EXPERIMENTATION

# Statistical Power, Sample Size, and Minimum Detectable Effect in A/B Testing

A practitioner's review of statistical power, the 80% convention, minimum detectable effect, and the sample-size formula that sets how much traffic an A/B test needs.

**Bhaskar Roy Sarkar**, cazyweb · June 28, 2026 ·

statistical power | sample size | minimum detectable effect | A/B testing | Type II error

Download PDF

ABSTRACT

Statistical power -- the probability that a test correctly rejects a false null hypothesis -- is the foundational design parameter for any A/B experiment. Yet practitioners routinely launch tests without calculating it, producing a chronic pattern of underpowered experiments that miss genuine conversion lifts, waste traffic, and generate a disproportionate share of false discoveries among the results that do reach significance. This paper reviews the core concepts of Type II error, the 80% power convention codified by Cohen, and the minimum detectable effect (MDE) as the practical planning instrument that links sample size to business impact. We derive the two-proportion sample-size formula from first principles, show how each input -- baseline conversion rate, MDE, significance threshold, and power target -- shifts the required sample, and translate the formula into test-duration estimates relevant to ecommerce teams. We also document the winner's curse effect that inflates apparent lifts in underpowered

studies, and connect the fixed-horizon sample-size calculation to the sequential and always-valid inference methods that safely accommodate continuous monitoring.

## 1. Introduction

A/B testing is the closest approximation ecommerce teams have to a controlled scientific experiment. A visitor is randomly assigned to a control or treatment experience; their behavior is recorded; and a statistical test adjudicates whether the observed difference in conversion rates is likely to reflect a genuine causal effect rather than chance. The logic is clean, but its guarantees depend on a planning step that practitioners frequently skip: deciding in advance how large the experiment must be.

This paper is a practitioner review of that planning step. It covers the definition and interpretation of statistical power [1], the two-proportion formula that translates a target power level into a required sample size [2], the minimum detectable effect as the primary dial for controlling that sample size [3], and the consequences -- both for traffic waste and for result reliability -- of neglecting these calculations [4]. It closes with a discussion of the relationship between fixed-horizon designs and the sequential methods that have become common on commercial experimentation platforms [5].

## 2. The Neyman-Pearson Framework and Two Types of Error

Modern hypothesis testing rests on a framework introduced by Neyman and Pearson in their 1933 paper on the most efficient tests of statistical hypotheses [6]. The framework distinguishes two ways a test can fail.

A **Type I error** (false positive, probability alpha) occurs when the test declares a difference statistically significant even though the null hypothesis of no difference is true. Setting alpha = 0.05 means accepting a 5% chance of this mistake in any single test.

A **Type II error** (false negative, probability beta) occurs when the test fails to detect a real difference. The complement of beta is **statistical power**: Power = 1 - beta. A test with 80% power has a 20% chance of missing a real effect of the specified size.

These two error probabilities are not independent. For a fixed sample size, reducing alpha (demanding stronger evidence) increases beta (making it harder to detect a real signal). For a fixed alpha, increasing the required sample size is the only way to reduce beta -- that is, to push power up [1].

## 3. The 80% Power Convention and Where It Comes From

The near-universal choice of 80% as the minimum acceptable power originates with Jacob Cohen. In his 1988 textbook, Cohen proposed that a 4:1 ratio of Type I to Type II error was a reasonable default when costs were uncertain, yielding a convention of alpha = 0.05 and power = 0.80 [1]. His 1992 primer in Psychological Bulletin condensed this reasoning into a practical reference and supplied sample-size tables for eight standard tests, cementing 80% as the field benchmark [7].

Cohen was explicit that 80% is a floor, not a ceiling, and that studies where the cost of a missed effect is high should target 90% or 95% power. In ecommerce CRO, where shipping a losing variant has a real revenue cost, teams with sufficient traffic are well-advised to target 90%. The cost is a larger sample -- roughly 75% more observations relative to an 80%-powered design at the same alpha and MDE [7].

What the 80% convention does not mean is that a study can declare a negative result as evidence that an effect does not exist. A test with 80% power still has a 1-in-5 chance of missing a real effect. A negative result from an underpowered study is largely uninformative.

## 4. The Minimum Detectable Effect

The minimum detectable effect (MDE) is the smallest true difference in conversion rates that a planned experiment would detect with the target power, given the chosen significance threshold [3]. It is a design choice, not a property of the data.

Choosing an MDE requires a business judgment: what is the smallest conversion-rate lift that would change a product decision? In ecommerce contexts, a 20% relative lift on a 3% baseline -- an absolute difference of 0.6 percentage points -- might represent a threshold below which the implementation cost of a new feature exceeds the revenue gain. Setting the MDE to 0.6 pp sets the experiment up to be decisive at that threshold.

