Significance tests

Part 5: Power and sample size

By: Noureddin Sadawi, PhD

University of London

Introduction

- ‘One step in statistical calculations for sample size is to ask “Will a hypothesis test actually reveal a difference between treatments A and B?”
- The outcome of a hypothesis test—the p-value—depends on what the real difference is between treatment A and treatment B.
- It also depends on the luck of the draw—who gets selected for the groups in the experiment.
- But it makes sense that the bigger the actual difference between treatments A and B, the greater the probability that our experiment will reveal it; and the smaller the difference, the more data will be needed to detect it.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key terms for power and sample size

- ‘**Effect size:** The minimum size of the effect that you hope to be able to detect in a statistical test, such as “a 20% improvement in click rates.”
- **Power:** The probability of detecting a given effect size with a given sample size.
- **Significance level:** The statistical significance level at which the test will be conducted.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example

- ‘Power is the probability of detecting a specified effect size with specified sample characteristics (size and variability).
- For example, we might say (hypothetically) that the probability of distinguishing between a .330 hitter and a .200 hitter in 25 at-bats is 0.75.
- The effect size here is a difference of .130.
- And “detecting” means that a hypothesis test will reject the null hypothesis of “no difference” and conclude there is a real effect.
- So the experiment of 25 at-bats (n = 25) for two hitters, with an effect size of 0.130, has (hypothetical) power of 0.75, or 75%.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Statistical software

- ‘You can see that there are several moving parts here, and it is easy to get tangled up in the numerous statistical assumptions and formulas that will be needed (to specify sample variability, effect size, sample size, alpha-level for the hypothesis test, etc., and to calculate power).
- Indeed, there is special-purpose statistical software to calculate power.
- Most data scientists will not need to go through all the formal steps needed to report power, for example, in a published paper.
- However, they may face occasions where they want to collect some data for an A/B test, and collecting or processing the data involves some cost.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Power analysis procedure

1. ‘Start with some hypothetical data that represents your best guess about the data that will result (perhaps based on prior data)—for example, a box with 20 ones and 80 zeros to represent a .200 hitter, or a box with some observations of “time spent on website”.
2. Create a second sample simply by adding the desired effect size to the first sample—for example, a second box with 33 ones and 67 zeros, or a second box with 25 seconds added to each initial “time spent on website”.
3. Draw a bootstrap sample of size *n* from each box.
4. Conduct a permutation (or formula-based) hypothesis test on the two bootstrap samples and record whether the difference between them is statistically significant.
5. Repeat the preceding two steps many times and determine how often the difference was significant—that’s the estimated power.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Sample size

- ‘The most common use of power calculations is to estimate how big a sample you will need.
- For example, suppose you are looking at click-through rates (clicks as a percentage of exposures), and testing a new ad against an existing ad.
- How many clicks do you need to accumulate in the study? If you are interested only in results that show a huge difference (say, a 50% difference), a relatively small sample might do the trick.
- If, on the other hand, even a minor difference would be of interest, then a much larger sample is needed.
- A standard approach is to establish a policy that a new ad must do better than an existing ad by some percentage, say, 10%; otherwise, the existing ad will remain in place.
- This goal, the “effect size,” then drives the sample size.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example

- ‘Suppose current click-through rates are about 1.1%, and you are seeking a 10% boost to 1.21%.
- So we have two boxes: box A with 1.1% ones (say, 110 ones and 9,890 zeros), and box B with 1.21% ones (say, 121 ones and 9,879 zeros).
- For starters, let’s try 300 draws from each box (this would be like 300 “impressions” for each ad).
- Suppose our first draw yields the following:

  Box A: 3 ones

  Box B: 5 ones’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example

- ‘Right away we can see that any hypothesis test would reveal this difference (5 versus 3) to be well within the range of chance variation.
- This combination of sample size (n = 300 in each group) and effect size (10% difference) is too small for any hypothesis test to reliably show a difference.
- So we can try increasing the sample size (let’s try 2,000 impressions), and require a larger improvement (50% instead of 10%).’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example

- ‘Suppose current click-through rates are still 1.1%, but we are now seeking a 50% boost to 1.65%.
- So we have two boxes: box A still with 1.1% ones (say, 110 ones and 9,890 zeros), and box B with 1.65% ones (say, 165 ones and 9,868 zeros).
- Now we’ll try 2,000 draws from each box.
- Suppose our first draw yields the following:

  Box A: 19 ones

  Box B: 34 ones’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example

- ‘A significance test on this difference (34–19) shows it still registers as “not significant” (though much closer to significance than the earlier difference of 5–3).
- To calculate power, we would need to repeat the previous procedure many times, or use statistical software that can calculate power, but our initial draw suggests to us that even detecting a 50% improvement will require several thousand ad impressions.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Summary

‘For calculating power or required sample size, there are four moving parts:

- Sample size.
- Effect size you want to detect.
- Significance level (alpha) at which the test will be conducted.
- Power.

Specify any three of them, and the fourth can be calculated. Most commonly, you would want to calculate sample size, so you must specify the other three.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Key ideas

- ‘Finding out how big a sample size you need requires thinking ahead to the statistical test you plan to conduct.
- You must specify the minimum size of the effect that you want to detect.
- You must also specify the required probability of detecting that effect size (power).
- Finally, you must specify the significance level (alpha) at which the test will be conducted.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).