Basics of A/B Experimentation, Part 1

Roger Longbotham Process Performance Management https://ppmdatascience.solutions

In this series I will be giving the basic knowledge one needs to successfully conduct A/B experiments online. Note that there are a number of topics beyond the basics that are helpful in increasing the trustworthiness and power of the experiments or for analysis in special cases.

In this White Paper I will be addressing the following topics:

- Why run experiments?
- Online experimentation lifecycle
- Choosing metrics
- Strategies for reducing bias and variability
- Measuring variability
- Significance testing

Why Run Experiments?

If you want to make a change to your website, what are the alternatives?

1. Make the change (or not) based on your (or someone's) opinion. This could be a design or usability expert, a product owner, a committee or a manager.
2. Collect data relevant to the change to inform the decision.
3. Run a valid, randomized controlled experiment to test the change. This will be an objective determination of how the users react to the change, then you can make an informed decision.

Example. In the early 2000s Greg Linden at Amazon had an idea: when someone placed an item into their shopping cart, show recommendations based on that item. (Ref: link). He mocked it up and showed it around. Most thought it had merit but a Senior Vice President forbid him from working further on the project – too risky. The argument FOR the feature was that you may be able to increase the average basket size and cross-sell into other categories. The argument the manager made AGAINST it was it would distract people from checking out, reducing conversion rate. At that time, conversion rate was king, so the SVP told him to drop the project.

Instead of totally dropping the project, Greg ran a test of the feature on the Amazon experimentation platform. The results of the test were unambiguous. It was a clear winner and Amazon was giving up a lot of sales by not having the feature on the site.

I have personally run into resistance from managers of websites who think their knowledge and experience is enough to make good decisions about changes. One manager proclaimed, "It's in our DNA!" I have run enough experiments to know that I, personally, cannot predict which feature will win and I have not run into anybody who can predict winners consistently.

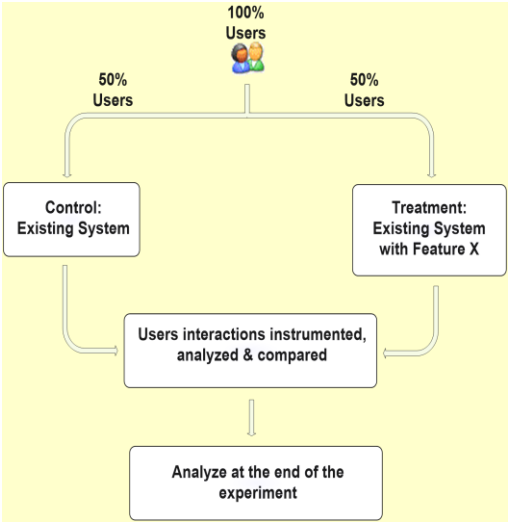
When I was running online experiments at Microsoft we collected statistics on which proposed changes make an improvement over the existing user experience. Note that these are ideas that were expected to make an improvement to the key metrics. From our data, one-third of these changes actually hurt the

key metrics, one-third made no difference (statistically) and one-third were helpful. If a site can sort through the proposed changes (with experimentation) and implement only those that are proved to be helpful they will make a lot more progress than if they relied on opinion and rolled out all those that were thought would work.

What about the second alternative? i.e. collect observational data and conduct a (perhaps sophisticated) analysis to determine if the change is helpful. There are two problems with this. For almost all proposed changes, this idea has never been tried before so there is no historical data to rely on. If the idea had been tried before there is going to be a large amount of uncontrolled variation that will affect the analysis. Online site traffic is very dynamic with large changes within a day, within a week, week to week and trending over time. No analysis is able to remove this non-stationarity as well as a randomized controlled experiment and in almost every case an observational analysis will have significant unknown biases giving misleading results.

Online Experimentation Lifecycle

First, the big picture:



The concept is simple

- Randomly split traffic between two (or more) versions to be compared
    - Version A is the baseline or Control
    - Version B is the change or Treatment
- Collect metrics of interest
- After the experiment has completed, analyze the data for the metrics for the two versions to determine if there is a statistically significant difference between them.

A randomized controlled experiment is the best way to prove causality, i.e. that the difference in the metrics that is observed is due to the change that was made and not just correlation. However, we must run statistical tests to confirm the observed difference in the versions is not due to chance.

