RAG KNOWLEDGE BASE

Explaining the Streamlit platform

1. **Experiment design tab:** calculating sample sizes before running an experiment

Purpose: Before launching an A/B test, we want to calculate how many users (in treatment vs control) we'll need to confidently detect a desired effect. With the right sample size, we can obtain the desired power of the experiment.

Statistical Concept: We follow the "static test" design approach.

In A/B testing, a **static test** means:

- We predefine the sample size before we start the experiment.
- We don't peek at the data and stop early.
- We run the experiment to completion based on that sample size.

That's why we need the sample size calculator: To figure out what number to predefine, based on the math behind hypothesis testing.

- We first determine the targeted sample size based on the **minimal detectable effect (MDE)**.
  To calculate how many users we need, we first decide the smallest improvement we want to be able to detect (MDE) and then calculate the sample size that will let us detect that effect with confidence.

- If we want to detect a certain uplift, we need to feed some numbers into our sample size calculator.

Inputs we need to provide:

- **Uplift**

  The improvement (or difference) we expect or hope to see in our treatment group compared to the control group in an A/B test.

  It answers the question: "How much better (or worse) is the new version compared to the old one?"

- **Estimate of the population mean** (baseline value)

This is our starting point, or current performance before any changes.

The baseline conversion rate (often denoted as p1) represents the current performance of our system before any changes are made.
“What percentage of users convert (i.e. take the desired action) under normal conditions, without the new change we want to test?”

Conversion rate = Total number of users exposed/Number of users who converted

- **Treatment share**

  This is the percentage of users you assign to the treatment group (the group that sees the new version).

  In a 50/50 A/B test: 50% of users go to control (old version) and 50% go to treatment (new version)

- **Type I error (α)**

  Probability of a false positive, concluding there’s an effect when there isn’t.
  It's the significance level.

- **Type II error (β)**

  Probability of a false negative (missing a real effect when one exists):
  β = 1 – Power

  example: Power = 0.80, so β = 0.20
  "I want an 80% chance of detecting the uplift if it’s real thus β = 0.2"

- **Effect relation to base metric / mean: 'absolute' or 'relative'**

  How we define the uplift.
  Are we giving the difference directly (absolute), or as a percent increase (relative)?

- **Key outcome metric type:** Binary or Continuous

Statistical tests to find the right sample size:

1. Two-Sample Proportion Z-Test

Purpose: Used to determine whether there is a statistically significant difference between the proportions of a binary outcome in two independent groups

Use case: Applied when the outcome variable is binary (e.g. converted/did not convert).

Hypotheses

- **Null hypothesis (H_0):** $p1=p2$
  There is no difference in proportions between treatment and control groups.
- **Alternative hypothesis (H_1):** $p1 \neq p2$
  There is a difference in proportions.

Assumptions

- Observations are independent in both groups.
- Sample sizes are large enough for the normal approximation to apply (generally: $np \geq 5$ and $n(1-p) \geq 5$)
- Binary outcome data (success/failure).
- Groups are randomly assigned and not overlapping.

2. Two-Sample T-Test (Equal variance)

Purpose: Used to test whether there is a statistically significant difference between the means of a continuous outcome in two independent groups.

Use case: Applied when the outcome variable is continuous (e.g. revenue, number of actions).

Hypotheses

- **Null hypothesis (H_0):** $\mu 1=\mu 2$
  The group means are equal.
- **Alternative hypothesis (H_1):** $\mu 1 \neq \mu 2$
  The group means differ.

Assumptions

- Outcome data in each group are normally distributed (especially important with small samples).
- Groups are independent and randomly assigned.
- The variances of the two groups are equal (this is called the pooled variance assumption).
- Continuous outcome variable.

The two-sample z-test (for binary outcomes) and the two-sample t-test (for continuous outcomes) each provide a mathematical formula to calculate the sample size per group based on the inputs we provide!

Formula for z-test:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot [p_1(1-p_1) + p_2(1-p_2)]}{(p_1 - p_2)^2}$$

Formula for t-test:

$$n = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_1 - \mu_2)^2}$$

This makes sense if:

- Our test only consists of a single point of data collection such as sending out 100k emails at once (we are running our experiment all at once, not in phases).
- We are willing to commit in advance to collect a certain sample size.

2. Monitoring tab

This section comes after our experiment has started or completed, it assumes we've already collected data. Its goal is to validate that our data is reliable and correctly structured, so we can trust any conclusions we draw from it.

What do variant 0, 1, and 2 represent?
In the context of A/B testing, variants refer to the different versions of the product, feature, or experience that are being compared during the experiment.

- **Variant 0** typically represents the control group, the default or current version shown to users.
- **Variant 1 and Variant 2** represent treatment groups, alternative versions that we are testing to evaluate whether they lead to improvements in key metrics (such as conversion or revenue).

In the Monitoring streamlit tab we see:

- **Table** that provides a precise count of users in each group and the total sample size used in the experiment.

- **Pie chart** summarizes the proportion of users assigned to each variant over the entire experiment period.

  It matters because we expect the user allocation to be approximately equal across all variants (roughly 33.3% each) in a well-randomized A/B test. Imbalances could indicate randomization failure or tracking issues.

- **Line plot** displays the cumulative number of users assigned to each variant (control and treatments) over time during the experiment.
  The x-axis shows trigger_dates, indicating when users were assigned.
  The y-axis shows the growing count of users in each group.
  Each line represents one of the three variants.

  We use it to verify that:

Users were assigned to variants consistently over time.

There were no major drops, surges, or delays in data collection for any group.

Randomization appears to have been uniform and stable throughout the test period.

- **Distribution test:** The Chi-squared-test that there is no association or relationship between the variant and the segment in an A/B test has p-values. (variant assignment should be independent of user segments)

  In other words: each variant (control/treatments) should include a balanced mix of users from all segments.

  Null Hypothesis (Ho): Variant assignment is independent of the segment
  Alternative Hypothesis (H1): Variant assignment depends on the segment

  Thus, if p-value > 0.05 we fail to reject Ho and thus there is sufficient indication that there is no dependence between the segment and variant assignments.

- **Heatmaps:** contribution heatmaps from a Chi-squared test of independence

  Each heatmap shows how much each cell (variant × segment) contributes to the overall chi-squared statistic for that segment.

  Rows = Variants (control and treatments)
  Columns = Segment values
  Color intensity = Degree of observed deviation from what we'd expect under perfect randomization (i.e. how much more or less users appeared in a given cell than expected)

  Lighter color: balanced assignment
  Darker color: some deviation from expected (possible imbalance)

  But the heatmaps must always be interpreted together with the p-value.

- **Dynamic and static variant per segment counts:** These plots allow us to visually check whether variant assignment was balanced within each subgroup across time.

  X-axis (trigger_dates): Time when users were assigned to a variant.

  Y-axis (variant_cnt): Cumulative number of users assigned.

  One subplot per segment category.

  One line per variant within each subplot.

Parallel and similarly sloped lines across variants indicate balanced randomization and steady data collection within each segment. If one line is much higher or steeper, it might mean one group got more users from that segment, which could affect results. Always check these plots along with statistical tests to be sure.

- **Distributions of segments (Histograms):**
  These plots are important because they help ensure that each variant group is made up of similar types of users. This allows us to trust that any differences we observe are due to the treatment itself, not who ended up in each group.

  If the bars for each segment category are similar in height across all variants, it means users were evenly distributed, randomization worked well.
  If one variant has a much taller or shorter bar, it may signal imbalance that could affect your results.

- **Distribution of outcome metrics:**

  Histograms: Show the frequency of values across bins. Help detect skewed distributions or outliers.

  Cumulative Distribution Functions (CDFs): Show the cumulative proportion of users below a certain value. Help compare entire distributions, not just averages.

  Quantile-Quantile (QQ) Plots: Compare quantiles of two distributions.
  Points on the identity line → similar distributions
  Points above/below → distributions differ

- **Testing for Normality of the Data:**
  In traditional frequentist A/B testing (like using a t-test), we assume the data is normally distributed.
  This matters because if the data is not normal, the test might give misleading results, especially for small samples or skewed distributions.

  To check normality, we use two tools:

  1. Q-Q Plot (Quantile-Quantile Plot)
  This plot compares the quantiles of your data to the quantiles of a theoretical normal distribution.
  If the data follows the identity line, it's roughly normal.
  If the points curve away from the line (especially at the tails) then the data is likely not normally distributed.

  2. Shapiro-Wilk Test

A formal statistical test for normality.
Null hypothesis (H_o): Data is normally distributed.
If the p-value is very small (< 0.05): We reject H_o and thus data is not normal.

- **Monitoring treatment effect / Sequential testing over time:**

  Helps us track how metrics, treatment effects, and statistical significance evolve over time as data accumulates.

  This is called sequential testing, and it allows us to stop early if the treatment effect becomes clearly significant or clearly insignificant (if the p-value in the following plot is below a predefined threshold such as alpha=0.05).

  Left column: metric average over time
  Plots the average value of the metric for each variant over time.
  Helps detect early shifts or trends in user behavior.

  Middle column: treatment effect over time
  Shows how the difference between two groups evolves.
  The shaded area is the confidence interval around that difference.
  The horizontal line at 0 means “no difference” between the two treatments.

  Right column: sequential p-value
  This shows the p-value for the ongoing test between Treatment 1 and Treatment 2 over time. It’s dynamically updated as more data comes in (changes over time).

  H_o (Null Hypothesis): There is no difference in the metric between Treatment 1 and Treatment 2.
  A p-value below 0.05 suggests a statistically significant difference between the two treatments.

3. Experiment evaluation tab

This section is where we formally test whether the treatments had a statistically significant effect on our outcome metrics (like conversion, revenue, or actions) using standard statistical tests.

It builds on the earlier steps:

- We've designed the experiment
- Checked the data quality and balance
  Now, we're ready to run hypothesis tests and interpret results.

Process:

We compare each treatment variant with the control group.

The statistic of interest is the difference in means for every specified metric.

Thus, **H_o (Null Hypothesis):** No difference between the means of control and treatment.

H_1 (Alternative Hypothesis): There is a difference.

If p-value < 0.05, we typically reject H_o and conclude there's a significant effect.

We also provide confidence intervals for metrics.

Which tests are used?

- Binary outcomes (e.g. conversion): use a **proportion z-test**
- Continuous outcomes (e.g. revenue, num_actions): use a **t-test**
  (Only valid if normality assumptions hold, if not other methods should be used)

Power/sensitivity improvement via CUPED

- **CUPED** stands for Controlled Pre-Experiment Data.
  It's a statistical technique used to reduce variance in your outcome metric by using pre-experiment information that's correlated with the outcome.

- In A/B testing, high variance can make it harder to detect small effects. CUPED helps by:

  Reducing noise in our outcome metric.
  Improving statistical power, meaning we're more likely to detect a real effect.
  Requiring a smaller sample size for the same level of confidence.
- CUPED uses **pre-experiment metrics** (e.g. past user behavior before the A/B test began) to adjust the outcome.

  These pre-experiment values must be:
  Unrelated to treatment assignment.
  Strongly correlated with the post-experiment outcome.
  The higher the correlation, the more effective CUPED is at reducing variance.

- **Example:**
  Suppose your outcome is revenue during the test.
  You include revenue from the month before the test as a pre-experiment metric.
  If users who spent more before also tend to spend more during the experiment, CUPED will adjust for that and remove unnecessary variation.

- **CUPED improves:**
  Sensitivity: Better at detecting small effects.
  Power: Higher chance of finding statistically significant results when they truly exist.
  Confidence Intervals: Tend to be narrower because variance is lower.

- Interpretation remains the same as traditional t-tests, but results are often more stable and precise.

- **What if the CUPED's effect significance is different from the Frequentist Evaluation's effect significance?**
  Generally, we can still consider the effect to be significant if CUPED is significant while Frequentist Evaluation is not.
  However, if it's the other way and CUPED's effect is not significant while Frequentist Evaluation's is, we have a reason to believe that there is some issue with CUPED that requires further investigation.

Segmentation with Wise pizza:

This is a technique used to:

- Detect unusual segments where the treatment effect differs significantly from the global average.
- Help us understand which combinations of user characteristics (segments) are driving the overall difference between control and treatment groups.

- Segments with strong positive or negative impact help us understand which user groups drove the overall result and can guide further investigation or targeting.

- How is different from the previous tests?
    - Runs separately within user segments (e.g. only for 10+ transfers, or New + 10- transfers).
    - Tests whether treatment works better or worse in specific subgroups.
    - Helps identify heterogeneous effects that would be hidden in global tests.
    - Multiple tests per metric, one per segment.
    - Example question it answers: “Did Treatment 1 improve conversion specifically for 10+ transfers users?”

- The **table** reports statistical test results for each segment (pizza slice) comparing the chosen Treatment to Control, for multiple outcome metrics (conversion, num_actions, revenue).

- The **plot** shows the chart is split into three main panels per segment:

  1. Impact on overall total (blue bars)
  This shows how much each segment contributes to the total difference in the outcome.
  It combines the segment size and segment effect (meaning even small segments can have large impact if their effect is big).
  Positive bar = segment is helping increase the metric
  Negative bar = segment is dragging the metric down

  2. Segment averages (red bars)
  This shows the average value of the metric just within the segment.
  The dashed red line is the global average across the whole dataset (test-control difference).
  You can quickly see if the segment is above or below average.

  3. Segment sizes (green bars)
  This shows how many observations (users) are in the segment.
  Helps you evaluate how reliable or impactful each segment might be.
  (Larger bars mean more users. A small but extreme segment may be less stable than a large one.)

- **Note on p-values:** Wise Pizza does not correct p-values for multiple comparisons. This means some results may appear significant by chance, especially when testing many segments.
  However, the method tries to reduce false positives by only surfacing segments with large and meaningful effects.

Evaluation for segments:

- For each variant vs control comparison in the selected segment, the **table** shows:
  - Mean values of each outcome metric for both groups (control and treatment).
  - The estimated treatment effect (both for the absolute and relative value)
    = mean of treatment − mean of control
  - The p-value, which tells us if this effect is statistically significant.
  - A confidence interval for the estimated effect.

- A p-value below 0.05 and a confidence interval that excludes 0 typically indicate a meaningful treatment effect in that segment.

4. Bayesian A/B test evaluation tab

Unlike frequentist A/B testing, which relies on p-values and fixed null hypotheses, Bayesian A/B testing uses probabilities to estimate how likely it is that a treatment is better, worse, or practically the same as the control.

What the Bayesian tab does:

- Fits a Bayesian model to our chosen outcome metric.
- Draws the posterior distribution of the treatment effect.
- Calculates probabilities for different conditions.

Posterior distribution plot:

- Shows the distribution of likely values for the treatment effect.
- The wider the curve, the more uncertainty.
- The center of the curve shows the most probable effect.
- The High Density Interval bar shows the 95% credible interval (like CI in frequentist).

Empirical CDF (ECDF)

- The ECDF shows the cumulative probability of the treatment effect.
- It gives insight into how likely the effect is to fall below or above certain thresholds.

Different probabilities we get:

- **Probability that the treatment effect > 0** (How likely it is that the treatment improved conversion).
- **Probability that the treatment effect < 0** (How likely it is that the treatment made things worse).
- **Probability that treatment effect is above or below a custom threshold.**

- **Probability that treatment effect is outside a custom interval.**

- **Probability that treatment effect is in ROPE (Region of Practical Equivalence)**

  The Region of Practical Equivalence (ROPE) defines a range of effect sizes that are considered too small to be practically meaningful.
  In this analysis, the Bayesian model shows a 0.00% probability that either Variant 1 or Variant 2 had an effect outside the ROPE. This suggests the treatment effects, even if present, are likely negligible (not large enough to be useful or actionable).

  Why this is useful:
- Just saying "p-value < 0.05" doesn't tell us if the change is actually big enough to justify rollout.
- ROPE focuses on practical significance, not just statistical noise.
- Here, even if we saw a >50% chance of improvement, the effect size is too small to matter.

Source: https://www.optimizely.com/optimization-glossary/ab-testing/

A/B testing

What is A/B testing?

A/B testing (also known as split testing or bucket testing) is a methodology for comparing two versions of a webpage or app against each other to determine which one performs better. It works by showing two variants of a page to users at random and using statistical analysis to determine which variation achieves better results for your conversion goals.

In practice, this is how A/B testing works:

- Creating two versions of a page - the original (control or A) and a modified version (variation or B)
- Randomly splitting your traffic between these versions
- Measuring user engagement through a dashboard
- Analyzing results to determine if the changes had positive, negative, or neutral effects

The changes you test can range from simple adjustments (like a headline or button) to complete page redesigns. By measuring the impact of each change, A/B testing turns website optimization from guesswork into data-informed decisions, shifting conversations from "we think" to "we know."

As visitors are served either the control or variation, their engagement with each experience is measured and collected in a dashboard and analyzed through a statistical engine. You can then determine whether changing the experience (variation or B) had a positive, negative or neutral effect against the baseline (control or A).

"*The concept of A/B testing is simple: show different variations of your website to different people and measure which variation is the most effective at turning them into customers."*

Dan Siroker and Pete Koomen (Book | A/B Testing: The most powerful way to turn clicks into customers)

Why you should A/B test

A/B testing allows individuals, teams, and companies to make careful changes to their user experiences while collecting data on the impact it makes. This allows them to construct hypotheses and to learn what elements and optimizations of their experiences impact user behavior the most. In another way, they can be proven wrong—their opinion about the best experience for a given goal can be proven wrong through an A/B test.

More than just answering a one-off question or settling a disagreement, A/B testing can be used to continually improve a given experience or improve a single goal like conversion rate optimization (CRO) over time.

Examples of A/B testing applications:

- **B2B lead generation:** If you're a technology company, you can improve your landing pages by testing changes to headlines, form fields, and CTAs. By testing one element at a time, you can identify which changes increase lead quality and conversion rates.
- **Campaign performance:** If you're a marketer running a product marketing campaign, you can optimize ad spend by testing both ad copy and landing pages. For example, testing different layouts helped identify which version converted visitors to customers most efficiently, reducing overall customer acquisition costs.
- **Product experience:** The product teams in your company can use A/B testing to validate assumptions, prioritize features that matter, and deliver products without risks. From onboarding flows to in-product notifications, testing helps optimize the user experience while maintaining clear goals and hypotheses.

A/B testing helps transform decision-making from opinion-based to data-driven, challenging the HiPPO slang (Highest Paid Person's Opinion).

As Dan Siroker notes, "It's about being humble... maybe we don't actually know what's best, let's look at data and use that to help guide us."

How to do A/B testing

The following is an A/B testing framework you can use to start running tests:

1. Collect data

- Use analytics tools like Google Analytics to identify opportunities
- Focus on high-traffic areas through heatmaps
- Look for pages with high drop-off rates

2. Set clear goals

- Define specific metrics to improve
- Establish measurement criteria
- Set target improvements

3. Create test hypothesis

- Form clear predictions
- Base ideas on existing data
- Prioritize by potential impact

4. Design variations

- Make specific, measurable changes
- Ensure proper tracking
- Test technical implementation

5. Run the experiment

- Split traffic randomly
- Monitor for issues
- Collect data systematically

6. Analyze results

- Check statistical significance
- Review all metrics
- Document learnings

If your variation wins, fantastic! Apply those insights across similar pages and continue iterating to build on your success. But remember - not every test will be a winner, and that's perfectly fine.

More on failed A/B tests: How test failures can lead you to success

In A/B testing, there are no true failures - only opportunities to learn. Every test, whether it shows positive, negative, or neutral results, provides valuable insights about your users and helps refine your testing strategy.

A/B testing examples

Here are two examples of A/B testing in action.

1. Homepage A/B test

Felix and Michiel from the digital team decided to add a paw-some surprise to our creative and messaging on our homepage.

Their target was to drive user engagement.

The answer in this case, the team found out, was a lot of woofs.

During the experiment, website visitors who pet the dog on the website's homepage, got a link to our "Evolution of Experimentation" report.

However, you'll only see the dog 50% of the time.

Result: *People exposed to the dog consumed the content 3x more than those who didn't see the dog*.

2. Pop-up to flop-up

Ronnie Cheung, Senior Strategy Consultant, Optimizely, wanted to introduce a facility detail pop-up on the map view as when users were clicking on pins on the map view, they would be taken to a PDP page which added an extra step to complete checkout.

- **Result:** Fewer users entered the checkout page
- **Takeaway:** Improve the pop-up information so that users can confidently proceed to checkout.

If you'd like to see more, here's an industry-specific list of A/B testing use cases and examples.

And for success stories, check out the big book of experimentation. It has 40+ case studies showing the challenges and the hypotheses used to solve them.

Creating a culture of A/B testing

Great digital marketing teams make sure to involve multiple departments in their experimentation program. When testing across different departments and touchpoints, you can increase the confidence level that the changes you're making to your marketing are statistically significant and making a positive impact on your bottom line.

Use cases include:

1. **A/B testing social media:** Post timing, content formats, Ad creative variations, audience targeting, campaign messaging
2. **A/B testing marketing:** Email campaigns, landing pages, Ad copy and creatives, call-to-action buttons, form designs
3. **Website A/B testing:** Navigation design, page layouts, content presentation, checkout processes, search functionality

But you can only scale your program if it adopts a test-and-learn mindset. Here's how to build a culture of experimentation:

1. Leadership buy-in

- Demonstrate value through early wins
- Share success stories
- Link results to business goals

2. Team empowerment

- Provide necessary tools
- Offer training
- Encourage hypothesis generation

3. Process integration

- Make testing part of the development workflow
- Create clear testing protocols
- Document and share learnings

6 measuring pillars for building a culture of experimentation

A/B testing metrics

A/B testing requires analytics that can track multiple metric types while connecting to your data warehouse for deeper insights.

To start, here's what you can measure:

- **Primary success metrics:** Conversion rate, click-through rate, revenue per visitor, average order value
- **Supporting indicators**: Time on page, Bounce rate, Pages per session, User journey patterns
- **Technical performance:** Load time, error rates, mobile responsiveness, browser compatibility

What really makes the difference is warehouse native analytics. It allows you to maintain full control over data location by keeping your test data in-house. Further, you can test against real business outcomes and enable automated cohort analysis. It provides seamless cross-channel testing with a single source of truth while maintaining strict data governance and compliance.

Understanding A/B test results

Test results vary based on your business type and goals. For example, while e-commerce sites focus on purchase metrics, B2B companies might prioritize lead generation metrics. Whatever your focus, start with clear goals before launching your test.

For example, if you're testing a CTA button, you'll see:

- Number of visitors who saw each version
- Clicks on each variant
- Conversion rate (percentage of visitors who clicked)
- Statistical significance of the difference

When running A/B tests and analyzing results, statistical significance tells you if your test results are reliable or just random chance.

When analyzing results:

- Compare against your baseline (A version)
- Look for a statistically significant uplift
- Consider the practical impact of the improvement
- Check if results align with other metrics

Segmenting A/B tests

Larger sites and apps often employ segmentation for their A/B tests. If your number of visitors is high enough, this is a valuable way to test changes for specific sets of visitors. A common segment used for A/B testing is splitting out new visitors versus return visitors. This allows you to test changes to elements that only apply to new visitors, like signup forms.

On the other hand, a common A/B testing mistake made is to create audiences for tests that are too small. So:

- Only segment when you have sufficient traffic
- Start with common segments (new vs. returning visitors)
- Ensure segment size supports statistical significance
- Avoid creating too many small segments that could lead to false positives

A/B testing & SEO

Google permits and encourages A/B testing and has stated that performing an A/B or multivariate test poses no inherent risk to your website's search rank. However, it is possible to jeopardize your search rank by abusing an A/B testing tool for purposes such as cloaking. Google has articulated some best practices to ensure that this doesn't happen:

1. **No cloaking:** Cloaking is the practice of showing search engines different content than a typical visitor would see. Cloaking can result in your site being demoted or even removed from the search results. To prevent cloaking, do not abuse visitor segmentation to display different content to Googlebot based on user-agent or IP address.
2. **Use rel="canonical":** If you run a split test with multiple URLs, you should use the rel="canonical" attribute to point the variations back to the original version of the page. Doing so will help prevent Googlebot from getting confused by multiple versions of the same page.
3. **Use 302 redirects instead of 301s:** If you run a test that redirect the original URL to a variation URL, use a 302 (temporary) redirect vs a 301 (permanent) redirect. This tells search engines such as Google that the redirect is temporary and that they should keep the original URL indexed rather than the test URL.

