<https://rpubs.com/superseer/ab_testing>

# A/B Testing

#### [*Niranjan Shetty*](http://niranjan.co/)

* [1 Introduction](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#introduction)
  + [1.1 What A/B testing isn’t good for](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#what-ab-testing-isnt-good-for)
  + [1.2 When did people first start to use A/B testing?](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#when-did-people-first-start-to-use-ab-testing)
  + [1.3 Note on metrics](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#note-on-metrics)
  + [1.4 Binomial Distribution](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#binomial-distribution)
  + [1.5 Confidence Interval](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#confidence-interval)
  + [1.6 Hypothesis Testing](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#hypothesis-testing)
  + [1.7 Comparing two samples](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#comparing-two-samples)
  + [1.8 Practical significance](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#practical-significance)
  + [1.9 Size vs Power trade-off](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#size-vs-power-trade-off)
* [2 Metrics](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#metrics)
  + [2.1 Characteristics of a metric:](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#characteristics-of-a-metric)
  + [2.2 Categories of metrics](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#categories-of-metrics)
  + [2.3 Absolute vs Relative](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#absolute-vs-relative)
  + [2.4 Calculating variability](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#calculating-variability)
  + [2.5 Non-parametric methods](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#non-parametric-methods)
* [3 Designing an Experiment](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#designing-an-experiment)
  + [3.1 Unit of Diversion](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#unit-of-diversion)
  + [3.2 Population vs Cohort](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#population-vs-cohort)
* [4 Analyzing Results](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#analyzing-results)
  + [4.1 Sanity Tests](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#sanity-tests)
  + [4.2 Analysis with a Single Metric](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#analysis-with-a-single-metric)
  + [4.3 Multiple checks](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#multiple-checks)
  + [4.4 Gotchas](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#gotchas)
* [5 Summary](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#summary)
* [6 References](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#references)
* [7 Glossary](https://rstudio-pubs-static.s3.amazonaws.com/201749_9fc280333a5c4f448687e1d99b9bdf76.html#glossary)

This document consists of notes based on the course ‘[A/B Testing](https://www.udacity.com/course/progress#!/c-ud257)’ on Udacity by [Diane Tang](http://research.google.com/pubs/author3770.html), [Carrie Grimes](http://research.google.com/pubs/author31907.html), and Caroline Buckey. These notes are not comprehensive. Please refer to the course page for details.

# 1 Introduction

A/B testing is a methodology for testing product changes. You split your users to two groups - the control group which sees the default feature, and an experimental group that sees the new features. Examples of A/B testing include Amazon personal recommendations, ranking change in Linkedin (e.g. whether to show a news item or a possible connection to a user). The goal of A/B testing is to design an experiment that is going to be robust and gives you repeatable results so that you can make a good decision about whether to launch that product or feature.

## 1.1 What A/B testing isn’t good for

A/B testing is not good for testing new experiences. It may result in change aversion (where users don’t like changes to the norm), or a novelty effect (where users see something new and test out everything).

The two things with new experiences is a) having a baseline and b) how much time needs to be allowed for the users to adapt to the new experience, so you can say what is going to be the plateaud experience and make a robust decision.

Finally, A/B testing cannot tell you if you are missing something.

In these cases, user logs can be used to develop hypothesis that can then be used in an A/B test. A/B testing gives broad quantitiative data, while other techniques such as user research, focus groups, human evaluation give you deep qualitative data

## 1.2 When did people first start to use A/B testing?

A/B testing originated in agriculture. In medicine, their version of A/B testing is clinical trials. The difference between clinical trails and online A/B testing is that in clinical trials you have fewer patients with a lot of information on them. In online A/B tests, you have millions of users but limited information on the users.

## 1.3 Note on metrics

Generally, rate is used when you want to measure the usability of the site, and probability when you want to measure the impact.

## 1.4 Binomial Distribution

For a binomial distribution with probability *p* p , the mean is given by *p* p and the standard deviation is *p*∗(1−*p*)/*N* − − − − − − − − − − −  √  p∗(1−p)/N where *N* N is the number of trials. A binomial distribution can be used when

1. The outcomes are of 2 types
2. Each event is independent of the other
3. Each event has an identical distribution (i.e. *p* p is the same for all)

## 1.5 Confidence Interval

A confidence interval indicates the range within which the mean is expected to fall in multiple trials of the experiment.

