by 0.5 seconds is meaningfully

different from raising 10% of them by 5 seconds. They compare 50'th and 90'th percentiles within the A and B pupulations. Comparing quantiles is more complicated than comparing means and bootstrapping is too slow at scale.

Some networks can be chopped up in to pieces that barely overlap at all and then treatments can be randomized to those pieces. This becomes very difficult in networks of people where some may have thousands of neighbors.

A second SUTVA violation arises in two-sided markets visualized as bipartite graphs. Think of links between advertisers and members. An experiment on one side of the graph can affect participants on the other and indirectly spill over to the first side. What that means is that, for instance, a difference observed between members in treatment and control groups might not end up as the real difference seen when making the change for all members.

15.2 Guest lecture by Michael Sklar

We had a lecture by Michael Sklar, a PhD candidate at Stanford working with Professor T.-L. Lai, entitled "Trial Design for Precision Medicine + Applications to Oncology". His lecture focused on the high and rising costs of pharmaceutical research in the US and how this is spurring the development of new complex experimental trial designs. There is an especially great need for new designs for cancer drugs because drug development for oncology has an unusually low success rate (3 percent versus 20 percent outside of oncology).

One method he described is the ***basket trial*** where for intance one drug is tested against multiple cancer type within one experiment. Another is the ***umbrella trial*** in which multiple drugs are tested against one cancer type. The third kind was the ***platform trial*** where, similarly to a bandit method, the protocol calls for algorithmic addition or exclusion of new treatment arms over time. A platform trial might also be a basket trial or an umbrella trial. The term ***master protocol*** is used to describe basket, umbrella and platform trials.

15.3 First hybrid

Sometimes we have observational and experimental data on the same phenomenon. It would be worthwhile to use them both together, especially if the resulting method is better than either of them on their own.

In other settings we might face resistance to doing an experiment. It may then still be possible to inject a small amount of randomness into a plan to gather data.

This lecture presented results from Rosenman et al. (2018) on merging a small experimental data set with a larger observational one. The motivating setting is that a large insurer or national health organization might have enormous observational records along with a small randomized clinical trial on the same disease.

One of the methods was based on a causal inference approach involving ***propensity scores***. The propensity $e(\boldsymbol{x}) = \Pr(W = 1 \,|\, \boldsymbol{x})$ is simply the chance of getting an experimental treatment given the covariates $\boldsymbol{x}$. See Imbens and Rubin (2015) for an explanation of how propensity methods can be used to estimate a causal claim as well as the additional assumptions one must make in order for the causal interpretation to be justified. One approach to estimating the causal effect of a treatment is to stratify a population based on their values $e_i = e(\boldsymbol{x}_i)$. The treatment effect in each stratum is estimated by the simple difference between average Y values for control and treated stratum members. The overall treatment effect is a weighted average of stratum values.

The first proposal in Rosenman et al. (2018) is to simply find ***counterfactual propensities*** $e(\boldsymbol{x}_i)$ for subjects i in the randomized trial. Those subjects are then added to the corresponding propensity strata of the observational data and contribute to the averages there. This is called the ***spike-in*** method. There are several other proposed methods some designed to fix possible biases in the spike-in method.

The ***Women's Health Initiative*** has data of this type relating hormone therapy to coronary heart disease (among other responses). It was a good test case for these methods because it had both observational and experimental studies of this issue. Furthermore, the experiment was large enough that it could be split into two subsets, with one of them held out to define the true treatment effect and the other combined with the observational data to estimate that effect. The spike-in method turned out to have less bias than simply using the large observational data set and less variance than using just a smallish experiment and less mean squared error than either study had on its own.

