Contents

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
- there are adversarial elements, e.g., bots and spammers
- the users may be quite different from the data scientists and engineers
- 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.

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 $n_1/n \overset{\cdot}{\sim} \mathcal{N}(1/2, n/4)$). 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

$$\mathbb{E}\biggl(\sum_{j=1}^J(\hat\tau_j - \tau_j)S_j\biggr) = \sum_{j=1}^J \int_{Z^{1-\alpha_j}\sigma_j}^{\infty} \frac{\hat\tau - \tau}{\sigma_j}\varphi\Bigl(\frac{\hat\tau_j - \tau_j}{\sigma_j}\Bigr)\,\mathrm{d}\hat\tau_j > 0,$$

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. 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.

The original title of their article was “On the Near Impossibility of Measuring the Returns to Advertising”.

Bibliography

Berger, P. D., Maurer, R. E., and Celli, G. B. (2018). *Experimental Design*. Springer, Cham, Switzerland, second edition.

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.

Kohavi, R., Tang, D., and Xu, Y. (2020). *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press.

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.

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.

Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media.

Vaver, J. and Koehler, J. (2011). Measuring ad effectiveness using geo experiments. Technical report, `https://research.google/pubs/pub38355/`.