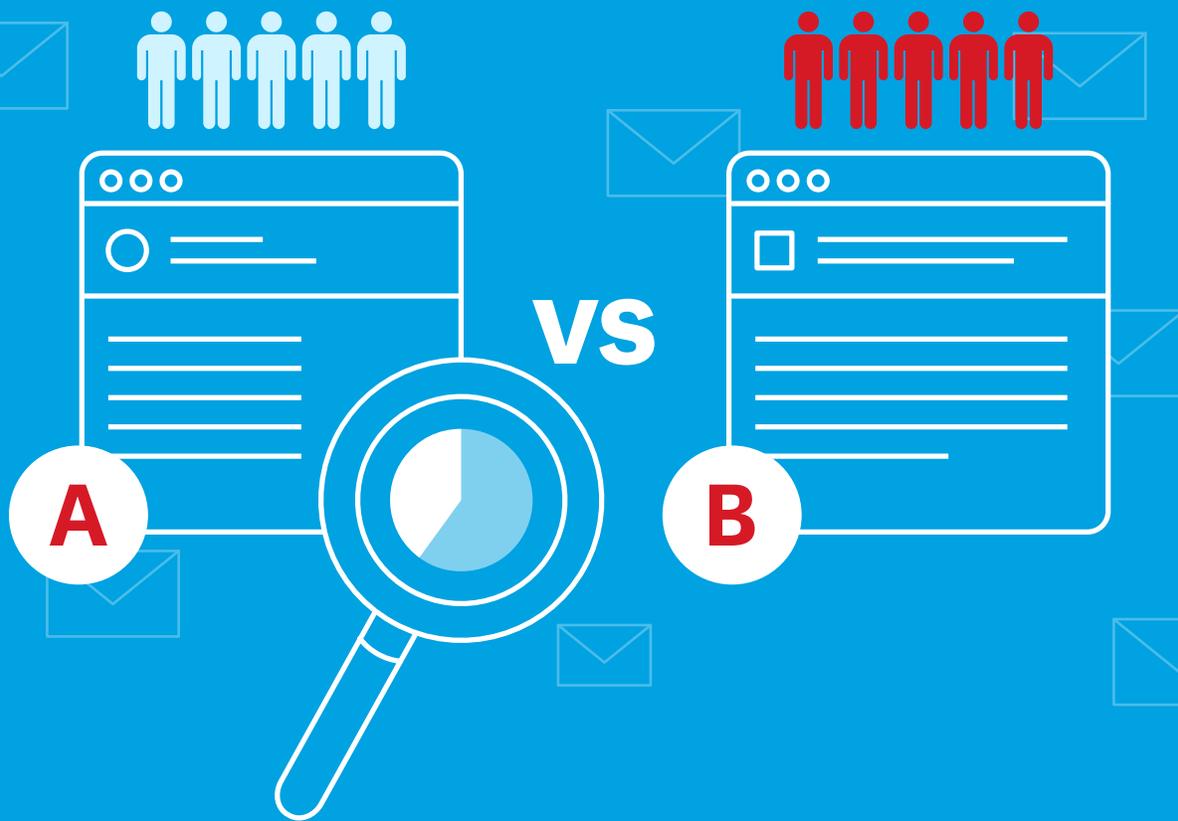


All About A/B Testing

How to Test, What to Test,
and How to Analyze Your Results



 Return Path



Introduction

A/B testing, also called split testing, is one of the best ways to optimize your email campaign. But it can be daunting to figure out. So daunting, in fact, that many marketers decide it's simply not worth the time and effort.

But think about this:

Without testing, how can you be sure your emails are performing as well as they could be?

Quite simply, you can't!

A/B testing is a sure-fire way to figure out what's working (and what's not) in your email marketing campaigns. It can be used to test almost any element of your email campaign: subject line, offer, copy length, imagery, layout, and so much more. When it's done well, the advantages of A/B testing are more than enough to offset the additional effort.

In this ebook, we'll give you the tools to become an A/B testing expert, including all the ins and outs of a well-executed A/B test.

Even the simplest of A/B tests can provide eye-opening results, and small changes can have a huge impact on the effectiveness of your email campaigns. The important thing is to just start testing! So let's get started now.

