Beyond A/B Testing

Getting More from Experiments

Bob Wilson (he/him/his)
Data Scientist @ Netflix

May 3, 2023

Outline

Sample Ratio Mismatch

- Scenario: A/B test with 50/50 test/control split.
- Observation: 52K people in test group, 48K in control.

Sample Ratio Mismatch

- Scenario: A/B test with 50/50 test/control split.
- Observation: 52K people in test group, 48K in control.
  - Highly implausible to have this imbalance (p=1e-36).
- Conclusion: Something went wrong; should investigate cause.

Guardrail Metrics

- Not the primary measure of interest in a test.
- Ensures test was conducted properly or surfaces unintended consequences.
- Sample Ratio Mismatch (SRM) is a guardrail metric that can highlight problems with the test.
    - (Kohavi, Tang, and Xu 2020)

Sample Ratio Mismatch as a Guardrail Metric

- Flawed Approach: test the null hypothesis the sample ratio matches the test design.
  - Rejecting this null hypothesis proves the test deviated from the design.
  - Flaw: failing to reject the null hypothesis does not prove the test matched the design.
- Asymmetric Roles of Null and Alternative Hypotheses:
  - Rejecting the null proves the alternative.
  - Failing to reject the null does not prove the null; it proves nothing.
  - Always set the alternative hypothesis to be the thing you want to prove.
- Want to prove sample ratio, q, matched test design.
  - Null Hypothesis: $q < 0.5 - \epsilon$ or $q > 0.5 + \epsilon$.
  - Alternative: $0.5 - \epsilon \leq q \leq 0.5 + \epsilon$.
  - Equivalence test.

Equivalence Testing

- Equivalence test at level α:
  - (Wellek 2010)
  - Calculate a $100(1 - 2 \cdot \alpha)\%$ confidence interval on the sample ratio.
  - Check if confidence interval entirely within $(0.5 - \epsilon, 0.5 + \epsilon)$.
  - Choose fairly large $\epsilon \approx 0.01$, say:
    - Power should be $\gg 80\%$, say 99% or 99.9%.
    - Small imbalances are fine; want to highlight big imbalances while avoiding false alarms.

"Do No Harm" Guardrails

- Every A/B test has one or more (ideally exactly one) evaluation criteria.
- Hopefully thing being tested improves the evaluation criteria.
- Any change may have unintended consequences.
  - Detect and surface as risks.
- Suggestions: engagement, retention, conversion rates, ...
- Don't care about *improving* these.
  - Want to avoid harming them.

Non-Inferiority Testing

- Goal: prove change did not harm guardrail metric.
  - Set alternative hypothesis to be the thing we want to prove.
- Null hypothesis is that change *did* harm metric.
  - Requires margin parameter, $\epsilon > 0$.
  - $H_0 : \text{effect} < -\epsilon$.
  - Set fairly large to avoid false alarms.
  - Want high power!

Summary So Far

- Want to reject null hypothesis that at least one guardrail violated.
  - Alternative: all guardrails satisfied.
- When all guardrails satisfied, statistical power is the probability we correctly conclude that.
  - Want power very high.
- Typical adjustment for multiple comparisons reduces power both for guardrail metrics and primary evaluation criteria.

Intersection-Union Test

- (Berger 1982)
  - Idea: test hypotheses sequentially at the nominal levels (as if there were no multiple comparisons), but if we fail to reject a null hypothesis, don't evaluate any other metrics (including primary evaluation criterion).
  - If we reject each individual null, conclude all guardrails satisfied and evaluate primary criterion.
  - Controls the overall Type-I error rate (Rosenbaum 2008).
- Justification: if something has gone wrong with the test, conclusions likely not valid or of secondary importance.
  - More complicated *closed testing* approaches allow for still testing the primary criteria, but with less power to adjust for multiple comparisons (Wiens and Dmitrienko 2005).

Key Takeaways

- Guardrail metrics increase confidence:
  - Test was administered properly.
  - No unintended side-effects.
- Use the right test for the objective.
  - Alternative hypothesis matches what we want to prove.
  - Null hypothesis is the logical complement.
- Closed testing procedures avoid inflating Type-I error rate or reducing power.

Further Reading

- Berger, Roger L. 1982. "Multiparameter Hypothesis Testing and Acceptance Sampling." *Technometrics* 24 (4). Taylor & Francis, Ltd., American Statistical Association, American Society for Quality: pg. 295–300.
- Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments*. Cambridge University Press.
- Rosenbaum, Paul R. 2008. "Testing Hypotheses in Order." *Biometrika* 95 (1). Oxford University Press, Biometrika Trust: pg. 248–52.
- Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. 2nd ed. CRC Press.
- Wiens, Brian L., and Alexei Dmitrienko. 2005. "The Fallback Procedure for Evaluating a Single Family of Hypotheses." *J Biopharm Stat* 15 (6): pg. 929–42.