A media company might want to increase readership, increase the amount of time readers spend on their site, and amplify their articles with social sharing. To achieve these goals, they might test variations on:

- Email sign-up modals
- Recommended content
- Social sharing buttons

A travel company may want to increase the number of successful bookings are completed on their website or mobile app, or may want to increase revenue from ancillary purchases. To improve these metrics, they may test variations of:

- Homepage search modals
- Search results page
- Ancillary product presentation

An e-commerce company might want to improve their customer experience, resulting in an increase in the number of completed checkouts, the average order value, or increase holiday sales. To accomplish this, they may A/B test:

- Homepage promotions
- Navigation elements
- Checkout funnel components

A technology company might want to increase the number of high-quality leads for their sales team, increase the number of free trial users, or attract a specific type of buyer. They might test:

- Lead form fields

- Free trial signup flow
- Homepage messaging and call-to-action

Three takeaways

You can apply to your A/B testing program:

1. You can't rationalize customer behavior.
2. No idea is too big, too clever, or too 'best practice' that it can't be tested.
3. Completely redesigning a website from scratch is not the way to go. Go specific, but start small.

Remember, testing is an incredibly valuable opportunity to learn how customers interact with your website. Start now with Optimizely Web Experimentation.

You may find it interesting: AI experimentation (from ideation to results)

Frequently asked questions about A/B testing

What is the minimum time to run an A/B test?

Tests usually run for 1-2 weeks to account for traffic patterns, but the exact duration depends on your traffic volume and desired confidence level.

What are some common mistakes to avoid in A/B testing?

Here are three A/B testing myths to avoid: Testing is just about optimization, more tests mean more impact, and analytics is simply data.

How do you determine the right sample size for an A/B test?

Use baseline conversion rate, minimum detectable improvement, and desired confidence level (typically 95%) through a sample size calculator.

Can A/B testing be applied to offline marketing strategies?

Yes, A/B testing works for offline strategies like direct mail, in-store displays, and sales scripts. Just maintain controlled conditions and sufficient data collection for meaningful results.

What's the difference between A/B testing and multivariate testing?

Unlike simple A/B tests, multivariate testing examines multiple variables simultaneously and gives you the combined impact of changes.

Source: https://vwo.com/ab-testing/

What is A/B testing?

A/B testing compares two versions of an app or webpage to identify the better performer. It's a method that helps you make decisions based on real data rather than just guessing. It compares options to learn what customers prefer. You can test website/app layouts, email subject lines, product designs, CTA button text, colors, etc.

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

The version that moves your business metric(s) in the positive direction is known as the 'winner.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

The metrics for conversion are unique to each website. For instance, in the case of eCommerce, it may be the sale of the products. Meanwhile, for B2B, it may be the generation of qualified leads.

A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc. If you're not A/B testing your website, you're surely losing out on a lot of potential business revenue.

Why should you consider A/B testing?

If B2B businesses today are unhappy with all the unqualified leads they get per month, eCommerce stores, on the other hand, are struggling with a high cart abandonment rate. Meanwhile, media and publishing houses are also dealing with low viewer engagement. These core conversion metrics are affected by some common problems like leaks in the conversion funnel, drop-offs on the payment page, etc.

Let's see why you should do A/B testing:

1. Solve visitor pain points

Visitors come to your website to achieve a specific goal that they have in mind. It may be to understand more about your product or service, buy a particular product, read/learn more about a specific topic, or simply browse. Whatever the visitor's goal may be, they may face some common pain points while achieving their goal. It can be a confusing copy or hard to find the CTA button like buy now, request a demo, etc.

Not being able to achieve their goals leads to a bad user experience. This increases friction and eventually impacts your conversion rates. Use data gathered through visitor behavior analysis tools such as heatmaps, Google Analytics, and website surveys to solve your visitors' pain points. This stands true for all businesses: eCommerce, travel, SaaS, education, media, and publishing.

2. Get better ROI from existing traffic

As most experienced optimizers have come to realize, the cost of acquiring quality traffic on your website is huge. A/B testing lets you make the most out of your existing traffic and helps you increase conversions without having to spend additional dollars on acquiring new traffic. A/B testing can give you high ROI as sometimes, even the minutest of changes on your website can result in a significant increase in overall business conversions.

3. Reduce bounce rate

One of the most important metrics to track to judge your website's performance is its bounce rate. There may be many reasons behind your website's high bounce rate, such as too many options to choose from, expectations mismatch, confusing navigation, use of too much technical jargon, and so on.

Since different websites serve different goals and cater to different segments of audiences, there is no one-size-fits-all solution to reducing bounce rates. However, running an A/B test can prove beneficial. With A/B testing, you can test multiple variations of an element of your website till you find the best possible version. This not only helps you find friction and visitor pain points but helps improve your website visitors' overall experience, making them spend more time on your site and even converting into a paying customer.

4. Make low-risk modifications

Make minor, incremental changes to your web page with A/B testing instead of getting the entire page redesigned. This can reduce the risk of jeopardizing your current conversion rate.

A/B testing lets you target your resources for maximum output with minimal modifications, resulting in an increased ROI. An example of that could be product description changes. You can perform an A/B test when you plan to remove or update your product descriptions. You do not know how your visitors are going to react to the change. By running an A/B test, you can analyze their reaction and ascertain which side the weighing scale may tilt.

Another example of low-risk modification can be the introduction of a new feature change. Before introducing a new feature, launching it as an A/B test can help you understand whether or not the new change that you're suggesting will please your website audience.

Implementing a change on your website without testing it may or may not pay off in both the short and long run. Testing and then making changes can make the outcome more certain.

5. Achieve statistically significant improvements

Since A/B testing is entirely data-driven with no room for guesswork, gut feelings, or instincts, you can quickly determine a "winner" and a "loser" based on statistically significant improvements in metrics like time spent on the page, number of demo requests, cart abandonment rate, click-through rate, and so on.

6. Redesign the website to increase future business gains

Redesigning can range from a minor CTA text or color tweak to particular web pages to completely revamping the website. The decision to implement one version or the other should always be data-driven when A/B testing. Do not quit testing with the design being finalized. As the new version goes live, test other web page elements to ensure that the most engaging version is served to the visitors.

What can you A/B test?

Your website's conversion funnel determines the fate of your business. Therefore, every piece of content that reaches your target audience via your website must be optimized to its maximum potential. This is especially true for elements that have the potential to influence the behavior of your website visitors and business conversion rate. When undertaking an optimization program, test the following key site elements (the list, however, is not exhaustive):

1. Headlines and subheadlines

A headline is practically the first thing that a visitor notices on a web page. It's also what defines their first and last impression, filling in the blanks on whether or not they'll go ahead and convert into paying customers. Hence, it's imperative to be extra cautious about your site's headlines and subheadlines. Ensure they're short, to-the-point, catchy, and convey your desired message in the first stance. Try A/B testing a few copies with different fonts and writing styles, and analyze which catches your visitors' attention the most and compels them to convert. You can also use VWO's AI-powered text generation system to generate recommendations for the existing copy on your website.

2. Body

The body or main textual content of your website should clearly state what the visitor is getting – what's in store for them. It should also resonate with your page's headline and subheadline. A well-written body can significantly increase the chances of turning your website into a conversion magnet.

While drafting your website's content, keep the following two parameters in mind:

- **Writing style:** Use the right tonality based on your target audience. Your copy should directly address the end-user and answer all their questions. It must contain key phrases that improve usability and stylistic elements that highlight important points.
- **Formatting:** Use relevant headlines and subheadlines, break the copy into small and easy paragraphs, and format it for skimmers using bullet points or lists.

Interestingly, experience optimizers can now take advantage of artificial intelligence to create website copies. GPT-3.5 Turbo or Generative Pre-trained Transformer 3, is an AI-powered neural network that has the ability to produce nearly flawless text content relevant to any given context. Built by OpenAI, GPT-3.5 Turbo uses machine learning to predict and draft content just like a human. The best part? You can now integrate OpenAI's GPT-3.5 Turbo with the VWO Testing account and create variations for your website copy and deploy them without the help of an expert writer or IT, respectively.

3. Subject lines

Email subject lines directly impact open rates. If a subscriber doesn't see anything they like, the email will likely wind up in their trash bin.

According to recent research, average open rates across more than a dozen industries range from 25 to 47 percent. Even if you're above average, only about half of your subscribers might open your emails.

A/B testing subject lines can increase your chances of getting people to click. Try questions versus statements, test power words against one another, and consider using subject lines with and without emojis.

Design and layout

Because everything seems so essential, businesses sometimes struggle with finding only the most essential elements to keep on their website. With A/B testing, this problem can be solved once and for all.

For example, as an eCommerce store, your product page is extremely important from a conversion perspective. One thing for sure is that with technological progress in its current stage, customers like to see everything in high definition before buying it. Therefore, your product page must be in its most optimized form in terms of design and layout.

Along with the copy, the page's design and layout include images (product images, offer images, etc.) and videos (product videos, demo videos, advertisements, etc.). Your product page should answer all of your visitor's questions without confusing them and without getting cluttered:

- **Provide clear information:** Based on the products you sell, find creative ways to provide all necessary context and accurate product descriptions so that prospective buyers do not get overwhelmed with an unorganized copy while looking for answers to their queries. Write clear copies and provide easily noticeable size charts, color options, etc.
- **Highlight customer reviews:** Add both good and bad reviews for your products. Negative reviews add credibility to your store.
- **Write simple content:** Avoid confusing potential buyers with complicated language in the quest to decorate your content. Keep it short, simple, and fun to read.
- **Create a sense of urgency:** Add tags like 'Only 2 Left In Stock', countdowns like 'Offer Ends in 2 Hours and 15 Minutes', or highlight exclusive discounts and festive offers, etc., to nudge the prospective buyers to purchase immediately.

Other important pages whose design needs to be on point are pages like the home page and landing page. Use A/B testing to discover the most optimized version of these critical pages. Test as many ideas as you can, such as adding plenty of white space and high-definition images, featuring product videos instead of images, and testing out different layouts.

Declutter your pages using insights from heatmaps, clickmaps, and scrollmaps to analyze dead clicks and identify distractions. The less cluttered your home page and landing pages, the more likely it is for your visitors to easily and quickly find what they're looking for.

Navigation

Another element of your website that you can optimize by A/B testing is your website's navigation. It is the most crucial element when it comes to delivering an excellent user experience. Make sure you have a clear plan for your website's structure and how different pages will be linked to each other and react within that structure.

Your website's navigation starts on the home page. The home page is the parent page from which all other pages emerge and link back to each other. Make sure your structure is such that visitors can easily find what they're looking for and do not get lost because of a broken navigation path. Each click should direct visitors to the desired page.

Mentioned below are some ideas to help you step up your navigation game:

- **Match visitor expectations** by placing your navigation bar in standard places like horizontal navigation on the top and vertical down the left to make your website easier to use.
- **Make your website's navigation predictable** by keeping similarly themed content in the same bucket or in related buckets to reduce your visitor's cognitive load. For example, as an eCommerce store, you may be selling a variety of earphones and headphones. Some of them may be wired, while others may be wireless or ear-pods. Bucket these in such a way that when a visitor looks for earphones or headphones, they find all these varieties in one place rather than having to search for each kind separately
- **Creating a fluid, easy-to-navigate website** by keeping its structure simple, predictable, and matching your visitors' expectations. This will not only increase the chances of getting more conversions but also create a delightful customer experience forcing visitors to come back to your website.

Forms

Forms are mediums through which prospective customers get in touch with you. They become even more important if they are part of your purchase funnel. Just as no two websites are the same, no two forms addressing different audiences are the same. While a small comprehensive form may work for some businesses, long forms might do wonders for their lead quality for other businesses.

You can figure out which style works for your audience the best by using research tools/methods like form analysis to determine the problem area in your form and work towards optimizing it.

CTA (Call-to-action)

The CTA is where all the real action takes place – whether or not visitors finish their purchases and convert if they fill out the sign-up form or not, and more such actions that have a direct bearing on your conversion rate. A/B testing enables you to test different CTA copies, their placement across the web page, toy with their size and color scheme, and so on. Such experimentation helps understand which variation has the potential to get the most conversions.

Social proof

Social proof may take the form of recommendations and reviews from experts in particular fields, from celebrities and customers themselves, or can come as testimonials, media mentions, awards and badges, certificates, and so on. The presence of these proofs validates the claims made by your website. A/B testing can help you determine whether or not adding social proof is a good idea. If it is a good idea, what kinds of social proof should you add, and how many should you add? You can test different types of social proofs, their layouts, and placements to understand which works best in your favor.

Content depth

Some website visitors prefer reading long-form content pieces that extensively cover even the minutest of details. Meanwhile, many others just like to skim through the page and deep dive only into the topics that are most relevant to them. In which category does your target audience fall?

A/B test content depth. Creating two pieces of the same content, one that's significantly longer than the other, provides more details. Analyze which compels your readers the most.

Understand that content depth impacts SEO and many other business metrics such as the conversion rate, page time spent, and bounce rate. A/B testing enables you to find the ideal balance between the two.

What are the different types of A/B tests?

Post learning about which web page elements to test to move your business metrics in the positive direction, let's move ahead and learn about the different kinds of testing methods along with their advantages.

Ideally, there are four basic testing methods – A/B testing, Split URL testing, Multivariate testing, and Multipage testing. We've already discussed the first kind, namely, A/B testing. Let's move on to the others.

Split URL testing

Many people in the testing arena confuse Split URL testing with A/B testing. However, the two are fundamentally very different. Split URL testing refers to an experimentation process wherein an entirely new version of an existing web page URL is tested to analyze which one performs better.

Typically, A/B testing is used when you wish to only test front-end changes on your website. On the other hand, Split URL testing is used when you wish to make significant changes to your existing page, especially in terms of design. You're not willing to touch the existing web page design for comparison purposes.

When you run a Split URL test, your website traffic is split between the control (original web page URL) and variations (new web page URL), and each of their respective conversion rates is measured to decide the winner.

Advantages of Split URL testing

- Ideal for trying out radical new designs while using the existing page design for comparative analysis.
- Recommended for running tests with non-UI changes, such as switching to a different database, optimizing your page's load time, etc.
- Change up web page workflows. Workflows dramatically affect business conversions, helping test new paths before implementing changes and determining if any of the sticking points were missed.
- A better and much-recommended testing method for dynamic content.

Multivariate testing (MVT)

Multivariate testing (MVT) refers to an experimentation method wherein variations of multiple-page variables are simultaneously tested to analyze which combination of variables performs the best out of all the possible permutations. It's more complicated than a regular A/B test and is best suited for advanced marketing, product, and development professionals.

Here's an example to give you a more comprehensive description of multivariate testing. Let's say you decide to test 2 versions, each of the hero image, call-to-action button color, and headlines of one of your landing pages. This means a total of 8 variations are created, which will be concurrently tested to find the winning variation.

Here's a simple formula to calculate the total number of versions in a multivariate test:

[No. of variations of element A] x [No. of variations of element B] x [No. of variations of element C]… = [Total No. of variations]

When conducted properly, multivariate testing can help eliminate the need to run multiple and sequential A/B tests on a web page with similar goals. Running concurrent tests with a greater number of variations helps you save time, money, and effort and come to a conclusion in the shortest possible time.

Advantages of Multivariate testing

Multivariate testing typically offers primary three benefits:

- Helps avoid the need to conduct several sequential A/B tests with the same goal and saves time since you can simultaneously track the performance of various tested page elements.
- Easily analyze and determine the contribution of each page element to the measured gains,
- Map all the interactions between all independent element variations (page headlines, banner images, etc.).

Multipage testing

Multipage testing is a form of experimentation where you can test changes to particular elements across multiple pages.

There are two ways to conduct a multipage test. One, you can either take all the pages of your sales funnel and create new versions of each, which makes your challenger the sales funnel, and you then test it against the control. This is called **Funnel Multipage testing**.

Two, you can test how the addition or removal of recurring element(s), such as security badges, testimonials, etc., can impact conversions across an entire funnel. This is called **Classic or Conventional Multipage testing**.

Advantages of Multipage testing

Similar to A/B testing, Multipage testing is easy to create and run and provides meaningful and reliable data with ease and in the shortest possible time.

The advantages of multipage testing are as follows:

1. It enables you to create consistent experiences for your target audience.
2. It helps your target audience see a consistent set of pages, no matter if it's the control or one of its variations.
3. It enables you to implement the same change on several pages to ensure that your website visitors don't get distracted and bounce off between different variations and designs when navigating through your website.

Which statistical approach to use to run an A/B test?

Post learning about four different types of A/B testing experimentation methods, it's equally important to understand which statistical approach to adopt to successfully run an A/B test and draw the right business conclusion.

Ideally, there are two types of statistical approaches used by A/B/n experimenters across the globe: Frequentist and Bayesian. Each of these approaches has its own pros and cons. However, we, at VWO, use, support, and promote the Bayesian approach.

The comparison between the two approaches given below will help you understand why.

Frequentist approach:

The frequentist approach of probability defines the probability of an event with relation to how frequently (hence the name) a particular event occurs in a large number of trials/data points. When applied to the world of A/B testing, one can see that anyone going with the frequentist approach would need more data (a function of more number of visitors tested and over longer durations) to come to the right conclusions. This is something that limits you in scaling up any A/B testing effort. According to the Frequentist approach, it is essential to define your A/B test's duration based on sample size to reach the right test conclusions. The tests are based on the fact that every experiment can be repeated infinite times.

Following this approach calls for a lot of attention to detail for every test that you run because for the same set of visitors, you'll be forced to run longer duration tests than the Bayesian approach. Hence, each test needs to be treated with extreme care because there are only a few tests that you can run in a given timeframe. Unlike Bayesian statistics, the Frequentist approach is less intuitive and often proves difficult to understand.

Bayesian approach:

As compared to the Frequentist approach, Bayesian statistics is a theory-based approach that deals with the Bayesian interpretation of probability, where probability is expressed as a degree of belief in an event. In other words, the more you know about an event, the better and faster you can predict the end outcomes. Rather than being a fixed value, probability under Bayesian statistics can change as new information is gathered. This belief may be based on past information such as the results of previous tests or other information about the event.

Frequentist Approach	**Bayesian Approach**
Frequentist Statistics follow the 'Probability as Long-Term Frequency' definition of probability.	Bayesian Statistics follow the notions of 'Probability as Degree of Belief' and 'Logical Probability.'
In this approach, you only use data from your current experiment. The frequentist solution is to conduct tests and draw conclusions.	In this approach, you use your prior knowledge from the previous experiments and try to incorporate that information into your current data. The Bayesian solution is to use existing data to draw conclusions.
Give an estimated mean (and standard deviation) of samples where A beats B but completely ignores the cases when B beats A.	It takes into account the possibility of A beating B and also calculates the range of the improvement you can expect.

Requires the test to run for a set period to get correct data from it but can't figure out how close or far A and B actually are. It fails to tell you the probability of A beating B.	Gives you more control over testing. You can now plan better, have a more accurate reason to end tests, and get into the nitty-gritty of how close or far apart A and B are.

Unlike the frequentist approach, the Bayesian approach provides actionable results almost 50% faster while focusing on statistical significance. At any given point, provided you have enough data at hand, the Bayesian approach tells you the probability of variation A having a lower conversion rate than variation B or the control. It does not have a defined time limit attached to it, nor does it require you to have an in-depth knowledge of statistics.

In the simplest of terms, the Bayesian approach is akin to how we approach things in everyday life. For instance, you misplaced your mobile phone in your house. As a frequentist, you would only use a GPS tracker to track it and only check the area the tracker is pointing to. While as a Bayesian, you will not only use a GPS tracker but also check all the places in the house you earlier found your misplaced phone. In the former, the event is considered a fixed value, while in the latter, all past and future knowledge are utilized to locate the phone.

To get a clearer understanding of the two statistical approaches, here's a comparison table just for you:

Once you've figured out which testing method and statistical approach you wish to use, it's time to learn the art and science of performing A/B tests on VWO's A/B testing platform.

How to perform an A/B test?

A/B testing offers a very systematic way of finding out what works and what doesn't work in any given marketing campaign. Most marketing efforts are geared toward driving more traffic. As traffic acquisition becomes more difficult and expensive, it becomes paramount to offer your users the best experience who comes to your website. This will help them achieve their goals and allow them to convert in the fastest and most efficient manner possible. A/B testing in marketing allows you to make the most out of your existing traffic and increase revenue inflow.

A structured A/B testing program can make marketing efforts more profitable by pinpointing the most crucial problem areas that need optimization. A/B testing is now moving away from being a standalone activity that is conducted once in a blue moon to a more structured and continuous activity, which should always be done through a well-defined CRO process. Broadly, it includes the following steps:

Step 1: Research

Before building an A/B testing plan, one needs to conduct thorough research on how the website is currently performing. You will have to collect data on everything related to how many users are coming onto the site, which pages drive the most traffic, the various conversion goals of different pages, etc. The A/B testing tools used here can include quantitative website analytics tools such as Google Analytics, Omniture, Mixpanel, etc., which can help you figure out your most visited pages, pages with most time spent, or pages with the highest bounce rate. For example, you may want to start by shortlisting pages that have the highest revenue potential or the highest daily traffic. Following this, you may want to dive deeper into the qualitative aspects of this traffic.

Heatmap tools are the leading technology used to determine where users are spending the most time, their scrolling behavior, etc. This can help you identify problem areas on your website. Another popular tool used to do more insightful research is website user surveys. Surveys can act as a direct conduit between your website team and the end user and often highlight issues that may be missed in aggregate data.

Further, qualitative insights can be derived from session recording tools that collect data on visitor behavior, which helps in identifying gaps in the user journey. In fact, session recording tools combined with form analysis surveys can uncover insights on why users may not be filling your form. It may be due to some fields that ask for personal information or users, maybe abandoning your forms for too long.

As we can see, both quantitative and qualitative research can help us prepare for the next step in the process, making actionable observations for the next steps.

Step 2: Observe and formulate hypothesis

Get closer to your business goals by logging research observations and creating data-backed hypotheses aimed at increasing conversions. Without these, your test campaign is like a directionless compass. The qualitative and quantitative research tools can only help you with gathering visitor behavior data. It is now your responsibility to analyze and make sense of that data. The best way to utilize every bit of data collated is to analyze it, make keen observations on them, and then draw websites and user insights to formulate data-backed hypotheses. Once you have a hypothesis ready, test it against various parameters such as how much confidence you have of it winning, its impact on macro goals, and how easy it is to set up, and so on.

While brainstorming new testing ideas, if you ever find yourself facing a creativity block, don't worry—VWO has a solution for you. You can now get AI-generated testing ideas within minutes.

The webpage URL you entered is scanned to show personalized testing ideas for that page. For instance, if you enter the URL of your pricing page, you will be presented with several relevant ideas supported by correct hypotheses, valid scientific principles, and feasible actionables. Additionally, you can generate testing ideas based on a specific goal for that page.

For example, if your objective is to 'Increase clicks on the contact sales team CTA,' you will see relevant ideas to help you achieve that goal. You can then add these testing ideas to VWO Plan and create a robust pipeline of tests to be carried out in the future. Swiftly jotting down recommendations and aligning them with specific goals can help save time and accelerate the testing process.

Step 3: Create variations

The next step in your testing program should be to create a variation based on your hypothesis, and A/B test it against the existing version (control). A variation is another version of your current version with changes that you want to test. You can test multiple variations against the control to see which one works best. Create a variation based on your hypothesis of what might work from a UX perspective. For example, enough people not filling forms? Does your form have too many fields? Does it ask for personal information? Maybe you can try a variation with a shorter form or another variation by omitting fields that ask for personal information.

Step 4: Run test

Before we get to this step, it's important to zero upon the type of testing method and approach you want to use. Once you've locked down on either one of these types and approaches based (refer to the above-written chapters) on your website's needs and business goals, kick off the test and wait for the stipulated time for achieving statistically significant results. Keep one thing in mind – no matter which method you choose, your testing method and statistical accuracy will determine the end results.

For example, one such condition is the timing of the test campaign. The timing and duration of the test have to be on point. Calculate the test duration keeping in mind your average daily and monthly visitors, estimated existing conversion rate, minimum improvement in conversion rate you expect, number of variations (including control), percentage of visitors included in the test, and so on.

