A/B Testing

Tom Keefe
2018-10-25

A/B testing: will A or B get more clicks?

Which of these are A/B tests?

New users of your website fills out a form on the site and receive a welcome email. On Thursday, Creative gives you a new email template to try. You send all Thursday welcome emails with this template and compare the click rate to Wednesday’s. Is this an A/B test?

When filling out the form on your site, half your audience supplies their first name. You start a test where the welcome email includes the first name if they give it, e.g. (Hi Tom!). Is this an A/B test?

You have a daily “recipe of the day” newsletter. After a month of mailing the newsletter, you split your audience 90-10 and start sending the 10% a different template. Is this an A/B test?

Which of these are A/B tests?

You work for an online store. Many users add items to their “cart” but don’t actually buy them. A common strategy is to email the user a reminder about the items in their cart. To test two different “abandoned cart” templates, you send 90% of Tuesday’s abandoned cart emails with template A, and 10% with template B. Is this an A/B test?

You have a daily “recipe of the day” newsletter. After a month of mailing the newsletter, you split your audience 50-50 and start sending half of them a different template. Is this an A/B test?

A simple A/B test

Suppose I have a successful cooking blog. Instant Pot contracts me to advertise Instant Pots to my audience. I draft two emails to my readers, and on Monday, I email template A to 500 users, and template B to another 500.

When I check the stats on Tuesday,

- 50 clicks from group A (10% conversion rate)
- 75 clicks from group B (15% conversion rate)

We check if this is a statistically significant difference with a chi-square test.

What not to do:

Suppose I send 500 emails with a 50/50 split for my A/B test. I peek at the results after 200 observations, and again at the end. There are 4 possible scenarios:

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
After 200 observations	Insignificant	Insignificant	Significant	Significant
After all 500 observations	Insignificant	Significant	Insignificant	Significant

We conclude:	Insignificant	Significant	Insignificant	Significant

Source: Evan Miller, "How not to run an A/B test."

What not to do:

Again, I peek after 200 observations. But this time, I stop the experiment if the result is significant.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
After 200 observations	Insignificant	Insignificant	Significant	Significant
After all 500 observations	Insignificant	Significant	*trial stopped*	*trial stopped*

We conclude:	Insignificant	Significant	Significant	Significant

What happened here?

Source: Evan Miller, “How not to run an A/B test.”

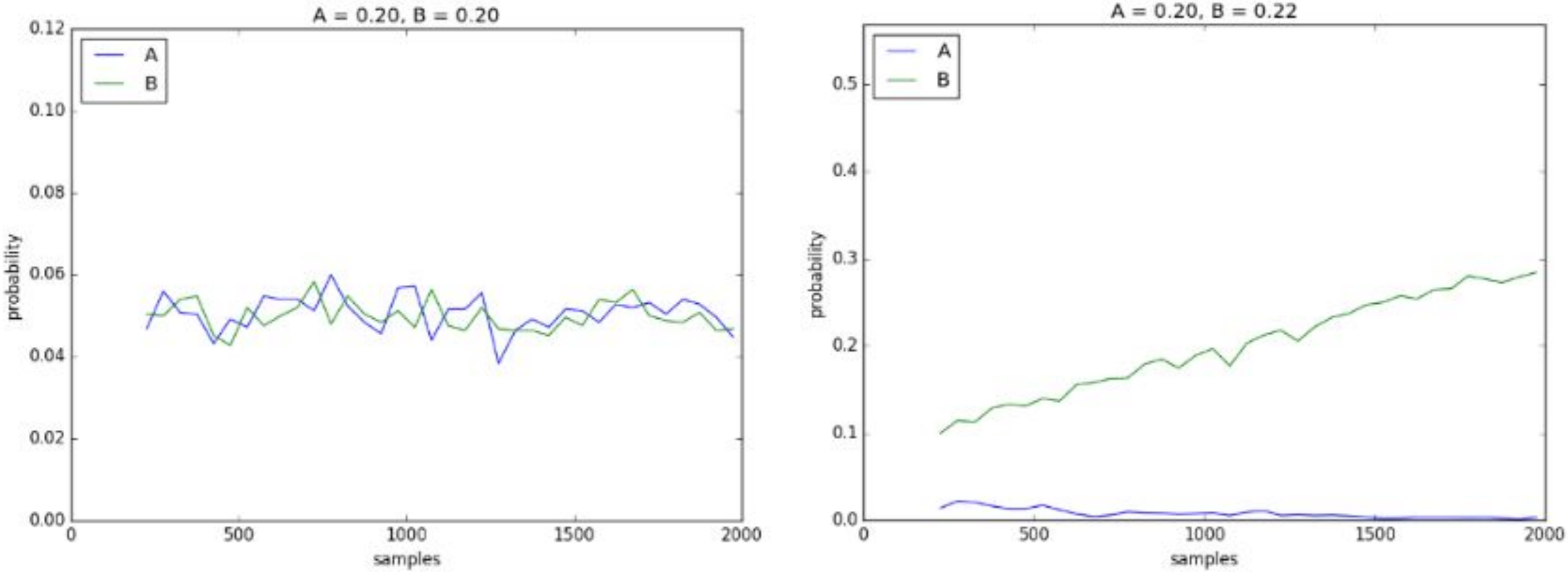
No peeking!

