Fundamentals of Data Science

Introduction to experiment design

Ramesh Johari

Running a randomized experiment

We've seen how we can use a hypothesis test to analyze the outcome of an experiment.

But how do we design the randomized experiment in the first place? In particular, how do we choose the *sample size* for the experiment?

This is one of the first topics in *experimental design*.

Simplifying assumptions

We make two assumptions in this section to make the presentation more transparent:

- We will assume perfect splitting, so that with a sample size of n observations we have $n_1 = n_0 = n/2$.
- We will assume that the variance of both potential outcomes is the same:

$$\mathrm{Var}(Y(1)) = \mathrm{Var}(Y(0)) = \sigma^2.$$

What are we trying to do?

An experiment needs to balance the following two goals:

- Find true treatment effects when they exist;
- But without falsely finding an effect when one doesn't exist.

The first goal is to *control false negatives* (high power).

The second goal is to *control false positives* (small size).

Note that larger sample sizes enable higher power, smaller size, or both.

A survey of the approach

Sample size selection typically proceeds as follows:

- Commit to the level of false positive probability you are willing to accept (e.g., no more than 5%).

A survey of the approach

Sample size selection typically proceeds as follows:

- Commit to the level of false positive probability you are willing to accept (e.g., no more than 5%).
- Commit to the smallest ATE you want to be able to detect; this is the minimum detectable effect (MDE).

A survey of the approach

Sample size selection typically proceeds as follows:

- Commit to the level of false positive probability you are willing to accept (e.g., no more than 5%).
- Commit to the smallest ATE you want to be able to detect; this is the minimum detectable effect (MDE).
- Commit to the power you require at the MDE (e.g., 80%).

A survey of the approach

Sample size selection typically proceeds as follows:

- Commit to the level of false positive probability you are willing to accept (e.g., no more than 5%).
- Commit to the smallest ATE you want to be able to detect; this is the minimum detectable effect (MDE).
- Commit to the power you require at the MDE (e.g., 80%).

Fixing these three quantities completely determines the sample size required. (This is sometimes called a *power calculation* or a *sample size calculation*.)

Review: Size and power of the Wald test

The Wald statistic is $T = \widehat{\mathsf{ATE}}/\widehat{\mathsf{SE}}$, where:[1]

$$\widehat{\mathsf{SE}} = \sqrt{\frac{2\hat{\sigma}^2}{n}}.$$

It is approximately distributed as $\mathcal{N}(\mathsf{ATE}/\widehat{\mathsf{SE}}, 1)$.

[1] Recall that we assumed $\sigma_1^2 = \sigma_0^2 = \sigma^2$.

Review: Size and power of the Wald test

The Wald statistic is $T = \widehat{\mathsf{ATE}}/\widehat{\mathsf{SE}}$, where:[1]

$$\widehat{\mathsf{SE}} = \sqrt{\frac{2\hat{\sigma}^2}{n}}.$$

It is approximately distributed as $\mathcal{N}(\mathsf{ATE}/\widehat{\mathsf{SE}}, 1)$.

- If we reject when $|T| \geq z_{\alpha/2}$, then the test has size α.

[1] Recall that we assumed $\sigma_1^2 = \sigma_0^2 = \sigma^2$.

Review: Size and power of the Wald test

The Wald statistic is $T = \widehat{\mathsf{ATE}}/\widehat{\mathsf{SE}}$, where:[1]

$$\widehat{\mathsf{SE}} = \sqrt{\frac{2\hat{\sigma}^2}{n}}.$$

It is approximately distributed as $\mathcal{N}(\mathsf{ATE}/\widehat{\mathsf{SE}}, 1)$.

- If we reject when $|T| \geq z_{\alpha/2}$, then the test has size α.
- The power of the test when the true treatment effect is $\mathsf{ATE} = \theta \neq 0$ is:

  $$\mathbb{P}(|T| \geq z_{\alpha/2} | \mathsf{ATE} = \theta).$$

  Note that with more data, the power increases, because $\widehat{\mathsf{SE}}$ drops. (If you want, this can be computed using the normal cdf.)

[1] Recall that we assumed $\sigma_1^2 = \sigma_0^2 = \sigma^2$.

Sample size calculation with the Wald test

When sample size increases, we can "detect" true treatment effects that are smaller and smaller.

In particular:

- Suppose we use the size α Wald test (e.g., $\alpha = 0.05$).

Sample size calculation with the Wald test

When sample size increases, we can "detect" true treatment effects that are smaller and smaller.

In particular:

- Suppose we use the size α Wald test (e.g., $\alpha = 0.05$).
- Suppose we fix the MDE we want to be able to detect.

Sample size calculation with the Wald test

When sample size increases, we can "detect" true treatment effects that are smaller and smaller.

In particular:

- Suppose we use the size α Wald test (e.g., $\alpha = 0.05$).
- Suppose we fix the MDE we want to be able to detect.
- Suppose we require power at least β (e.g., $\beta = 0.80$) for a true treatment effect that is at least the MDE.

Sample size calculation with the Wald test

When sample size increases, we can "detect" true treatment effects that are smaller and smaller.

In particular:

- Suppose we use the size α Wald test (e.g., $\alpha = 0.05$).
- Suppose we fix the MDE we want to be able to detect.
- Suppose we require power at least β (e.g., $\beta = 0.80$) for a true treatment effect that is at least the MDE.
- This will determine the sample size n we need for the experiment.

Note that fixing any three of the four quantities α, β, MDE, and n determines the fourth!

Sample size calculation with the Wald test: A picture

Let's suppose we use $\alpha = 0.05$ and $\beta = 0.80$.

We work out the relationship between n and the MDE.

Key takeaway

So we find the following calcuation for the relationship between n and MDE, given $\alpha = 0.05$ and $\beta = 0.80$:

$$n = \frac{2 \times (2.8)^2 \hat{\sigma}^2}{MDE^2}.$$

The single most important intuition from the preceding analysis is this:

> *The standard error is inversely proportional to* $\sqrt{n}$*, and this means the required sample size* n *(for a given power and size) scales* inverse quadratically *with the* MDE*.*

So, for example, detecting an MDE that is half as big will require a sample size that is *four* times as large!

A final thought: No peeking!

Suppose you designed an experiment following the previous approach.

But now, instead of waiting until the sample size n is reached, you examine the p-value on an ongoing basis, and reject the null if you ever see it drop below α.

What would this do to your inference from the experiment?