Table of Contents

- Chapter 1**
What is A/B Testing?.....4

- Chapter 2**
Deciding What to Test.....5

- Chapter 3**
Setting Up and Running
Your A/B Test.....7

- Chapter 4**
Analyzing Your Results.....9

- Chapter 5**
A/B Testing Best Practices.....11

- Chapter 6.**
Master Class: Understanding
Statistical Significance.....13

- Chapter 7**
Master Class: Measuring the
Business Impact of Your Test.....14

Chapter 1:

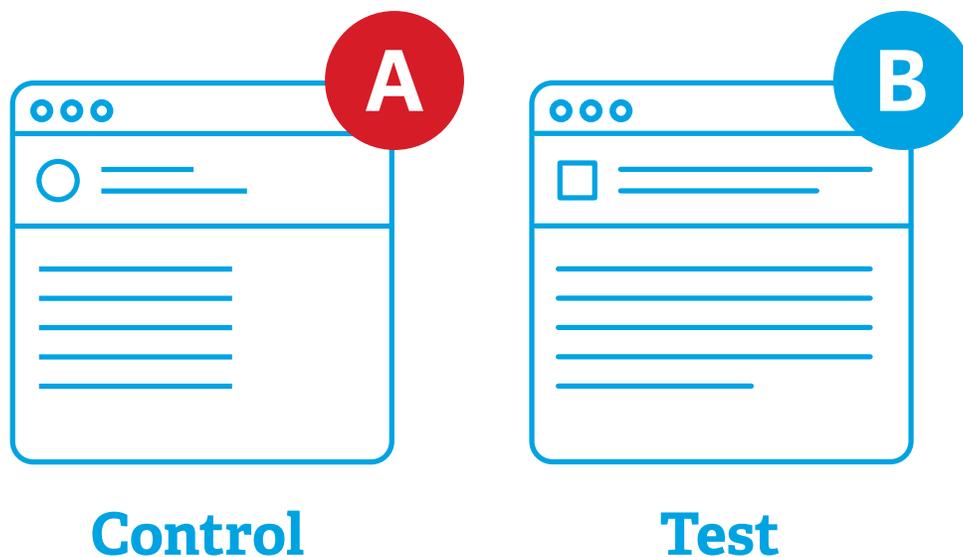
What is A/B Testing?

A/B testing is an email optimization technique that can provide valuable insights into various elements of your email marketing. At the most basic level, an A/B test compares two different variations of a single email campaign to determine which one performs better.

To perform an A/B test, you will first need to decide which aspect of your email campaign you want to test. Then you'll create two versions of the same email:

- Version A (the control) is the "standard" version of the email—often something you've already been using and tracking its performance
- Version B (the test) contains the variation you want to test

Each version of the campaign gets sent to a segment of your audience. By analyzing the results of both the "A" and "B" version of your campaign, you can determine a "winner." This winning version is then sent to the remaining subscribers on your list.



Chapter 2:

Deciding What to Test

The first step in setting up an A/B test is deciding what to test—and the possibilities are virtually endless. If you can change it, you can test it! But that doesn't mean you should spend months testing every aspect of your email campaign. Instead, focus on the things that are most important, or those likely to have the greatest impact.

Keep in mind, different tests will impact different parts of the conversion process. So you may want to start by evaluating the strengths and weaknesses of your current campaign performance. If your open rate is low, testing different subject lines is a great idea. On the other hand, if you're looking to increase your click-through rate, you'd want to test elements like the offer or call to action (CTA).

At a high level, the elements you might want to consider include:

- Subject line (alter the length, wording, etc.)
- Personalization (e.g., "Mr. Smith" vs. "Kyle")
- Headline (vary the size, color, style, or wording)
- Content (longer vs. shorter, text vs. visuals)
- Layout (single column vs. two column, or different placement for different elements)
- Imagery
- Offer (Example: "Save 20%" vs. "Get free shipping")
- Call to action (Example: "Buy Now!" vs. "See Plans & Pricing")

Need some inspiration? Here are 50 fresh ideas to get your creative juices flowing.



50 Testing Ideas to Get You Started

1. Add personalization to your subject line.
2. Use numbers instead of text in your subject line.
3. Reword your subject line to create a sense of urgency.
4. Test direct vs. "teaser" copy in your subject line.
5. Try motivational copy ("act now") vs. loss aversion ("don't miss out") in your subject line.
6. Feature different newsletter articles in your subject line.