15.4 Second hybrid

The second hybrid method from that lecture was about the tie-breaker design as analyzed by Owen and Varian (2020). That design inserts some randomness into a ***regression discontinuity design*** or RDD. In an RDD we have an assigment variable x with a threshold t. Subjects with $x_i > t$ get the treatment while subjects with $x_i \leqslant t$ get the control. In an observational setting we might suppose that subjects with x_i barely larger than t are almost the same as subjects with x_i barely smaller than t at least in terms of how they would respond to the treatment. An RDD then looks for a discontinuity in the regression function $\mu(t) = \mathbb{E}(Y \,|\, x)$ at the point $x = t$. The size of the discontinuity may have a causal interpretation. See Imbens and Rubin (2015) for more.

In a tie-breaker design there are potentiallytwo thresholds A and B with $A \leqslant t \leqslant B$. If $x_i \leqslant A$ then subject i gets control. If $x_i \geqslant B$ then subject i gets treatment. If $A < x_i < B$ then subject i gets the treatment with probability 1/2. Tie-breaker designs have been used to award scholarships (!).

The paper Owen and Varian (2020) was motivated by loyalty reward programs that companies might offer to their best customers. For instance they might offer an upgrade to the top 10% of customers ranked in some way appro-

priate to the business. In a tie-breaker they could offer it instead to the top 5% of customers and randomly to half of the next 10% of customers.

The analysis in Owen and Varian (2020) shows that statistical efficiency is monotonically increasing in the amount of experimentation. Of course there is a cost issue preventing one from just making all of the awards at random. The paper analyzes that tradeoff. It also shows that there is no benefit to making the award probability take values other than 0%, 50% or 100%, perhaps on a sliding scale.

16

Wrap-up

The final lecture of the class had several of the students presenting their final projects. These were about understanding or tuning household tasks like cuisine or entertaining young children or morning wakeup rituals or hobbies such as horticulture. It was very nice to see a range of design ideas. From a survey of the class it seemed that fractional factorials and analysis of covariance ideas turned out to be most widely used.

Prior to those examples was a short note summarizing the topics of the course. That was preceded by a brief overview of the statistics problem in general.

16.1 What statistics is about

At a very high level view, our primary challenge in statistics is to say something about numbers we don't have using numbers we do have. In prediction settings we want to know about future values of Y for some $\boldsymbol{x}$, using past $(\boldsymbol{x}_i, Y_i)$ data. Other settings manifest differently but we are still using known values to say something about unknowns. Phrased this way, our task seems at first like it might be impossible.

We connect our knowns to unknowns by choosing a model that we can think of as generating both kinds of data as depicted in the left panel of Figure 16.1. The issue of ***external validity*** that we frequently raised involves the model not changing between those two settings. ***Internal validity*** is then about the model holding for the data we do have. In statistical inference, we reverse the direction of one of the arrows, letting us learn something about the model from the known data. This is a problem of ***inductive inference*** describing the general model from the specific known data. Using what we know about the

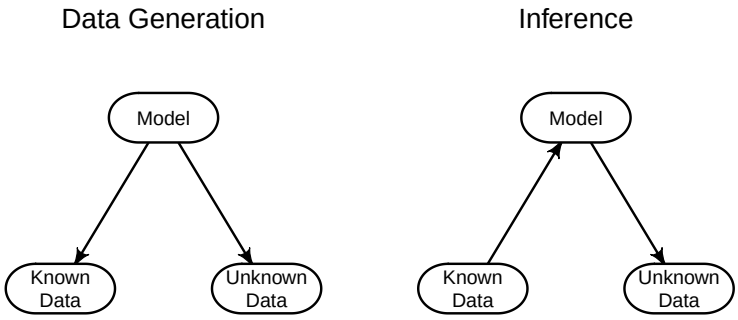


Figure 16.1: The left figure shows how we envision a one and the same statistical model produces both our known data and some unknown data of interest. In inference we reverse the arrow from the model to the known data. Then the known data tell us something about the model (with some quantified uncertainty). From that we can derive consequences about the unknown data.

model, ***deductive inference*** lets us derive consequences for the unknown data. Induction leaves us with some uncertainty about the model. When we derive something about the unknown data we can propagate that uncertainty.

One could argue that it is logically impossible to learn the general case from specific observations. For a survey of the problem of induction in philosophy, see Henderson (2018). We do it anyway, accepting that the certainty possible in mathematics may not be available in other settings.