Use our Bayesian Calculator to calculate the duration for which you should run your A/B tests for achieving statistically significant results.

Step 5: Analyse results and deploy changes

Even though this is the last step in finding your campaign winner, analysis of the results is extremely important. Because A/B testing calls for continuous data gathering and analysis, it is in this step that your entire journey unravels. Once your test concludes, analyze the test results by considering metrics like percentage increase, confidence level, direct and indirect impact on other metrics, etc. After you have considered these numbers, if the test succeeds, deploy the winning variation. If the test remains inconclusive, draw insights from it, and implement these in your subsequent tests.

A/B testing lets you systematically work through each part of your website to improve conversions.

How to make an A/B testing calendar – plan & prioritize

A/B testing should never be considered an isolated optimization exercise. It's a part of a wider holistic CRO program and should be treated as such. An effective optimization program typically has two parts, namely, plan and prioritize. Waking up one day and deciding to test your website is not how things are done in CRO. A good amount of brainstorming, along with real-time visitor data, is the only way to go about it.

In plain words, you begin by analyzing existing website data and gathering visitor behavior data, then move on to preparing a backlog of action items based on them, further prioritizing each of these items, running tests, and then drawing insights for the future. Eventually, when, as experience optimizers, you conduct enough ad-hoc based tests, you would want to scale your A/B testing program to make it more structured.

The first step to doing this is by making an A/B testing calendar. A good testing calendar or a good CRO program will take you through 4 stages:

Stage 1: Measure

This stage is the planning stage of your A/B testing program. It includes measuring your website's performance in terms of how visitors are reacting to it. In this stage, you should be able to figure out what is happening on your website, why it is happening, and how visitors are reacting to it. Everything that goes on in your website should correspond to your business goals. So before everything else, you need to be sure what your business goal/s is (are). Tools like Google Analytics

can help you measure your goals. Once you have clearly defined goals, set up GA for your website and define your key performance indicators.

Let's take an online mobile phone cover store as an example. The business goal for this store is to increase revenue by increasing online orders and sales. The KPI set to track this goal would then be the number of phone covers sold.

This stage, however, does not simply end with defining website goals and KPIs. It also includes understanding your visitors. We have already discussed the various tools that can be used to gather visitor behavior data. Once data is collected, log in observations and start planning your campaign from there. Better data means higher sales.

Once the business goals are defined, KPIs set, and website data and visitor behavior data analyzed, it is time to prepare a backlog.

Backlog: "an accumulation of tasks unperformed or materials not processed."

Your backlog should be an exhaustive list of all the elements on the website that you decide to test based on the data you analyzed. With a data-backed backlog ready, the next step is formulating a hypothesis for each backlog item. With the data gathered in this stage and its analysis, you will now have enough context of what happens on your website and why. Formulate a hypothesis based on them.

For example, after analyzing the data gathered using quantitative and qualitative research tools in the 1st stage, you come to the conclusion that not having multiple payment options led to maximum prospect customers dropping off on the checkout page. So you hypothesize that "adding multiple payment options will help reduce drop off on the checkout page."

In short, by the end of this stage, you will know the whats and whys of your website.

Stage 2: Prioritize

The next stage involves prioritizing your test opportunities. Prioritizing helps you scientifically sort multiple hypotheses. By now, you should be fully equipped with website data, visitor data and be clear on your goals. With the backlog, you prepared in the first stage and the hypothesis ready for each candidate, you are halfway there on your optimization roadmap. Now comes the main task of this stage: prioritizing.

In stage 2, you should be fully equipped to identify problem areas of your website and leaks in your funnel. But not every action area has equal business potential. So it becomes imperative to weigh out your backlog candidates before picking the ones you want to test. There are a few things to be kept in mind while prioritizing items for your test campaign like the potential for improvement, page value and cost, the importance of the page from a business perspective, traffic on the page, and so on.

But how can you ensure that no subjectivity finds its way in your prioritization framework? Can you be 100% objective at all times? As humans, we give loads of importance to gut feelings, personal opinions, ideas, and values because these are the things that help us in our everyday lives. But, CRO is not everyday life. It is a scientific process that needs you to be objective and make sound data-backed decisions and choices. The best way to weed out these subjectivities is by adopting a prioritization framework.

There are many prioritization frameworks that even experts employ to make sense of their huge backlogs. On this pillar page, you will learn about the most popular frameworks that experience optimizers use – the CIE prioritization framework, the PIE prioritization framework, and the LIFT Model.

1. CIE Prioritization Framework

In the CIE framework, there are three parameters on which you must rate your test on a scale of 1 to 5:

- Confidence: On a scale of 1 to 5 – 1 being the lowest and 5 being the highest – select how confident you are about achieving the expected improvement through the hypothesis.
- Importance: On a scale of 1 to 5 – 1 being the lowest, and 5 being the highest – select how crucial the test (for which the hypothesis is created) is.
- Ease: On a scale of 1 to 5 – 1 being the most difficult, and 5 being the easiest – select the complexity of the test. Rate how difficult it will be to implement the changes identified for the test.

Before you rate your hypotheses, consider these 3 things:

A. How confident are you of achieving the uplift?

Prototyping the user persona, you are targeting can help you determine the potential of a hypothesis. With a sound understanding of your audience, you can make an educated assumption on whether the hypothesis will address the users' apprehensions and doubts and nudge them to convert or not.

B. How valuable is the traffic you are running this test for?

Your website may be attracting visitors in large numbers, but not all visitors become buyers. Not all convert. For example, a hypothesis built around the checkout page holds a higher importance than the one built around the product features page. This is because visitors on the checkout page are way deep in your conversion funnel and have a higher chance to convert rather than visitors on your product features page.

C. How easy is it to implement this test?

Next comes determining the ease of implementing your test. Try to answer some questions: Would it need a lot of strategizing on your part to implement the hypothesis? What is the effort needed in designing and developing the solution proposed by the hypothesis? Can the changes suggested in the hypothesis be implemented using just the Visual Editor, or does it warrant adding custom code? It is only after you have answered all these and other such questions should you rate your backlog candidate on the easing parameter.

2. PIE Prioritization Framework

The PIE framework was developed to answer the question, "Where should I test first?". The whole aim of the prioritization stage in your A/B testing journey is to find the answer to this very question. The PIE framework talks about 3 criteria that you should consider while choosing what to test when: potential, importance, and ease.

Potential means a page's ability to improve. The planning stage should equip you with all the data you need to determine this.

Importance refers to a page's value: how much traffic comes to the page. If you have identified a problem page, but there is no traffic on that page, then that page is of less importance when compared to other pages with higher traffic.

The third and final criteria is ease. Ease defines how difficult it is to run a test on a particular page or element. One way to determine ease of testing a page is using tools like landing page analyzer to determine the current state of your landing pages, estimate the number and scale of change it would require, and prioritize which ones to do or whether to do it at all. This is important from the perspective of resources. Many businesses drop the idea of undertaking and A/B testing campaign because of the lack of resources. These resources are of 2 kinds:

A. Human resource

Even though businesses have been using CRO and A/B testing for many years, it is only recently that the two concepts gained a front stage. Because of this, a large segment of the market does not have a dedicated optimization team, and when they do, it is usually limited to a handful of people. This is where a planned optimization calendar comes in handy. With a properly planned and prioritized backlog, a small CRO team can focus its limited resources on high stake items.

B. Tools:

As popular as CRO and A/B testing are getting, so are hundreds of A/B testing tools– both low end and high. Without the perspective of an expert, if businesses were to pick one out of the lot, say the cheapest one, and start A/B testing every single item on the backlog, they will reach no statistically significant conclusion. There are 2 reasons for this: one, testing without prioritization is bound to fail and not reap any business profits. Two, not all tools are of the same quality.

Some tools may be costlier, but they are either integrated with good qualitative and quantitative research tools or are brilliant standalone tools making them more than capable of producing statistically significant results. While the other lot may be cheaper and lure businesses during capital crunch and with a huge backlog, these tools will only be an investment loss to them without any benefits. Prioritization will help you make sense of your backlog and dedicate whatever little resources you have to a profitable testing candidate.

Backlog candidates should be marked on how hard they are to test based on technical and economic ease. You can quantify each potential candidate as a business opportunity based on the above criteria and choose the highest scorer. For example, like an eCommerce business, you may want to test your homepage, product page, checkout page, and thank you (rating) page. Now according to the PIE framework, you line these up and mark them potential, importance and ease:

	Potential	Importance	Ease	PIE(P+I+E/3)
Homepage	10	9	7	8.6
Product page	8	10	9	9
Checkout page	9	10	9	9.3
Thank you page	8	6	10	8

*marked out of a total of 10 points per criteria.

3. The LIFT Model

The LIFT Model is another popular conversion optimization framework that helps you analyze web and mobile experiences, and develop good A/B test hypotheses. It draws on the 6 conversion factors to evaluate experiences from the perspective of your page visitor: Value Proposition, Clarity, Relevance, Distraction, Urgency, and Anxiety.

With prioritization, you can have your A/B testing calendar ready for execution for at least 6 to 12 months. This will not only give you time, and a heads-up to prepare for the test but also plan around your resources.

Stage 3: A/B test

The third and most crucial stage is the testing stage. After the prioritization stage, you will have all the required data and a prioritized backlog. Once you have formulated hypotheses that align to your goal and prioritized them, create variations, and flag off the test. While your test is running, make sure it meets every requirement to produce statistically significant results before closure, like testing on accurate traffic, not testing too many elements together, testing for the correct amount of duration, and so on.

Stage 4: Repeat

This stage is all about learning from your past and current test and applying them in future tests. Once your test runs for the stipulated amount of time, stop the test and start analyzing the data thus gathered. The first thing you will realize is one of the many versions that were being tested had performed better than all others and won. It's time for you and your team to now figure out why that happened. There can be 3 outcomes of your test:

- Your variation or one of your variations will have won with statistical significance.
- Your control was the better version and won over the variation/s.
- Your test failed and produced insignificant results. Determine the significance of your test results with the help of tools like the A/B test significance calculator.

In the first two scenarios, do not stop testing just because you have a winner. Make improvements to that version and keep testing. In the third scenario, recall all the steps and identify where you went wrong in the process and re-do the test after rectifying the mistake.

Here is a downloadable A/B testing calendar sample for your reference. To use this spreadsheet, click on the 'File' option in the main menu and then click on 'Make a copy.'

File > Make a copy

			Week											
URL	Element to be tested	Variation to be tested	Jan-Week 1	Jan-Week 2	Jan-Week 3	Jan-Week 4	Feb-Week 1	Feb-Week 2	Feb-Week 3	Feb-Week 4	Mar-Week 1	Mar-Week 2	Mar-Week 3	Mar-Week 4
vwo.com/ab-testing	CTA	Change CTA color from Green to Red	Test Concluded											
vwo.com/product	Pop-up	Implement exit-intent popup				Test Concluded								
vwo.com/plans	Headline	New headline: Find a Plan That Best Suits Your Needs					Test Concluded							
vwo.com/case-studies	Headline	New headline: Find a Case Study That Works for You							Test Concluded					
vwo.com	Layout	New Layout: Move testimonial section just below the masthead banner								Test Underway				
vwo.com/resources/webinars/scaling-testing-program/	CTA	New CTA text: Watch Now									Test Underway			
vwo.com/request-demo	Layout	New Layout: Add video testimonial above the fold										Test Yet to Begin		
vwo.com/blog	Headline	New headline: Join 50,000 other marketers who read VWO blog											Test Yet to Begin	

When scaling your A/B testing program, keep in mind the following points:

A. Revisiting previously concluded test:

With a prioritized calendar in place, your optimization team will have a clear vision of what they will test next and which test needs to be run when. Once you have tested each element or most elements in the backlog, revisit each successful as well as failed campaigns. Analyze the test results and determine whether there is enough data to justify running another version of the test. If there is, then run the test again – with necessary edits and modifications.

B. Increasing testing frequency:

While you should always be cautious of testing too many elements together, increasing your testing frequency is essential in scaling your testing program. Your optimization team will have to plan it in such a way that none of the tests affect others or your website's performance. One way to do this is by running tests simultaneously on different web pages of your website or testing elements of the same web page at different time periods. This will not only increase your testing frequency but also, none of the tests will affect others. For instance, you can simultaneously test one element each of your homepage, checkout page, and sign-up page at one time and other elements (1 element at a time) of these pages after the current test concludes.

C. Spacing out your test:

This flows from the previous point. If you look at the calendar above, you will see that not more than two tests overlap each other at any given week. In a quest to increase your testing frequency, do not compromise with your website's overall conversion rate. If you have two or more critical elements to be tested on the same web page, space the two out. As pointed earlier, testing too many elements of a web page together makes it difficult to pinpoint which element influenced the success or failure of the test most.

Let's say, for example, you want to test one of your ad's landing pages. You lock in on testing the CTA to increase sign-ups and banners to decrease the bounce rate and increase time spent. For the CTA, based on your data, you decide to change the copy. For the banner, you decide to test a video against a static image. You deploy both tests at the same time, and at the conclusion, both your goals were met. The problem here is that data showed that while sign-ups did increase from the new CTA, the

video (apart from reducing the bounce rate and increasing average time spent on the page) too helped in this. Most of the people who watched the video also ended up signing up.

The problem now is that, because you did not space the two tests, it became impossible to tell which element contributed most to the sign-up increase. Had you timed the two tests better, much more significant insights could have been gathered?

D. Tracking multiple metrics:

You usually measure an A/B test's performance based on a single conversion goal and put all your trust on that goal to help you find the winning variation. But sometimes, the winning variation affects other website goals as well. The example above is applicable here too. The video, in addition to reducing bounce rate and increasing time spent, also contributed to increased sign-ups. To scale your A/B testing program, track multiple metrics so that you can draw more benefits with less effort.

Having a thoroughly built calendar helps to streamline things to a great extent. VWO has an inbuilt calendar-making feature known as the Kanban board that helps track your tests' progress at various stages.

What are the mistakes to avoid while A/B testing?

A/B testing is one of the most effective ways to move business metrics in a positive direction and increase the inward flow of revenue. However, as stated above, A/B testing demands planning, patience, and precision. Making silly mistakes can cost your business time and money, which you can't afford. To help you avoid making blunders, here's a list of some of the most common mistakes to remember when running an A/B test:

Mistake #1: Not planning your optimization Roadmap

A. Invalid hypothesis:

In A/B testing, a hypothesis is formulated before conducting a test. All the next steps depend on it: what should be changed, why should it be changed, what the expected outcome is, and so on. If you start with the wrong hypothesis, the probability of the test succeeding decreases.

B. Taking others' word for it:

Sure, someone else changed their sign-up flow and saw a 30% uplift in conversions. But it is their test result, based on their traffic, their hypothesis, and their goals. Here's why you should not implement someone else's test results as is onto your website: no two websites are the same – what worked for them might not work for you. Their traffic will be different; their target audience might be different; their optimization method may have been different than yours, and so on.

Mistake #2: Testing too many elements together

Industry experts caution against running too many tests at the same time. Testing too many elements of a website together makes it difficult to pinpoint which element influenced the test's success or failure the most. The more the elements tested, the more needs to be the traffic on that page to justify statistically significant testing. Thus, prioritization of tests is indispensable for successful A/B testing.

Mistake #3: Ignoring statistical significance

If gut feelings or personal opinions find a way into hypothesis formulation or while you are setting the A/B test goals, it is most likely to fail. Irrespective of everything, whether the test succeeds or fails, you must let it run through its entire course so that it reaches its statistical significance.

For a reason, that test results, no matter good or bad, will give you valuable insights and help you plan your upcoming test in a better manner.

You can get more information about the different types of errors while dealing with the maths of A/B testing.

Mistake #4: Using unbalanced traffic

Businesses often end up testing unbalanced traffic. A/B testing should be done with the appropriate traffic to get significant results. Using lower or higher traffic than required for testing increases the chances of your campaign failing or generating inconclusive results.

Mistake #5: Testing for incorrect duration

Based on your traffic and goals, run A/B tests for a certain length of time to achieve statistical significance. Running a test for too long or too short a period can result in the test failing or producing insignificant results. Because one version of your website appears to be winning within the first few days of starting the test does not mean that you should call it off before time and declare a winner. Letting a campaign run for too long is also a common blunder that businesses commit. The duration for which you need to run your test depends on various factors like existing traffic, existing conversion rate, expected improvement, etc.

Learn how long you should run your test.

Mistake #6: Failing to follow an iterative process

A/B testing is an iterative process, with each test building upon the results of the previous tests. Businesses give up on A/B testing after their first test fails. But to improve the chances of your next test succeeding, you should draw insights from your last tests while planning and deploying your next test. This increases the probability of your test succeeding with statistically significant results.

Additionally, do not stop testing after a successful one. Test each element repetitively to produce the most optimized version of it even if they are a product of a successful campaign.

Mistake #7: Failing to consider external factors

Tests should be run in comparable periods to produce meaningful results. It is wrong to compare website traffic on the days when it gets the highest traffic to the days when it witnesses the lowest traffic because of external factors such as sales, holidays, and so on. Because the comparison here is not made between likes, the chances of reaching an insignificant conclusion increase. Use VWO's A/B Test Significance Calculator to know if the results your test achieved were significant or not.

Mistake #8: Using the wrong tools

With A/B testing gaining popularity, multiple low-cost tools have also come up. Not all of these tools are equally good. Some tools drastically slow down your site, while others are not closely integrated with necessary qualitative tools (heatmaps, session recordings, and so on), leading to data deterioration. A/B testing with such faulty tools can risk your test's success from the start.

Mistake #9: Sticking to plain vanilla A/B testing method

While most experience optimizers recommend that you must start your experimentation journey by running small A/B tests on your website to get the hang of the entire process. But, in the long run, sticking to plain vanilla A/B testing methods won't work wonders for your organization. For instance, if you are planning to revamp one of your website's pages entirely, you ought to make use of split testing. Meanwhile, if you wish to test a series of permutations of CTA buttons, their color, the text, and the image of your page's banner, you must use multivariate testing.

What are the challenges of A/B testing?

The ROI from A/B testing can be huge and positive. It helps you direct your marketing efforts to the most valuable elements by pinpointing exact problem areas. But every once in a while, as an experience optimizer, you may face some challenges when deciding to undertake A/B testing. The 6 primary challenges are as follows:

Challenge #1: Deciding what to test

You can't just wake up one day and decide to test certain elements of your choice. A bitter reality that experience optimizers are now coming to realize is that not all small changes that are easy to implement are always the best when you consider your business goals and often fail to prove significant. The same goes for complex tests. This is where website data and visitor analysis data come into play. These data points help you overcome the challenge of 'not knowing what to test' out of your unending backlog by generally pointing to the elements which may have the most impact on your conversion rates or by directing you to pages with the highest traffic.

Challenge #2: Formulating hypotheses

In great resonance with the first challenge is the second challenge: formulating a hypothesis. This is where the importance of having scientific data at your disposal comes in handy. If you are testing without proper data, you might as well be gambling away your business. With the help of data gathered in the first step (i.e., research) of A/B testing, you need to discover where the problems lie with your site and come up with a hypothesis. This will not be possible unless you follow a well-structured and planned A/B testing program.

Challenge #3: Locking in on sample size

Not many experience optimizers are statisticians. We often make the mistake of calling conclusive results too quickly because we are more often than not after quick results. As experience optimizers, we need to learn about sample sizes, in particular, how large should our testing sample size be based on our web page's traffic.

Challenge #4: Analyzing test results

With A/B testing, you will witness success and failure at each step. This challenge, however, is pertinent to both successful and failed tests:

1. Successful campaigns:

It's great that you ran two tests, and both of them were successful in producing statistically significant results. What next? Yes, deploying the winner, but what after that? What experience optimizers often fail to do or find difficult is interpreting test results. Interpreting test results after they conclude is extremely important to understand why the test succeeded. A fundamental question to be asked is – why? Why did customers behave the way they did? Why did they react a certain way with one version and not with the other versions? What visitor insights did you gather,

and how can you use them? Many experience optimizers often struggle or fail to answer these questions, which not only help you make sense of the current test but also provide inputs for future tests.

2. Failed campaigns:

Sometimes, experience optimizers don't even look back at failed tests. They either have a hard time dealing with them, for example, while telling the team about the failed tests or have no clue what to do with them. No failed test is unsuccessful unless you fail to draw learnings from them. Failed campaigns should be treated like pillars that would ultimately lead you to success. The data gathered during the entire A/B testing process, even if in the end, the test failed, is like an unopened pandora box. It contains a plethora of valuable data and insights that can give you a head start for your next test.

Additionally, with the lack of proper knowledge on how to analyze the gathered data, the chances of data corruption increase manifold. For example: without having a process in place, there will be no end to scrolling through heatmaps data or sessions recording data. Meanwhile, if you are using different tools for these, then the chances of data leakage while attempting to integrate them also increase. You may also fail to draw any significant insights while wandering directionless through data and just drown under them.

Challenge #5: Maintaining a testing culture

One of the most crucial characteristics of optimization programs like CRO and A/B testing is that it is an iterative process. This is also one of the major obstacles that businesses and experience optimizers face. For your optimization efforts to be fruitful in the long run, they should form a cycle that roughly starts with research and ends in research.

This challenge is not just a matter of putting in effort or about having the required knowledge. Sometimes due to resource crunch, businesses rarely or intermittently use A/B testing and fail to develop a proper testing culture.

Challenge #6: Changing experiment settings in the middle of an A/B test

When you launch an experiment, you must commit to it completely. Try and not to change your experiment settings, edit or omit your test goals, or play with the design of the control or the variation while the test is running. Moreso, do not try and change the traffic allocations to variations as well because doing so will not only alter the sampling size of your returning visitors but massively skew your test results as well.

So, given all these challenges, is A/B testing worth undertaking?

From all the evidence and data available on A/B testing, even after these challenges, A/B testing generates great ROI. From a marketing perspective, A/B testing takes the guesswork out of the optimization process. Strategic marketing decisions become data-driven, making it easier to craft an ideal marketing strategy for a website with well-defined ends. Without an A/B testing program, your marketing team will simply test elements at random or based on gut feelings and preferences. Such data-less testing is bound to fail.

If you start strong with a good website and visitor data analysis, the first three challenges can easily be solved. With the extensive website and visitor data at your disposal, you can prioritize your backlog, and you won't even have to decide on what to test. The data will do all the talking. With such quality data coupled with your business expertise, formulating a working hypothesis becomes

just a matter of going through the available data and deciding what changes will be best for your end goal. To overcome the third challenge, you can calculate the apt sample size for your testing campaign with the help of many tools available today.

The last two challenges are related to how you approach A/B testing. If you treat A/B testing like an iterative process, half of the fourth challenge may not even be on your plate. And the other half can be solved by hiring experts in the field or by getting trained on how to analyze research data and results correctly. The right approach to tackle the last challenge is to channel your resources on the most business-critical elements and plan your testing program in a way that, with the limited resource, you can build a testing culture.

A/B testing and SEO

As far as implications of SEO on A/B testing are concerned, Google has cleared the air on their blog post titled "Website Testing And Google Search. The important bits from that post are summarized below:

No cloaking

Cloaking – showing one set of content to humans, and a different set to Googlebot – is against our Webmaster Guidelines, whether you're running a test or not. Make sure that you're not deciding whether to serve the test or which content variant to serve, based on user-agent. An example of this would always be serving the original content when you see the user-agent "Googlebot." Remember that infringing our Guidelines can get your site demoted or even removed from Google search results – probably not the desired outcome of your test.

Only use 302 redirects

If you're running an A/B test that redirects users from the original URL to a variation URL, use a 302 (temporary) redirect, not a 301 (permanent) redirect. This tells the search engines that this redirect is temporary – it will only be in place as long as you're running the experiment – and that they should keep the original URL in their index rather than replacing it with the target of the redirect (the test page). JavaScript-based redirects also got a green light from Google.

Run experiments for the appropriate duration

The amount of time required for a reliable test will vary depending on factors like your conversion rates, and how much traffic your website gets. A good testing tool should tell you when you've gathered enough data to be able to draw reliable conclusions. Once you have concluded the test, you should update your site with the desired variation(s) and remove all elements of the test as soon as possible, such as alternate URLs or testing scripts and markup.