For e.g., consider *p* ^  p^ - the proportion of users that click, where *N* N is the number of users. Let us assume a binomial distribution (this requires *Np* ^ >5 Np^>5 and *N*(1−*p* ^ )>5 N(1−p^)>5 ). The margin of error is given by

*m*=*z*∗*se* m=z∗se

*m*=*z*∗*p* ^ .(1−*p* ^ )*N*  − − − − − − − −  √  m=z∗p^.(1−p^)N

For a 95% confidence interval, z = 1.96.

## 1.6 Hypothesis Testing

The null hypothesis states that the difference between the control and experiment is due to chance. If *p* *cont*  pcont and *p* *test*  ptest are the control and test probabilities, then according to the null hypothesis

*H* 0 :*p* *exp* −*p* *cont* =0 H0:pexp−pcont=0

The alternate hypothesis is that

*H* 1 :*p* *exp* −*p* *cont* ≠0 H1:pexp−pcont≠0

## 1.7 Comparing two samples

For comparing two samples, we calculate the pooled standard error. For e.g., suppose *X* *cont*  Xcont and *N* *cont*  Ncont are the control number of users that click, and the total number of users in the control group. Let *X* *exp*  Xexp and *N* *exp*  Nexp be the values for the experiment. The pooled probability is given by

*p* ^  *pool* =*X* *cont* +*X* *exp* *N* *cont* +*N* *test*   p^pool=Xcont+XexpNcont+Ntest

*SE* *pool* =*p* ^  *pool* ∗(1−*p* ^  *pool* )∗(1*N* *cont*  +1*N* *test*  ) − − − − − − − − − − − − − − − − − − − − − − − − − − − −  √  SEpool=p^pool∗(1−p^pool)∗(1Ncont+1Ntest)

*d* ^ =*p* ^  *exp* −*p* ^  *cont*  d^=p^exp−p^cont

*H* 0 :*d*=0 where *d* ^ ∼*N*(0,*SE* *pool* ) H0:d=0 where d^∼N(0,SEpool)

If *d* ^ >1.96∗*SE* *pool*  d^>1.96∗SEpool or *d* ^ <−1.96∗*SE* *pool*  d^<−1.96∗SEpool then we can reject the null hypothesis and state that our difference represents a statistically significant difference

## 1.8 Practical significance

Practical significance is the level of change that you would expect to see from a business standpoint for the change to be valuable. What is considered practically significant can vary by field. In medicine, one would expect a 5,10 or 15% improvement for the result to be considered practically significant. At Google, for example, a 1-2% improvement in click through probability is practically significant.

The statistical significance bar is often lower than the practical significance bar, so that if the outcome is practically significance, it is also statistically significant.

## 1.9 Size vs Power trade-off

One of the decisions is to determine the number of data points needed to get a statistically significant result. This is called statistical power. Power has an inverse trade-off with size. The smaller the change you want to detect or the increased confidence you want to have in the result, means you have to run a larger experiment.

As you increase the number of samples, the confidence interval moves closer to the mean

*α*=*P*(reject null | null true) α=P(reject null | null true)

*β*=*P*(fail to reject null | null false) β=P(fail to reject null | null false)

1−*β* 1−β is referred to as the sensitivity of the experiment, or statistical power. People often choose high sensitivity, typically around 80%.

For a small sample, *α* α is low and *β* β is high. For a large sample *α* α remains the same but *β* β goes down (i.e. sensitivity increases). A good online calculator for determing the number of samples is [here](http://www.evanmiller.org/ab-testing/sample-size.html). As you change one of the parameters, your sample size will change as well. For example:

* If you increase the baseline click through probability (under 0.5) then this increases the standard error, and therefore, you need a higher number of samples
* If you increase the practical significance level, you require a fewer number of samples since larger changes are easier to detect
* If you increase the confidence level, you want to be more certain that you are rejecting the null. At the same sensivitiy, this would require increasing the number of samples
* If you want to increase the sensitivity, you need to collect more samples

The example below demonstrates the analysis of an A/B test:

N\_cont = 10072 # Control samples (pageviews)

N\_exp = 9886 # Test samples (pageviews)