A famous observation from George Box is that all models are wrong, but some are useful. Nobody, not even Box, could give us a list of which one is useful when. As applied statisticians, it is our responsibility to do that in any case that we handle. There are settings where we believe that we can get a usable answer from an incorrect answer. Sometimes we know that small errors in the model will yield only small errors in our inferences. This is a notion of ***robustness***. In other settings we can get consequences from our model that can be tested later in new settings. This is a notion of ***validation***, as if we were doing "guess and check". Then, even if the model had errors we can get a measure of how well it performs.

There are approaches to inference that de-emphasize or perhaps remove the model. We can imagine the path being like the right hand side of Figure 16.2 that avoids the model entirely and then unlocks more computational possibilities. There may well be an implicit model there, such as unknown $(\boldsymbol{x}, Y)$ pairs being IID from the same distribution that produced the known ones. IID sampling would provide a justification for using cross-validation or the bootstrap. For a discussion of the role of models in statistics and whether they are really necessary, see Breiman (2001).

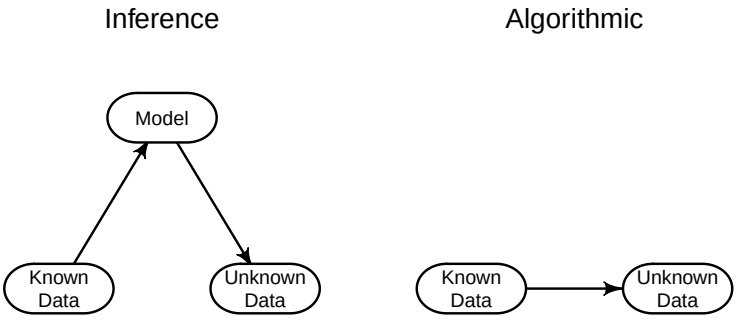


Figure 16.2: Sketch of an algorithmic approach to learning about unknown data from known data. There is only a remnant of a model.

The setting we considered in this course relates to the usual inference problem as shown in Figure 16.3. We were down in the lower left hand corner looking at how to make the data that would then be fed into the inferential machinery.

16.2 Principles from experimental design

Statistical inference from data faces several obstacles:
1) cost of data,
2) confounding of causes,
3) correlation of predictors,
4) noise,
5) interactions,
6) missing variables, and
7) external validity.
In the face of these obstacles, experimental design offers the following techniques:
a) Randomization,
b) Blocking and balancing,
c) Factorials,
d) Fractional replication,
e) Covariate adjustments (ancova),
f) Adaptation (bandits and sequential DOE),
g) Nesting and split-plots,
h) Random effects models, and
i) Replication.

Obstacle 1, the cost of data, is often forgotten in data analysis because, once the data are available that cost is not pertinent to its analysis. It is a sunk cost. Cost is important in experimental design and many of the designs we saw were chosen to reduce that cost. For instance in fractional factorial experiments,

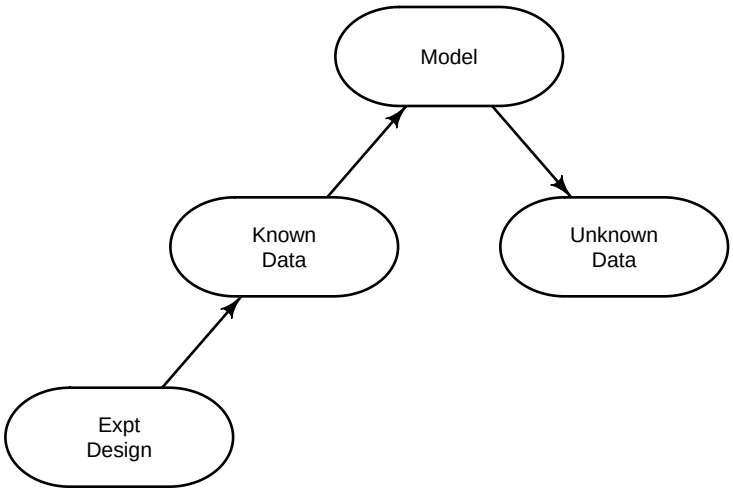


Figure 16.3: The place held by experimental design in statistical inference.