Use rel="canonical" links

Google suggests using rel="canonical" link attribute on all alternate URLs for you to be able to highlight that the original URL is actually the preferred one. This suggestion stems from the fact that rel="canonical" more closely matches your intent in this situation when compared to other methods like no index meta tag. For instance, if you are testing variations of your product page, you don't want search engines not to index your product page. You just want them to understand that all the test URLs are close duplicates or variations on the original URL and should be grouped together, with the original URL as the hero. Sometimes, in these instances, using no index rather than rel="canonical" in such a situation can sometimes have unexpected bad effects.

A/B testing examples

A/B testing in Media & Publishing Industry

Some goals of a media and publishing business may be to increase readership and audience, to increase subscriptions, to increase time spent on their website by visitors, or to boost video views and other content pieces with social sharing and so on. You may try testing variations of email sign-up modals, recommended content, social sharing buttons, highlighting subscription offers, and other promotional options.

Any of us who is a Netflix user can vouch for their streaming experience. But not everyone knows how they manage to make it so good. Here's how – Netflix follows a structured and rigorous A/B testing program to deliver what other businesses struggle to deliver even today despite many efforts – a great user experience. Every change that Netflix makes to its website goes through an intense A/B testing process before getting deployed. One example to show how they do it is the use of personalization.

Netflix uses personalization extensively for its homepage. Based on each user's profile, Netflix personalizes the homepage to provide the best user experience to each user. They decide how many rows go on the homepage and which shows/movies go into the rows based on the users streaming history and preferences.

They follow the same exercise with media title pages as well. Within these pages, Netflix personalizes what titles are we most likely to watch, the thumbnails we see on them, what title text entices us to click, or if social proof helps make our decision easier, and so on. And this is just the tip of the iceberg.

A/B Testing in eCommerce Industry

Through A/B testing, online stores can increase the average order value, optimize their checkout funnel, reduce cart abandonment rate, and so on. You may try testing: the way shipping cost is displayed and where, if, and how the free shipping feature is highlighted, text and color tweaks on the payment page or checkout page, the visibility of reviews or ratings, etc.

In the eCommerce industry, Amazon is at the forefront in conversion optimization partly due to the scale they operate at and partly due to their immense dedication to providing the best customer experience. Amongst the many revolutionary practices they brought to the eCommerce industry, the most prolific one has been their '1-Click Ordering'. Introduced in the late 1990s after much testing and analysis, 1-Click Ordering lets users make purchases without having to use the shopping cart at all.

Once users enter their default billing card details and shipping address, all they need to do is click on the button and wait for the ordered products to get delivered. Users don't have to enter their billing and shipping details again while placing any orders. With the 1-Click Ordering, it became impossible for users to ignore the ease of purchase and go to another store. This change had such a huge business impact that Amazon got it patented (now expired) in 1999. In fact, in 2000, even Apple bought a license for the same to be used in their online store.

People working to optimize Amazon's website do not have sudden 'Eureka' moments for every change they make. It is through continuous and structured A/B testing that Amazon is able to deliver the kind of user experience that it does. Every change on the website is first tested on their audience and then deployed. If you were to notice Amazon's purchase funnel, you would realize that even

though the funnel more or less replicates other websites' purchase funnels, each and every element in it is fully optimized, and matches the audience's expectations.

Every page, starting from the homepage to the payment page, only contains the essential details and leads to the exact next step required to push the users further into the conversion funnel. Additionally, using extensive user insights and website data, each step is simplified to their maximum possible potential to match their users' expectations.

Take their omnipresent shopping cart, for example.

There is a small cart icon at the top right of Amazon's homepage that stays visible no matter which page of the website you are on.

The icon is not just a shortcut to the cart or reminder for added products. In its current version, it offers 5 options:

- Continue shopping (if there are no products added to the cart)
- Learn about today's deals (if there are no products added to the cart)
- Wish List (if there are no products added to the cart
- Proceed to checkout (when there are products in the cart)
- Sign in to turn on 1-Click Checkout (when there are products in the cart)

With one click on the tiny icon offering so many options, the user's cognitive load is reduced, and they have a great user experience. As can be seen in the above screenshot, the same cart page also suggests similar products so that customers can navigate back into the website and continue shopping. All this is achieved with one weapon: A/B Testing.

A/B Testing in Travel Industry

Increase the number of successful bookings on your website or mobile app, your revenue from ancillary purchases, and much more through A/B testing. You may try testing your home page search modals, search results page, ancillary product presentation, your checkout progress bar, and so on.

In the travel industry, Booking.com easily surpasses all other eCommerce businesses when it comes to using A/B testing for their optimization needs. They test like it's nobody's business. From the day of its inception, Booking.com has treated A/B testing as the treadmill that introduces a flywheel effect for revenue. The scale at which Booking.com A/B tests is unmatched, especially when it comes to testing their copy. While you are reading this, there are nearly 1000 A/B tests running on Booking.com's website.

Even though Booking.com has been A/B testing for more than a decade now, they still think there is more that they can do to improve user experience. And this is what makes Booking.com the ace in the game. Since the company started, Booking.com incorporated A/B testing into its everyday work process. They have increased their testing velocity to its current rate by eliminating HiPPOs and giving priority to data before anything else. And to increase the testing velocity, even more, all of Booking.com's employees were allowed to run tests on ideas they thought could help grow the business.

This example will demonstrate the lengths to which Booking.com can go to optimize their users' interaction with the website. Booking.com decided to broaden its reach in 2017 by offering rental

properties for vacations alongside hotels. This led to Booking.com partnering with Outbrain, a native advertising platform, to help grow their global property owner registration.

Within the first few days of the launch, the team at Booking.com realized that even though a lot of property owners completed the first sign-up step, they got stuck in the next steps. At this time, pages built for the paid search of their native campaigns were used for the sign-up process.

Both the teams decided to work together and created three versions of landing page copy for Booking.com. Additional details like social proof, awards, and recognitions, user rewards, etc. were added to the variations.

The test ran for two weeks and produced a 25% uplift in owner registration. The test results also showed a significant decrease in the cost of each registration.

A/B Testing in B2B/SaaS Industry

Generate high-quality leads for your sales team, increase the number of free trial requests, attract your target buyers, and perform other such actions by testing and polishing important elements of your demand generation engine. To get to these goals, marketing teams put up the most relevant content on their website, send out ads to prospect buyers, conduct webinars, put up special sales, and much more. But all their effort would go to waste if the landing page which clients are directed to is not fully optimized to give the best user experience.

The aim of SaaS A/B testing is to provide the best user experience and to improve conversions. You can try testing your lead form components, free trial sign-up flow, homepage messaging, CTA text, social proof on the home page, and so on.

POSist, a leading SaaS-based restaurant management platform with more than 5,000 customers at over 100 locations across six countries, wanted to increase their demo requests.

Their website homepage and Contact Us page are the most important pages in their funnel. The team at POSist wanted to reduce drop-off on these pages. To achieve this, the team created two variations of the homepage as well as two variations of the Contact Us page to be tested. Let's take a look at the changes made to the homepage. This is what the control looked like:

The team at POSist hypothesized that adding more relevant and conversion-focused content to the website will improve user experience, as well as generate higher conversions. So they created two variations to be tested against the control. This is what the variations looked like:

Control was first tested against Variation 1, and the winner was Variation 1. To further improve the page, variation one was then tested against variation two, and the winner was variation 2. The new variation increased page visits by about 5%.

Conclusion

After reading this comprehensive piece on A/B testing, you should now be fully equipped to plan your own optimization roadmap. Follow each step involved diligently and be wary of all major and minor mistakes that you can commit if you do not give data the importance it deserves. A/B testing is invaluable when it comes to improving your website's conversion rates.

If done with complete dedication, and with the knowledge you now have, A/B testing can reduce a lot of risks involved when undertaking an optimization program. It will also help you significantly improve your website's UX by eliminating all weak links and finding the most optimized version of your website.

Source: https://www.dynamicyield.com/lesson/bayesian-testing/

Frequentist vs. Bayesian approach in A/B testing

The industry is moving toward the Bayesian framework as it is a simpler, less restrictive, more reliable, and more intuitive approach to A/B testing.

Idan Michaeli

Data Science and Predictive Modeling Expert

Summarize this articleHere's what you need to know:

- A/B testing is shifting from frequentist to Bayesian methods due to its simplicity, flexibility, reliability, and intuitiveness.
- The key difference lies in interpreting probability: frequentists view it as event likelihood based on repeated trials, while Bayesians see it as a belief measure updated with new data and prior knowledge.
- Frequentist A/B testing relies on hypothesis testing with a predetermined sample size and p-values to determine statistical significance.
- Bayesian A/B testing continuously updates beliefs about conversion rates based on collected data, enabling more flexible and data-driven decisions.
- Advantages of Bayesian testing include no need for a fixed sample size, the ability to peek at data, and intuitive results like the probability of one variation outperforming another.
- Despite its complexity, Bayesian A/B testing is becoming more accessible thanks to user-friendly statistical engines.

The days where marketers made changes to their websites based on gut-feelings alone are far behind us. We are now deep in the A/B testing era, basing our decisions on as much empirical data as possible.

To enable that, the community has looked for A/B testing tools to help us make informed decisions based on collected data. Or, in more exact terms: to soundly generalize from observed data and gain insight into the future.

The aim of this post is to discuss the evolution these tools are currently undergoing, from the basic "frequentist" testing method used in the past (and still commonly used today) to the new Bayesian testing method which the industry is moving toward.

Underlying these two approaches is a different viewpoint on what probability is. I will attempt to highlight that difference and the practical implications it has, but without delving into the hard-core math.

What is Hypothesis Testing?

At the dawn of the A/B testing era, statisticians provided a very basic framework for statistical inference in an A/B testing scenario. Commonly known as "Hypothesis Testing," the procedure goes as follows:

1. Start with the existing version of the web page or the tested element within it. That existing version is now termed the "baseline" (or variation A).
2. Set up the alternative variation, a.k.a the "treatment" (or variation B).
3. Calculate the required sample size in advance using a calculator such as this one. This calculation is based on the baseline's current conversion rate (which must be already known), the minimum difference in performance you wish to detect, and the desired Statistical Power (i.e. in rough terms, how reliable that detection should be, as higher reliability requires a greater sample size.)
4. Launch the test and let it run (without peeking at the data) until the required sample size per variation is reached. Seriously, no peeking allowed! (Here's why). Yes, I know. Often people do not calculate sample size in advance and cannot resist the urge to check out the data. But in truth, if you look and see significant results before reaching a proper sample size, it can cause you to jump to conclusions which can significantly degrade the reliability of the results. So no peeking!
5. Now that we have the samples, we can observe the performance of each variation and calculate whether the stronger performing variation is, in fact, better than its competitor in a statistically significant fashion. Again, a calculator such as this one may help, emitting the p-value, a.k.a "confidence." But, what does p-value actually stand for?

The P-Value Misinterpretation

People tend to think that the p-value represents the probability that variation B is really better than A. However, this is a common misinterpretation; It actually has to do with the hypothesis that is at the base of Hypothesis Testing.

What p-value tests for is not the "optimistic" case in which our alternative variation B is really better than the baseline. Instead, we start with a pessimistic hypothesis (called the "Null Hypothesis") stating that the newly introduced variation B is not any better than the existing baseline A, and that the observed differences represent no more than random noise.

We then try to reject this hypothesis by calculating how rare our empirical findings are if the above Null Hypothesis is correct. The p-value represents *that probability*.

If the p-value is below a certain threshold (often taken as 0.05), we can state that our finding allows us to reject the Null Hypothesis and thus declare variation B as the winner. In addition, this Hypothesis Testing framework provides a way to calculate a confidence interval, which is aimed at getting a sense of how confident we are that the measured values will last for the long haul.

Take a moment and notice how convoluted and unintuitive this whole framework is.

It requires a known baseline and reaching a predefined sample size before we're allowed to look at the data and draw conclusions. These conclusions are based on metrics that, to paraphrase Inigo Montoya in the "Princess Bride," do not mean what most people think they mean.

This begs the question: can we do something to make things simpler, less restrictive, more reliable and more intuitive? The answer is ***yes***.

Frequentist vs. Bayesian

In the field of statistical inference, there are two very different, yet mainstream, schools of thought: the frequentist approach, under which the framework of Hypothesis Testing was developed, and the Bayesian approach, which I'd like to introduce to you now.

The difference between these two rival schools can be explained through the different interpretation each gives to the term probability. The intro given here is adapted from this series of blog posts.

Let's take a concrete case-in-point: say we are interested in discovering the average height of American citizens nowadays. For a frequentist, this number is unknown but fixed. This is a natural intuitive view, as you can imagine that if you go through all American citizens one by one, measure their height and average the list, you will get the actual number.

However, since you do not have access to all American citizens, you take a sample of, say, a thousand citizens, measure and average their height to produce a point estimate, and then calculate the estimate of your error. The point is that the frequentist looks at the average height as a single unknown number.

A Bayesian statistician, however, would have an entirely different take on the situation. A Bayesian would look at the average height of an American citizen not as a fixed number, but instead as an unknown distribution (you might imagine here a "bell" shaped normal distribution).

Bayesian Statistics and Probability: How it Breaks Down

Initially, the Bayesian statistician has some basic prior knowledge which is being assumed: for example, that the average height is somewhere between 50cm and 250cm.

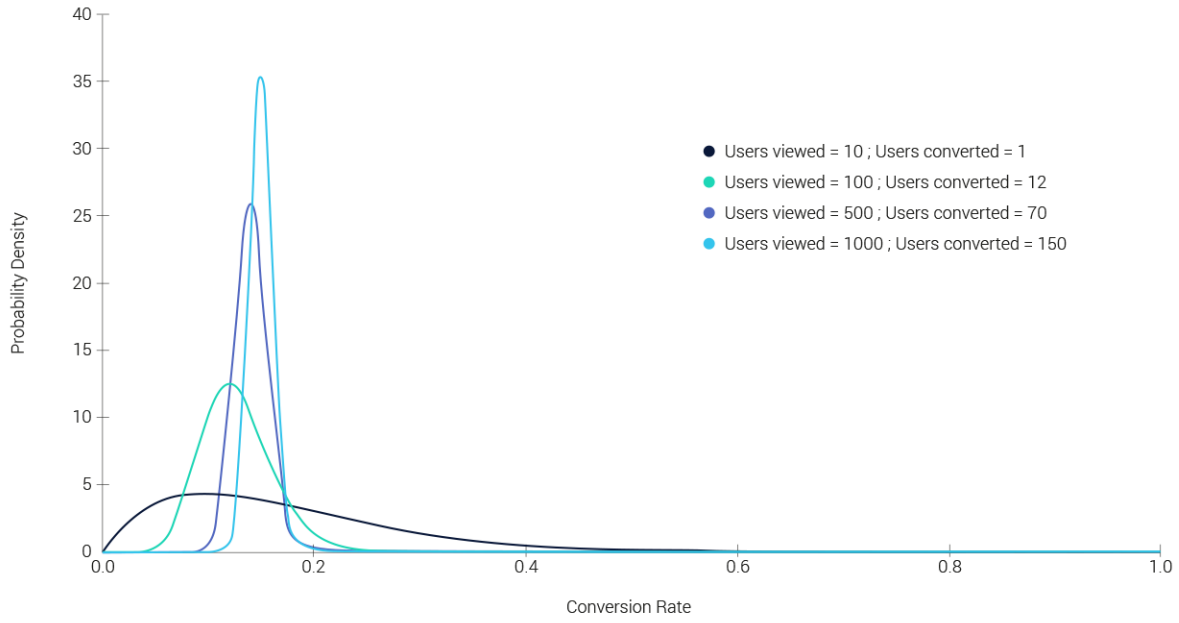
Then, the Bayesian begins to measure heights of specific American citizens, and with each measurement updates the distribution to become a bit more "bell-shaped" around the average height measured so far. As more data is collected, the "bell" becomes sharper and more concentrated around the measured average height.

For Bayesians, probabilities are fundamentally related to their knowledge about an event. This means, for example, that in a Bayesian view, we can meaningfully talk about the probability that the true conversion rate lies in a given range, and that probability codifies our knowledge of the value based on prior information and/or available data.

For Bayesians, the concept of probability is extended to cover degrees of certainty about any given statement on reality. However, in a strict frequentist view, it is meaningless to talk about the probability of the true conversion rate. For frequentists, the true conversion rate is by definition a single fixed number, and to talk about a probability distribution for a fixed number is mathematically nonsensical.

The same logic applies when seeking to measure the conversion rate of a web-based purchase funnel. Sure, probability can certainly be estimated in a frequentist fashion by measuring the ratio of how many times a conversion was made out of a huge number of trials. But this is not fundamental to the Bayesian, who can stop the test at any point and calculate probabilities from data.

To illustrate the convergence process of the distribution as more data is collected, here is a plot based on test data. Notice how the bell shape becomes sharper (more certain) as data streams in:



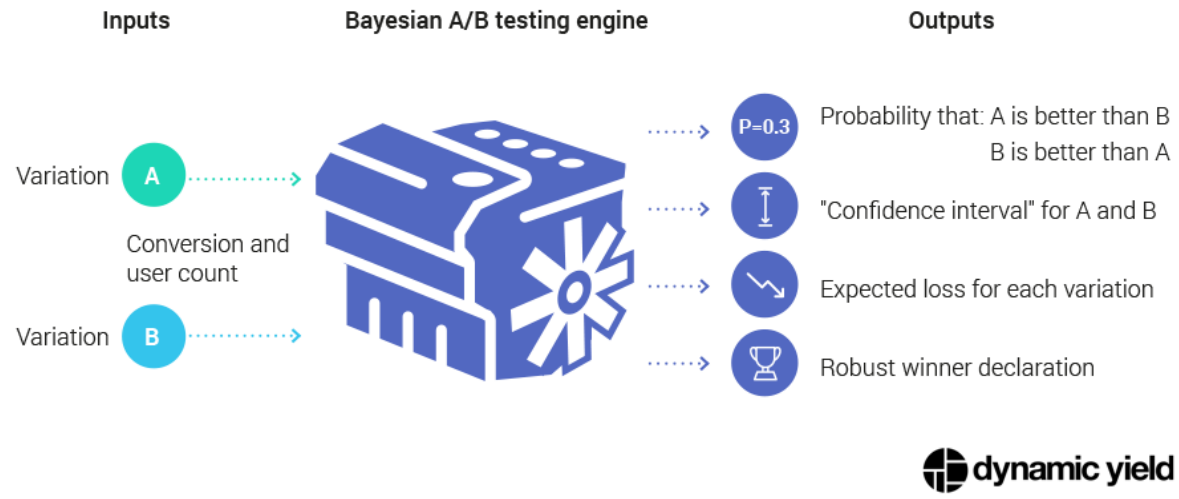
The surprising thing is that this arguably subtle difference in philosophy between these schools leads in practice to vastly different approaches to the statistical analysis of data.

A Bayesian Framework for A/B Testing

The math behind the Bayesian framework is quite complex so I will not get into it here. In fact, I would argue that the fact that the math is more complicated than can be computed with a simple calculator or Microsoft Excel is a dominant factor in the slow adoption of this method in the industry.

The framework involves some daunting terms such as Prior, Posterior, Bayes Theorem, Beta and Gamma distributions, Monte Carlo Integration, and more.

However, if we look at the Bayesian statistical engine as a “mathematical black box”, we will see that the inputs to it, and more importantly the outputs from it, are actually quite simple and intuitive. An input/output diagram of the engine would look something like this:



Although we’re not delving into specifics, you can immediately see that the Bayesian engine provides answers to more direct questions such as:

What is the probability that A is better than B? (contrast this with the contrived p-value mentioned earlier)

If I declare B as a winner, and it is not really better, how much should I expect to lose in terms of conversion rate?

Also, this engine can emit a new kind of "confidence interval" metric, mathematically termed Highest Posterior Density Region (HPDR). This provides an answer to a pretty intuitive requirement, namely: give me the conversion rate boundaries (interval) in which the true conversion rate falls with (say) 95% probability.

If we want to be exact about the meaning of the oft-used frequentist confidence interval, it means roughly the following: "if we would have repeated this test many times, and would have calculated a different confidence interval for each case, then in 95% of the times the actual conversion rate would fall within this interval." How is that for intuitiveness?

Let's summarize how the two frameworks compare:

	Hypothesis Testing	**Bayesian A/B Testing**
Knowledge of Baseline Performance	Required	Not Required
Intuitiveness	Less, as p-value is a convoluted term	More, as we directly calculate the probability of A being better than B
Sample size	Pre-defined	No need to pre-define
Peeking at the data while the test runs	Not allowed	Allowed (with caution)
Quick to make decisions	Less, as it has more restrictive assumptions on distributions	More, as it has less restrictive assumptions
Representing uncertainty	Confidence Interval (again, a convoluted interpretation which is often misunderstood)	Highest Posterior Density Region – highly intuitive interpretation
Declaring a winner	When sample size is reached and p-value is below a certain threshold	When either "probability to be best" goes above a threshold or the expected loss is below a threshold (in which case a "tie" can be declared between multiple variations)

I should note, though, that generally speaking, a frequentist A/B testing framework which performs as well as the Bayesian framework described here *is* possible, but further development would be needed above what is usually implemented.

Conclusion

To conclude, the industry is moving toward the Bayesian framework as it is a simpler, less restrictive, more reliable and more intuitive approach to A/B testing.

P.S. Here at Dynamic Yield, we have made the move to a Bayesian statistical engine not only for binary objectives such as goal conversion rate and CTR but also for non-binary objectives such as Revenue Per User (which merits a future blog post in its own right). With this new engine, our customers now benefit from a quicker and more robust statistical engine.

Source: https://medium.com/@ibtesamahmex/beginners-guide-to-bayesian-ab-testing-22f40988d5e6

Frequentist vs Bayesian

First things first, statistics gives you a way to reason under uncertainty and this uncertainty can be quantified through probability. For the two school of thoughts — their fundamental approach to uncertainty itself is different and therefore their interpretation of probability.

Frequentists look at probabilities objectively and consider them unknown but fixed. You can calculate this unknown probability by repeated trials. For example you can know whether a coin is fair by repeatedly tossing it and writing down the outcomes. At the end of your let's say 100 trials you can calculate the probability of heads and also the uncertainty associated with this estimate(confidence interval).

Bayesians look at the world more subjectively and interpret probability as a measure of belief. They also believe(pun intended ;-) in updating their prior beliefs when encountering new information. In the event of some observed data Bayesians are more flexible and let you incorporate prior knowledge in probability assessments, whereas a frequentist will just look at the observed data and want to know the "truth"(fixed but unknown probability). Probability for Bayesians is a random variable.

Just for fun, picture frequentists as your strict parents who always want to know the truth and only the truth, and picture Bayesians like that cool aunt who is ready to update her beliefs and change. To sum it up, Bayesians are way more flexible.

Now, coming back from this philosophical discussion to AB testing, let me list down why you would want to prefer Bayesian AB testing over the vanilla(frequentist) AB testing

- *you can incorporate prior information such as domain knowledge.*

- *you need less data to come to a conclusion.*

- *It can give you a more useful result in the end than just a yes/no answer by giving the probability of superiority and expected loss.*

Now, if you are convinced let's just dive right into the steps to do Bayesian AB testing along with a basic way to implement it in Python.

Steps to implement Bayesian AB testing

1. Define the hypothesis

Clearly state the null hypothesis (e.g., the two versions of a website are equal in terms of conversion rate) and the alternative hypothesis (e.g., one version has a higher conversion rate than the other).

2. Collect data

Randomly assign visitors to the different versions of your website (A and B) and collect relevant data, such as the number of visitors and the number of conversions for each version. Until now the process is very similar to a Frequentist AB test.

3. Choose a prior distribution

Select a prior distribution that reflects your beliefs about the conversion rates of the two versions before seeing any data. The selection of the prior distribution depends on your likelihood distribution(distribution of observed data). If we're flipping a coin, the binomial distribution shows what *n* number of coin flips would look like with probability *p* of being heads.

However, often *p* itself has a distribution. The distribution of *p* is the **conjugate prior distribution**. As you can see, conjugate priors are the source of our data's likelihood distribution.

Since this is one of the most important steps in Bayesian AB testing, we'll spend some more time talking about how to choose a prior distribution according to our prior beliefs.

