Notes on Randomized Experiments

Benjamin Recht

November 4, 2025

1 Potential Outcomes

Here's a super fast introuction to randomized experiments in the potential outcomes framework. The model (usually attributed to Neyman [1923] and Rubin [1974]) asserts that we have n *units* on which we can intervene. These are abstractions of our experimental subjects. They could be people in a clinical trial or plots of land in an agricultural experiment. Each unit can either receive a treatment or not. Our model assumes that if unit i receives a treatment, then its outcome would be $y_i(1)$. If it doesn't receive a treatment, its outcome would be $y_i(0)$. If we define T_i to be a binary variable, equal to 1 if the unit receives the treatment and 0 if the unit does not receive the treatment, then the outcome of the unit is

$$y_i = T_i y_i(1) + (1 - T_i) y_i(0) \,.$$

This is so far just a mathematical expression of our model. Let us now use it in the context of a randomized experiment.

2 Randomized Experiments

For any unit, we'd be interested in determining the *treatment effect*

$$\tau_i = y_i(1) - y_i(0) \,.$$

But since we can only observe one of the two outcomes, there's no way to compute this quantity for any unit. Randomization gives us an approach to estimate the average treatment effect on multiple units.

The basic idea goes back to your first course on probability. Suppose that you select a random subset I of the integers $[n] = \{1, 2, \ldots, n\}$ of size $n/2$. You assign the units indexed by I to treatment and observe their outcomes. The only thing that determines their assignment is this random selection procedure. Then by the law of large numbers

$$\frac{2}{n} \sum_{i \in I} y_i(1) \approx \frac{1}{n} \sum_{k=1}^{n} y_k(1) \,.$$

By the exact same reasoning, the units in I^c are randomly assigned to control. Hence we'd have

$$\frac{2}{n}\sum_{j\in I^c} y_j(1) \approx \frac{1}{n}\sum_{k=1}^{n} y_k(1)\,.$$

But this means

$$\frac{2}{n}\sum_{i\in I} y_i(1) \approx -\frac{2}{n}\sum_{j\in I^c} y_j(1) \approx \frac{1}{n}\sum_{k=1}^{n} y_k(1) - y_k(0) = \frac{1}{n}\sum_{k=1}^{n} \tau_k\,.$$

So while we can't estimate the treatment effect for any unit, we can estimate the *average* of the treatment effects of all of the units.

The important thing here is that the only reason any unit was assigned to treatment and control was a randomized assignment. Further, this model assumes that the treatment of any unit i only affects unit i and does not affect any other unit. This might not be the case in an agricultural experiment if the plots are right next to each other.

3 Bernoulli experiments

In this section, we make the last section rigorous, using a slightly different randomized assignment that is easier to analyze. Let τ denote the *average treatment effect* of all of the units

$$\tau = \frac{1}{n}\sum_{k=1}^{n} \tau_k = \frac{1}{n}\sum_{k=1}^{n} y_k(1) - y_k(0)\,.$$

We will estimate τ through a randomized algorithm.

Let the treatment assignment T_i be assigned at random, equal to 1 with probability 1/2 and 0 with probability 1/2. That is, T is a Bernoulli random variable with $p = 1/2$. Then we can define the *Horvitz–Thompson estimator* as

$$\widehat{\tau} = \frac{2}{n}\sum_{\{i\,:\,T_i=1\}} y_i(1) - \frac{2}{n}\sum_{\{j\,:\,T_j=0\}} y_j(0)$$

In terms of the variable T, it's convenient to rewrite this expression as

$$\widehat{\tau} = \frac{2}{n}\sum_{i=1}^{n} T_i y_i(1) - (1-T_i) y_i(0)\,.$$

$\widehat{\tau}$ is an unbiased estimator of τ. This is because $\mathbb{E}[T_i y_i(1)] = \frac{1}{2} y_i(1)$ for all units. This is only true because we assumed that T_i was assigned randomly and that the only thing that effected the outcome of unit i was its treatment assignment. But under these assumptions, randomized algorithms give us unbiased estimates of the average treatment effect.