we would purposely sacrifice statistical correctness (e.g., an unbiased variance estimate) in order to study more variables at a fixed cost. Nesting and split-cost designs take advantage of the fact that some experimental factors are cheaper to change than others. Adaptive sampling via bandits is cost driven. It can reduce the cost incurred on the experimental units themselves. Sequential experiments also counter the cost of continuing to experiment once the better treatment has been clearly identified.

Confounding of potential causes was one of the primary motivations for the use of randomization. Randomization reduces the risk that some important cause other than the treatment is perfectly or even strongly associated with the treatment. Missing or unmeasured variables interfere with causal claims. We can think of them as continuously varying quantities that cause similar problems to confounding. Randomization ensures that those missing variables cannot be strongly associated with the treatment.

Correlated predictors can be problematic in regression settings because they make it harder to tell which variable is important. Many of the designs we studied produced perfectly uncorrelated predictors.

Regression methods average away the noise. In optimal designs we found ways to minimize the effect of noise on the variance of regression coefficients. Replication raises the sample size and thus helps us reduce noise. We also used blocking methods to get better comparisons of "like with like" by arranging that treatment and control would both be applied to similar experimental units,

with similarity defined by one or more categorical variables. The analysis of covariance was useful to balance out impacts of continuous variables.

Interactions severely complicate interpretation of the effect of variables. We looked at factorial designs that allowed us to estimate those interaction effects.

External validity is critical problem for causal inferences. Having an experiment contain a wide variety of settings helps to improve its external validity. Random effect models also provide greater external validity.

Bibliography

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135.

An, J. and Owen, A. (2001). Quasi-regression. *Journal of complexity*, 17(4):588–607.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press, Princeton, NJ.

Atkinson, A., Donev, A., and Tobias, R. (2007). *Optimum experimental designs, with SAS*. Oxford University Press.

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. (2019). Botorch: Programmable Bayesian optimization in PyTorch. Technical report, arXiv:1910.06403.

Berger, P. D., Maurer, R. E., and Celli, G. B. (2018). *Experimental Design*. Springer, Cham, Switzerland, second edition.

Blatman, G. and Sudret, B. (2010). Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering & System Safety*, 95(11):1216–1229.

Booth, K. H. and Cox, D. R. (1962). Some systematic supersaturated designs. *Technometrics*, 4(4):489–495.

Bose, R. C. (1938). On the application of the properties of Galois fields to the problem of construction of hyper-Graeco-Latin squares. *Sankhyā: The Indian Journal of Statistics*, pages 323–338.

Box, George, E., Hunter, J. S., and Hunter, W. G. (2005). *Statistics for experimenters: design, innovation, and discovery*. Wiley, New York.

Box, G. E. (1957). Evolutionary operation: A method for increasing industrial productivity. *Journal of the Royal Statistical Society: Series C*, 6(2):81–101.

Box, G. E. and Behnken, D. W. (1960). Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475.

Box, G. E. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons, New York.

Box, G. E., Hunter, W. H., and Hunter, S. (1978). *Statistics for experimenters*, volume 664. John Wiley and sons, New York.

Box, G. E. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B*, 13(1):1–38.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231.

Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics*, pages 69–79.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Budne, T. A. (1959). The application of random balance designs. *Technometrics*, 1(2):139–155.

Byrne, D. M. and Taguchi, S. (1987). The taguchi approach to parameter design. *Quality progress*, 20(12):19–26.

Caflisch, R. E., Morokoff, W., and Owen, A. B. (1997). Valuation of mortgage backed securities using Brownian bridges to reduce effective dimension. *Journal of Computational Finance*, 1(1):27–46.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21(2):191–208.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons, New York.