1. **Non-informative prior**: If you have no prior knowledge or strong beliefs about the conversion rates, you can use a non-informative prior, such as a uniform distribution. These priors are designed to be non-committal and let the data speak for themselves. *Caution:* in the absence of a prior your Bayesian AB test is just like a frequentist AB test. You can also use beta distribution and take parameters from the control group, basically asserting the prior belief that there is no difference between test and control.
2. **Informative prior**: If you have some prior knowledge about the conversion rates based on previous experiments, historical data, or domain expertise, you can use an informative prior. This could be a beta distribution with parameters based on your prior knowledge. You could also use the conversion rates observed in a previous A/B test as the parameters of a beta distribution for your prior.
3. **Hierarchical prior**: In some cases, you may want to use a hierarchical prior, where the parameters of the prior distribution are themselves drawn from a distribution. This can be useful when you have different levels of prior information for different segments of your population. If you are just getting started you can ignore this and just use the same prior for all segments of your data.

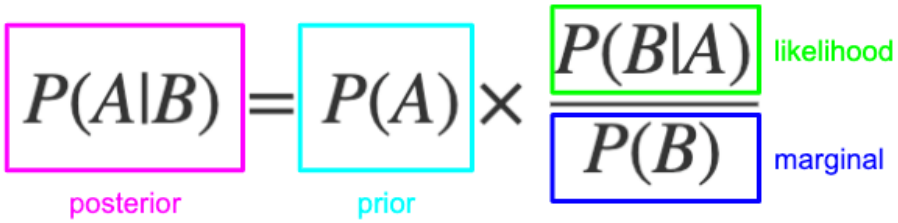
The choice of the prior distribution also depends on the type of your data. You can reference the below table for it.

Type of variable	Likelihood distribution	Conjugate Prior (dist)
Binary(yes/no)	Binomial distribution	Beta distribution
Multiple categories	Multinomial distribution	Dirichlet distribution
Continuous variable	Exponential distribution	Gamma distribution

Now that we have the priors lets move to the next step.

4. Update the prior with data aka calculate the posterior

Let's revisit the Bayes theorem once to see how we can calculate the posterior or update our prior with new data.



Source

In this equation we have the prior and the likelihood(observed data) but not the marginal probability and it's usually unfeasible to compute. Fortunately, there's a way out. Instead of computing the posterior probability we can sample from it. You can do this using a method like Markov Chain Monte Carlo (MCMC) or by drawing random samples from the Beta distributions. We won't go into the details of MCMC here but let's develop an intuition on how to do this.

Let's continue with the example of two versions of a webpage A and B . A is the original version and in B we have changed the colour of a button. We want to do an AB test and see whether B is better than A. We don't have much prior information about the conversion rates of A and B, therefore we'll take a non-informative prior such as a uniform distribution or a Beta distribution with alpha and beta as 1(equivalent to uniform distribution).

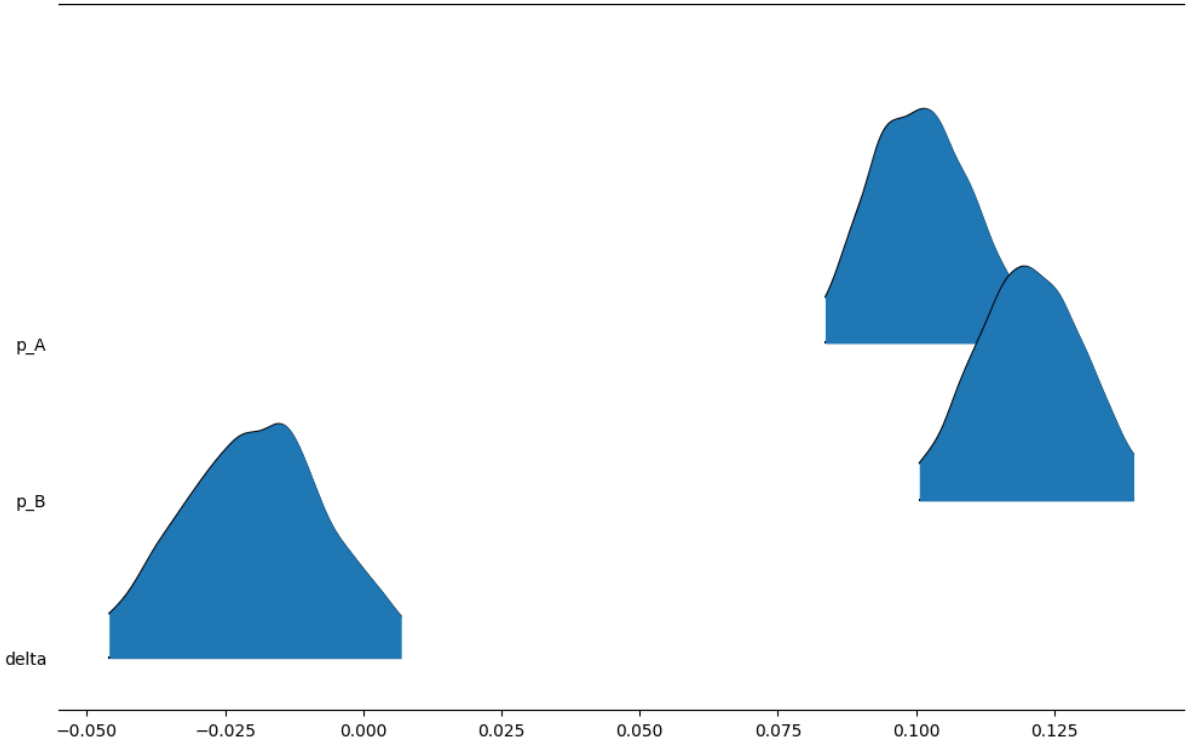
If 1000 users each are shown designs A and B, let's say 100 users convert(click on the button) in version A while 120 users convert in version B. This observed data can be modelled as a Binomial distribution. Let's try to code this using PyMC3.

```
import pymc as pm

# Observed data
n_A ,n_B = 1000, 1000
obs_A = 100
obs_B = 120

with pm.Model() as model:
    # Prior distributions for the probabilities p_A and p_B
    p_A = pm.Beta('p_A', alpha=1, beta=1)
    p_B = pm.Beta('p_B', alpha=1, beta=1)

    # Deterministic delta function to calculate the difference in p_A and p_B
    delta = pm.Deterministic('delta', p_A - p_B)

    # Observed data is modeled as a Binomial distribution
    obs_A = pm.Binomial('obs_A', n=n_A, p=p_A, observed=obs_A)
    obs_B = pm.Binomial('obs_B', n=n_B, p=p_B, observed=obs_B)

    # Perform Markov Chain Monte Carlo sampling
    trace = pm.sample(draws=1000)
```

Given the prior and likelihood distributions we use pm.sample() to estimate the posterior distributions of the click-through rates and the difference between them. It uses MCMC to sample from the posterior distribution. The basic intuition behind MCMC is that it produces random walks through the posterior, visiting regions with a large posterior proportionally more often. That's how it draws samples from the posterior distribution.

Now, let's visualise the posterior distribution of A, B and the delta between them.

pm.plot_forest(trace, kind='ridgeplot',var_names=['p_A', 'p_B', 'delta'],combined=True)



- The plot displays the density of data along a single axis, with the x-axis representing the variable being analyzed(conversion rate here) and the y-axis representing the density of data at each value of the variable.
- The shape and spread of p_A and p_B can provide us with insights into the uncertainty and differences in the conversion rates between the two variants. p_A is centered around 0.1 as in the data and p_B is centered around 0.12
- The distribution of delta will directly show the difference in conversion rates between the two variants. Since, delta is p_A- p_B negative value of delta indicates that variant B has a higher conversion rate than variant A, while a positive value indicates the opposite
- A larger absolute value of delta implies a more substantial difference between the two variants. In our case since most of the delta is below 0 (0 lies outside the 95% confidence interval) we can be reasonably confident that Design B has a higher conversion rate.

5. Calculate the probability of Superiority

- For each iteration i.e. draw of the MCMC sampler, calculate the difference between the parameters of interest (e.g., conversion rates) for the two variants.

- Count the number of iterations where one variant has a higher value than the other. Divide the count by the total number of iterations to obtain the probability of superiority.
- A probability of superiority close to 1 indicates that one variant is highly likely to have a higher value than the other.
- A probability close to 0.5 suggests that there is no clear superiority between the two variants.

```
import numpy as np
np.mean(trace['posterior']['p_B'] > trace['posterior']['p_A'])
```

xarray.DataArray

```
array(0.9175)
```

For our case there is a 91% probability that version B is better than version A. This information is quite useful for us when taking a decision whether to roll out version B or not.

Conclusion

We started this article by talking about Bayesian mindset, contrasting it with frequentist mindset. Bayesian A/B testing incorporates prior knowledge and flexibility in updating beliefs based on new information, allowing for more nuanced results beyond a simple yes/no outcome.

This prior knowledge also lets us work with less data along with the ability to quantify uncertainty using probability of superiority rather than the rigid mindset of frequentist AB testing.

I aimed to give a step-by-step guide to Bayesian A/B testing that someone getting started can easily follow, including how to sample from the posterior, interpret the results and calculate the probability of superiority.

Source: https://easystats.github.io/bayestestR/articles/region_of_practical_equivalence.html

What is the *ROPE?*

Unlike a frequentist approach, Bayesian inference is not based on statistical significance, where effects are tested against "zero". Indeed, the Bayesian framework offers a probabilistic view of the parameters, allowing assessment of the uncertainty related to them. Thus, rather than concluding that an effect is present when it simply differs from zero, we would conclude that the probability of being outside a specific range that can be considered as **"practically no effect"** (*i.e.*, a negligible magnitude) is sufficient. This range is called the **region of practical equivalence (ROPE)**.

Indeed, statistically, the probability of a posterior distribution being different from 0 does not make much sense (the probability of it being different from a single point being infinite). Therefore, the idea underlining ROPE is to let the user define an area around the null value enclosing values that are **equivalent to the null** value for practical purposes (J. Kruschke, 2014; J. K. Kruschke, 2010; J. K. Kruschke, Aguinis, & Joo, 2012).

Equivalence Test

The ROPE, being a region corresponding to a "null" hypothesis, is used for the **equivalence test**, to test whether a parameter is **significant** (in the sense of *important* enough to be cared about). This test is usually based on the **"HDI+ROPE decision rule"** (J. Kruschke, 2014; J. K. Kruschke & Liddell, 2018) to check whether parameter values should be accepted or rejected against an explicitly formulated "null hypothesis" (*i.e.*, a ROPE). In other words, it checks the percentage of Credible Interval (CI) that is the null region (the ROPE). If this percentage is sufficiently low, the null hypothesis is rejected. If this percentage is sufficiently high, the null hypothesis is accepted.

Credible interval in ROPE *vs* full posterior in ROPE

Using the ROPE and the HDI as Credible Interval, Kruschke (2018) suggests using the percentage of the 95% HDI that falls within the ROPE as a decision rule. However, as the 89% HDI is considered a better choice (J. Kruschke, 2014; R. McElreath, 2014; Richard McElreath, 2018), bayestestR provides by default the percentage of the 89% HDI that falls within the ROPE.

However, *simulation studies data* suggest that using the percentage of the full posterior distribution, instead of a CI, might be more sensitive (especially do delineate highly significant effects). Thus, we recommend that the user considers using the **full ROPE** percentage (by setting ci = 1), which will return the portion of the entire posterior distribution in the ROPE.

What percentage in ROPE to accept or to reject?

If the HDI is completely outside the ROPE, the "null hypothesis" for this parameter is "rejected". If the ROPE completely covers the HDI, *i.e.*, all most credible values of a parameter are inside the region of practical equivalence, the null hypothesis is accepted. Else, it's unclear whether the null hypothesis should be accepted or rejected.

If the **full ROPE** is used (*i.e.*, 100% of the HDI), then the null hypothesis is rejected or accepted if the percentage of the posterior within the ROPE is smaller than to 2.5% or greater than 97.5%. Desirable results are low proportions inside the ROPE (the closer to zero the better).

How to define the ROPE range?

Kruschke (2018) suggests that the ROPE could be set, by default, to a range from -0.1 to 0.1 of a standardized parameter (negligible effect size according to Cohen, 1988).

- For **linear models (lm)**, this can be generalised to:[−0.1∗SDy,0.1∗SDy][−0.1∗*SDy*,0.1∗*SDy*].
- For **logistic models**, the parameters expressed in log odds ratio can be converted to standardized difference through the formula:π/3π/3(see the **effectsize** package, resulting in a range of -0.18 to -0.18. For other models with binary outcome, it is strongly recommended to manually specify the rope argument. Currently, the same default is applied that for logistic models.
- For **t-tests**, the standard deviation of the response is used, similarly to linear models (see above).
- For **correlations**, -0.05, 0.05 is used, *i.e.*, half the value of a negligible correlation as suggested by Cohen's (1988) rules of thumb.
- For all other models, -0.1, 0.1 is used to determine the ROPE limits, but it is strongly advised to specify it manually.

Sensitivity to parameter's scale

It is important to consider **the unit (*i.e.*, the scale) of the predictors** when using an index based on the ROPE, as the correct interpretation of the ROPE as representing a region of practical equivalence to zero is dependent on the scale of the predictors. Indeed, unlike other indices (such as the pd), the percentage in **ROPE** depend on the unit of its parameter. In other words, as the ROPE represents a fixed portion of the response's scale, its proximity with a coefficient depends on the scale of the coefficient itself.

For instance, if we consider a simple regression growth ~ time, modelling the development of **Wookies babies**, a negligible change (the ROPE) is less than **54 cm**. If our time variable is **expressed in days**, we will find that the coefficient (representing the growth **by day**) is of about **10 cm** (*the median of the posterior of the coefficient is 10*). Which we would consider as **negligible**. However, if we decide to express the time variable **in years**, the coefficient will be scaled by this transformation (as it will now represent the growth **by year**). The coefficient will now be around **3550** cm (10 * 355), which we would now consider as **significant**.

```
library(rstanarm)

library(bayestestR)

library(see)


data <- iris # Use the iris data

model <- stan_glm(Sepal.Length ~ Sepal.Width, data = data) # Fit model

# Compute indices

pd <- p_direction(model)

percentage_in_rope <- rope(model, ci = 1)


# Visualise the pd
```

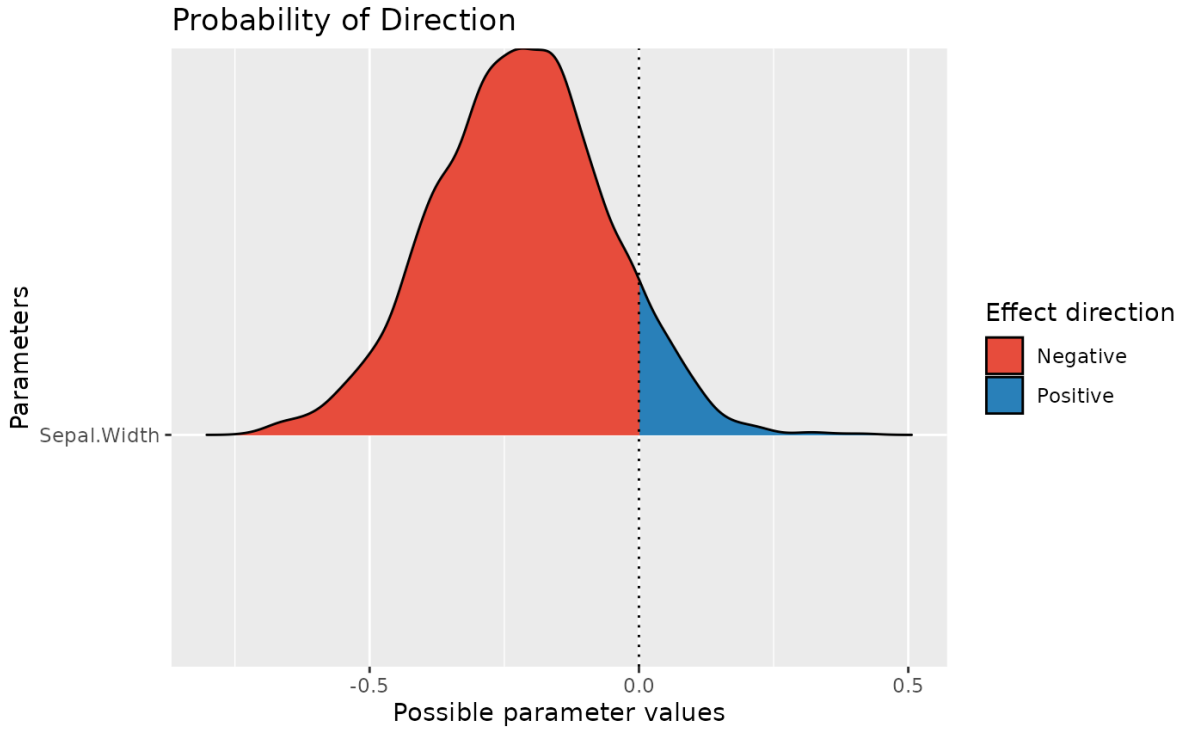
plot(pd)

pd

Visualise the percentage in ROPE

plot(percentage_in_rope)

percentage_in_rope



> Probability of Direction

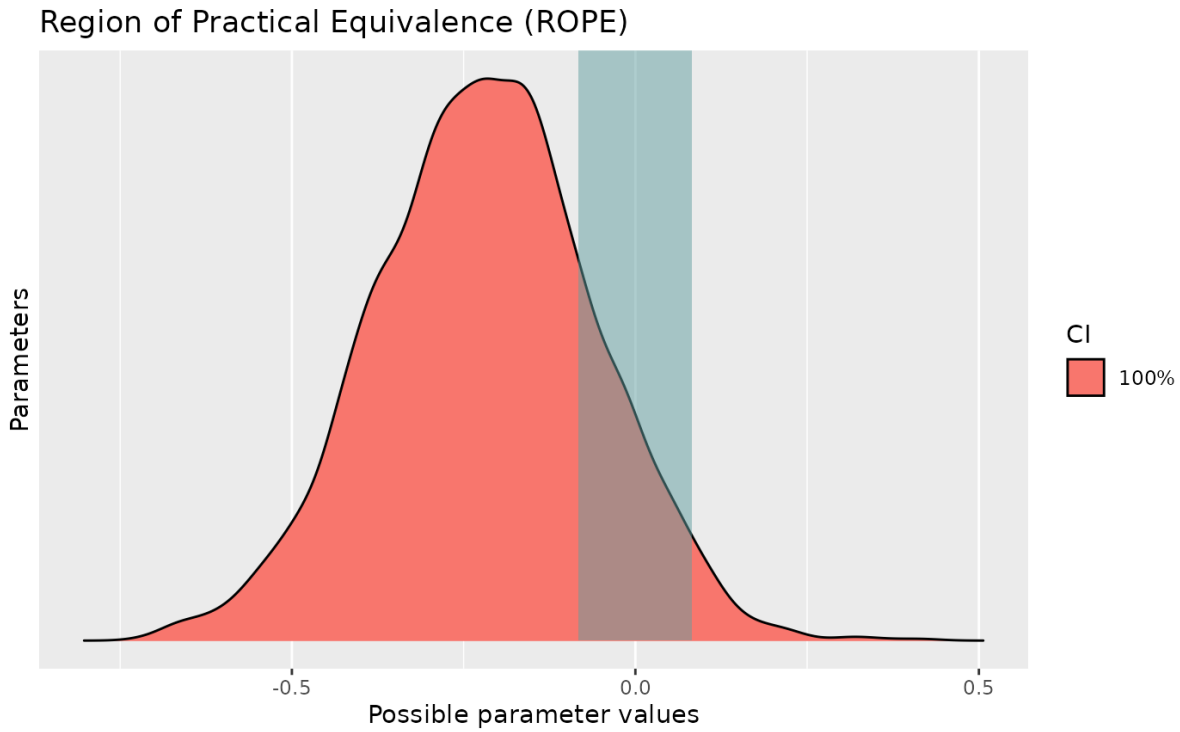
>

> Parameter | pd

> ----------

> (Intercept) | 100%

> Sepal.Width | 91.65%



> # Proportion of samples inside the ROPE [-0.08, 0.08]:

>

> Parameter | inside ROPE

> ------------------------

> (Intercept) | 0.00 %

> Sepal.Width | 16.28 %

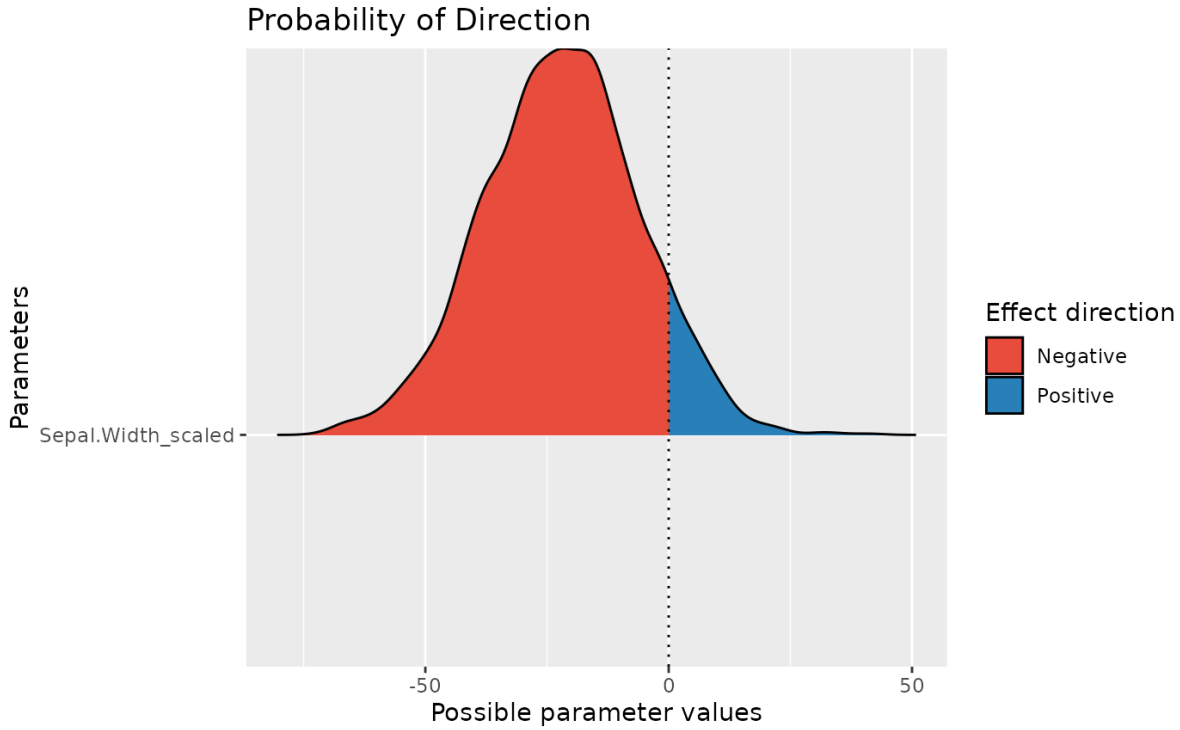
We can see that the *pd* and the percentage in ROPE of the linear relationship between **Sepal.Length** and **Sepal.Width** are respectively of about 92.95% and 15.95%, corresponding to an **uncertain** and **not significant** effect. What happen if we scale our predictor?

```
data$Sepal.Width_scaled <- data$Sepal.Width / 100 # Divide predictor by 100

model <- stan_glm(Sepal.Length ~ Sepal.Width_scaled, data = data) # Fit model

# Compute indices

pd <- p_direction(model)

percentage_in_rope <- rope(model, ci = 1)


# Visualise the pd

plot(pd)

pd
```

Visualise the percentage in ROPE

plot(percentage_in_rope)

percentage_in_rope



> Probability of Direction

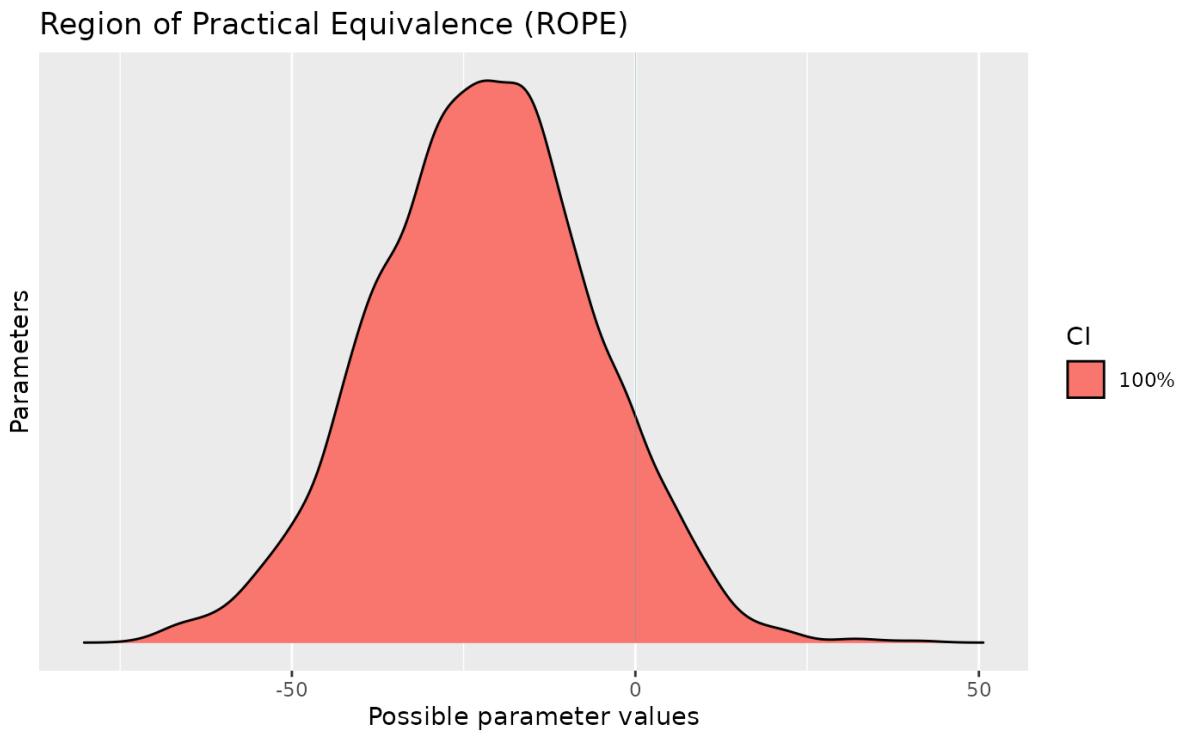
>

> Parameter | pd

> ----------

> (Intercept) | 100%

> Sepal.Width_scaled | 91.65%



> # Proportion of samples inside the ROPE [-0.08, 0.08]:

>

> Parameter | inside ROPE

> ---

> (Intercept) | 0.00 %

> Sepal.Width_scaled | 0.10 %

As you can see, by simply dividing the predictor by 100, we **drastically** changed the conclusion related to the **percentage in ROPE** (which became very close to 0): the effect could now be **interpreted as being significant**. Thus, we recommend paying close attention to the unit of the predictors when selecting the ROPE range (*e.g.*, what coefficient would correspond to a small effect?), and when reporting or reading ROPE results.

