CXL institute

Statistical Power & Sample Size Calculations

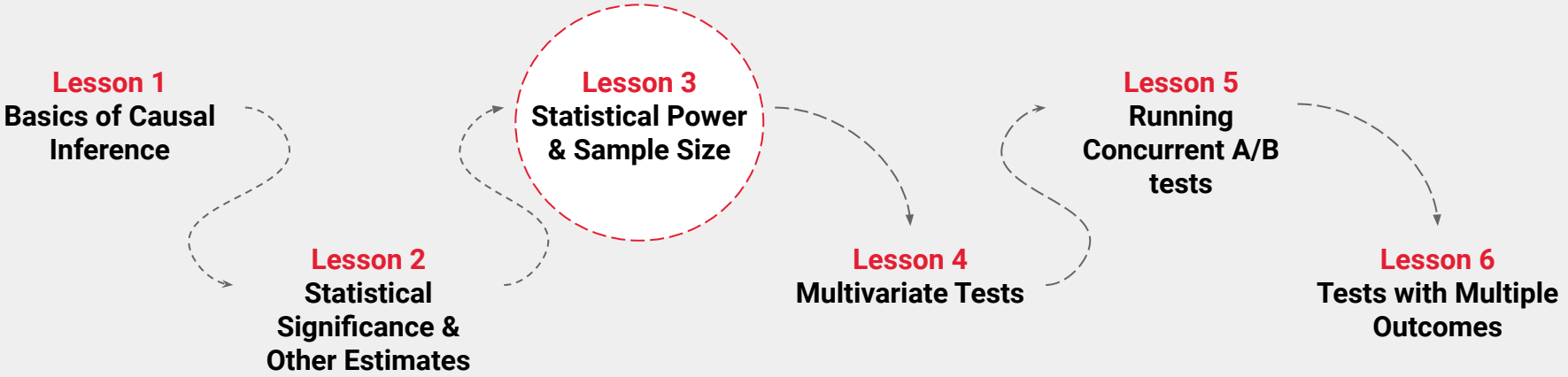
Georgi Georgiev · Applied Statistician · Analytics-Toolkit.com

Author of “Statistical Methods in Online A/B Testing”

The Journey Of This Course

Part 1 | **BASIC CONCEPTS**

Part 2 | **ADVANCED APPLICATIONS**



CXL institute

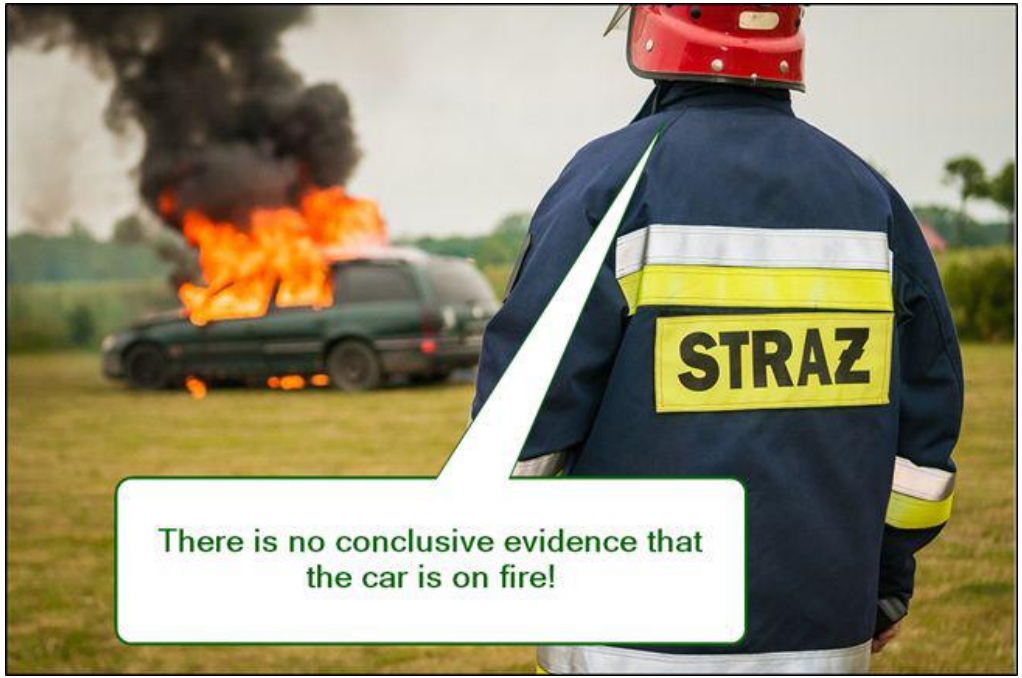
Lesson Objectives:

1.) Understand type II errors, the concept of statistical power of a test, and how to account for them in planning an A/B test
2.) Learn how to determine the required sample size for a simple A/B test
3.) Understand the relationship between power, significance threshold, minimum effect of interest, and sample size
4.) Be able to plan a fixed-sample A/B test so it achieves a target power at a specified minimum effect of interest

Two types of errors in A/B testing

Judgement of Null Hypothesis	**Null Hypothesis (H_0) is Valid/True**	**Null Hypothesis (H_0) is Invalid/False**
Reject	Type I Error (false positive)	Correct Inference (true positive)
Fail to Reject	Correct Inference (true negative)	Type II Error (false negative)

False Negatives



Type II Error Rate of a Statistical Test

- The type II error rate is the probability of observing a non-significant p-value at a certain threshold α if a true effect of a certain magnitude μ_1 is in fact present.

$$\beta(T_{(\alpha)}; \mu_1) = P(d(X) \leq c_{(\alpha)};\ \mu = \mu_1)$$

- A function of the testing procedure and μ_1
- Denoted by β
- By definition the less important kind of error

Statistical Power of a Statistical Test

- **The statistical power of a statistical test is defined as the probability of observing a p-value statistically significant at a certain threshold α if a true effect of a certain magnitude μ_1 is in fact present.**

$$POW(T_{(\alpha)}; \mu_1) = P(d(X) > c_{(\alpha)};\ \mu = \mu_1)$$

- **It is the inverse of the type II error rate: POW = 1 - β.**
- **Denoted by POW().**
- **A function of the testing procedure and μ_1 .**

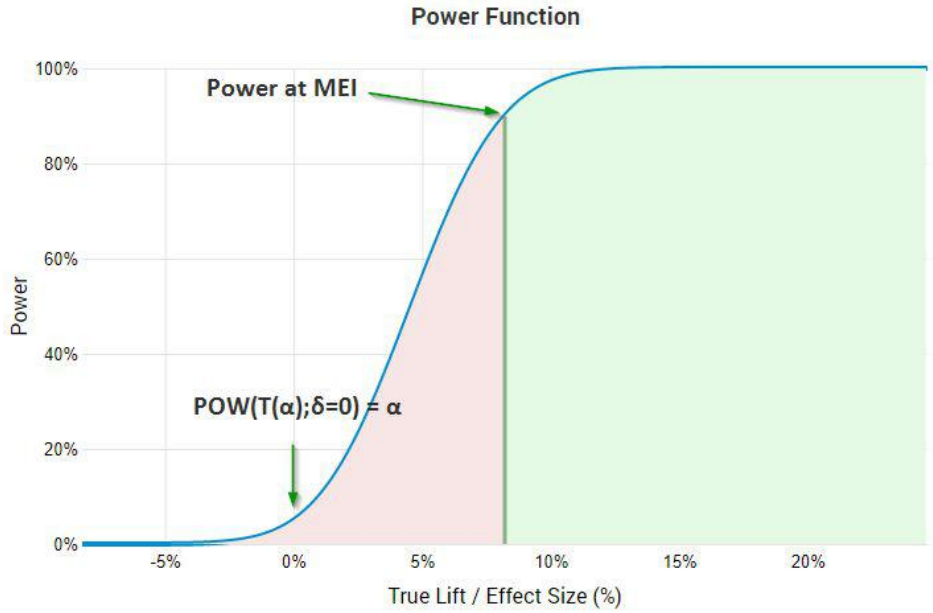
Statistical Power for absolute difference in proportions

$$POW(T_{(\alpha)};\delta) = \Phi\left(\frac{\sqrt{n_1}\cdot\delta}{\sigma_{pooled}} - Z_{1-\alpha}\right)$$