Constantine, P. G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM, Philadelphia.

Conway, J. H. and Sloane, N. J. A. (2013). *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media.

Cornell, J. A. (20002). *Experiments with mixtures: designs, models, and the analysis of mixture data*. John Wiley & Sons, New York, third edition.

Cox, D. R. (1958). *Planning of experiments*. Wiley, New York.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1(4):311–341.

Diaconis, P. (1988). Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics IV, in two volumes*, volume 1, pages 163–176.

Diaconis, P. and Gangolli, A. (1995). Rectangular arrays with fixed margins. In Aldous, D. Diaconis, P. S. J. and Steele, J. M., editors, *Discrete probability and algorithms*, volume 72. Springer, New York.

Dick, J. and Pillichshammer, F. (2010). *Digital sequences, discrepancy and quasi-Monte Carlo integration*. Cambridge University Press, Cambridge.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9(3):586–596.

Fang, K.-T., Li, R., and Sudjianto, A. (2006). *Design and modeling for computer experiments*. CRC press, Boca Raton, FL.

Fisher, R. and Mackenzie, W. (1923). Studies in crop variation: The manurial response of different potato varieties. *Journal of Agricultural Sciences*, 13:311–320.

Fleiss, J. L. (1986). *Design and analysis of clinical experiments*. John Wiley & Sons, New York.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer, New York.

Georgiou, S. D. (2014). Supersaturated designs: A review of their construction and analysis. *Journal of Statistical Planning and Inference*, 144:92–109.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B*, 41(2):148–164.

Hamada, M. and Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica*, pages 1–28.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal arrays: theory and applications*. Springer Science & Business Media, New York.

Henderson, L. (2018). The problem of induction. In *The Stanford Encyclopedia of Philosophy*. `https://plato.stanford.edu/entries/induction-problem/`.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325.

Imbens, G. (2019). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Technical report, National Bureau of Economic Research.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jiang, T. and Owen, A. B. (2003). Quasi-regression with shrinkage. *Mathematics and Computers in Simulation*, 62(3-6):231–241.

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of statistical planning and inference*, 26(2):131–148.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189–206):1.

Jones, B. and Kenward, M. G. (2014). *Design and analysis of cross-over trials*. CRC press.

Jones, B. and Nachtsheim, C. J. (2009). Split-plot designs: What, why, and how. *Journal of quality technology*, 41(4):340–361.

Khuri, A. I. and Mukhopadhyay, S. (2010). Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(2):128–149.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Kirk, R. E. (2013). *Experimental design: Procedures for the Behavioral Sciences*. SAGE Publications, Inc., Thousand Oaks, CA.

Koehler, J. R. and Owen, A. B. (1996). Computer experiments. *Design and analysis of experiments*, 13:261–308.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181.

Kohavi, R., Tang, D., and Xu, Y. (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.

Krahmer, F. and Ward, R. (2011). New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer-Verlag, New York.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lee, M. R. (2017). *Prediction and Dimension Reduction Methods in Computer Experiments*. PhD thesis, Stanford University.

Lee, M. R. and Shen, M. (2018). Winner's curse: Bias estimation for total effects of features in online controlled experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 491–499.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4):469–473.

Lewis, R. A. and Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973.

Li, X. and Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Lin, D. K. (1993). A new class of supersaturated designs. *Technometrics*, 35(1):28–31.

Lin, D. K. (1995). Generating systematic supersaturated designs. *Technometrics*, 37(2):213–225.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Mase, M., Owen, A. B., and Seiler, B. (2019). Explaining black box decisions by shapley cohort refinement. Technical report, arXiv:1911.00467.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

Miller, R. G. (1997). *Beyond ANOVA: basics of applied statistics*. CRC press, Boca Raton, FL.

Molnar, C. (2018). *Interpretable machine learning: A Guide for Making Black Box Models Explainable*. Leanpub.

Montgomery, D. C. (1997). *Design and analysis of experiments*. John Wiley & sons, 4 edition.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*, volume 29. Oxford University Press, USA.