Multicollinearity: Non-independent covariates

When **parameters show strong correlations**, *i.e.*, when covariates are not independent, the joint parameter distributions may shift towards or away from the ROPE. Collinearity invalidates ROPE and hypothesis testing based on univariate marginals, as the probabilities are conditional on independence. Most problematic are parameters that only have partial overlap with the ROPE region. In case of collinearity, the (joint) distributions of these parameters may either get an increased or decreased ROPE, which means that inferences based on ROPE are inappropriate (J. Kruschke, 2014).

The equivalence_test() and rope() functions perform a simple check for pairwise correlations between parameters, but as there can be collinearity between more than two variables, a first step to check the assumptions of this hypothesis testing is to look at different pair plots. An even more sophisticated check is the projection predictive variable selection (Piironen & Vehtari, 2017).

Source: https://help.vwo.com/hc/en-us/articles/30447193353241-What-is-the-Region-of-Practical-Equivalence-ROPE

Mathematically speaking, even a very small improvement over the baseline can be considered and given enough visitors, it will be declared statistically significant. However, from a business perspective, not all improvements are worth deploying, and we have found that experimenters often ignore such infinitesimal changes. The case is the same for any equivalent decline, as well.

Such a region where any minuscule improvement or decline is regarded as equivalent to the baseline's conversion rate is called the Region of Practical Equivalence (ROPE). Businesses can configure the ROPE for their campaigns individually for all metrics, depending on their scenario.

Defining Your ROPE: Choosing the Right Region of Practical Equivalence

Assume that your baseline conversion rate (aka the baseline average) is 40%. You have decided that any conversion rate between 38% and 42% will be regarded as equivalent to that of the baseline. For this, ROPE is calculated as follows:

Step 1: Obtain the difference between these individual extremes and the baseline average. Here, the difference is ± 2.

Step 2: Divide the difference by the baseline average.

Now, ROPE = ± 2 / (40%) = ± 5

You can set ± 5 as the ROPE for your campaign in this case.

Unlocking Insights: The Value of ROPE in Decision-Making

There are three main ways in which VWO uses ROPE to deliver value to customers. Note that a higher value of ROPE will lead to an increased intensity of these benefits but also has some tradeoffs that are discussed in the next section.

- **Declaring when to stop a variation:** Using ROPE, VWO can give an early decision to disable a variation if it detects that the variation does not have the potential to be better than the baseline.
- **Minimize False Positives:** Without the concept of ROPE, we might mistake random fluctuations in data for significant improvements. ROPE acts as a safeguard against false positives, helping us focus on changes that truly matter.
- **Saving Visitors on Average:** Including ROPE helps accelerate the decision to disable underperforming variations and campaigns, which can lead to overall resource savings. Since campaigns with significant winners are typically less common than those that don't produce a clear winner, using ROPE often results in conserving visitors on average.

Considerations for ROPE

Determining the width of your Region of Practical Equivalence (ROPE) is a nuanced decision that depends on various factors, including the nature of your experiment, the desired level of confidence, and the practical significance of the changes you seek.

- **Default Values:** A conservative ROPE value of ±1% is a good default to start with, as it will save you a good amount of the False Positive Rate. VWO has configured this by default.

- **Understand Your Business Context:** Consider the unique aspects of your business and the target metric. Some businesses might have smaller margins of acceptable change, while others may tolerate larger values. Your understanding of what constitutes a meaningful change in your specific context will influence the value of your ROPE.
- **Iterative Refinement**: A/B testing is an iterative process. Initially, you might start with a broader ROPE and then refine it based on the results of your experiments. This adaptive approach ensures that your ROPE aligns more closely with the actual impact on your business.

In total, ROPE is a vital feature in the optimization toolkit for making informed decisions based on statistically significant and practically relevant data. After all, the essence lies not just in making changes but in implementing those that truly matter for a more optimized and seamless online experience.

Source: https://en.wikipedia.org/wiki/Chi-squared_test

Chi-squared distribution, showing χ2 on the x-axis and p-value (right tail probability) on the y-axis.

A chi-squared test (also chi-square or χ2 test) is a statistical hypothesis test used in the analysis of contingency tables when the sample sizes are large. In simpler terms, this test is primarily used to examine whether two categorical variables (two dimensions of the contingency table) are independent in influencing the test statistic (values within the table).[1] The test is valid when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table. For contingency tables with smaller sample sizes, a Fisher's exact test is used instead.

In the standard applications of this test, the observations are classified into mutually exclusive classes. If the null hypothesis that there are no differences between the classes in the population is true, the test statistic computed from the observations follows a χ2 frequency distribution. The purpose of the test is to evaluate how likely the observed frequencies would be assuming the null hypothesis is true.

Test statistics that follow a χ2 distribution occur when the observations are independent. There are also χ2 tests for testing the null hypothesis of independence of a pair of random variables based on observations of the pairs.

Chi-squared tests often refers to tests for which the distribution of the test statistic approaches the χ2 distribution asymptotically, meaning that the sampling distribution (if the null hypothesis is true) of the test statistic approximates a chi-squared distribution more and more closely as sample sizes increase.

History

In the 19th century, statistical analytical methods were mainly applied in biological data analysis and it was customary for researchers to assume that observations followed a normal distribution, such as Sir George Airy and Mansfield Merriman, whose works were criticized by Karl Pearson in his 1900 paper.[2]

At the end of the 19th century, Pearson noticed the existence of significant skewness within some biological observations. In order to model the observations regardless of being normal or skewed, Pearson, in a series of articles published from 1893 to 1916,[3][4][5][6] devised the Pearson distribution, a family of continuous probability distributions, which includes the normal distribution and many skewed distributions, and proposed a method of statistical analysis consisting of using the Pearson distribution to model the observation and performing a test of goodness of fit to determine how well the model really fits to the observations.

Pearson's chi-squared test

See also: Pearson's chi-squared test

In 1900, Pearson published a paper[2] on the χ2 test which is considered to be one of the foundations of modern statistics.[7] In this paper, Pearson investigated a test of goodness of fit.

Suppose that n observations in a random sample from a population are classified into k mutually exclusive classes with respective observed numbers of observations xi (for i = 1,2,…,k), and a null

hypothesis gives the probability pi that an observation falls into the ith class. So we have the expected numbers mi = npi for all i, where

∑

i

=

1

k

p

i

=

1

∑

i

=

1

k

m

i

=

n

∑

i

=

1

k

p

i

=

n

{\displaystyle {\begin{aligned}&\sum _{i=1}^{k}{p_{i}}=1\\[8pt]&\sum _{i=1}^{k}{m_{i}}=n\sum _{i=1}^{k}{p_{i}}=n\end{aligned}}}

Pearson proposed that, under the circumstance of the null hypothesis being correct, as n → ∞ the limiting distribution of the quantity given below is the χ2 distribution.

X

2

=

∑

i

=

1

k

(

x

i

−

m

i

)

2

m

i

=

∑

i

=

1

k

x

i

2

m

i

–

n

{\displaystyle X^{2}=\sum _{i=1}^{k}{\frac {(x_{i}-m_{i})^{2}}{m_{i}}}=\sum _{i=1}^{k}{{\frac {x_{i}^{2}}{m_{i}}}-n}}

Pearson dealt first with the case in which the expected numbers mi are large enough known numbers in all cells assuming every observation xi may be taken as normally distributed, and reached the result that, in the limit as n becomes large, X2 follows the χ2 distribution with k − 1 degrees of freedom.

However, Pearson next considered the case in which the expected numbers depended on the parameters that had to be estimated from the sample, and suggested that, with the notation of mi being the true expected numbers and m'i being the estimated expected numbers, the difference

X

2

–

X

'

2

=

∑

i

=

1

k

x

i

2

m

i

–

∑

i

=

1

k

x

i

2

m

i

,

{\displaystyle X^{2}-{X'}^{2}=\sum _{i=1}^{k}{\frac {x_{i}^{2}}{m_{i}}}-\sum _{i=1}^{k}{\frac {x_{i}^{2}}{m'_{i}}}}

will usually be positive and small enough to be omitted. In a conclusion, Pearson argued that if we regarded X′2 as also distributed as χ2 distribution with k − 1 degrees of freedom, the error in this approximation would not affect practical decisions. This conclusion caused some controversy in practical applications and was not settled for 20 years until Fisher's 1922 and 1924 papers.[8][9]

Other examples of chi-squared tests

One test statistic that follows a chi-squared distribution exactly is the test that the variance of a normally distributed population has a given value based on a sample variance. Such tests are uncommon in practice because the true variance of the population is usually unknown. However, there are several statistical tests where the chi-squared distribution is approximately valid:

Fisher's exact test

For an exact test used in place of the 2 × 2 chi-squared test for independence when all the row and column totals were fixed by design, see Fisher's exact test. When the row or column margins (or both) are random variables (as in most common research designs) this tends to be overly conservative and underpowered.[10]

Binomial test

For an exact test used in place of the 2 × 1 chi-squared test for goodness of fit, see binomial test.

Other chi-squared tests

Cochran–Mantel–Haenszel chi-squared test.

McNemar's test, used in certain 2 × 2 tables with pairing

Tukey's test of additivity

The portmanteau test in time-series analysis, testing for the presence of autocorrelation

Likelihood-ratio tests in general statistical modelling, for testing whether there is evidence of the need to move from a simple model to a more complicated one (where the simple model is nested within the complicated one).

Yates's correction for continuity

Main article: Yates's correction for continuity

Using the chi-squared distribution to interpret Pearson's chi-squared statistic requires one to assume that the discrete probability of observed binomial frequencies in the table can be approximated by the continuous chi-squared distribution. This assumption is not quite correct and introduces some error.

To reduce the error in approximation, Frank Yates suggested a correction for continuity that adjusts the formula for Pearson's chi-squared test by subtracting 0.5 from the absolute difference between each observed value and its expected value in a 2 × 2 contingency table.[11] This reduces the chi-squared value obtained and thus increases its p-value.

Chi-squared test for variance in a normal population

If a sample of size n is taken from a population having a normal distribution, then there is a result (see distribution of the sample variance) which allows a test to be made of whether the variance of the population has a pre-determined value. For example, a manufacturing process might have been in stable condition for a long period, allowing a value for the variance to be determined essentially without error. Suppose that a variant of the process is being tested, giving rise to a small sample of n product items whose variation is to be tested. The test statistic T in this instance could be set to be the sum of squares about the sample mean, divided by the nominal value for the variance (i.e. the value to be tested as holding). Then T has a chi-squared distribution with n − 1 degrees of freedom. For example, if the sample size is 21, the acceptance region for T with a significance level of 5% is between 9.59 and 34.17.

Example chi-squared test for categorical data

Suppose there is a city of 1,000,000 residents with four neighborhoods: A, B, C, and D. A random sample of 650 residents of the city is taken and their occupation is recorded as "white collar", "blue collar", or "no collar". The null hypothesis is that each person's neighborhood of residence is independent of the person's occupational classification. The data are tabulated as:

A B C D Total

White collar 90 60 104 95 349

Blue collar 30 50 51 20 151

No collar 30 40 45 35 150

Total 150 150 200 150 650

Let us take the sample living in neighborhood A, 150, to estimate what proportion of the whole 1,000,000 live in neighborhood A. Similarly we take

349

/

650

to estimate what proportion of the 1,000,000 are white-collar workers. By the assumption of independence under the hypothesis we should "expect" the number of white-collar workers in neighborhood A to be

150

×

349

650

≈

80.54

{\displaystyle 150\times {\frac {349}{650}}\approx 80.54}

Then in that "cell" of the table, we have

(

observed

−

expected

)

2

expected

=

(

90

−

80.54

)

2

80.54

≈

1.11

{\displaystyle {\frac {\left({\text{observed}}-{\text{expected}}\right)^{2}}{\text{expected}}}={\frac {\left(90-80.54\right)^{2}}{80.54}}\approx 1.11}

The sum of these quantities over all of the cells is the test statistic; in this case,

≈

24.57

{\displaystyle \approx 24.57}. Under the null hypothesis, this sum has approximately a chi-squared distribution whose number of degrees of freedom is

(

number of rows

−

1

)

(

number of columns

−

1

)

=

(

3

−

1

)

(

4

−

1

)

=

6

{\displaystyle ({\text{number of rows}}-1)({\text{number of columns}}-1)=(3-1)(4-1)=6}

If the test statistic is improbably large according to that chi-squared distribution, then one rejects the null hypothesis of independence.

A related issue is a test of homogeneity. Suppose that instead of giving every resident of each of the four neighborhoods an equal chance of inclusion in the sample, we decide in advance how many residents of each neighborhood to include. Then each resident has the same chance of being chosen as do all residents of the same neighborhood, but residents of different neighborhoods would have different probabilities of being chosen if the four sample sizes are not proportional to the populations of the four neighborhoods. In such a case, we would be testing "homogeneity" rather than "independence". The question is whether the proportions of blue-collar, white-collar, and no-collar workers in the four neighborhoods are the same. However, the test is done in the same way.

Applications

In cryptanalysis, the chi-squared test is used to compare the distribution of plaintext and (possibly) decrypted ciphertext. The lowest value of the test means that the decryption was successful with high probability.[12][13] This method can be generalized for solving modern cryptographic problems.[14]

In bioinformatics, the chi-squared test is used to compare the distribution of certain properties of genes (e.g., genomic content, mutation rate, interaction network clustering, etc.) belonging to different categories (e.g., disease genes, essential genes, genes on a certain chromosome etc.).[15][16]

Source: https://www.optimizely.com/optimization-glossary/conversion-rate/

What is a conversion rate?

Conversion rates are a percentage typically used in digital marketing to evaluate performance of website traffic, marketing campaigns and conversions. To calculate a conversion rate, take the number of conversions divided by the total number of visitors. For example, if an ecommerce site receives 200 visitors in a month and has 50 sales, the conversion rate would be 50 divided by 200, or 25%.

A conversion can refer to any desired action that you want the user to take. This can include anything from a click on a button (CTA) to making a purchase and becoming a new customer. Websites and apps often have multiple conversion goals, and each will have its own conversion rate.

Why conversion rates are important

Tracking conversion rates allows you to measure the performance of your web pages and apps. Understanding what percentage of your users are completing the goals that drive your business allows you to gauge the success of your site or app and identify areas for improvement.

Improving your conversion rate also allows you to get more sales with the same amount of traffic. If you are spending $1,000 a month on advertising to drive 500 visitors to your site, if you double your conversion rate you essentially double the value of your ad spend. You can then cut back on your ad spend and get the same benefit as you were getting before, or invest the additional revenue into new ad programs.

Many factors can impact your conversion rate or cause it to go up and down. Something as simple as introducing new messaging, or doing search engine optimization (SEO) can make conversion rates fluctuate. While higher conversion rates are generally considered better, the more advantage you're taking of the traffic you have, not all sources of traffic are created equal and can still contribute to more new customers even though their conversion rates aren't as high. For example, organic traffic has higher conversion rates than display ads because people searching for something directly typically show more intent than people clicking a banner advertisement.

Measuring different kinds of conversion rates

- General conversion rates can be based on form fills, downloads, clicks
- Ecommerce conversion rates are add to shopping cart clicks, purchases
- Organic search conversion rates can be measured using blog articles read (scroll tracking) divided by search traffic
- Social media conversion rates might be calculated using direct messages divided by followers

- Click-through rate is tracked using clicks on a banner or advertisement divided by impressions

What factors impact conversion rates

A lot of factors can impact good conversion rates, including but not limited to:

- Source of app and website visitors, depending on channels and mediums
- Types of conversions like form fills or purchases
- Region, in some countries online purchasing is more popular than in others
- Messaging on the landing pages
- How well optimized your website or app is
- Device types like mobile devices, desktop or tablet
- User experience, the better the experience, typically the higher the conversion rates

That last one, user experience, is an important factor. Adding elements to your website or app that might seem like conversion rate improvements can hit a ceiling, where you're taking conversions from one action and converting them elsewhere. So always keep an eye on overall conversion rates as well as individual action's conversion rates. You want to end up with net-positive improvements, adding to the overall conversions, not taking away only to convert elsewhere.

Some elements we've found to have high conversion rates but can negatively impact user experience are:

- Popups, increasing page conversion rates but reducing them elsewhere
- Pervasive and intrusive interstitials, disturbing the visitor as they're reading something like a blog
- Dark patterns, where misleading messaging is used to trick visitors into converting

How to measure conversion rate

There's multiple ways to do conversion tracking, but the generally accepted practice is to take:

total number of visitors / conversions

Let's break that down. 'Visitors' in this case are all the people visiting your website, where as 'conversions' are the total amount of completed actions on said website. Dividing one by the other gives you a percentage, also known as the *conversion rate*.

Typical conversions are purchases, form fills, add to shopping cart, clicking a call-to-action or any worthy key performance indicators (KPIs) for your business like lead generation.

Some more complex websites might not have the goal to convert all visitors to the same type of conversion and need to adjust their marketing strategy. For example, if you have a large website with support, legal, a blog and other sections that don't contain any forms, you can exclude metrics from the total visitors to get a more true conversion rate.

Keep in mind that although you can filter conversion rates by narrowing their scope, that also makes you lose some visibility into overall conversion rates. Higher conversion rates might make other

conversion rates go down in other areas. For instance, we tested improving the chatbot on optimizely.com, increasing it's conversion rates, but in the end it turned out it was cannibalizing on conversions for our other forms. Essentially optimizing the chatbot stole attention from the landing page and pushed people to convert elsewhere.

A more tailored and true conversion rate might look like:

Page views for "products" / shopping cart call-to-action (CTA) clicks

Landing page visits / form fills to download an ebook

To get started measuring your website's conversion rate, use a tool like Google Analytics to set up a conversion rate tracking dashboard so its easier to monitor over time. Most web tracking tools come with many different types of conversion tracking events out of the box, and should start recoding visitor data as soon as they're installed.

How to improve your conversion rate

The process of identifying conversion goals, calculating their conversion rates, and optimizing your site or app to achieve higher conversion rates is known as conversion rate optimization or CRO. Conversion optimization is done by formulating hypotheses for why visitors aren't converting and coming up with ideas for improving conversions, then testing those ideas through a process called A/B testing, in which two versions of a page are tested against each other to see which one performs better.

Start by identifying what your current average conversion rates are and comparing against a benchmark conversion rate. This can either be industry, device or technology specific, so it's good practice to take the average for your business.

Then, take multiple data sources into account beyond the typical web metrics if possible. Consider taking heatmaps, ecommerce data, CRM data and other inputs to determine areas of improvement. These can often have surprising learnings you can't glean from just looking at web metrics.

By continually identifying new conversion goals, identifying areas where your conversion rate can be improved, and implementing improvements to website's templates, you can continuously improve the performance of your website or app and boost conversions with minimal additional traffic. Converting more potential customers into business.

Conversion rate optimization in action

A real world example of CRO is captured in a case study from ComScore, a web analytics company that provides marketing data to enterprises. The company started by setting a conversion goal from their product page (leads generated), determined the conversion rate, then set up an experiment with different ideas for improving the conversion rate of the page.

The hypothesis that they came up with was that they suspected including testimonials on the page would increase visitor trust and lead to more conversions. They also tested a version of the page that included both testimonials and the logos of the companies providing the testimonials.

Through this A/B/n test, they found that the version of their page with testimonials and logos performed 69% better than their original page. This is a clear example of how a company was able to improve their conversion rate through testing and have a measurable impact on their business.

Boost your conversion rate with Optimizely experimentation

Optimizely is the leading platform for A/B testing and conversion rate optimization. Installing Optimizely is incredibly easy, and requires just installing a single snippet of JavaScript on your site.

Once Optimizely is enabled, the visual editor allows you to make changes to your website or app without any coding or developers required. Launching experiments is as simple as the click of the button, and Optimizely will automatically display visitors different versions of your site to visitors.

Once an experiment is set up, Optimizely's advance stats engine will tell you when a test has reached statistical significance, so that you can confidently report whether a change performs better or worse than the original.

Source: https://www.optimizely.com/optimization-glossary/conversion-rate-optimization/

What is conversion rate optimization?

Conversion rate optimization (CRO) is the process of increasing the percentage of conversions from a website or mobile app through desired action. It involves:

- Generating ideas for improving site/app elements
- Validating hypotheses through A/B testing and multivariate testing
- Enhancing user experience to boost conversions

Looking at lessons learned from 127,000 experiments, under-prioritized metrics like search rate can increase conversions.

Why is conversion rate optimization important?

By having a conversion rate optimization strategy, you can:

- Increase revenue per visitor
- Lower customer acquisition costs
- Get more value from existing visitors/users
- Acquire more customers and grow your business

Example: If a landing page has an average conversion rate of 10% and receives 2000 visitors a month, then the page will generate 200 conversions per month. If the conversion rate can be improved to 15% by optimizing different elements on the page, the number of conversions generated jumps by 50% to 300 per month.

In digital marketing, there is always room for improvement when it comes to website conversion rate, and the best companies are constantly iterating and improving their sites and apps to create a better experience for their users and grow conversions.

An effective CRO strategy relies on several key elements:

1. **User research**
   Understanding your audience's needs and behaviors.
2. **Website analytics**
   Gathering and analyzing user data in real-time.