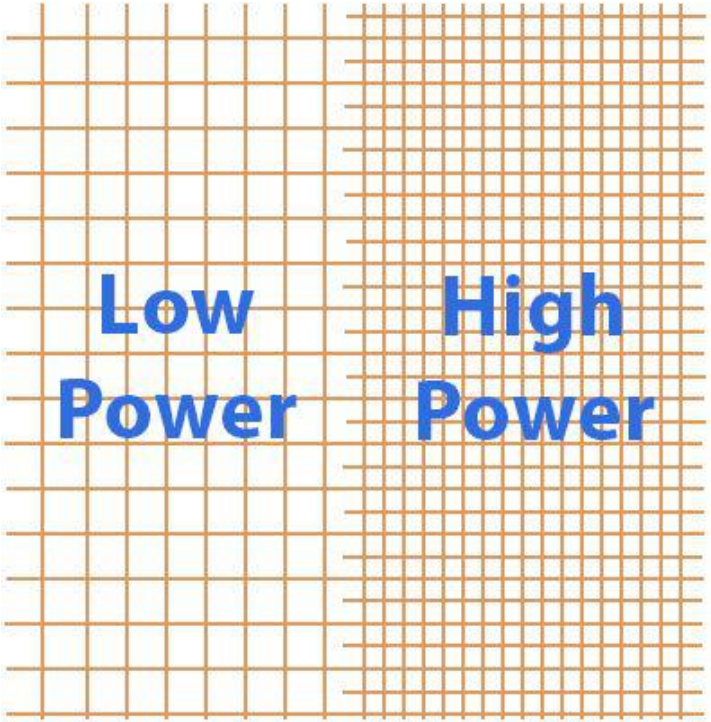
$$\sigma_{pooled} = \sqrt{\frac{p_1\cdot(1-p_1)}{r} + p_2\cdot(1-p_2)}$$

$r = n_1/n_2$
Φ - cumulative distribution function
δ - minimum effect of interest

Power as Function of Minimum Effect of Interest



Statistical Power: a Metaphor



Large Effect
Medium Effect
Small Effect

The lower-powered net will only "catch" big effects, while missing medium and small effects most of the time.

The higher-powered net will catch both big and medium effects, but will often miss small effects.

Analytics-Toolkit.com

Evidence of absence?