One-Sided Noncompliance

- How does sample ratio mismatch occur?
  - One possibility: server outages.
- Example: Feature Gate.
  - Release new feature to some people and not others, at random.
  - Requires coordination between client and server.
  - Server outages can cause *one-sided noncompliance:* people receiving a treatment that doesn't match their intended group assignment.
- Other examples of one-sided noncompliance:
  - Encouraging people to use new feature.
  - Marketing/Holdout split for ROI measurement.
    - People in Marketing group may not actually see ad.

As-Treated and Per Protocol Analyses are Flawed

- As-Treated:
  - Ignore treatment assignment.
  - Compare behavior of people who use feature with people who don't use feature.
- Per Protocol:
  - Ignore people in holdout who use feature,
  - or people in treatment who don't use feature/see ad.
- As-Treated and Per Protocol provide biased estimate of treatment effect.
  - Features correlated with noncompliance and outcome confound comparison.
  - (Imbens and Rubin 2015, sec. 23.9)

Intent-to-Treat Analysis

- Intent to Treat: Ignore noncompliance.
  - Treatment/control split is randomized, so no bias.
- Measures impact of treatment assignment, not feature use or ad exposure
  - For minor noncompliance, basically the same.
- Intent to Treat measures the wrong thing, but what it measures is unbiased.

Potential Exposures

- Let $Z_i(0) = 1$ if individual i would be exposed to feature if assigned to holdout group and 0 otherwise.
- Let $Z_i(1) = 1$ if individual i would be exposed to feature if assigned to treatment group and 0 otherwise.
  - One-sided noncompliance $\Rightarrow Z_i(0) = 0$ for everyone or $Z_i(1) = 1$ for everyone (depending on type of noncompliance).
- Know counterfactual exposure for one group.
  - Can only speculate about counterfactual exposure for other group.

Compliers and Non-Compliers

- If $Z_i(0) \neq Z_i(1)$, call person i a "complier".
  - Otherwise, $Z_i(0) = Z_i(1)$ and call person i a "non-complier".
  - We know the categories for one group but not the other.
- By random assignment, the proportion of compliers and non-compliers roughly equal in both treatment and holdout groups.

Potential Outcomes

- Let $Y_i(0)$ be the outcome we would observe for person i if assigned to holdout.
- Let $Y_i(1)$ be the outcome we would observe for person i if assigned to treatment.
- We observe the potential outcome corresponding to the group assignment.
  - We can only speculate about the counterfactual outcome.
- Effect of treatment on person i is $Y_i(1) - Y_i(0)$.
  - Individual treatment effects are unobservable!
  - Can only estimate average treatment effects across multiple people.

Instrumental Variables

- Intent to Treat estimate averages over compliers and non-compliers:

$$\text{ITT} = \text{Impact}_{\text{compliers}} \cdot \pi_{\text{compliers}} + \text{Impact}_{\text{non-compliers}} \cdot \pi_{\text{non-compliers}}.$$

$$\pi_{\text{compliers}} + \pi_{\text{non-compliers}} = 1.$$

- Suppose we believe $\text{Impact}_{\text{non-compliers}} = 0$.

$$\Rightarrow \text{ITT} = \text{Impact}_{\text{compliers}} \cdot \pi_{\text{compliers}}$$

$$\Rightarrow \text{Impact}_{\text{compliers}} = \frac{\text{ITT}}{\pi_{\text{compliers}}}.$$

Partial Identification

- Cannot infer average treatment effect since cannot infer effect on non-compliers.
- If outcome bounded (say, binary), effect on non-compliers is bounded (say, between -1 and +1).
- Manski-based uncertainty interval on average treatment effect (Manski 2003):

$$\text{Impact}_{\text{compliers}} \cdot \pi_{\text{compliers}} \pm \pi_{\text{non-compliers}}.$$

Dose-Response Models

- Suppose noncompliance is temporary.
  - $\Rightarrow$ effect on non-compliers non-zero, but still small.
  - Insight: treatment effect proportional to exposure:

$$Y_i(1) - Y_i(0) = \beta \cdot (Z_i(1) - Z_i(0)) \tag{1}$$

  - Works for 2-sided non-compliance as well.
- Randomization Inference:
  - (Rosenbaum 2020, sec. 18.4)
  - Rearrange (1):

$$Y_i(1) - \beta \cdot Z_i(1) = Y_i(0) - \beta \cdot Z_i(0).$$

  - To test null hypothesis $\beta = \beta_0$, form adjusted responses in each group and test equality of distributions (t-test, Wilcoxon rank sum, ...).

Key Takeaways

- Easy to adjust for minor noncompliance
- Don't use As-Treated or Per Protocol!
- Intent to Treat measures the wrong thing, but what it measures is unbiased
- If we assume the impact on non-compliers is zero, can estimate effect on compliers.
  - Average treatment effect *partially identified.*
- More nuanced models for usage-impact possible.