3. **User experience (UX) design**
   Creating intuitive, enjoyable user interactions.
4. **Landing page optimization**
   Refining entry points for maximum impact.
5. **Copywriting**
   Crafting persuasive, action-oriented content.
6. **Page load speed**
   Ensuring quick loading times across devices.
7. **Trust building**
   Incorporating elements that boost credibility.
8. **Conversion funnel analysis**
   Identifying and addressing drop-off points.
9. **Mobile optimization**
   Providing excellent experiences on all devices.

Conversion rate optimization examples

We've got two examples from real practitioners to prove conversion rate optimization can help you learn interesting things.

Example 1: Real book cover vs. Abstract version

Joe Geoghan, Senior Visual Brand Design Specialist, Optimizely, wanted to test the real cover vs. an abstract version of the cover for The Big Book of Experimentation in an email body. Assuming the real cover would win, it was the cover used in most of the emails.

Result: The abstract version still ended up winning

Takeaway: Both cover illustrations were too small to be legible. The abstract version was concise and showed you exactly what you were getting into. In design, clarity matters.

Example 2: Predicting the next best action

Charlotte Golding and her team at Virgin Media wanted to predict the Next Best Action (NBA) so they could design personalized experiences for their customers. They assumed customer would only have specific requests like improving the network in their area or upgrading their existing broadband, etc.

Result: The team found that the same customer would come in with different requests each day. One day, they were looking for customer care and the next day, they just wanted to upgrade. This wasn't initially factored in the NBA but after the experiment, the team had to optimize their model to better understand on which next best action to show to a customer.

Takeaway: Customers can come to your website about a different thing every day. Don't just put them in a single personalized experience and expect the same results. Optimize the model regularly.

Remember, any marketing strategy relies on a variety of techniques, each targeting different aspects of the user experience. Here are a few conversion rate optimization techniques:

1. **Optimize Call-to-Action:** Craft compelling, action-oriented CTA buttons with strategic placement and contrasting colors.
2. **Improve user experience (UX):** Simplify navigation, improve page load times, and ensure mobile responsiveness.
3. **Test:** Systematically compare variations of page elements to identify top performers.
4. **Personalize:** Tailor messaging and offers based on user behavior, preferences, or demographics.
5. **Include social proof:** Leverage customer testimonials, reviews, social media threads, and usage statistics to build trust.
6. **Build trust:** Display security badges, certifications, and clear policies to alleviate user concerns.

Establishing conversion metrics

Conversion rate optimization begins by first identifying what the conversion goals are for any given web page or app screen. The success metrics of your website or mobile app will depend on the type of business you're in, and what your goals are.

For example, if you sell products online via ecommerce channels, a conversion for you may be the number of purchases or the number of website visitors that add a product to their shopping cart. If you sell products or services to businesses, you might be measuring the number of leads your website collects or the number of white paper downloads.

Some common conversion goals organized by usability and industry type:

- **Media:** Pageviews, ad views, newsletter subscriptions, recommended content engagement
- **Ecommerce:** Product sales, add-to-carts, shopping cart completion rate, e-mail newsletter sign-ups
- **Travel:** Booking conversions, ancillary purchases, social shares
- **B2B:** Leads generated, deals closed

Once you have established the conversion metrics for your digital interactions with your audience, you can increase conversion rates by improving your digital customer experiences.

Conversion rate optimization process

Once your conversion metrics have been identified, here's a simple data-driven process you want to follow for converting site visitors:

- Identify your conversion goals
- Analyze your current sales funnel
- Focus on high-traffic or underperforming pages
- Develop hypotheses for improvements
- Test your hypotheses

- Analyze results and implement winning changes
- Continuously iterate and improve

You can start by optimizing pages that receive the greatest amount of traffic. By focusing on these pages, you will be able to see the results of your changes faster and have a larger impact on your business.

Other potential places to start include your highest-value pages that are underperforming compared to the rest of your site. Again, improving these areas can have the greatest immediate impact on your conversion goals.

For example, a clothing retailer may find that their page for hats receives a lot of traffic but has a conversion rate that is much lower than the rest of the site. By improving the conversion rate of that page, the retailer might see a big improvement in sales.

CRO best practices

When it comes to CRO, great results aren't possible without specific action and experimentation. Here are some of the best CRO practices you can use to get started.

1. Research your target audience and website traffic. Understand their pain points.
2. Test clear Call-to-Action (CTA)**.** Do not rush your visitors. Anticipate how ready they're to buy and send them to the next step accordingly.
3. Don't add too much information on every page. Each page should lead to a clear next step.
4. Optimize for mobile devices. Ensure all functionalities and CTAs work.
5. Reduce load time for your slow-loading web pages to reduce bounce rates.
6. Use trust signals like customer testimonials, case studies, social proof, industry badges, etc.
7. Personalize content and product recommendations based on user behavior.
8. Identify areas of a page that are most (or least) engaging with heatmaps and improve wherever required.

Not all ideas improve your website conversion rate

There are tonnes of ideas folks want to implement on their website, all of which seem like a great idea at the time. Most teams come up with benchmarks and ideas, push them to production, and then try and measure the results through a CRO test. However, only 12% of experiments run actually produce a winning result.

Because they've already invested in the process, the focus is on making it work. But what if the wrong ideas were being tested from the start?

Change gears a bit. Testing isn't just about finding winners. This is a legacy way of thinking about CRO.

Experimentation is about learning. The only way your optimization efforts 'fail' is if you fail to learn from it. For example, changing a call-to-action and product page design and messaging is a great way to see how potential customers interact on the page. Some even prefer seeing the pricing upfront.

Focus on using data at every step (Google Analytics functionality can help you).

Source: https://www.statsig.com/blog/cuped

CUPED Explained

CUPED is slowly becoming a common term in online experimentation since its coining by Microsoft in 2013.

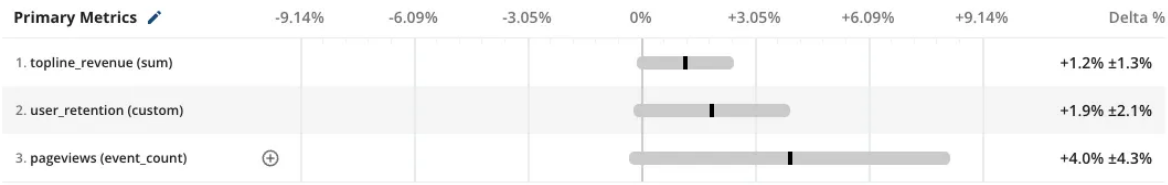
Meaning *Controlled-experiment Using Pre-Experiment Data*, CUPED is frequently cited as—*and used as*—one of the most powerful algorithmic tools for increasing the speed and accuracy of experimentation programs.

In this article, we'll:

- Cover the background of CUPED
- Illustrate the core concepts behind CUPED
- Show how you can leverage this tool to run faster and less biased experiments

What CUPED solves:

As an experiment matures and hits its target date for readout, it's not uncommon to see a result that seems to be ***only barely*** outside the range where it would be treated as statistically significant. In a frequentist world, this isn't sufficient evidence that your change caused a change in user behavior.



If there was a real effect, you needed more **sample size** to increase your chances of getting a statistically significant result. In an experiment, the standard error or "noise" goes down with the square root of your sample size. However, sample size is an expensive resource, usually proportional to the enrollment window of your experiment.

Waiting for more samples delays your ability to make an informed decision, and it doesn't guarantee you'll observe a statistically significant result when there is a real effect.

Even at companies with immense scale like Facebook and Amazon, people have to deal with the pain of waiting for experiments to enroll users and mature because they're usually looking for relatively small effects.

Consider this: *A 0.1% increase to revenue at Facebook is worth upwards of $100 million per year!*

For smaller companies, small effect sizes can become infeasible to measure. It would just take too long to get the sample needed to reliably observe a statistically significant change in their target metric.

Because of this cost, a number of methods have been developed in order to **decrease the standard error** for the **same metric and sample size.**

CUPED is an extremely popular implementation that uses pre-experiment data to explain away some of the variance in the result data.

The statistical concept behind CUPED

Like many things in experimentation, the core concept behind CUPED is simple, but its implementation can be tricky (and expensive!).

The guiding principle of CUPED is that **not all variance in an experiment is random.** In fact, a lot of the differences in user outcomes are based on **pre-existing factors that have nothing to do with the experiment.**

Let's talk about this for a minute:

Say we want to run a test to see if people run slower with weights attached to them. From a physics perspective, the answer seems pretty obvious. We might record data like this:

Person	Test Group	Experiment Mile Time
Sally	Weights	6:40
Dave	No Weights	7:10
Jane	Weights	8:20
Bob	No Weights	9:00

If we average out our results, we might clearly see the expected effect, but we might not; there's a lot of variance and overlap in the observed mile times. It should be pretty clear, however, that *how fast the runners already were* might be an underlying factor. What if we asked them to run a mile a week ago to establish a **baseline**?

Person	Group	Baseline Mile Time	Experiment Mile Time	Change
Sally	Weights	6:30	6:40	**+10**
Dave	No Weights	7:08	7:10	**+2**
Jane	Weights	7:30	8:20	**+50**
Bob	No Weights	9:15	9:00	**-15**

In the context of their "typical" mile time, this effect should be much clearer! We've implicitly switched from caring about their raw "mile time" into caring about **the difference from what we'd expect!**

By doing this, we've also "explained" some of the noise and variance in the experiment metric. Before, we saw a difference of 140 seconds between the fastest and slowest runner. Now, we've

reduced the range in our metric to 65 seconds; this lower range should mean that the variance we'd use to calculate confidence intervals and p-values will be lower.

This is conceptually very similar the original implementation of CUPED; we use the pre-experiment data for a metric to normalize the post-experimental values. How **much** we normalize is based on how well the pre-experiment data predicts the experiment data - we'll dive into this later.

Bias correction

Because experimental groups are randomly assigned, there's a chance that the two groups randomly have different **baseline** run times. If you're unlucky, that difference could even be statistically significant. This means that even if the weights did nothing, you might conclude that there's a difference between the two groups.

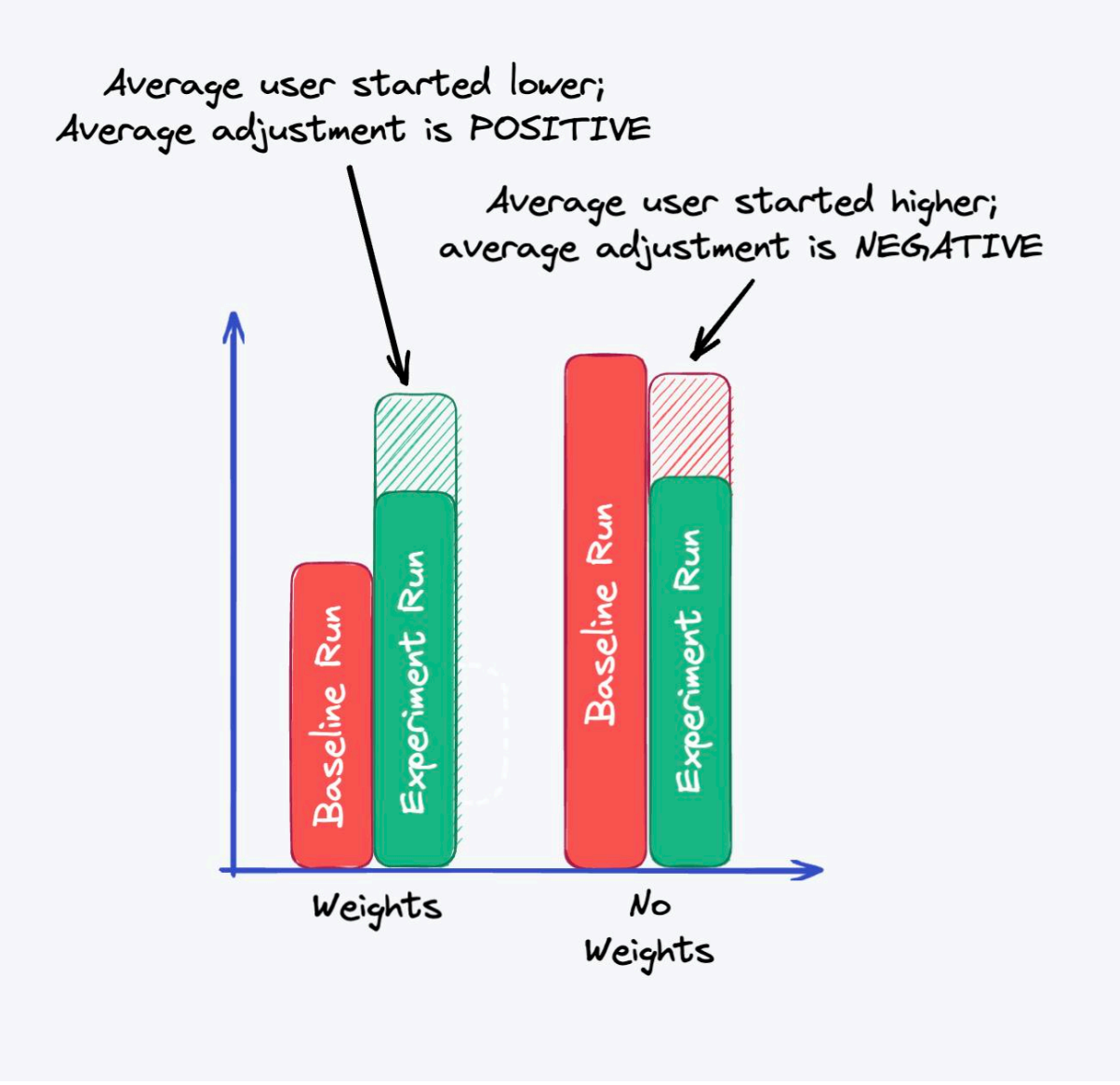
If you have access to that baseline data, it'd be possible to conclude that there was a pre-existing difference and be wary of the results. In the example below, it's pretty obvious that the difference in the groups **before** the test would make the results extremely skewed:

You might note that you can see that the weighted runners' times went up, and the unweighted runners' times went down. This relative change *does* match our expectation. Would it be possible to infer that there *is* an effect here? Correcting this data with CUPED can help!

Correction

Conceptually, if one group has a **faster average baseline**, their experiment results will also be faster. When we apply a CUPED correction, the **faster group's metric will be adjusted downwards relative to the slower group.**

In this example, the post-adjustment averages might move something like this, pushing the weights group's experiment value higher than the control group. We could follow up with a statistical test to understand if the difference in adjusted values is **statistically significant**.



Stratification

Some variants of CUPED are 'non-parametric' or 'bucketed'. What this usually means is that (in this example) we would split users into groups based on their pre-experiment run times, and measure metrics relative to the average metric value of that group.

For example, consider the data below - this is for the bucket of users who ran between a 6:30 and 6:40 mile in the baseline:

Group	Prev Mile Time Bucket	Avg Bucket Time	Experiment Mile Time	Adjusted Mile Time
Weights	6:30-6:40	6:42	6:50	+8
No Weights	6:30-6:40	6:42	6:35	-7
Weights	6:30-6:40	6:42	7:02	+20

Other variables

More complex implementations of CUPED don't just rely on a single historical data point for the same metric. They can pull in other information as well, as long as it's independent of the experiment group the user is in.

In the example above, we could add age group as a factor in the experimentation. This has relatively little to do with our experiment, but could be a major factor in people's mile times! By including this as a factor in CUPED, we can reduce even more variance.

Group	Baseline Mile Time	Age	Experiment Mile Time	Change
Weights	6:30	25	6:40	**+10**
No Weights	7:08	32	7:10	**+2**
Weights	7:30	22	8:20	**+50**
No Weights	9:15	45	9:00	**-15**

Using CUPED in practice

In practice, we can't just subtract out a user's prior values from their experimental values. The reason for this is also conceptually simple—people's past behavior isn't always a perfect predictor for their future behavior.

A mental model for the math we'll use

Before we go further, it's useful to understand the relationship between experimentation and regression (the ordinary-least-squares or "OLS" regression you'd run in excel.)

A **T-test** for a given metric is mathematically equivalent to running a regression where the dependent variable is your metric and the independent variable is a user's experiment group. To demonstrate

this, I generated some data for the example experiment above, where users' paces are based on a randomly-assigned baseline pace and if they're in the test group.

The population statistics for this are:

Population statistics mean time (s): 547.6 test mean (s): 551.3 control mean (s): 543.8 test n: 100 control n: 100

Let's compare the outputs of running a T-test and running an OLS where we use the 1-or-0 test flag as the independent variable.

T-test:

scipy.stats.ttest_ind(test, ctr)

se calculated using student's SE equation with population variance

effect size: 7.4617

pvalue: 0.116

t-stat: 1.577

standard error: 4.730

OLS:

```
==============================================================
            coef    std err          t      P>|t|
--------------------------------------------------------------
const       543.8446     3.345    162.594      0.000
test          7.4617     4.730      1.577      0.116
==============================================================
```

Comparing these, we notice a lot of similarities:

- The effect size in our T-test (the delta between test and control) is exactly the same as the "test" variable's coefficient in the OLS regression.
- The standard error for the coefficient is the same as the standard error for our T-test.
- The p-value for the "test" variable coefficient is the same as for our t-test!

In short, our standard T-test is basically a regression against a 1-or-0 variable!

When we want to make regressions more accurate, we might add relevant explanatory variables. We can do the same for our test; again, this is the core concept behind CUPED.

Let's include baseline pace as a factor in our regression. We should expect this to change the regression quite a bit, since it's such a powerful explanatory variable—*and it does*.

==========

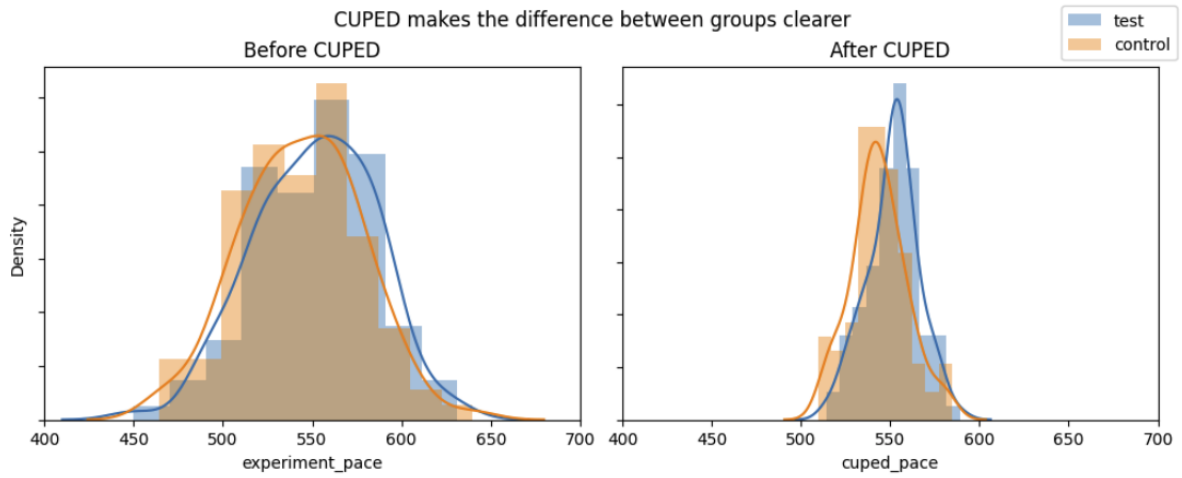
	coef	std err	t	P>|t|
const	8.8023	19.273	0.457	0.648
test	7.6892	2.134	3.602	0.000
baseline_pace	0.9897	0.036	27.846	0.000

==========

Let's review:

- The "test" variable's coefficient (the estimate of the experiment effect) didn't change much. That's **expected** - unless there was a significant difference between the groups **before** the experiment we should get a similar estimate of the experiment effect.
- The standard error (and accordingly p-value) went down from **4.73** to **2.13.** This is because a lot of the noise we previously attributed to our test variable **wasn't random:** Tt was coming from users having different baselines, which we're now accounting for!
- Our p-value goes from 0.116 to 0.000 because of the decreased Standard Error. The result, which was previously not statistically significant, is now clearly significant.

Using CUPED with the baseline pace achieves nearly-identical results. To visualize the reduction in Variance/Standard Error, I plotted the distribution of user paces from this sample dataset before and after I applied CUPED:



When we apply CUPED, we see a large reduction in variance and p-value, just like in the regression results. Using the pre-experiment data reduced the variance, p-value, and the data we would need to consistently see this result.

CUPED math and implementation

For more details on this, please refer to the **2013 Microsoft white paper.** *We've used many formulas that appear in that paper here.*

To reduce variance by using other variables, we'll need to make adjustments such that we end up with an **unbiased estimator** of group means that we'll use in our calculations. An unbiased estimator simply means that the **expected value** of the estimator is equal to the **true value** of the parameter we're estimating.

In practice, this means we need to pick an adjustment that is independent of which test group a user is assigned to.

For the original, simplest implementation of CUPED we'll refer to our pre-experiment values as X and our experiment values as Y. We'll adjust Y to get a covariate-adjusted Ycv according to the formula below:

$$Y\hat{}cv=Y—-\theta X—+\theta E(X)$$

Here, **θ** could be any derived constant. What this equation means it that, for any **θ**, we can take two steps:

- Multiply the pre-experiment population mean by **θ** and add it to each user's result
- Subtract from each user's result **θ** multiplied by their pre-experiment value

This gives us an unbiased estimator Ycv which factors in the covariate into our estimates. We can calculate the variance of the new estimator term:

$$var(Y\hat{}cv)=var(Y\hat{}-\theta X\hat{})=var(Y\hat{}-\theta X)n=1n(var(Y)+\theta 2var(X)-2\theta cov(X,Y))$$

This is the variance of our adjusted estimator for Y. This variance turns out to be the smallest for:

$$\theta=cov(Y,X)var(X)$$

This is the term we'd use to calculate the **slope in an OLS regression**! This is also the term we'll up using in our data transformation - we take all the data in the experiment and calculate this theta. The final variance for our estimator is

$$var(Y\hat{}cv)=var(Y^{-})(1-\rho 2)$$

where **ρ** is the correlation between X and Y. The correlation between the pre-experiment and post-experiment data is directly linked to how much the variance is reduced. Note that since **ρ** is bounded between [-1, 1], this new variance will always be less than or equal to the original variance.

In practice

To create a data pipeline for the basic form of CUPED, you need to carry out the following steps. With X referring to pre-experiment data points and Y points referring to experiment data:

- Calculate the covariance between Y and X as well as the variance and mean of X. Use this to calculate **θ** per the formula above.

    - This requires that users without pre or post-experiment data are included as 0s if they are to be included in the adjustment

- For each user, calculate the user's individual pre-experiment value. It's common to choose to not apply an adjustment for users who are not eligible for pre-experiment data (for example new users) - this is effectively a one-level striation.
- Join the population statistics to the user-level data
- Calculate user's adjusted terms as $Y + \theta * (\text{population mean of } X) - \theta X$
- Run and interpret your statistical analysis as you normally would, using the adjusted metrics as your inputs

Implications from the CUPED math (above):

There are many covariates we could use for variance reduction; the main requirement is that it is independent of the experiment group which the user is assigned to. Generally, data from before an experiment is safest.

We commonly use the same **metric** from before the experiment as a covariate because in practice it's usually a very effective predictor, and it makes intuitive sense in most cases.

We should calculate the group statistics for the pre-experiment/post-experiment data across the entire experiment population—not on a per-group basis—because it's possible there's an interaction effect between the treatment and the pre-exposure data. For example, *users who run faster may be better equipped to run with weights*, and so the correlation between the pre and post-periods would be different than for slower users.

New Users won't have pre-experiment data. An experiment with no pre-experiment data won't be able to leverage CUPED. In these cases, the best bet is to use covariates like demographics if possible.

If an experiment has some new users and some established users, you can use CUPED and split the population by another binary covariate: D*o they have pre-experiment data or not?* Functionally, this means you just apply CUPED only on users with pre-experiment data as discussed above.

CUPED best practices

- CUPED is most effective on existing user experiments where you have access to user's historical data. For new users experiments, stratification or other covariates like demographics can be useful, but you won't be able to leverage as rich of a covariate.
- CUPED needs historical data to work; this means that you need to make sure your metric data goes back to before the start of the pre-experiment data window.
- CUPED's ability to adjust values is based on how correlated a metric is with its past value for the same user. Some metrics will be very stable for the same user and allow for large adjustments; some are noisy over time for the same user, and you won't see much of a difference in the adjusted values.