Plan the experiment

The Experimentation Lifecycle begins with a page or product owner, a usability expert or manager who has an idea that has the potential to improve some aspect of the website. In most cases, the change is expected to improve one or more key metrics. (I have run experiments where the change is not expected to positively effect user experience. One example is a change to the underlying software for the site that would allow new capabilities. We did not expect a positive change to key metrics and we hoped the change would not negatively impact any key metrics.) The experimenter needs to determine which key metrics are expected to be improved. Will the change increase revenue? Click-throughs? Conversion rate? Other metric? There is a very good statistical reason for doing this. I once ran a seminar for experimenters at Amazon and one engineer wanted to peruse the hundreds of metrics that were produced during an experiment and make the decision after viewing the result. Here's the problem with that. The truth of whether there is a difference or not is unknown to the experimenter so we have to run statistical hypothesis tests to help us decide if there is a difference. Statistical

comparisons are set up with the test showing "no difference" 95% of the time if there is truly no difference between the Treatment and Control but there will be a statistically significant difference 5% of the time even though there is no difference. (Note that the 5% false positive rate can be changed by the experimenter, but 5% is the most common value used.) If I view hundreds of metrics after the experiment there will be a large number that are positive and statistically significant even when there is no difference between the two versions. For this reason we should determine our primary decision-making metric in advance to avoid making a decision based on random variation in the metrics. (Otherwise we may be guilty of "cherry-picking" the results.) Generally, we will want to roll out the Treatment only if the chosen key metric is statistically significantly better than the Control and if it does not negatively impact other key metrics. (More detail on this in Part 2 of the series.)

A change may be a visible change to the user (although each user would only see one version of the page) or it could be a change that is not visible. For example, at Amazon we regularly ran experiments to tune the parameters of the recommendation algorithm.

Most experiments should not have all users in them. The experimenter needs to determine which users may be affected by the change and only allow those users into the experiment. The most common way to do this is to collect data on everyone and then filter the data for analysis to include only users who were determined could have been affected.

Finally, one needs to get an estimate on how long to run the experiment. This will be covered in Part 3.

Run the experiment

Monitor the results of the experiment while it is running but with some caveats. First, after the first day or two you don't need to check the results more than once or twice a day. If the experimental Treatment is VERY negative (and very statistically significant) you can shut down the experiment and trouble shoot. Check all diagnostic metrics and look for other clues as to what may have gone wrong with the instrumentation or other aspect of the setup. It's also possible the proposed change was not such a good idea. One should also consider the possibility that a change in how the idea is implemented could improve the results. Another experiment would be needed to test this.

One should be very cautious about ending an experiment early because the Treatment is seen to be positive and statistically significant. Unless the Treatment is extremely statistically significant and there are no warning signals from diagnostic or guardrail metrics (discussed in Part 2) the experiment should continue for the planned run length.

Analyze and follow-up

Once the experiment has run for the planned length of time conduct a full analysis of the experiment. The determination of whether the Treatment is successful or not depends on 1) the primary metric identified earlier being positive and statistically significant, 2) if other key metrics are negative and statistically significant, and 3) whether there are any warning signals from diagnostic or guardrail metrics. One should always do a full analysis and not trust *prima facia* results since unexpected biases can often affect the results.

If the Treatment is determined to be successful, roll out the change. If the treatment is neutral (i.e. primary metric is not statistically significant) the experimenter could decide to continue running the experiment longer if it is positive or stop the experiment, trouble shoot and run a variation on the

Treatment that may perform better. If the Treatment is negative, either move on to test another idea or trouble shoot and test a variation on the Treatment. In any case, you can conduct an analysis for subsets of the visitors to see if the Treatment had a differential impact on them. This can give you clues as to why the Treatment is successful or not. It can also help give you insights to your visitors' behavior and other ideas that may improve your key metrics.

If code was added to enable the experiment to run, the experimenter should always remove the code after the experiment. In addition, best practice is to do a short write-up on the experiment: objective, Variant(s) with screenshots, experiment parameters and results. This is quite useful to enable organizational learning about what works and what doesn't work.

Choosing Metrics

A more detailed coverage of metrics will be undertaken in Part 2, but it is appropriate to give some initial guidance at this early stage.