X\_cont = 974 # Control clicks

X\_exp = 1242 # Exp. clicks

p\_pool = (X\_cont + X\_exp)/(N\_cont+N\_exp)

se\_pool = sqrt(p\_pool\*(1-p\_pool)\*(1/N\_cont + 1/N\_exp))

p\_cont = X\_cont/N\_cont

p\_exp = X\_exp/N\_exp

d = p\_exp - p\_cont

d

## [1] 0.02892847

m = 1.96\*se\_pool

m

## [1] 0.008717974

cf\_min = d-m

cf\_max = d+m

d\_min = 0.02 # Minimum practical significance value for difference

cf\_min

## [1] 0.0202105

cf\_max

## [1] 0.03764645

In the above example, since the minimum confidence limit is greater than 0 and the practical significance level of 0.02, we conclude that it is highly probable that click through probability is higher than 0.02 and is significant. Based on this, one would launch the new version

# 2 Metrics

Two types of checks

* Invariant checking: Metrics that shouldn’t change between your test and control
* Evaluation: High level business metrics, user experience with the product

How do we go about making a definition of a metric (for sanity checking)?

1. High level concept of metrics (e.g active users, CTR)
2. Details (e.g. how do you define user activity)
3. Take a set of metrics and summarize them into a single metric (e.g. overall evaluation criterion (OEC))

For evaluation, you can choose either one metric or a whole suite of metrics. If you have multiple metric, you can combine them into one metric, such as an objective function, or an Overall Evaluation Criterion (OEC) - a term that Microsoft uses.

The last situation is how generally applicable the metric is. If you are running a suite of A/B tests, it is preferable to have a metric that works across the entire suite.

User funnel indicates a series of steps taken by users through the site. It is called a funnel because every subsequent stage has fewer users than the stage above. Each stage is a metric - total count, rate, and probability (i.e. a unique user progressed down).

Some metrics may be difficult to measure because

* they don’t have access to data and/or
* it takes too long to collect

External data can be used. 3 categories of companies that gather data

1. Companies that collect data (e.g. Comscore, Nielsen)
2. Companies that conduct surveys (e.g. Pew)
3. Academic papers

The above can help you benchmark your own metrics against the industry

Internal data can be used as well. You could do

1. Retrospective analysis: Look at historic data to look at changes and see the evaluation
2. Surveys and User experience research: This helps you develop ideas on what you want to research

The problem with these studies is that they show you correlation, and not causation, compared to running an experiment.

Talk to your colleagues about what ideas they think make sense for metrics.

You can gather additional data by

1. User Experience Research (UER) - high depth on a few users. This is good for brainstorming. You can also use special equipment in a UER (e.g. eye movement camera) that you cannot use on your website. You may want to validate the results using retrospective analysis
2. Focus groups: Medium depth and # of participants. Get feedback on hypotheticals, but may run into the issue of groupthink
3. Surveys have low depth but high # of participants: Useful for metrics you cannot directly measure. Can’t directly compare with other metrics since population for survey and internal metrics may be different.

Example of a metric:

High level metric - Click-through probability

1. Def #1 (Cookie probability): For each , number of cookies that click divided by number of cookies
2. Def #2 (Pageview probability): Number of pageviews with a click within divided by number of pageviews
3. Def #3 (Rate): Number of clicks divided by number of pageviews

You may have to filter out spam and fraud to de-bias the data. One way to figure out if you are biasing or de-biasing the data by filtering, is to slice your data and then calculate the metric for each slice after filterig. If you are affecting any slide disproportionately, then you may be biasing your data with filtering

To remove any weekly effects when looking say at total active cookies over time, use week-over-week i.e. divide current data by data from a week ago. Alternately, one can use year-over-year.

## 2.1 Characteristics of a metric:

1. Sensitivity and Robustness: Whether the metric is sensitive to changes you care about, and is robust to changes you don’t care about (e.g. mean is sensitive to outliers, median is robust but not sensitive to changes to small group of users). This can be measured by using prior experiments to see if the metric moves in a way that intuitively make sense. Another alternative is to do A/A tests to see if the metric picks up any spurious differences
2. Distribution: Obtained by doing a distribution on the retrospective data

## 2.2 Categories of metrics

4 categories of metrics

1. Sums and counts (e.g. # of users who visited a page)
2. Distributional metrics - means, medians, percentiles
3. Probabilities and rates
4. Ratios

## 2.3 Absolute vs Relative

The simplest way to compare metrics for test and control is to take a difference.

If you are running a lot of experiments you want to use the relative difference i.e the percentage change. The main advantage of computing the percentage change is that you only have to choose one practical significance boundary to get stability over time. If you are running a lot of experiments over time, your metrics are probably changing over time. Using relative difference helps here by having to use one practical significance boundary rather than change it as the system changes. The main disadvantage is variability, relative differences such as ratios are not as well behaved as absolute differences

## 2.4 Calculating variability

We want to check the variability of a metric to later determine the sizing of the experiment and to analyze confidence intervals and draw conclusions. If we have a metric that varies a lot, then the practical significance level that we are looking for may not be feasible.