- The calculations for significance assume that the sample size was fixed in advance.
- If you peek and stop the experiment early, you're overestimating the p-value.

Q: What if you did a significance test after every single observation?

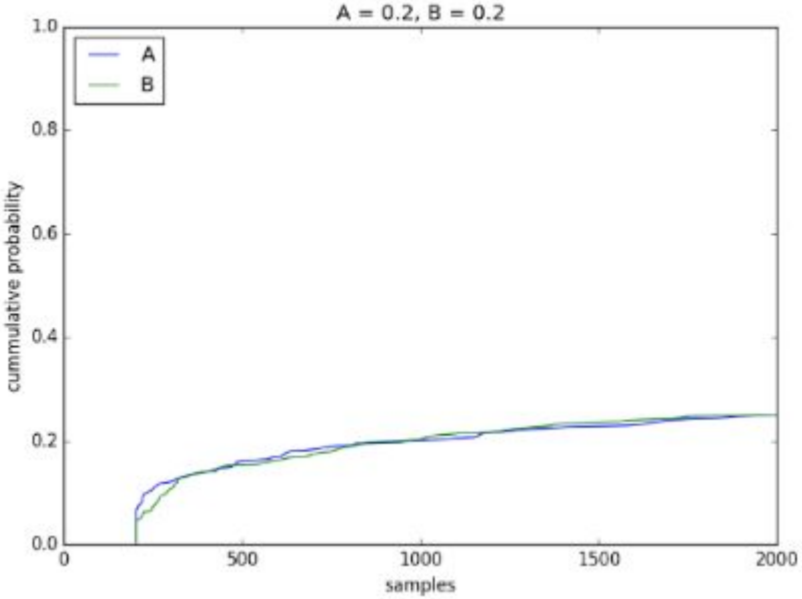
Source: Evan Miller, "How not to run an A/B test."

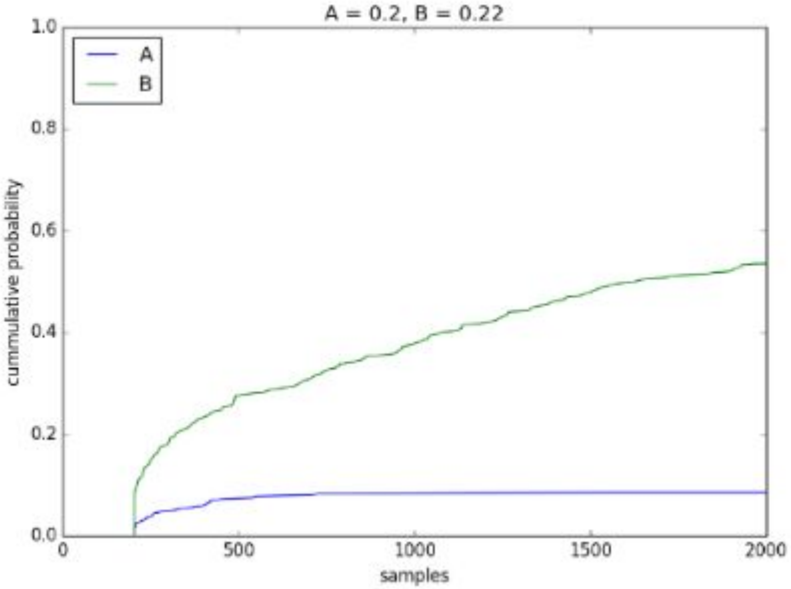
Probability of concluding A or B is better, for increasing sample sizes



Source: Paul Draper, “The Fatal Flaw of A/B Tests: Peeking”

And when we peek after each trial





Source: Paul Draper, "The Fatal Flaw of A/B Tests: Peeking"

Power

- To have valid significance estimates, we need to set the sample size in advance.
- How do we know what size sample to use?
- We need to “back out” the sample size from effect we want to detect.
- Example from last slide:
  - Baseline conversion rate: 20%
  - Want to detect if template B gives at least 22% (a 10% relative increase)
- But now we need to pick another parameter: power.
- Power is the probability of correctly detecting that template B is 10% better, if indeed it is.

How to correctly stop an A/B test early

Q: Why would we do this?

Various methods exist. They’re all very mathy. One solution is called the multi-armed bandit.

The multi-armed bandit problem

The one arm bandit.

The multi-armed bandit problem

How should I choose which levers to pull to maximize my earnings?

The multi-armed bandit problem

- Explore vs. exploit: Want to *exploit* the best known slot machine, but also have to *explore* the available choices.
- Flexible method for testing many options at once while maximizing your earnings.
- The Bayesian multi-armed bandit never actually decides on a best option, but cleverly decides whether to exploit the best known slot machine or explore another.
- After every trial, it updates its belief about the payout probability of that lever.
- Note: it peeks after every trial!