Each site should have a set of standard key metrics that track how well the site is doing with visitors. These may be known as Key Performance Indicators (KPIs) and should always be consistent with the overall goals or objectives of the site.

Type of site / Example	Business Objective
Content / CNN, MSN, Foxnews,	Attract visitors, sell ads
Ecommerce / Amazon, Taobao, Etsy	Sell first or third party items
Support / support.microsoft.com, eHow	Support product, sell ads
Gaming / Xbox, Zynga	Subscription, purchases
Marketing / Lexus, Tourism bureau	Send people to buy offline or online
Social Networking / facebook, qq, linkedin	Attract visitors, sell ads, upgrade
Community / Technet, AMA	Sell ads, support offline community
Search / Google, Baidu, Bing	Attract visitors, sell ads, send to other site
Services / VOIP, ISP, data storage, travel	Subscription, sell services, commission

Figure 1. Examples of Business Objectives for Different Types of Websites

As mentioned earlier, each experimenter should choose one key metric to be the primary determinant of success for an experiment. I call this the Primary Metric (PM) for the experiment. Of course, different experiments may have different PMs but it is important that the PM be identified prior to starting the experiment. It is often the case that the thing you would like to measure as a determinant of success is not measurable, in which case you may use a surrogate or proxy (more on this in Part 2).
There are some metrics that are fairly universal to websites:

1. A metric for the amount of traffic to the site, usually measured by number of unique visitors to the site in a given time period (by day, week, month, etc.)
2. Metrics for the activity of visitors.

a. **Engagement**. In order to measure how engaged a visitor may have been, most sites define a period when the visitor seems to be engaged as a session. Typically, a **session** is defined as a series of activities of a user on the site until there is a 30 minute gap in activity. If the user returns after this 30 minute gap, a new session is started. Some metrics based on sessions are
    i. Length of session (in seconds, minutes, etc.)
    ii. Number of actions of a particular type during a session (clicks, pageviews, categories visited, number of items purchased, etc.)
    iii. Number of sessions by a user over a period of time (per week, month or during experiment)

    In addition to measuring engagement as activities per session, most sites measure activities of a user over a time period or for the duration of the experiment.

b. **Goals attained**. If there are specific goals for visitors, whether a visitor attains each goal can be a key metric. Some examples are
    i. If visitor signed up for the newsletter,
    ii. If visitor downloaded the featured content,
    iii. If visitor gave their email address,
    iv. If visitor left a product review.

Key metrics may also be specific for different types of websites:

Type of site / Example	Sample Primary Metrics (KPIs)
Content / CNN, MSN, Foxnews	# of visitors, engagement, ad clicks
Ecommerce / Amazon, Taobao, Etsy	# of units sold; $, €, ¥; loyalty
Support / support.microsoft.com, eHow	Survey results, # visitors, ad clicks
Gaming / Zynga, Xbox	Purchases, subscriptions sold/renewed
Marketing / Lexus, Tourism bureau	# visitors, engagement, clicks to sites
Social Networking / facebook, qq, linkedin	# visitors, engagement, ad clicks, # upgrades
Community / Technet, ASA	# visitors, engagement, ad clicks
Search / Google, Baidu, Bing	# visitors, engagement, ad clicks
Services / VOIP, ISP, data storage, travel	Subscriptions, services sold, commissions

Figure 2. Examples of Key Metrics for Different Types of Websites

Not all metrics that are important for a website may be useful for comparing the variants in an experiment. For example, most sites are keenly interested in increasing (and hence measuring) number of visitors. However, the count of number of visitors is not a good metric to compare variants. Since we are randomizing users into each variant (see next section for more on randomization) the number of users in each variant is determined by the percentage assigned in the planning stage and usually has nothing to do with how well the variant is performing. In fact, since most sites use randomization by

user, the metrics of interest for comparison are metrics like number of clicks per user, revenue per user, sessions per user, clicks per session, etc.

Diagnostic metrics, guardrail metrics and other metrics you may need are discussed in Part 2.