To calculate the confidence interval, you need

* Variance (or standard deviation)
* Distribution

For a binomial distribution, your estimated variance is *p*(1−*p*)/*N* p(1−p)/N . Estimated variance of the mean is *σ* 2 /*N* σ2/N . If the underlying data is normal, then the median will be normal. But if the underlying data is not normal, the median may not be normal. The mean is typically normally distributed irrespective of the distribution of the underlying data due to Central Limit Theorem. Difference in counts may be normally distributed. The variance of the difference will be the sum of the variances of each of the 2 counts. Rates tend to have a poisson distribution, and the variance for a poisson is just the mean. For ratios of test over control, both the mean and the variance depends on the distribution of the test and control metrics.

e.g.

x <- c(87029, 113407, 84843, 104994, 99327, 92052, 60684)

stder <- sd(x)/sqrt(length(x))

conf95\_min = mean(x) -1.96\*stder

conf95\_max = mean(x) + 1.96\*stder

conf95\_min

## [1] 79157.54

conf95\_max

## [1] 104367

## 2.5 Non-parametric methods

This is a way to analyze data without making an assumption about the distribution. At Google, it was observed that the analytical estimates of variance was often under-estimated, and therefore they have resorted to use empirical measurements based on A/A test to evaluate variance. If you see a lot of variability in a metric in an A/A test, it is probably too sensitive to be used. Rather than do several multiple A/A tests, one way is to do a large A/A test, and then do bootstrap to generate small groups and test the variability.

With A/A tests, we can

1. Compare result to what you expect (sanity check)
2. Estimate variance empirically and use your assumption about the distribution to calculate confidence
3. Directly estimate confidence interval without making any assumption of the data

In summary, different metrics have different variability. The variability may be high for certain metrics which makes them useless even if they make business or product sense. Computing the variability of a metric is tricky and one needs to take a lot of care.

For a lot of analysts, a majority of the time is spent is validating and choosing a metric compared to actually running the experiment. Being able to standardize the definitions was critical in the test. When measuring latency, are you talking about when the first byte loads and when a last byte loads. Also, for latency, the mean may not change at all. The signals (e.g. slow/fast connections or browsers) causes lumps in the distribution, and no central measure works. One needs to look at the right percentile metric. The key thing is that you are building intuition, you have to understand data, and the business, and work with the engineers to understand how the data is being captured.

# 3 Designing an Experiment

1. Choose subject: What are the units in the population you are going to run the test on? (unit of diversion)
2. Choose population: What population are you going to use (US only?)
3. Size
4. Duration

Typically you want to assign people and not events since the same user may see different changes. If you use a person, you typically use a cookie which may change by platform. The alternative then is to use a user id

## 3.1 Unit of Diversion

Commonly used units of diversion are:

1. User identifier (id): Typically the username or email address used on the website. It is typically stable and unchanging. If user id is used as a unit of diversion, then it is either in the test group or the control group. User ID is personally identifiable
2. Anonymous id: This is usually an anonymous identifier such as a cookie. It changes with browser or device. People may often refresh their cookies every time they log in. It is difficult to refresh a cookie on an app or a phone compared to the computer.
3. Event: An event is a page load that can change for each user. This is used typically for changes that is not user facing.

Lesser used units of diversion are

1. Device id: Typically available for mobile devices. It is tied to a specific device and cannot be changed by the user.
2. IP address: The ip address is location specific, but may change as the user changes location (e.g. testing on infrastructure change to test impact on latency)

3 main considerations in selecting an appropriate unit of diversion

1. Consistency
2. Ethical
3. Variability

Variability is higher when it is calculated empirically than when calculated analytically. This is because the unit of analysis (i.e. the denominator in the metric) is different from the unit of variability.

E.g. If unit of diversion is a query, then coverage (= #queries with ads/ # queries) will have lower variability compared to using a cookie as a unit of diversion. This is because when a query is used, the unit of diversion matches the unit of analysis (which is the denominator of the metric i.e. query)

In the medical industry, users are paired with each other based on location, demographics. However, given how little information there is on users on the internet, this is not widely practiced.

e.g. Consider an experiment where you are analyzing data for a particular region (NZ), and for the rest of the world. For the global data that includez NZ and the rest of the world, what is the pooled standard error?