The MDE trades off against sample size in a critical way: required sample size scales with the **inverse square** of the MDE [8]. Halving the MDE -- asking the experiment to detect a subtler effect -- quadruples the required sample. This quadratic relationship is the central practical insight in experimental design for CRO. It means that ambitious sensitivity targets rapidly become infeasible for sites with limited daily traffic.

## 5. The Two-Proportion Sample-Size Formula

For an A/B test comparing two conversion rates, the required per-group sample size is derived from the normal approximation to the binomial. The standard formula, as codified in Fleiss, Levin, and Paik [2] and widely reproduced in the applied statistics literature [9], is:

n = ( z_(1-alpha/2) * sqrt(2 * p_bar * (1 - p_bar)) + z_(1-beta) * sqrt(p_c*(1-p_c) + p_t*(1-p_t)) )^2 / (p_t - p_c)^2

where p_c is the baseline (control) conversion rate, p_t = p_c + MDE is the alternative rate being tested against, p_bar = (p_c + p_t)/2 is the pooled proportion under the null, z_(1-alpha/2) is the z-score for the chosen significance level (1.96 for a two-sided test at alpha = 0.05), and z_(1-beta) is the z-score for the power level (0.84 for 80% power, 1.28 for 90% power).

The formula produces the per-group sample; the total required sample is 2n for an equal-allocation two-variant test.

### 5.1 What Each Input Does

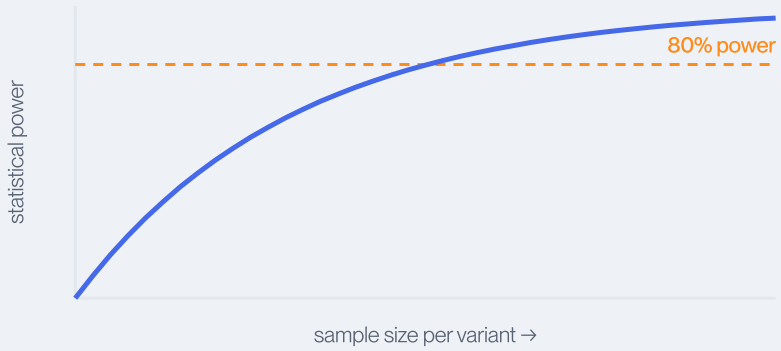
**Baseline conversion rate (p_c).** The variance of a Bernoulli random variable is p(1-p), which is maximised at p = 0.50. This means experiments on high-baseline metrics (e.g., add-to-cart rate of 40%) require more observations per group than experiments on low-baseline metrics at the same absolute MDE. Kohavi, Tang, and Xu emphasise that practitioners should always work with the actual baseline measured in production, not a guessed value, because errors in p_c propagate directly into the sample-size estimate [3].

**MDE.** As noted above, sample size scales with 1/MDE^2. The MDE is the most powerful lever in the practitioner's hands. Kohavi and colleagues recommend expressing MDE as a relative percentage of the baseline and requiring teams to justify why that threshold is commercially meaningful before a test launches [10].

**Alpha.** Tightening the significance threshold from 0.05 to 0.01 increases the required z-score from 1.96 to 2.58 and roughly doubles the sample size at 80% power. Teams running many concurrent tests should consider controlling the false discovery rate across tests, which changes how alpha is set [11].

**Power (1 - beta).** Increasing from 80% to 90% (z from 0.84 to 1.28) increases sample size by approximately 27%; moving to 95% (z = 1.645) increases it by roughly 60% relative to 80% [7].

A quick approximation due to Lehr provides a useful sanity check: for alpha = 0.05 (two-sided) and 80% power, the per-group sample size is approximately 16 * s^2 / d^2, where s is the standard deviation and d is the effect size [8]. For proportions, this simplifies to the same quadratic dependence on MDE.




For a fixed effect size, statistical power rises with the sample size per variant: more traffic means a higher chance of detecting a real difference. 80% power is the common target, and under-powered tests miss real wins.

## 6. Translating Sample Size to Test Duration

The required sample size becomes a required test duration once the available daily traffic to the experiment is known [3]. If an ecommerce product page receives T unique visitors per day and the A/B test splits them equally, each variant accumulates T/2 visitors per day. A test requiring n per group will run for at least 2n/T days.