7. Change your pre-header text.
8. Try a different "From" name.
9. Vary the size and color of your headline.
10. Change the order of your content.
11. Adjust the wording of your body copy.
12. Vary copy length.
13. Alter the tone of your copy—formal and corporate vs. casual and friendly.
14. Try a "letter from the CEO" format vs. your normal layout.
15. Adjust the balance of editorial vs. promotional copy.
16. Test bulleted lists vs. paragraphs of copy.
17. Add personalization to body copy.
18. Change the balance of copy in the email vs. the landing page.
19. Adjust your color scheme.
20. Use color to highlight different parts of the email.
21. Test a multi-colored design vs. monochromatic.
22. Vary the size of your images.
23. Try people images vs. product images.
24. Test photo images vs. illustrations.
25. Personalize images based on customer profile or preferences.
26. Add (or move) social sharing buttons.
27. Streamline your layout.
28. Try a different font.
29. Test image-heavy layout vs. text-heavy.
30. Embed a video or animated gif.
31. Focus on benefits vs. product features.
32. Highlight a single product vs. many.
33. Rewrite product descriptions.
34. Promote products without offering a discount.
35. Try pricing and discount variations (% off vs. \$ off).
36. Offer free shipping instead of a purchase discount.
37. Personalize offers based on previous purchases.
38. Include product reviews or ratings.
39. Add a "free trial" offer.
40. Increase (or decrease) the number of links.
41. Change the size of your CTA button.
42. Get creative with your CTA wording.
43. Feature a single CTA vs. multiple different offers.
44. Present the CTA once vs. several times.
45. Change the order of your links or CTAs.
46. Vary the time you send your emails.
47. Personalize send time vs. sending all at once.
48. Send on a different day of the week.
49. Increase (or decrease) send frequency.
50. Reached the end of this list? Start over at the top!

Chapter 3

Setting Up and Running Your A/B Test

In this chapter, we'll walk you through the A/B testing process step-by-step. Since we want to be as specific as possible, let's assume you want to run one of the more common A/B tests: [subject line testing](#).

The basic steps for a subject line test would look like this:

1. Create a hypothesis.

Before you begin testing, it's critical to start with a hypothesis—in other words, knowing what you are testing and what the desired outcome is. The hypothesis helps to determine how large your test and control groups need to be, how long you need to run the test, and whether a test has conclusive results or not.

For example, your hypothesis might be that the new (test) subject line will increase read rate by five percentage points over the old (control) subject line. In this hypothesis you have stated that you are basing the success of the test on read rate, and that the increase needed for that test to be successful is five percent.

This ensures you will not only be able to successfully choose the size of your test and control groups (which in turn will help to determine whether your test results are statistically significant), but also clearly defines whether the test was successful (whether or not the stated goal was reached).

2. Use your hypothesis to determine sample size, then create test and control groups.

In this test, you are testing read rate and will therefore use your historical rate to decide [how many people need to be in the test and control groups](#). Say you have a historical read rate of 12% for your current (control) subject line. Your hypothesis states that you want to see an increase of five percentage points, but do you also want to see if a smaller change is statistically significant?

The smaller the change you want to detect, the larger your sample size will need to be. Say you want to detect a change of two percentage points or more (+/-2%); you can use [Return Path's Sample Size Calculator](#) to determine the size of your test and control groups. According to our tool, for a list of 100,000 people with a confidence level of 95% and a 2% margin of error, the size of both your test and control groups will be 1,936 people.

Finally, you need to create a random sample of 1,936 people for both the test and control groups. Don't forget to ensure the read rates of the two random groups are the same.

3. Create test variation and send.

As we discussed in Chapter 1, the control version of your email is typically a known quantity—the “standard” version of the subject line, which you’ve already been using and tracking performance. Now you need to create a “test” version of the subject line. This might include a variation on your normal wording, different keywords, personalization, numbers vs. text, etc. (If you need help coming up with ideas, check out [our Subject Line Optimizer tool](#).)