For the other, you have

N\_cont = 50000 + 6021

X\_cont = 2500 + 302

N\_exp = 50000 + 5979

X\_exp = 2500 + 374

p\_cont = X\_cont/N\_cont

p\_cont

## [1] 0.05001696

p\_exp = X\_exp/N\_exp

p\_exp

## [1] 0.05134068

p\_pool = (X\_cont + X\_exp)/(N\_cont + N\_exp)

se\_pool = sqrt(p\_pool\*(1-p\_pool)\*(1/N\_cont + 1/N\_exp))

se\_pool

## [1] 0.00131081

Since *abs*(*p* *cont* −*p* *exp* )<1.96∗*se* *p* *ool* abs(pcont−pexp)<1.96∗sepool , the global difference is not statistically significant

## 3.2 Population vs Cohort

A cohort is like an entering class for an analysis. A cohort may make more sense to look at a population when:

1. Looking for leaning effects
2. Examining user retention
3. Want to increase user activity
4. Anything that requires the user to be established

Practical considerations in experimental design

1. Duration
2. When to run the experiment
3. Fraction of the traffic to send to the experiment

Two different types of learning effects

1. Change aversion
2. Knowledge effect

When users first encounter a change they will react, but will eventually adapt to a change.

# 4 Analyzing Results

## 4.1 Sanity Tests

One of the first things to do once you finish collecting experimental data is to analyze the invariants. This is done by calculating the values for one or more invariants on the test and control group, and check if the difference is statistically significant. For e.g. if the values for an invariant (say total # of cookies) are x and y, then calculate the *se* se as 0.5∗0.5*x*+*y*  − − − − −  √  0.5∗0.5x+y , since one would expect the same number of cookies in both groups. Then calculate the margin as 1.96∗*se* 1.96∗se . If the margin is greater than *x*/(*x*+*y*)−*y*/(*x*+*y*) x/(x+y)−y/(x+y) , then the difference of the invariant is insignificant. However if the difference is greater than the margin, then the difference is insignifiant and needs to be investigated further

An example is provided below:

control\_event\_ct = c(2451,2475,2394,2482,2374,1704,1468)

test\_event\_ct = c(2404,2507,2376,2444,2504,1612,1465)

control\_total = sum(control\_event\_ct)

test\_total= sum(test\_event\_ct)

p\_cont = control\_total/(control\_total+test\_total)

p\_test = test\_total/(control\_total + test\_total)

p\_cont

## [1] 0.5005871

p\_test

## [1] 0.4994129

se = sqrt(0.5\*0.5/(control\_total+test\_total))

margin = 1.96\*se

p\_cf\_min = 0.5 - margin

p\_cf\_max = 0.5 + margin

p\_cf\_min

## [1] 0.4944032

p\_cf\_max

## [1] 0.5055968

The most common reasons for sanity checks failing is data capture. Other reasons could be experimental set-up, for e.g., where there is a filter on the test but not on the control

## 4.2 Analysis with a Single Metric

# Data provided from test

Xs\_cont = c(196, 200, 200, 216, 212, 185, 225, 187, 205, 211, 192, 196, 223, 192)

Ns\_cont = c(2029, 1991, 1951, 1985, 1973, 2021, 2041, 1980, 1951, 1988, 1977, 2019, 2035, 2007)

Xs\_exp = c(179, 208, 205, 175, 191, 291, 278, 216, 225, 207, 205, 200, 297, 299)

Ns\_exp = c(1971, 2009, 2049, 2015, 2027, 1979, 1959, 2020, 2049, 2012, 2023, 1981, 1965, 1993)

Xs\_cont\_sum = sum(Xs\_cont)

Ns\_cont\_sum = sum(Ns\_cont)

Xs\_exp\_sum = sum(Xs\_exp)

Ns\_exp\_sum = sum(Ns\_exp)

p\_cont = Xs\_cont\_sum/Ns\_cont\_sum

p\_exp = Xs\_exp\_sum/Ns\_exp\_sum

# Empirical standard error and count provided

empirical\_se = 0.0062

empirical\_ct = 5000

se = (sqrt(1/Ns\_cont\_sum + 1/Ns\_exp\_sum))\*empirical\_se/sqrt(1/empirical\_ct + 1/empirical\_ct)

# Calculating the cf for the difference

d = p\_exp-p\_cont

margin = se\*1.96

d\_c95min = d - margin

d\_c95max = d + margin

# Sign test

diff\_sign = Xs\_exp/Ns\_exp - Xs\_cont/Ns\_cont

pos\_diff = sum()

One thing to be wary of is Simpson’s paradox, where the effect in aggregate may indicate one trend, and at a granular level may show an opposite trend.