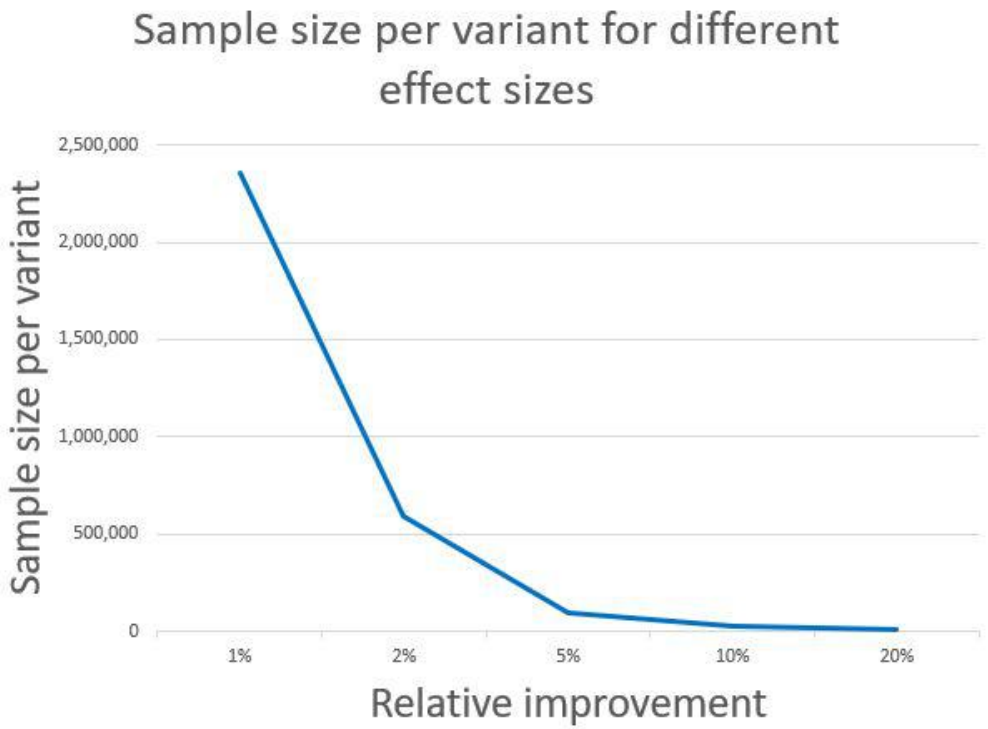
Relationship with other test parameters

Test Parameter	Change Direction	Effect on Power (all else fixed)
Minimum Effect of Interest	⇧	⇧
	⇩	⇩
Type I Error (α)	⇧	⇧
	⇩	⇩
Sample Size	⇧	⇧
	⇩	⇩

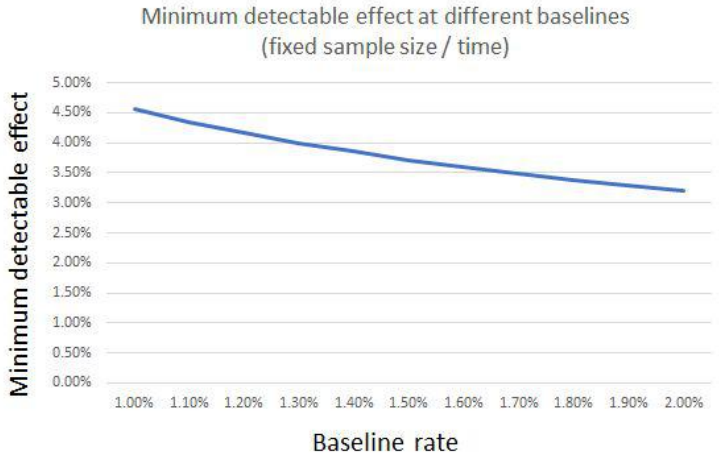
Power and minimum effect of interest

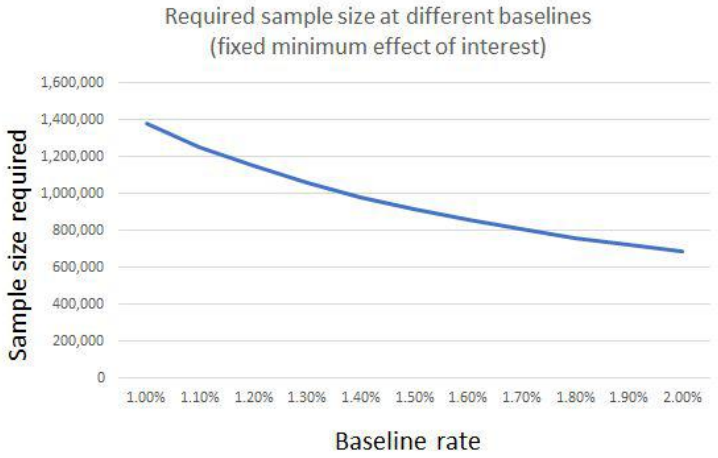
Relative Improvement of Interest	Sample Size Per Variant, Required for 80% Power
20%	6,424
10%	24,601
5%	96,195
2%	592,902
1%	2,360,495