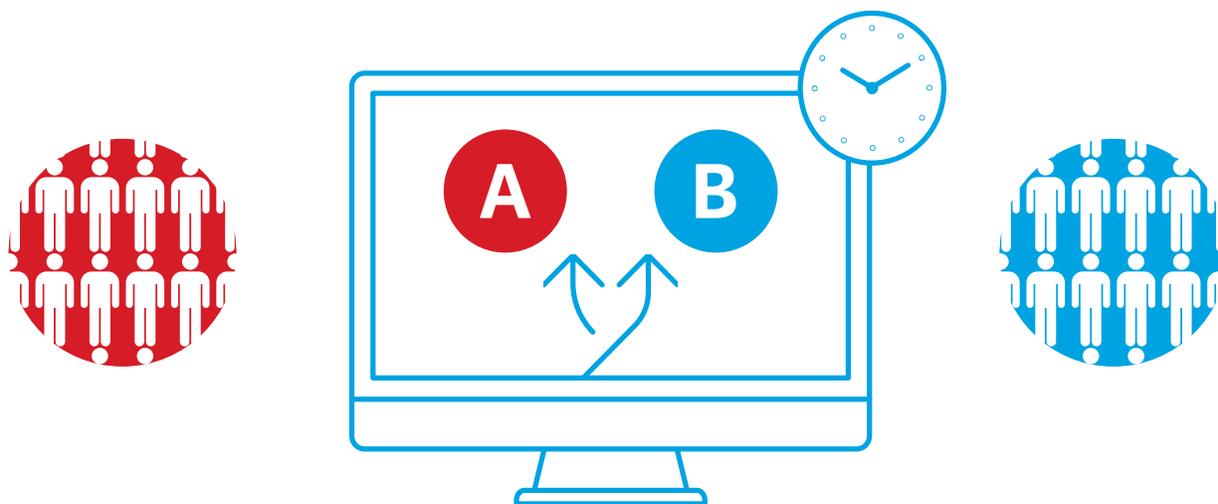
Once you have your test groups and two versions of the email ready to go, it’s time to hit send.

4. Decide how long to run the test.

This question depends a lot on what you are testing and your hypothesis. In our example, you are running a subject line test, and your hypothesis is that the new subject line will increase your read rate by 5%. So essentially you would send to the test and control groups, and see whether the read rate for the test group is in fact five percentage points higher than the control group. There are three possible results: failure, success, or inconclusive. The odds of someone reading your email decline sharply after about three to five days, so you can safely begin your analysis at this point.

Other types of tests might need to run until you have enough impressions to make a decision, or until the metric in question has normalized. For example, testing a change in send frequency requires you to observe a long-term effect, because after people adjust to the new frequency, you’ll need to determine whether they behave the same or differently than they did before the test.

The important takeaway here is that you need to understand the test you’re running and allow enough time for the campaign to run its course before you analyze the results of your test.



Chapter 4

Analyzing Your Results

You've built your sample groups, you've created your control and sample emails, and you've run your test. Now the real fun begins!

In this chapter, we'll help you determine what your test results mean and what to do next.

1. Assess the statistical significance of your results.

The first step in analyzing test result is to determine whether they are statistically significant. The question of "statistical significance" is really about determining whether a test's results are meaningful and repeatable.

If your control read rate is 12% and the test read rate comes back at 13%, the results are too close to be meaningful because you set up your test to detect a change of +/- two percentage points. In other words, the test is not statistically significant and your results would be inconclusive even if one version performed better than the other.

Say instead that the control read rate is 12% and the test read rate is 15%. In this case, the test version performed significantly better than the control version—and because you set up your test to detect a change of at least two percentage points, the test is statistically significant. Keep in mind, however, that statistical significance simply tells you whether the test provides a conclusive result, not whether the test was successful.

For an in-depth look at statistical significance, jump to Chapter 6, ["Master class: Understanding Statistical Significance."](#)

2. Pick a winner.

Assuming your test results are statistically significant, the next step is to determine the success or failure of your test. To do this, you'll need to look back at your original hypothesis.

In this case, your hypothesis was that the read rate for the test subject line would be at least 5% better than the read rate for the control subject line. In our sample test with a control read rate of 12%, a test read rate of 17% or better would clearly indicate a successful test. A test read rate below 17% would indicate either failure or possibly an inconclusive test.

3. Determine whether the impact is meaningful.

What does a “success,” “failure,” or “inconclusive” result mean for your company goals? What if your test only improved subject line read rates by three percentage points? A change of three percentage points is statistically significant according to your test hypothesis, but the test would be a failure because the test read rate increased by less than your stated goal of five percent. Is an increase of 3% significant enough to push that new subject line to the rest of your list? If it is, should your hypothesis be restated?