## 4.3 Multiple checks

The more things you test, the more likely you are to see significant difference just by chance. This is a problem, but since it is not repeatable for the same metric across multiple attempts, there is a way out. One can do multiple runs of the experiment, or alternately bootstrap. There is another technique called multiple comparison that adjusts your significance levels that accounts for how many metrics or tests you are doing.

For e.g. if you had 10 metrics where you used a 95% confidence interval for each metric, what is the probability that one of the metrics will show up as a false positive?

p1 = 0.99

p\_nofp = p1^10

p\_fp = 1-p\_nofp

p\_fp

## [1] 0.09561792

As you increase the number of metrics, you can use a higher confidence level to overcome false positives.

A different method used in practice is Bonferroni correction. It has the advantages of being simple, makes no assumptions, and guaranteed to give *α* *overall*  αoverall as low as you have specified.

To use it, calculate

*α* *individual* =*α* *overall* *n*  αindividual=αoveralln

For e.g. if you want *α* *overall*  αoverall to be 0.05 and there are 5 metrics then *α* *individual*  αindividual will be 0.05/3=0.01666 0.05/3=0.01666

Bonferroni methods may be very conservative. Alternatives include [closed testing procedure](https://en.wikipedia.org/wiki/Closed_testing_procedure), [Boole-Bonferroni bound](https://en.wikipedia.org/wiki/Boole%27s_inequality) and [Holm-Bonferroni method](https://en.wikipedia.org/wiki/Holm%E2%80%93Bonferroni_method). The *α* *overall*  αoverall above is often referred to as the familywise error rate (FWER). Another measure is the [contol false discovery rate](https://en.wikipedia.org/wiki/False_discovery_rate) (FDR) defined as the (# false positives)/(#rejections). CDR makes sense if you have a large number (200) metrics.

An alternative to using multiple metrics is to use an ‘Overall Evaluation Criterion’ (OEC).

## 4.4 Gotchas

Effect may ramp out as you implement the change. There could be seasonal effects. For e.g. students on summer break have very different behavior than when they come back. Similarly during black friday and other holidays. One of the ways is to leave a small sample out as a hold-out to track them over time.

# 5 Summary

* Decision to launch change should be guided by business reasons and not just
* Opportunity cost of launching the change (engineering costs, user experience), based on not launching it

# 6 References

1. [Additional Techniques for Brainstorming and Validating Metrics](https://storage.googleapis.com/supplemental_media/udacityu/3954679115/additional_techniques.pdf?GoogleAccessId=1069728276824-fdhtlb98k1m9qrmdgj4jgc7gjp2l1lsm@developer.gserviceaccount.com&Expires=1438057124&Signature=aisqm5m1VJX0%2Bk6ZKBlSAm6JxjzRbOVmr%2BGey%2BkDtTVYL%2BJzfI133AG%2BsnTFhmrr/b9xOWsf/6oBY%2B0Fy6WRanrBICYKMkJ/xpXroG1zSdp39GzHYpzA30H7btkoWMheAbmHqVS36gaWmr4PXyW9dNPKK2tq/QnJT04vnkjMsgU%3D)
2. [Comparing 2 populations: Binomial and Poisson](http://www.stat.wisc.edu/~wardrop/courses/371chapter9b.pdf)
3. [Tests for Two Poisson Means](http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/PASS/Tests_for_Two_Poisson_Means.pdf)
4. [Final Project Template](https://docs.google.com/document/d/16OX2KDSHI9mSCriyGIATpRGscIW2JmByMd0ITqKYvNg/edit)
5. [Large Scale Validation and Analysis of Interleaved Search Evaluation](http://www.cs.cornell.edu/People/tj/publications/chapelle_etal_12a.pdf)
6. [Overlapping Experimental Infrastructure: More, Better, Faster Experimentation](http://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/36500.pdf)

# 7 Glossary

change aversion: users averse to change reduce usage

interleaved experiment: an experiment in which you expose the same user to both A and B at the same time

interuser experiment: an experiment in which you expose users to either A or B.

intra-user experiment: is where you expose the same user to an experiment being on and of at different times.

novelty effect: users see something new and test out everything

retrospective analysis: look at historical data, observe changes, and conduct an evaluation

userflow: Shows the flow of users through the site, often referred to as a customer funnel.