Myers, R. H., Khuri, A. I., and Carter, W. H. (1989). Response surface methodology: 1966–l988. *Technometrics*, 31(2):137–157.

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons, New York, fourth edition.

Nair, V. N., Abraham, B., MacKay, J., Box, G., Kacker, R. N., Lorenzen, T. J., Lucas, J. M., Myers, R. H., Vining, G. G., Nelder, J. A., Phadke, M. S., Sacks, J., Welch, W. J., Shoemaker, A. C., Tsui, K. L., Taguchi, S., and Wu, C. F. J. (1992). Taguchi's parameter design: a panel discussion. *Technometrics*, 34(2):127–161.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. S.I.A.M., Philadelphia, PA.

Owen, A. B. (1992). Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, 2(2):439–452.

Owen, A. B. (2017). Confidence intervals with control of the sign error in low power settings. Technical report, Stanford University.

Owen, A. B. (2020). Monte Carlo theory, methods and examples. `https://statweb.stanford.edu/~owen/mc/`.

Owen, A. B. and Launay, T. (2016). Multibrand geographic experiments. Technical report, arXiv:1612.00503.

Owen, A. B. and Prieur, C. (2017). On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002.

Owen, A. B. and Varian, H. (2020). Optimizing the tie-breaker regression discontinuity design. *Electronic Journal of Statistics*, 14(2):4004–4027.

Paley, R. E. (1933). On orthogonal matrices. *Journal of Mathematics and Physics*, 12(1-4):311–320.

Park, J.-S. (1994). Optimal Latin-hypercube designs for computer experiments. *Journal of statistical planning and inference*, 39(1):95–111.

Patterson, H. D. (1954). The errors of lattice sampling. *Journal of the Royal Statistical Society, Series B*, 16(1):140–149.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

Phadke, M. S. (1989). *Quality engineering using robust design*. Prentice Hall, New York.

Phoa, F. K., Pan, Y.-H., and Xu, H. (2009). Analysis of supersaturated designs via the Dantzig selector. *Journal of Statistical Planning and Inference*, 139(7):2362–2372.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, New York. ACM.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Duxbury, third edition.

Ritter, K. (1995). *Average case analysis of numerical problems*.

Rosenman, E., Owen, A. B., Baiocchi, M., and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. Technical report, arXiv:1804.07863.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1).

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, pages 409–423.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd, New York.

Santner, T. J., Williams, B. J., and Notz, W. I. (2018). *The design and analysis of computer experiments*. Springer, New York, second edition.

Satterthwaite, F. (1959). Random balance experimentation. *Technometrics*, 1(2):111–137.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114.

Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance components*. John Wiley & Sons, New York.

Shirley, P. S. (1991). *Physically based lighting calculations for computer graphics*. PhD thesis, University of Illinois at Urbana-Champaign.

Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85.

Sobol', I. M. (1969). *Multidimensional Quadrature Formulas and Haar Functions*. Nauka, Moscow. (In Russian).

Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414.

Spall, J. C. (2003). *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, Hoboken, NJ.

Stein, M. L. (1989). Design and analysis of computer experiments: Comment. *Statistical Science*, 4(4):432–433.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Surjanovic, S. and Bingham, D. (2013). Virtual library of simulation experiments: test functions and datasets. `https://www.sfu.ca/~ssurjano/`.

Tang, B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American statistical association*, 88(424):1392–1397.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Vaver, J. and Koehler, J. (2011). Measuring ad effectiveness using geo experiments. Technical report, `https://research.google/pubs/pub38355/`.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons.

Yandell, B. S. (1997). *Practical data analysis for designed experiments*. CRC Press.

Yates, F. (1937). The design and analysis of factorial experiments. Technical Report 35, Imperial Bureau of Soil Science.

Youden, W., Kempthorne, O., Tukey, J. W., Box, G., and Hunter, J. (1959). Discussion of the papers of Messrs. Satterthwaite and Budne. *Technometrics*, 1(2):157–184.