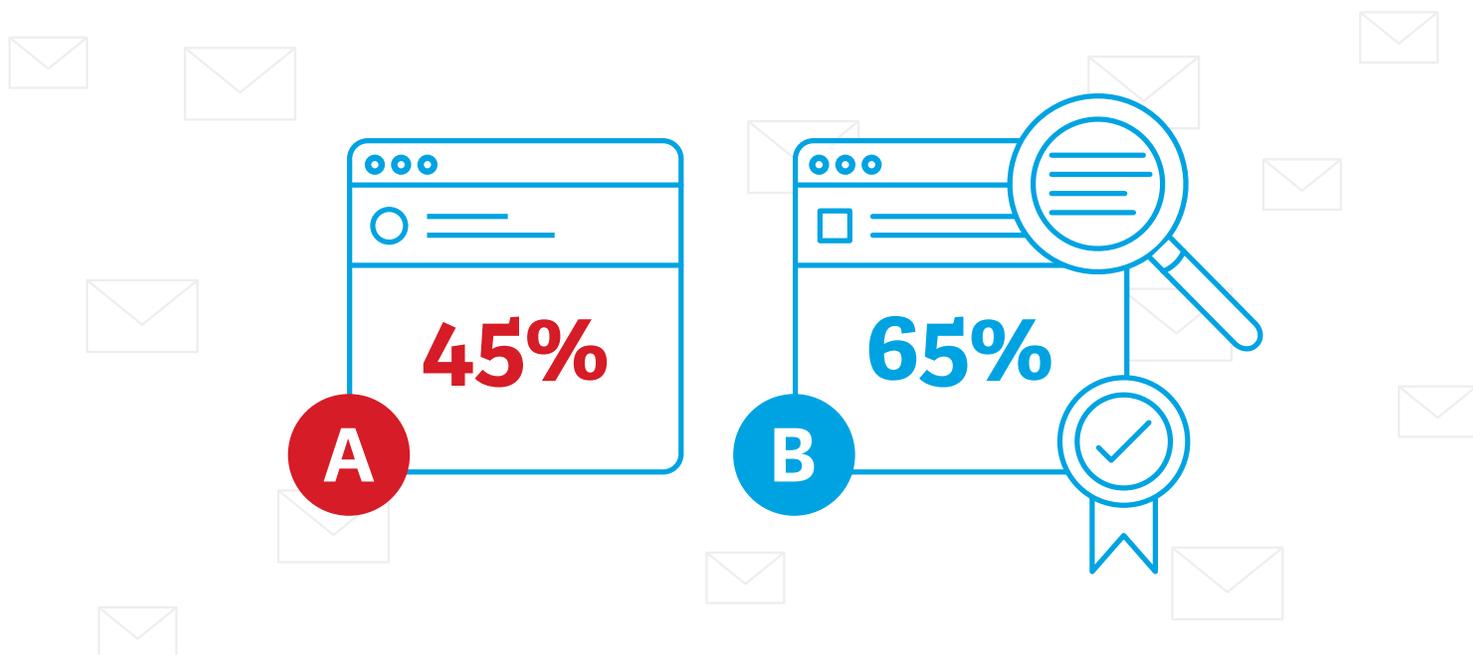
Clearly there’s no right or wrong answer here. The point is that you need to consider your results carefully rather than just rubber-stamping the test as a “success” or a “failure.” Even a small improvement can make a big difference, if it’s based on valid test results.

For an in-depth look at measuring impact, jump to Chapter 7, [“Master Class: Measuring the Business Impact of Your Results.”](#)

4. Draw conclusions and move forward.

If your test is a success, you can push the test version to the remainder of your list. If the test is a failure, you can send the control version instead. But does that mean that your testing work is done? Not even close! One of the best (and worst) things about A/B testing is there’s always something more you can test.

A/B test results are a great tool for email optimization, but it’s important to remember that campaign success may decline over time as you get new subscribers, seasons or trends change, etc. Capturing and keeping your subscribers’ attention is a constant battle, so keep testing and trying new things! With constant improvement, you’ll always be able to find new and different ways to push your ROI higher.



Chapter 5

A/B Testing Best Practices

Successful A/B testing depends on consistency and control. You're going to make important decisions based on your test results, so you need your data to be as accurate as possible. That requires careful planning and execution.