Strategies for Reducing Bias and Variability

Two of the key characteristics of a metric or test result are bias and variability. We will discuss ways to measure variability of our metrics in the next section, but variability will always be present. The key is to reduce it as much as possible. This is one way to improve the power of the experiment (more on power in Part 3). An estimate is **biased** if, on average, it is either higher than or lower than the thing it's estimating. Ideally, we want to eliminate bias altogether to have a completely trustworthy experiment. There are many things the experimenter and the experimentation platform can do to eliminate bias but one must continually be on guard to detect things that can be causing bias.

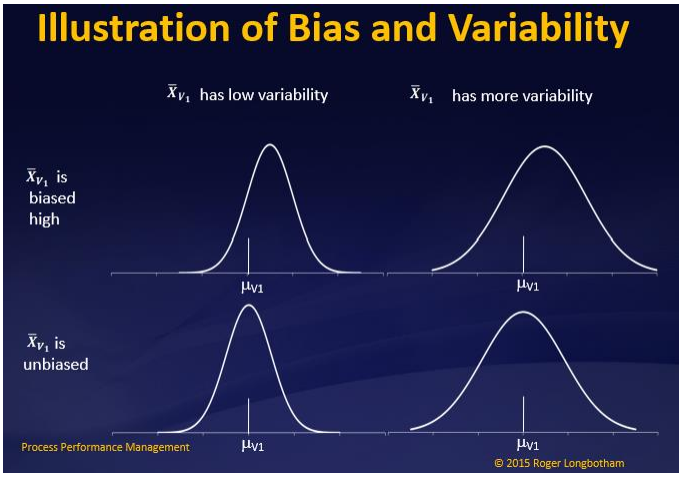


Figure 3. Graphical Depiction of Bias and Variability

Explanation of Figure 3.

There are four scenarios depicted. Each shows the true value of the quantity being estimated and the distribution of the estimate. In the upper row, both estimates are biased by the same amount, but the one on the right has more variability. In the lower row, both estimates are unbiased. We want an estimate that is unbiased and has low variability, like the lower left estimate.

Reducing or eliminating bias

There are a large number of things that can be biasing your results. I will be discussing some of the strategies we have used to eliminate bias. However, there is always the possibility of bias creeping into the experiment that these strategies don't address. Many or most of these will be detected by the diagnostic metrics (in Part 2) but one must always be vigilant for additional biases that are not detected by these metrics.

A basic design principle for AB online experimentation is that all variants should be as alike as possible except for the application of the treatment or change that is intended. Any unintended difference between the variants is a source of bias for your results. Some examples of unintended differences that I have observed:

- An update to the site that is not applied equally to all variants
- Another experiment that is running on the same population that has a differential effect on the variants. (One should always do visual checks of experiments running at the same time on the same users that there is no visual design interaction that could be a problem.) This is known as a "statistical" interaction even though the changes may not be interacting in a real way.

- If someone else is using only some of the variants in the test for something else, such as to test or launch a feature.

Figure 4 shows the hourly clickthrough rates for the Control and Treatment for an experiment over a four day period. Note that the two variants have basically the same clickthrough rate throughout this period except for a seven hour period. Our investigation discovered that a headline was different in the two variants for this period. The idea being tested had nothing to do with headlines so the headline difference provided a substantial bias indicating the Treatment was outperforming the Control when the test should have shown no difference.