That $\widehat{\tau}$ is unbiased is helpful to us, but if we want to act on an experiment, we need to know how precise this estimate is. In order to estimate precision, we can compute the variance of the estimator. Observe that

$$\widehat{\tau} - \tau = \frac{1}{n}\sum_{i=1}^{n}(2T_i - 1)(y_i(1) + y_i(0))\,.$$

From this, we can compute the variance of $\widehat{\tau}$.

$$\begin{aligned}
\mathrm{var}(\widehat{\tau}) &= \mathbb{E}\left[\left(\frac{2}{n}\sum_{i=1}^{n}(2T_i - 1)(y_i(1) + y_i(0))\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(2T_i - 1)(2T_j - 1)(y_i(1) + y_i(0))(y_j(1) + y_j(0))\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[(2T_i - 1)(2T_j - 1)\right](y_i(1) + y_i(0))(y_j(1) + y_j(0)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}(y_i(1) + y_i(0))^2
\end{aligned}$$

If the outcomes are bounded (e.g., if they are binary), the variance of the estimator $\widehat{ATE}$ is $O(1/n)$. This suggests that for large sample sizes, the variance of the Horvitz-Thompson estimator is small. In the next section, let's work through some examples of how small it might be.

But before that, let's quickly note that we can't estimate the variance of the Horvitz-Thompson estimator without knowing both potential outcomes $y_i(1)$ and $y_i(0)$. But we can provide an estimator for an upper bound of the variance as follows. Since $(y_i(1) + y_i(0))^2 \leq 2y_i(1)^2 + 2y_i(0)^2$, we can estimate

$$\widehat{\sigma}^2 = \frac{4}{n}\sum_{i=1}^{n} T_i y_i(1)^2 + (1 - T_i)y_i(0)^2$$

Then we have

$$\mathbb{E}[\widehat{\sigma}^2] = \frac{2}{n}\sum_{i=1}^{n} y_i(1)^2 + y_i(0)^2 \geq \mathrm{var}(\widehat{\tau})\,.$$

4 Statistical Significance?

Ugh. I hate this section so much. I've said this many times now, but if we want to have a high precision, we should demand that our variance be small. But let's try to unpack what people mean when they talk about statistical significance.

For this section, define the constant $\mathcal{C}_\alpha$ as the number which satisfies

$$\Pr[x \geq \mathcal{C}_\alpha] = \alpha$$

when $x \sim \mathcal{N}(0, 1)$.

z-score. The z-score, also called the signal to noise ratio, is the estimated effect size divided by the estimated standard deviation:

$$\widehat{z} = \frac{\widehat{\tau}}{\widehat{\sigma}}.$$

We'd like this number to have large magnitude. When people refer to "5 sigma events," they mean experiments where the z-score is greater than 5.

Confidence intervals. If we assume that the true distribution is Gaussian, then a confidence interval at the level α is given by

$$[\widehat{\tau} - \mathcal{C}_{\alpha/2}\widehat{\sigma}, \widehat{\tau} + \mathcal{C}_{\alpha/2}\widehat{\sigma}].$$

The idea here is that since you are taking a random measurement, if you repeated the measurement many times, only α fraction of the time would you measure something outside of this interval.

p-values Similarly, assuming the distribution is Gaussian, the $p - value$ associated with an estimate is

$$\Pr[x \geq |\hat{z}|] + \Pr[x \leq |\hat{z}|]$$

when $x \sim \mathcal{N}(0, 1)$.

Statistical signifcance . Something people say when the p-value is small. Usually $p < 0.05$ is good enough for these folks...

4.1 Power Calculations

A power calculation is something you're supposed to do in advance to guess how large a sample should be. Recall that power is the same as the true positive rate. In the case of randomized experiments, this is

$$\Pr[\text{you declare statistical significance} \mid \text{the true effect size } \tau \text{ is large}]$$

Here's a simple rule of thumb. Assume the process is Gaussian. Assume that you know that the variance of τ is at most σ^2/n when there are n units. If we normalize everything, then z will be a draw from a normal distribution $\mathcal{N}(0, 1)$.

$$\Pr[\text{you declare statistical significance} \mid \text{the true effect size } \tau \text{ is large}] = 1 - 2\Pr[z \leq C_{\alpha/2} - \frac{\sqrt{n}\tau}{\sigma}]$$

Setting this quantity to equal β, the power, this is true if

$$\frac{\sqrt{n}\tau}{\sigma} \geq C_{\alpha/2} + C_{(1-\beta)/2}$$

For $\alpha = 0.05$ and $\beta = 0.8$, this is roughly saying that

$$n \geq \frac{16\sigma^2}{\tau^2}$$

should suffice to achieve the desired power. This rule just says that if you want a large z-score, you need to pick a large n. How large is given by this handwavy argument about Gaussians. But the bottom line is that n must exceed σ^2/τ^2 by a large factor.

References

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.