Here are a few tips and best practices to keep in mind when planning and executing your A/B tests. By following these guidelines, you can be confident that your tests will provide sound, reliable results.

Test early. Test often.

When considering a change to an email campaign, testing should be among your first steps. There's no sense going too far down the road with a new design or strategy until you find out whether it's likely to improve your results.

Focus on frequently sent emails.

A/B testing is best for optimizing emails you send on a regular basis, like newsletters, welcome emails, standard promotions, and abandoned cart emails. Even small improvements to emails that you send all the time (to a large audience) can yield huge results. Don't waste your testing efforts on something like a holiday email—you'll have to wait a whole year to apply your learnings!

A larger test will provide more reliable results.

It only makes sense—the larger your sample size, the more likely that the test results will be predictive of your entire list. If you need help figuring out the right sample size for your test, check out [Return Path's Sample Size Calculator tool](#).

Make sure your test groups are random.

If you're creating your test and control groups manually, make sure the samples you choose are truly random. Grabbing an arbitrary chunk of email addresses from your list may seem random, but subscribers listed consecutively may share the same geography (if your list is sorted by address) or ethnicity (if it's sorted by last name), and this has the potential to skew your results. Instead, download your list as a .csv file and use Excel's "random sort" feature.

Always send both versions of your test simultaneously.

It's critical that both your test group and your control group receive your email at the same time, in order to eliminate the potential for skewed results based on timing. Even a couple of hours can make a difference (consider an email received at 5 a.m. versus one that hits the inbox at 7 a.m.), so simultaneous sending is critical.

Be patient.

Some tests can be run in just a few days, but others (like frequency tests) will take much longer to yield reliable results. Cutting the test off early just means there's more room for error—but the same can be said for letting it run too long. Most tests require between three and five days, but it really depends on your audience. A thorough understanding of their email habits (for example, how long it takes for your newsletter opens to drop off) is your best weapon in determining when your test is complete.

Trust your results.

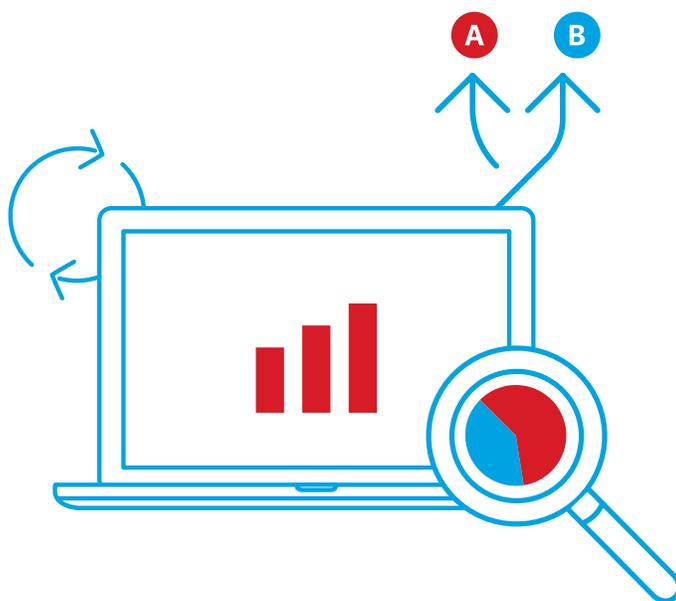
Resist the urge to listen to your gut instincts if the test results tell you something different. You're running controlled tests for reason, and empirical data beats opinions every time. If you truly doubt the test results (or you think the test might have somehow gotten skewed), run the test a second time.

Put your learnings into practice.

If your test yields both statistically significant results and a clear winner, your first step is to send the winning email to the rest of your list. But don't stop there! Now that you have some insight into what works with your audience, you can apply the change to other email campaigns as well. You might also try additional tests on the same variable, to see if you can push your results even further.

Keep testing!

One test is good, but there's always something more to learn. Your audience will never stop evolving, and neither should your email program! If you were successful in bumping up your open rate, maybe you work on click-throughs next. There's no such thing as a perfect email campaign, but each email you send can be better than the last.



Chapter 6

Master Class: Understanding Statistical Significance

Are these results statistically significant? This is a question every marketer should ask if they want to understand the impact of their findings. Whether or not your results are statistically significant will determine whether you move forward with the analysis, or conclude that you can't tell the difference between test and control results.