Power and minimum effect of interest



Minimum reliably detectable effect and sample size required at different baselines





Power and significance threshold

Significance Level	Sample Size Per Variant, Required for 80% Power
85%	141,645
90%	181,032
95%	248,286
98%	366,660
99%	403,037

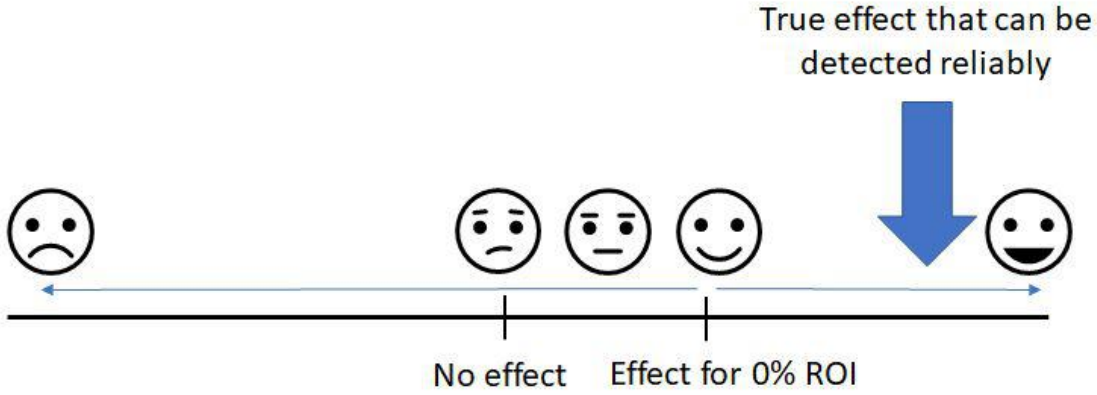
Power and baseline conversion rate

Baseline conversion rate	Sample Size Per Variant, Required for 80% Power
20%	20,149
10%	45,498
5%	96,195
2%	248,286
1%	501,771

Underpowered Tests

- The test has low power, ***relative*** to:

The minimum effect at which the test would be profitable



Examples of running underpowered tests

- **More common than you'd think: a sample of 115 A/B tests revealed 70% of them were underpowered.**

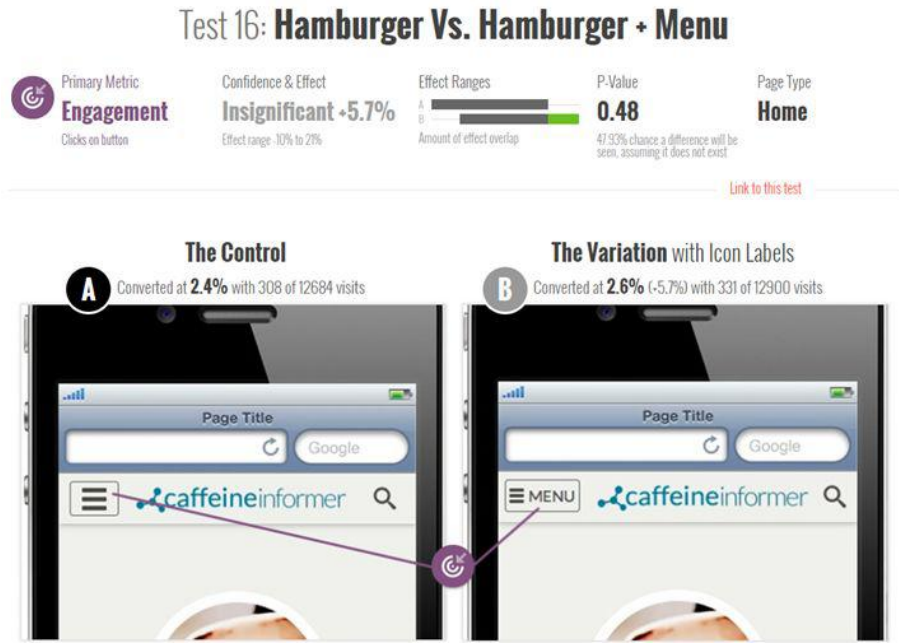
- **An actual example:**

  "At <company>, we recently stopped actively A/B testing […] we ran more than 70 conversion tests before making this decision and had only 3 significant winners."