Practitioners should apply several adjustments to this baseline:

- **Business cycles.** Weekly seasonality means a test that starts on a Tuesday captures a different customer mix than one that starts on a Saturday. Kohavi and colleagues recommend running for at least one or two full week cycles to avoid day-of-week confounding [12].
- **Traffic eligibility.** Not all page visitors may qualify for the experiment. If the test is on the checkout page and only 15% of sessions reach checkout, the effective daily sample is 15% of total site traffic.
- **Multiple variants.** Adding a third or fourth variant splits the same traffic further, multiplying duration proportionally and reducing the power of each pairwise comparison.

- **Novelty effects.** A new design element may attract inflated engagement in week one simply because it is new. Testing for a novelty effect and excluding the first few days of data can matter for long-running tests [10].

## 7. The Cost of Underpowered Tests

Understanding the consequences of running underpowered tests motivates the planning discipline. There are two distinct harms.

### 7.1 Missing Real Wins

An underpowered test has a high probability of producing a non-significant result even when the true effect equals or exceeds the target MDE. A test designed for 50% power -- a common outcome when sample sizes are set by gut feel -- will miss a real effect more often than it finds one. That missed lift compounds over the program lifetime: every cycle where a genuine improvement is declared neutral represents foregone revenue [13].

### 7.2 The Winner's Curse

The second harm is subtler. When a test is underpowered and does produce a significant result, that result is almost certainly an overestimate of the true effect. The mechanism is selection: to achieve statistical significance with too few observations, the observed difference must be unusually large -- larger than the true underlying effect. This phenomenon, called the winner's curse or post-selection bias, was documented in neuroscience by Button and colleagues [13] and has direct implications for CRO programs that run small tests and then roll out features based on the observed effect size.

Simmons, Nelson, and Simonsohn demonstrated in the psychology literature that flexible sample sizes combined with underpowered designs can push false-positive rates far above the nominal alpha -- in simulated conditions reaching 61% [11]. CRO practitioners face an analogous risk when they let test duration drift based on whether the dashboard looks promising.

Ioannidis showed formally that in any research field where most hypotheses tested are false and studies are underpowered, the majority of statistically significant findings will be false positives [4]. Ecommerce experimentation is not immune: the proportion of variants with a true effect is not known, and when it is low, even a nominally calibrated test program will accumulate more false discoveries than practitioners assume.

## 8. Peeking and Sequential Methods

The fixed-horizon sample-size calculation described above assumes that the significance test is applied exactly once, at a pre-specified sample size. In practice, practitioners check dashboards continuously. This behavior -- known as peeking -- invalidates the frequentist guarantees because the decision to stop is correlated with the observed test statistic.

Miller demonstrated in an influential 2010 blog post that peeking ten times during a test at a nominal alpha of 0.05 produces an actual false-positive rate closer to 19% [14]. Johari, Koomen, Pekelis, and Walsh formalized this result and showed that worst-case Type I error can approach the nominal level from far above when optional stopping is allowed with no correction [5]. Empirical analysis of production Optimizely tests found that a substantial fraction of declared significant results reflected peeking behavior rather than genuine treatment effects; deploying sequential always-valid inference reduced this false-positive inflation [15].

The solution is not to ban monitoring but to use inference procedures designed for continuous monitoring. The always-valid inference framework of Johari and colleagues [5] and the mSPRT (mixture sequential probability ratio test) family provide confidence sequences and p-value processes that remain valid at any data-dependent stopping time, at the cost of a modest increase in expected sample size relative to the fixed-horizon design.

For teams not yet ready to adopt sequential methods, Kohavi, Tang, and Xu recommend a simpler discipline: commit to a sample size before the test

launches, document that commitment, and treat the pre-specified endpoint as a hard stop [3].

## 9. Practical Recommendations for Ecommerce CRO Teams

The following summary translates the technical discussion into a design checklist for practitioners.

**Before launch:** Measure the baseline conversion rate from the last 30 days of production traffic. Define the MDE as the minimum relative lift that would justify shipping the variant. Run the two-proportion formula to determine required per-group sample size. Divide by daily per-variant traffic to get minimum test duration. Schedule the test end date before launch and record it.

**During the test:** Resist interpreting interim results. If continuous monitoring is operationally unavoidable, use a platform that implements sequential testing (mSPRT, always-valid confidence sequences) rather than fixed-horizon p-values.

**At the end:** Analyze exactly once at the pre-specified endpoint. Report the observed lift together with its confidence interval, not just the p-value. If the result is non-significant, report the observed power -- a non-significant result from a 90%-powered test means something very different from one in a 50%-powered test.

**Effect-size realism:** On a 3% checkout conversion rate, detecting a 5% relative lift (0.15 pp absolute) requires roughly 130,000 visitors per variant at 80% power and alpha = 0.05. Teams with under 10,000 daily sessions should restrict themselves to tests with MDEs above 15-20% relative, or consolidate traffic by testing higher-funnel pages.