To determine statistical significance, some researchers will look solely at their p-value, which calculates odds of the results being significant. P-values are generally set at .05 or .01 and are used to reject the null hypothesis. The null hypothesis is created at the beginning of your test states that the two approaches being tested will perform the same.

An example of a null hypothesis for a subject line test would state that the test subject line will perform the same as the control subject line. In setting up the test, you would select a p-value of .05 which is used to determine how large the sample size will need to be in order to say the results from your test and control group are different, thus rejecting the null hypothesis.

What does all this mean? As it turns out, very little. Ronald Fisher, the statistician who introduced p-values, claimed p-values are not the end-all be-all for testing. Rather, they are a starting point to make sure you start your test on the correct track. They help make sure your sample sizes are correct, so when you are assessing different rates you can confirm that the result of your test group and the result of your control group are different.

As a general rule of thumb, a p-value of .05 means you can expect the odds of the null hypothesis being true are about 1 in 20.

What p-values and statistical significance can tell you:

1. There is a difference between the test and the control.

What p-values and statistical significance can't tell you:

1. Which performed better, the test or control
2. How large the difference is between the test and control
3. Whether the difference between the test and control is large enough to be important
4. Any results about the magnitude of your test

To understand differences such as magnitude and relative importance, you need to look at effect size and confidence intervals, which we'll discuss in Chapter 7, "[Master Class: Measuring the Business Impact of Your Test.](#)"

Chapter 7

Master Class: Measuring the Business Impact of Your Test

In order to accurately assess your test results, it's important to understand the difference between test and control groups, as well as the impact of the test. This is done by using confidence intervals and effect size.

In Chapter 3, we discussed the importance of creating a good hypothesis. The same is true when it comes to interpreting your test results. This is where confidence intervals and effect size come into play. Confidence intervals are used as a check so you can understand the range of what the mean rates of your test and control group lie in. These are directly related to how your sample size is calculated and the alpha you choose.

The effect size gives the magnitude of the change. In a hypothesis, when you state how much change you want to see, this is a direct reference to the effect size. The effect size ultimately helps you decide whether or not the test was successful, and whether you should move forward with pushing the test to your entire list.

In short, the larger the effect size, the more impactful the results. There are many different calculations for effect size, but a common measure for a test with a binary outcome (such as "read" or "did not read") is the odds ratio.

The formula for odds ratio is:

$(a/b)/(c/d)$

So the calculation for the odds ratio of a subject line test based on number of reads would be:

(the number of reads of the test / the number of reads of the control)

divided by

(the number of non-reads of the test / the number of non-reads of the control)

If the odds ratio equals one, then the test and control perform the same. If the odds ratio is greater than one, then the test group is x times more likely to read the email than the control group. If the odds ratio is less than one, the result is not directly interpreted, except to say that the test performed worse.

In order to get a more interpretable result when the odds ratio is less than one, you would have to reverse the test and control positions in the equation, putting the control group numbers in the numerator and the test group numbers in the denominator. Then the odds ratio would read, the control group is x times more likely to read the email than the test group.

As an example, say you performed a subject line test. The results showed that 300 subscribers read the test subject line and 1200 subscribers did not. For the control group, 200 subscribers read the control subject line and 1300 subscribers did not. Thus the odds ratio is:

$$(300/200)/(1200/1300)$$

This means the test email is 1.6 times more likely to be read.

As you become more versed in testing and understanding the statistics behind it, you will be able to create a better hypothesis that not only addresses business needs but can address things such as confidence intervals and effect size. This will make not only running tests but interpreting and creating actionable results a breeze.

It is always important to remember that testing never ends! The more you test, the more you will be able to find out what improves and doesn't improve your email campaign.

A Smarter Way to Use Data

We are the world's leading
email data solutions provider.



Email Optimization

The right message, at the right time, to the right inbox means better relationships, greater reach and increased revenue. Email Optimization enables enhanced insights for better deliverability and more meaningful engagements.



Email Fraud Protection

The cost of a cyber attack goes beyond dollars and cents, it damages the integrity of a brand. Email Fraud Protection uses advanced fraud profiling data to respond to, and prevent, cyber attacks with greater speed.



Consumer Insight

The inbox provides a unique real-time view of consumer behavior - from brand affinity to detailed purchase records. Consumer Insight provides in-depth data across millions of global consumers enabling smarter decisions and better business results.



returnpath.com