  Poor tests or poor statistical design?

Examples of running underpowered tests

Why run such a test?

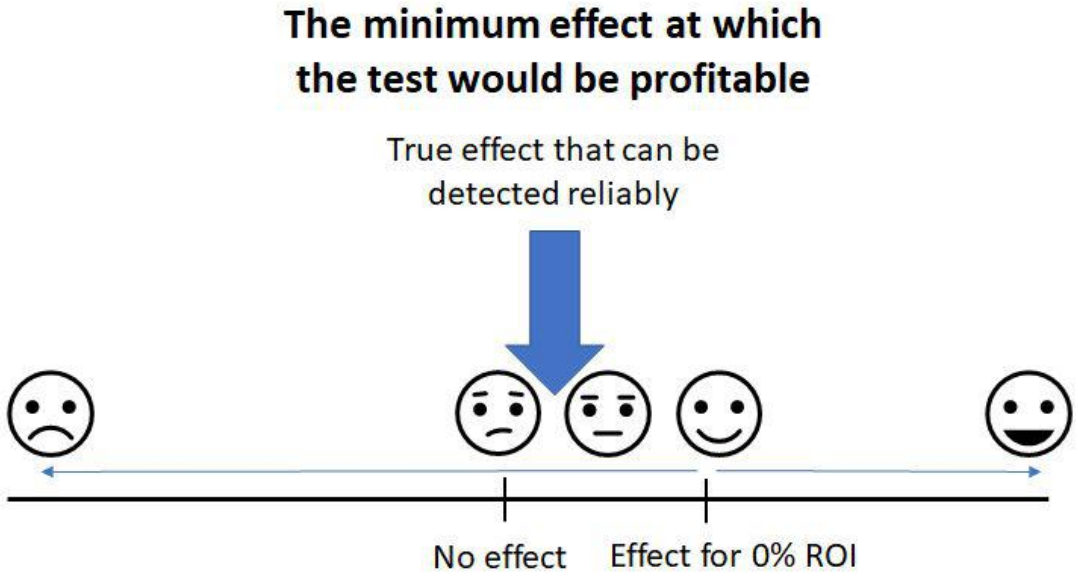


Examples of running underpowered tests

Why run such a test?

Overpowered Tests

- The test has high power, ***relative*** to:



A properly powered test

- **Is based on wisely chosen minimum effect of interest (effect size is both feasible and meaningful in the business sense).**
- **Has high enough probability to detect the chosen MEI so that you would not be sorry if you missed it.**
- **Does not include more users than needed to establish the effect with the required significance threshold.**

CXL institute

Assignment:

- You have 2 weeks free trial during which to test a new chat bot software. The website gets 100,000 users per week.
- The cost of the software, additional development, tracking etc. make a 0.0002 increase in CR (1% lift) interesting from a business perspective.
- The purchase is accompanied by some expenses and long-term investments so a 95% confidence threshold (0.05 significance threshold) is deemed acceptable.

Can the test be run in a reasonable manner given the above parameters? If not, will it be underpowered or overpowered?

CXL institute

Lesson Recap:

1.) We can now account for type II errors in A/B tests
2.) Understanding the relationship between sample size and other test parameters means we can plan tests more intelligently
3.) Recognizing underpowered and overpowered tests can help prevent running tests with unsuitable parameters

CXL institute

Lesson Resources:

- https://www.analytics-toolkit.com/statistical-calculators/ (free trial, sample size and power curves)
- https://www.gigacalculator.com/calculators/power-sample-size-calculator.php (free, some limitations)
- R power.prop.test built-in function (free, absolute difference of proportions calculations only, steeper learning curve) https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/power.prop.test