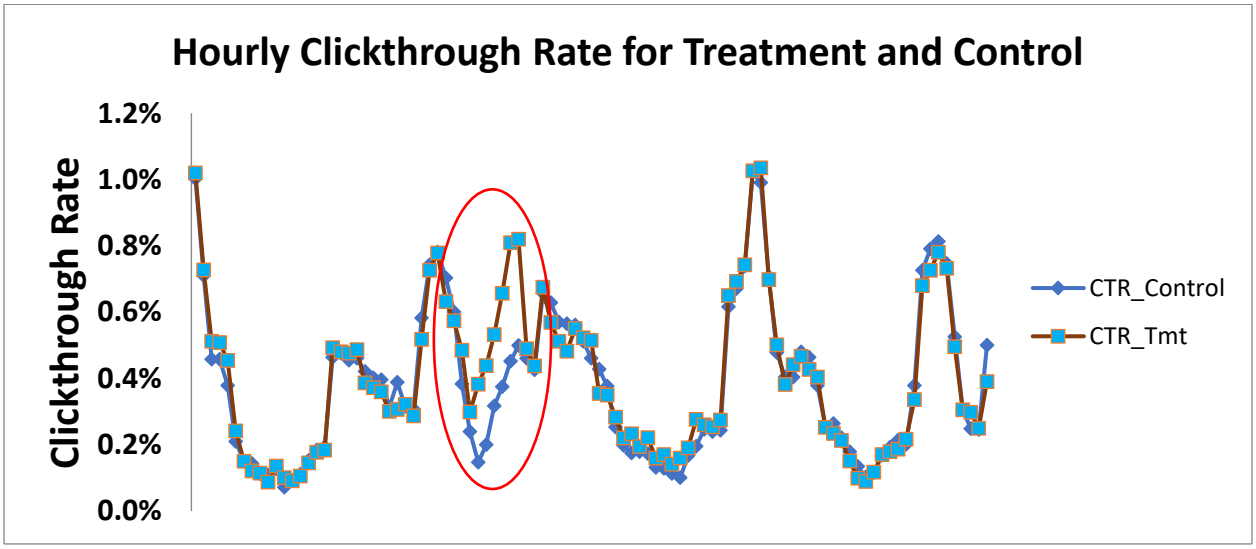


Figure 4. Comparison of Clickthrough rate for Treatment and Control over 4 day period

The first strategy to remove bias is to run all variants being tested concurrently without changing percentage in each variant. Figure 4 illustrates why this is so important. Traffic to websites has tremendous variability within a day, within a week and week to week. This difference in amount of traffic would be a large bias if the variants were not running at the same time. Typically, you would want the percentage in each variant to remain the same throughout the experiment. If this is not the case (such as when ramping up a Treatment, see Part 3) care must be taken in running and analysis of the experiment to make sure no bias is introduced.

Other strategies to remove bias include when and how long to run the experiment. Good practice is to run an experiment for two or more complete weeks. Some variants may have a different effect on certain days of the week (e.g. weekend versus weekday) so, in order to get an unbiased estimate of the long-term effect, full weeks should be included in the experiment. Also, it is possible that one week will not be representative of a long-term effect, so multiple weeks is preferred. One should also be careful of special events or seasonality. For example, an event such as the Olympics, a sales promotion or a holiday such as Christmas may have a differential effect on one of the variants that would not be sustained long-term. This would bias your estimate of the long-term impact of the Treatments.

Finally, we use randomization to remove bias that cannot be removed through careful planning. The basic principle is to randomly assign each visitor to the variants when they first arrive during the experiment. For example, if you want 50% in the Control and 25% in each of the two Treatments you will randomly assign users in those proportions. Of course, you shouldn't expect *exactly* 50%, 25%, 25% in the three variants, but the percentages should be very close to these and a basic diagnostic metric is to alarm if the percentages are not statistically close enough.

> *"Randomization is too important to be left to chance"*
>
> *Robert Coveyou, ORNL*

One must be careful to find a good randomization algorithm. Most algorithms are good in random assignment of users to an individual experiment but fall short if multiple experiments are running at the same time with the same users. Some of these algorithms demonstrate a correlation between two concurrent experiments in that a user that is in the Control of one experiment has a higher (or lower) probability of being in the Control of a different experiment. This is a potential source of bias for both experiments.

The most common way to randomize is by user. (In some experiments it may be preferable to randomize by page view, search or session.) One reason we usually randomize by user is that we want to know how users respond to the variants over the duration of the experiment. Another reason is we want users to have a consistent experience during the experiment. If users were randomly assigned a different variant each time they visited the site the experience may be disconcerting (depending on the change). Each user is randomly assigned to a variant the first time they come to the site during an experiment and are then given the same variant assignment each time they return to the site. One way to do this is to place the variant assignment in the user's cookie. The positive aspects of this are: 1) it gives users a consistent user experience for the duration of the experiment, 2) it allows us to calculate user level metrics (such as page views per user, sessions per user, revenue per user, etc.) which helps us understand how the variants affect user behavior over the duration of the experiment. On the other hand, the person will appear to be a different user if they clear cookies or if they use a different browser or device. If the user doesn't allow cookies, they may not be in the experiment at all.