Further Reading

- Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Katsimerou, Christina. 2021. "Leveraging Proxy Variables for Causal Inference." `https://booking.ai/leveraging-proxy-variables-for-causal-inference-9e42781`
- Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. Springer Series in Statistics.
- Rosenbaum, Paul R. 2020. *Design of Observational Studies*. 2nd ed. Springer Series in Statistics.
- Wilson, Bob. 2023. "Tests with One-Sided Noncompliance." `https://www.adventuresinwhy.com/post/instrumental_variables/`

What are Heterogeneous Treatment Effects?

- Humans are complex, diverse creatures.
- Different people may react very differently to the same treatment or experience.
- Heterogeneous Treatment Effect (HTE) estimation attempts to model these differences.
    - Also known as Uplift Modeling or Conditional Average Treatment Effects (CATE).

Why are HTEs Important?

- Chemotherapy is beneficial for (some) people with cancer.
  - Harmful for people without cancer!
  - Knowing who is likely to benefit from a treatment is essential.
- Consider two new vitamins in development:
  - Vitamin 1 is slightly beneficial for everyone.
  - Vitamin 2 is extremely beneficial for some and harmful for others, but on average it's slightly beneficial.
  - Vitamin 2 broadly dangerous, but safe when targeted appropriately.
- Understanding HTEs highlights risks associated with a treatment.
  - Permits targeting treatment at those who benefit the most.

Use Cases for HTEs

- Targeted discounts:
  - Discounts are counterproductive for people who were planning to purchase already.
  - Target discounts at people whose uplift in conversion rate makes up for the decrease in revenue.
- Personalized ads:
  - Different ads will resonate best with different people.
  - Digital platforms use bandit algos to target the right ad to the right person.

Machine Learning and HTE Estimation

- Suppose we observed the treatment effect for each individual as well as various covariates like age, gender, hobbies, . . .
  - Use ML to fit a model predicting treatment effect based on observed features.
- Individual treatment effects are unobservable.
  - Instead, we run an A/B test with treatment and control arms,
  - compare outcomes between groups to estimate average effect,
  - and incorporate covariates to assess HTE.

Two-Model Approach

- Simple, intuitive, but not efficient strategy:
  - Fit a model using covariates, x, to predict outcome in treatment group, $\hat{f}_1(x)$.
  - Fit another model for control group, $\hat{f}_0(x)$.
  - $g(x) := \hat{f}_1(x) - \hat{f}_0(x)$ predicts the treatment effect for a person with covariates x.
- Why it doesn't work well:
  - Errors in two models can magnify each other.
  - More sophisticated approaches train models in tandem to prevent this (Künzel et al. 2019; Kennedy 2022).

Summarizing Treatment Effects

- Most approaches estimate treatment effect for each individual.
- Decision trees helpful for summarizing and visualizing results.
  - First node most important predictor of treatment effect.
- Inference must account for model selection.
  - Data splitting: fit models on half of data; apply on other half (Athey and Imbens 2016).
- Closed testing helpful for multiple comparisons!
  - (Rosenbaum 2008)
  - Test first node at level 0.05.
    - $p > 0.05 \Rightarrow \text{stop}.$
  - Test nodes at layer $N + 1$ at level 0.05×2^{-N}.
    - Abandon paths when we get an insignificant p-value.
  - Continue down tree until significance budget fully spent.

HTEs not Causal

- Example: marketing measurement with holdout group.
  - Strong impact on purchase behavior for men.
  - Weak impact for women.
- Claiming ad had bigger impact on men is true but potentially misleading.
  - Suppose ad equally effective for male and female sports fans, equally effective for male and female non-fans.
  - Ad more effective for sports fans than non-fans.
  - Suppose ad was targeted to male sports fans and female non-fans.
  - Gender is a red-herring, sports fandom (and weird targeting) is the real explanation.
- HTEs do not have a causal interpretation, even when based on a perfectly-executed A/B test.

Causal Interactions

- (VanderWeele 2015) calls HTEs with a causal interpretation, "causal interactions".
- Option 1: Double Randomization
  - Randomize treatment and one or more modifiers.
  - Variation in treatment effect across modifiers is causal thanks to randomization.
  - Only possible for modifiers that can be controlled.
- Option 2: Observational study
  - Treatment still randomized.
  - Analyze HTEs as before, but include all confounders as features.
  - May not observe all confounders!
  - Conclusions sensitive to functional form.

Key Takeaways

- People will react differently to the same treatment.
  - A treatment may be beneficial in some cases and harmful in others (e.g. chemotherapy).
- HTEs helpful when there is a cost associated with treatment.
  - Target treatment only where there is sufficient benefit to justify the cost.
  - Examples: discounts, marketing.
- HTEs do not have a causal interpretation.
  - Must use observational techniques to infer causal aspects.

Further Reading

- Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27) pg. 7353—60.
- Kennedy, Edward H. 2022. "Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects."
- Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10) pg. 4156—65.
- VanderWeele, Tyler J. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

A Reading List for Observational Studies

- Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press.
- Rosenbaum, Paul R. 2019. *Observation & Experiment: An Introduction to Causal Inference*. Harvard University Press.
- Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Cambridge University Press.
- Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.