## 10. Conclusion

Statistical power is not a bureaucratic formality; it is the specification of what a test is capable of learning. A test with insufficient power is capable of learning very little: it will miss genuine improvements and, when it does find significance, will overstate the magnitude of the effect. The two-proportion formula, combined with a commercially grounded MDE, gives CRO practitioners a clear-eyed budget for how much traffic each experiment requires. Respecting that budget -- and pairing it with sequential inference when continuous monitoring is needed -- is the foundation of a trustworthy experimentation program.

---

## References


1. Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). *Lawrence Erlbaum Associates*. https://www.routledge.com/Statistical-Power-Analysis-for-the-Behavioral-Sciences/Cohen/p/book/9780805802832
2. Fleiss, J. L., Levin, B., and Paik, M. C. (2003). Statistical Methods for Rates and Proportions (3rd ed.). *Wiley*. https://onlinelibrary.wiley.com/doi/book/10.1002/0471445428
3. Kohavi, R., Tang, D., and Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. *Cambridge University Press*. https://experimentguide.com/
4. Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.0020124
5. Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2022). Always Valid Inference: Continuous Monitoring of A/B Tests. *Operations Research*. https://doi.org/10.1287/opre.2021.2135
6. Neyman, J., and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*. https://doi.org/10.1098/rsta.1933.0009
7. Cohen, J. (1992). A Power Primer. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.112.1.155
8. Lehr, R. (1992). Sixteen S-squared over D-squared: A Relation for Crude Sample Size Estimates. *Statistics in Medicine*. https://doi.org/10.1002/sim.4780110811
9. National Institute of Standards and Technology (2012). NIST/SEMATECH e-Handbook of Statistical Methods, Section 7.2.4.2: Sample Sizes Required. *NIST*. https://www.itl.nist.gov/div898/handbook/prc/section2/prc242.htm

10. Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled Experiments on the Web: Survey and Practical Guide. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-008-0114-1

11. Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*. https://doi.org/10.1177/0956797611417632

12. Kohavi, R. and Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. *Encyclopedia of Machine Learning and Data Mining, Springer*. https://doi.org/10.1007/978-1-4899-7687-1_891

13. Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience*. https://doi.org/10.1038/nrn3475

14. Miller, E. (2010). How Not To Run an A/B Test. *evanmiller.org*. https://www.evanmiller.org/how-not-to-run-an-ab-test.html

15. Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2017). Peeking at A/B Tests: Why It Matters, and What to Do About It. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3097983.3097992

## Cite this

**APA** Copy

```
Bhaskar Roy Sarkar (2026). Statistical Power, Sample Size, and Minimum
Detectable Effect in A/B Testing. cazyweb Research.
https://cazyweb.com/research/statistical-power-sample-size-ab-testing
```

**BIBTEX** Copy

```
@techreport{statistical-power-sample-size-ab-testing,
  author = {Bhaskar Roy Sarkar},
  title = {Statistical Power, Sample Size, and Minimum Detectable
Effect in A/B Testing},
  institution = {cazyweb Research},
  year = {2026},
  url = {https://cazyweb.com/research/statistical-power-sample-size-
ab-testing}
}
```

**RELATED READING**

[How long should you run an A/B test?]

[How to prioritize A/B tests]

[Always-valid inference in conversion testing]

Want a team to run experiments like this for you? See how we help

## Let's build something that converts.

Start a project with us, or join the bench of specialists and agencies we deliver with.

**Start a project** ↗

cazyweb

The engineering and growth team behind leading brands and companies across the US and Europe.

### CRO and ecommerce insights

Occasional, genuinely useful notes. No fluff.

you@company.com Subscribe →

Work email only. No spam. Unsubscribe any time. Privacy.

## Solutions

- Product Optimization
- Experimentation Program
- cazyweb Platform

## For Ecommerce

- CRO Program
- Email & Retention
- Paid Ads
- Landing Pages
- Revenue Calculator FREE
- A/B Test Calculator FREE
- CRO Agency
- Shopify CRO
- Ecommerce Growth
- Website Performance

## Company

- About
- For Brands
- For Agencies
- Careers HIRING
- Contact

## Explore

- Services
- Pricing
- Performance program
- Subscription
- Website + CRO from $99
- Case Studies
- Blog
- CRO Ideas
- Apps
- Learn

**CazyWeb IT Solutions**
GSTIN: 19JTLPS8476P2ZZ
hello@cazyweb.com

© 2026 cazyweb. All rights reserved.