Reducing variability

Perhaps it is obvious, but too much variability (or noise) in the experiment will hide any difference between variants that you want to see. Reducing variability will help increase the power of the experiment (i.e. the ability to see a difference that is meaningful.) Some of the approaches above for reducing bias will also help with reducing variability. In addition to those the following will help reduce noise:

- Limiting the analysis to only users who can experience a difference due to the variants,
- Removing robots (non-human visitors) from the analysis,
- Transformation of some metrics

The first of these has already been discussed. There are a number of methods for identifying robots among the users and whether (and which) metric transformations should be used depends on the distribution of the metrics themselves. Transformations will be discussed in Part 3 of this series.

Measuring Variability

The most common measure of variability of a metric is the standard deviation. Using the metric, number of clicks per user as an example, the formula for standard deviation is

$$\hat{\sigma} = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}}$$

Where x_i is a single value of the metric (e.g. number of clicks for a specific user) and $\bar{x}$ is the average of the individual values. n is the number of individual values observed (for this example, the number of users). Although this is the formula that is normally used to define standard deviation, there are more computationally efficient ways of arriving at the result in different situations (for large datasets or where the data are arriving over time and the standard deviation needs to be updated regularly.)

If the metric is a binary metric (e.g. 1s and 0s, or whether something happens or not) and we can assume no correlation between occurrances, the definitional formula simplifies considerably to

$$\hat{\sigma} = \sqrt{p * (1-p)}$$

where p is the proportion of ones.

Significance Testing

When comparing two variants, e.g. Treatment versus Control, for a metric we have a simple set-up. We start with the assumption that the Treatment had no effect on this metric and see if we can reject that notion. More formally, as a statistical hypothesis test:

H_0: $\mu_C = \mu_T$

H_A: $\mu_C \neq \mu_T$

If we do not reject the null hypothesis, H_0, we say the Treatment made no detectable difference and take action as if there were no difference between Treatment and Control (for this metric). If H_0 is rejected we need to determine if the change is positive or negative. (Generally, we will have a large number of metrics to assess during and after an experiment. For this reason, a dashboard should be set up with color coding to show if each metric is not significant or if it is significant with a positive or negative impact. Additionally, IMO you should set up each metric so that a positive direction is desirable and a negative direction is not. This simplifies and speeds up metric assessment.)

There are a number of ways to carry out the hypothesis test above. The most common is to assume a Normal distribution and do a two sample t-test (or test of two proportions if it is a binary metric.) In some situations one may be able to achieve greater power by using transformations of the data or different test approach.

When conducting these hypothesis tests, one must determine the level at which you would reject the null hypothesis when it is true. In statistics, this probability is called alpha, α. The most common level used for alpha is 0.05. The interpretation of alpha is "If the null hypothesis it true, i.e. the mean of the Treatment is the same as the mean of the Control (or the Treatment has no effect on average), we will make a mistake and reject the null hypothesis 5% of the time if we were to repeat this test many times on independent data." Rejecting the null hypothesis when it is true is called a Type I error.

One may ask, “Why should we make any mistakes at all?” or “Why are we setting our test up to make mistakes?”

The fact is, we do not know what truth is (i.e. whether the two means are equal or not), we only have some data that sheds some light on it. The (statistical) evidence we have can be used to limit the probability of a mistake but cannot eliminate it.

There is actually a second type of error we can make. This is not rejecting H_0 when it is false. i.e. there truly is a difference in the two means (or the Treatment really does have an effect on this metric) but we do not detect it. This is a Type II error. For a given data set and test procedure, if we reduce one type of error (I or II), the other will increase. The Type I error rate of 5% is often seen as an acceptable level of risk for this type of error in order to balance the Type II error rate. Once we have determined the Type I error rate we will use in our experiments, we can “size” the experiment to reduce the Type II error rate to an acceptable level. (By “sizing” we mean number of visitors in the variants as well as how long the test runs.) The topic of Power, sample size and length of test will be addressed in Part 3 of this series.

TOPICS FOR BASICS OF A/B EXPERIMENTATION, PART 2:

- More on metrics
- Confidence Intervals and presentation of results
- Analysis unit
- Analysis of ratios
- Initial experiment diagnostic checks
- A/A tests

TOPICS FOR BASICS OF A/B EXPERIMENTATION, PART 3:

- What is Power and how do I get more?
- How long should I plan on running this experiment?
- When and how to ramp up an experiment
- Ending an experiment early
- Platform options
- Ethical considerations