Source: https://www.optimizely.com/insights/blog/cuped-in-ab-testing-and-experimentation/

CUPED: Reducing variance in A/B testing isn't new but most are getting it wrong

by Misha Datsenko

CUPED stands for Controlled-experiment Using Pre-Existing Data, a statistical approach to reduce variance by using historical user behavior before the test begins.

TIME...

is the culprit of why well-designed experiments sometimes do not reach statistical significance.

Many A/B tests end up in the "inconclusive" graveyard, hovering just below the significance threshold. Between slow data collection and high-variance metrics, detecting real effects in your website redesign or pricing strategy can be frustratingly elusive.

What if you could tighten your confidence intervals and increase the statistical power of your experiments using data you already have?

That's where CUPED comes in. ***It stands for Controlled-experiment Using Pre-Existing Data, a statistical approach to reduce variance.***

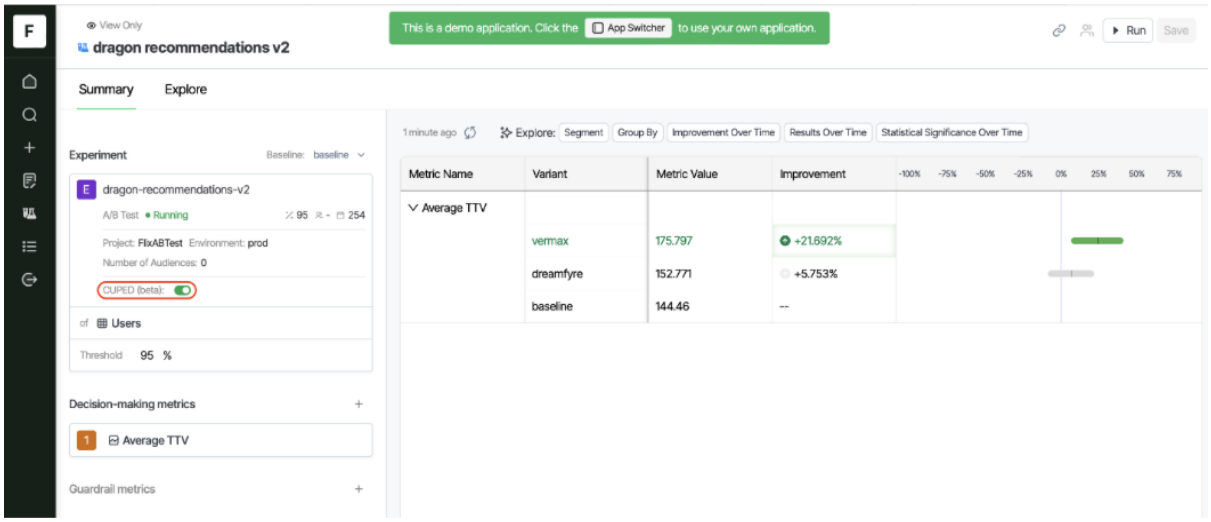


Image source: Optimizely

And the best part? You can use it too! Let's see how.

The experimentation efficiency gap

The challenge of experimentation timing varies by industry:

- E-commerce sites typically run tests for several weeks
- SaaS products often require longer test periods
- Media sites may see faster results due to higher traffic volumes

And that's if they conclude at all. Many simply get abandoned when results stay inconclusive.

Why does this happen? Three main reasons:

1. **High variance in metrics data:** Engagement metrics naturally fluctuate a lot between users, making it hard to spot true effects.
2. **Limited traffic:** Not every company is Google. Most businesses struggle to get enough users through an experiment.
3. **Opportunity cost:** Every week an experiment runs is another week you're delaying decisions and potential improvements. Longer experiments mean fewer tests you can run in a given timeframe, slowing down your overall learning velocity and product evolution.

High variance in a metric requires a larger sample size to reach statistical significance, which can take weeks consideringthe visitor traffic. When metrics fluctuate widely between users, you need more data to distinguish true difference from a pure chance.

CUPED makes a critical difference by reducing this variance using pre-experiment data. This allows you to reach statistical significance with smaller sample sizes, extracting clearer signals from the data you already have instead of simply collecting more.

This transforms a painfully slow learning cycle into a more efficient experimentation program. This is the experimentation efficiency gap that CUPED helps bridge.

Let's dive deeper into where CUPED came from and how it works.

How CUPED turns your existing data into faster wins

Microsoft Research published a paper in 2013 that introduced CUPED: ***Controlled-experiment Using Pre-Existing Data***.

A statistical method that makes your A/B tests more efficient by using data you already have.

Early adopters at Microsoft reported significant improvements in their testing capabilities. Companies like Netflix and Airbnb have since implemented similar approaches with impressive results.

What makes CUPED different is its elegant simplicity. It uses pre-experiment data as a covariate to reduce variance in your metrics.

If you want to measure how a new feature affects user spending, wouldn't it be helpful to account for how much those users were spending before your experiment?

That's exactly how CUPED filters out the noise so you can see the signal more clearly.

To truly appreciate CUPED's value, we need to understand its nemesis aka variance.

Variance is why two seemingly identical users can have wildly different behaviors:

- One spends $10 on your site
- Another spends $150
- And you're trying to detect a 5% improvement in average order value

See the problem?

With naturally high variance metrics like revenue or engagement, small treatment effects get buried under mountains of statistical noise. It's like trying to hear a whisper at a rock concert.

CUPED works by incorporating pre-experiment data as a covariate in your analysis.

This tightening of confidence intervals is what makes CUPED so effective. Same data and the same effect size, but suddenly you can see it.

Now that we understand how CUPED works, let's look at where it delivers the most impact.

Not all metrics benefit equally from CUPED...

Here's what you need to know:

1. Special considerations: Revenue metrics

Revenue metrics often have extremely high variance. Some users might spend $5 while others spend $500.

When applied to revenue metrics, CUPED looks for a correlation between past spending and current spending. Thus, CUPED will not be effective for new users, where we don't have past spending data.

A common implementation mistake is using covariates that are influenced by the treatment, which can lead to biased results. A best practice is to choose covariates that are measured before the experiment starts.

2. When to use CUPED

Best for: High-variance numeric metrics

- Revenue per visitor
- Average order value
- Session duration

These metrics see the biggest improvement with CUPED because they typically have:

- High natural variance between users
- Strong correlation between pre-experiment and during-experiment values

Less effective for: Binary conversion metrics

- Conversion rate (yes/no)
- Clickthrough rate (click/no click)

How to flip the CUPED switch in Optimizely

Optimizely makes using CUPED straightforward:

- **Compatible metrics:** Works with **numeric metrics** (revenue, engagement counts) but not binary conversion metrics
- **Pre-experiment data:** Uses pre-experiment values of your target metrics as covariates
- **Supported in Optimizely Analytics**: Functions on **Snowflake**, **BigQuery,** and **Databricks**
- **Implementation:** Simple toggle in experiment settings, no complex calculations needed

- **Data requirements:** Needs historical data for analyzed metrics; no effect on new metrics without history
- **Expected outcome:** Reduces variance, potentially cutting sample size requirements for metrics correlated to historical behavior

Here's how it looks with and without CUPED.

Without CUPED

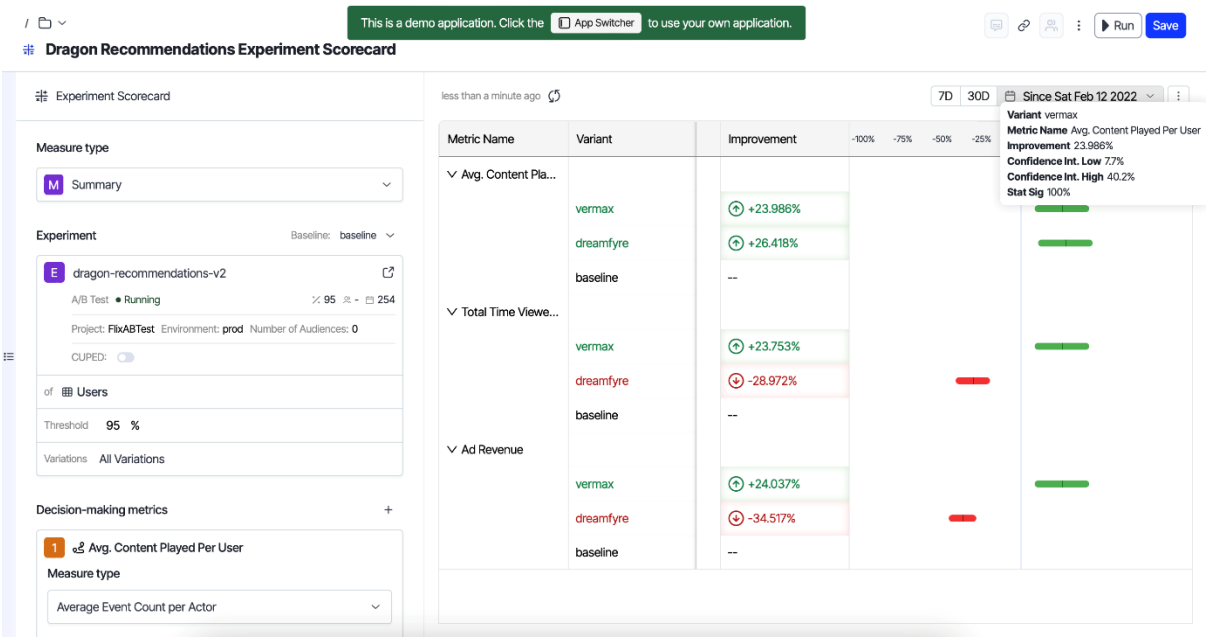


Image source: Optimizely

Now, with CUPED, there's a difference in the length of the confidence interval.

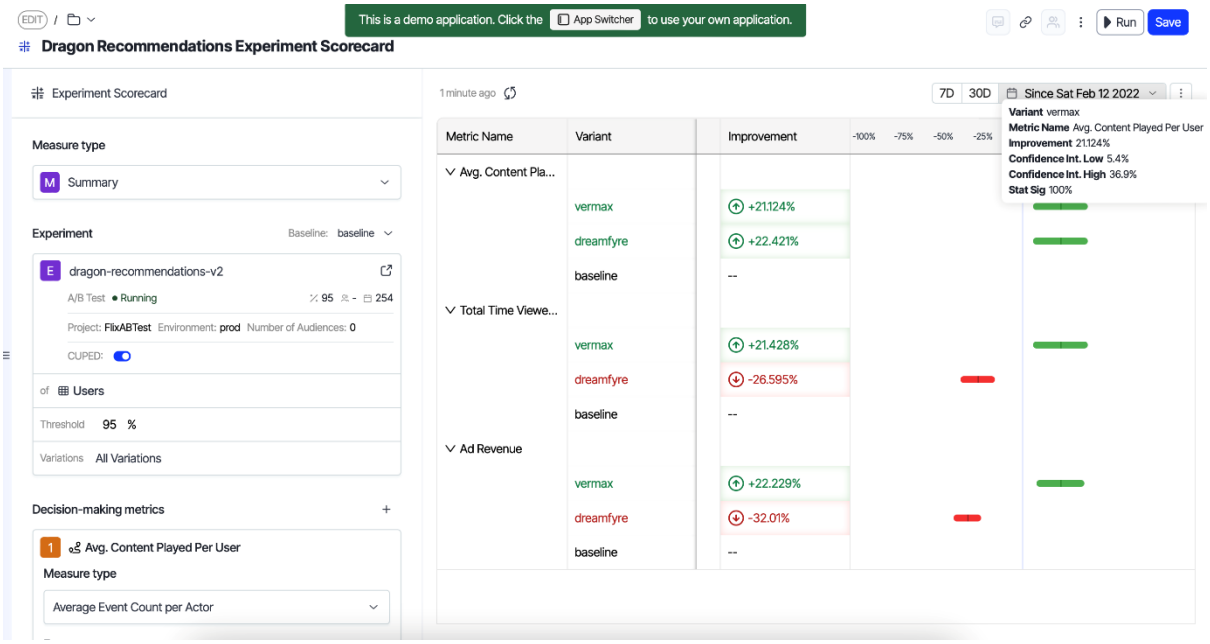


Image source: Optimizely

Three takeaways...

The future of experimentation isn't just about running more tests, it's about running smarter tests. CUPED is your first step in that direction.

1. **Increased efficiency:** You have a higher chance of seeing significant results with the same sample size.
2. **Not all metrics benefit equally:** Focus CUPED implementation on high-variance numeric metrics where you'll see the biggest gains.
3. **Implementation complexity varies:** There are different ways of implementing CUPED and different covariates that can be chosen. Optimizely's implementation uses historical metric data which fits the majority of our customer's use cases.

Ready to run smarter tests?

Start by identifying one high-variance metric in your experimentation program. Run a side-by-side comparison between your traditional analysis and a CUPED-enhanced test.

You'll likely see tighter confidence intervals, clearer results, and potentially reach statistical significance for a test that would have been inconclusive otherwise.

The path to more efficient experimentation begins with this simple step. Your future self will wonder how you ever tested without it.

Source: https://www.graphapp.ai/blog/understanding-cuped-variance-reduction-a-comprehensive-guide

Understanding CUPED Variance Reduction: A Comprehensive Guide

CUPED, or Controlled Experiments Using Pre-Experiment Data, is a sophisticated statistical technique designed to enhance the efficiency of A/B testing and experimental analysis. By leveraging historical data and covariates, CUPED aims to reduce variance in experimental outcomes, leading to more reliable and actionable insights. This guide delves into the core concepts, mathematical foundations, practical benefits, and implementation strategies of CUPED, catering specifically to software developers and analysts looking to optimize their data-driven initiatives.

Introduction to CUPED Variance Reduction

Variance reduction is a crucial aspect of data analysis that directly affects the reliability of conclusions drawn from experimental data. CUPED offers a structured approach to improving statistical estimates by utilizing pre-experiment data. This method stands out because it not only analyzes current experiment data but also historic information, allowing for a richer contextual understanding.

What is CUPED Variance Reduction?

CUPED is essentially a statistical modeling technique that incorporates pre-experiment metrics to adjust and refine the estimates obtained from ongoing experiments. By doing so, CUPED mitigates the effects of noise—unwanted variability—in the data, ultimately enhancing the statistical power and precision of the experimental results. The adjustments made by CUPED often lead to detecting true treatment effects that may otherwise be obscured by randomness.

Importance of CUPED in Data Analysis

In software development and data-heavy industries, being able to draw reliable conclusions from experiments is vital. CUPED empowers analysts by providing a robust mechanism for variance reduction, allowing teams to make confident decisions based on their data. By integrating historical data, it helps to contextualize the experiments, leading to a deeper understanding of user behavior and product performance.

Moreover, employing CUPED can significantly reduce the sample sizes needed for experiments. Smaller samples mean quicker turnarounds for experiments, enabling teams to iterate rapidly and enhance user experiences more effectively. This efficiency is particularly beneficial in fast-paced environments where time-to-market can be a critical factor in maintaining a competitive edge.

Additionally, CUPED's ability to leverage historical data not only improves the accuracy of current experiments but also fosters a culture of data-driven decision-making within organizations. As teams become more adept at utilizing past metrics, they can identify trends and patterns that inform future strategies. This continuous learning loop enhances the overall quality of insights derived from data, making it an invaluable tool for businesses aiming to optimize their operations and better serve their customers.

The Mathematics Behind CUPED Variance Reduction

Understanding the mathematical framework of CUPED is essential for implementing it effectively. At its core, CUPED employs regression techniques that use pre-existing information as a covariate. This mathematical underpinning is what enhances the precision of experimental results. By leveraging

historical data, CUPED helps in reducing the noise that often clouds the clarity of experimental findings, thereby allowing for more accurate interpretations and decisions based on the data.

Understanding the CUPED Algorithm

The CUPED algorithm typically starts with the formulation that involves adjusting treated and control group metrics based on pre-experiment data. The key component is the use of linear regression to model the relationship between the outcome variable and the covariates. This relationship is harnessed to create a modified outcome variable that has reduced variance. The process not only improves the signal-to-noise ratio but also allows for a more nuanced understanding of how different factors interplay within the experimental framework.

Once the adjustment is made, analysts can apply standard techniques to analyze the experimental results, benefitting from the statistical improvements without altering the fundamental A/B testing methodology. This seamless integration of CUPED into existing workflows means that organizations can adopt this powerful tool without overhauling their entire analytical process, making it a practical choice for many data-driven teams.

The Role of Covariates in CUPED

Covariates play a pivotal role in the CUPED model. These are the measurable factors from prior data that are believed to influence the outcome of the experiment. By controlling for these covariates, CUPED explicitly acknowledges the pre-existing characteristics of the data, allowing the analysis to focus on the effect of the treatment rather than the effects of variabilities. This focus helps in isolating the true impact of the intervention, which is critical for making informed business decisions.

Choosing appropriate covariates is crucial; they must be strongly predictive of the primary metric and relevant to the experiment. Careful consideration of covariate selection can lead to significant improvements in the quality of the conclusions drawn from the data. Additionally, the inclusion of multiple covariates can help in uncovering complex interactions that may not be evident when analyzing the data in isolation. This depth of analysis can ultimately lead to more robust insights, enabling organizations to refine their strategies and optimize outcomes based on empirical evidence.

Benefits of Using CUPED Variance Reduction

Implementing CUPED yields a myriad of benefits that can enhance a team's analytical capabilities and speed up iterative processes in software development.

Improving Experimental Efficiency

The most immediate benefit of using CUPED is its potential to improve experimental efficiency. By reducing the variability in outcomes, CUPED allows one to achieve the same level of statistical significance with a smaller sample size. In the fast-paced world of software development, this can be a game changer, as it results in quicker testing cycles and faster iterations.

With increased efficiency, resources can be allocated more strategically, enabling teams to focus on high-impact experiments that directly inform product development and user experience design. This not only accelerates the pace of innovation but also enhances the team's ability to pivot quickly in response to user feedback or market changes, ensuring that the product remains relevant and competitive.

Reducing Noise in Data Analysis

Another compelling benefit of CUPED is its effectiveness in reducing noise within the data. This reduction is crucial for organizations trying to draw meaningful insights from user interactions, particularly in cases where user behaviors are erratic and influenced by various external factors. By filtering out this noise, CUPED provides clearer signals regarding user responses to different treatments.

The clarity gained from implementing CUPED not only aids in decision-making but also allows teams to base their strategies on credible data rather than assumptions or speculative analysis. Furthermore, this enhanced data quality can lead to more robust predictive models, enabling teams to forecast user behavior with greater accuracy. As a result, organizations can tailor their marketing efforts and product features more effectively, ultimately driving user engagement and satisfaction.

Implementing CUPED Variance Reduction

When it comes to putting CUPED into practice, having a structured plan is vital. The implementation process is straightforward but requires careful attention to detail to ensure that results are valid and actionable.

Steps to Apply CUPED in Your Data

1. **Select Covariates:** Identify relevant historical metrics that correlate with your experimental outcome.
2. **Prepare Data:** Clean and collate both your pre-experiment and experimental datasets.
3. **Model Adjustment:** Use regression techniques to adjust your experimental outcome based on the selected covariates.
4. **Conduct Analysis:** Analyze the adjusted outcome using standard statistical methods and interpret the results.

By following these steps, organizations can systematically apply CUPED to their analysis processes, effectively leveraging the power of pre-experiment data to enhance their experimental insights. The careful selection of covariates is particularly crucial, as this step lays the foundation for the entire analysis. It is advisable to conduct exploratory data analysis (EDA) on historical data to uncover patterns and relationships that may not be immediately apparent. This can provide a more robust basis for selecting covariates that will yield the most significant variance reduction.

Common Mistakes and How to Avoid Them

While implementing CUPED can be beneficial, it is essential to be aware of common pitfalls that can undermine its effectiveness. One common mistake is selecting covariates that lack a strong correlation with the outcome variable, which can lead to ineffective adjustments.

Another issue is neglecting the impact of multicollinearity among covariates, where selected variables are highly correlated with each other. This situation can obscure the analysis and lead to misleading conclusions. To mitigate this risk, employing techniques such as variance inflation factor (VIF) analysis can help identify and address multicollinearity before proceeding with regression modeling.

Finally, it is crucial to maintain the integrity of the randomized experiment. CUPED should enhance your analytical efforts, not replace the fundamental principles of experimentation. This means ensuring that the randomization process remains intact and that any adjustments made do not

introduce bias into the experimental results. By adhering to these principles, organizations can maximize the benefits of CUPED while safeguarding the validity of their experimental findings.

Advanced Concepts in CUPED Variance Reduction

As individuals become more versed in CUPED, exploring advanced concepts can provide additional layers of insight and application.

Dealing with Non-Normal Distributions

While CUPED operates effectively under the assumption of normality, real-world data can often deviate from this ideal. Advanced techniques, such as transformations or robust statistical methods, can be employed to handle non-normal distributions effectively.

Recognizing the distribution of your data ahead of applying CUPED is crucial. Tailoring the approach based on the nature of the data can ensure that the applied techniques still lead to valid and reliable results. For instance, applying logarithmic or square root transformations can help stabilize variance and make the data more amenable to analysis. Additionally, employing robust methods like bootstrapping can provide a safeguard against the influence of outliers, ensuring that the results remain robust even in the presence of non-normality.

CUPED in Multivariate Analysis

In scenarios where multiple variables need to be analyzed simultaneously, CUPED can still shine. Techniques such as multivariate regression can be adapted to incorporate multiple covariates, allowing for a comprehensive understanding of how various factors interact and affect the outcome.

By embracing this multivariate approach, analysts can gain deeper insights into complex systems typically seen in user interactions and product responses, pushing the boundaries of standard analytical practices. Furthermore, integrating machine learning techniques with CUPED can enhance predictive modeling capabilities. For example, using algorithms like random forests or gradient boosting can help uncover intricate relationships among variables, providing a richer context for interpreting the effects of interventions. This synergy between CUPED and advanced modeling techniques can lead to more nuanced decision-making processes, ultimately driving better business outcomes.

Conclusion: The Future of CUPED Variance Reduction

As data analyses continue to evolve, CUPED is poised to play a significant role in enhancing decision-making and strategy development in various fields, particularly in technology-driven environments. Its unique ability to harness pre-experiment data makes it a powerful tool for software developers and data analysts alike.

Emerging Trends in Variance Reduction Techniques

Moving forward, we can expect continued advancements in variance reduction methodologies, including further refinements to CUPED itself. As new data analytics technologies and machine learning practices emerge, these innovations may offer even greater enhancements to traditional methods like CUPED.

Additionally, the integration of real-time data analytics with CUPED applications may become prevalent, allowing for instantaneous adjustments and analyses that could already be relevant before the completion of experimentation.

Final Thoughts on CUPED Variance Reduction

The journey into understanding CUPED variance reduction provides invaluable insights into effectively analyzing and interpreting experimental data. By blending robust statistical methods with practical application, CUPED stands as an essential technique for any serious data analyst.

Arming oneself with knowledge of CUPED not only enhances one's analytical toolkit but also greatly improves the overall efficiency of data-driven decision-making processes. Embrace CUPED, and turn historical data into a beacon for future insights and success.

Source: https://capturly.com/blog/131-simple-ways-to-use-heatmaps-for-a-b-testing/

What are heatmaps?

Before we get into anything, let's see exactly what heatmaps are. Heatmaps are used in many industries. You may have seen it in geographical data visualization and financial, business, or weather analysis.

Heatmaps are a useful way to see how visitors behave on your website. They show you which parts of your website get the most attention, using colors to make them easy to understand. By analyzing a heatmap, you can learn how engaged users are with your website and whether your design is working well.

There are 3 types of website heatmaps:

Click heatmaps

Want to know where your website users are clicking? Use click heatmaps! These powerful tools show which parts of your website are popular and which are ignored. Analyze user behavior, optimize CTAs and menu positions, and increase your conversion rate. Don't overlook the power of click heatmaps for a better user experience!

Scroll heatmaps

Wonder why useful information on your website doesn't reach customers? Scroll heatmaps show how far visitors scroll down web pages and the percentage of people who reach specific sections. Analyzing the data can help you get more people to buy your product by putting the important stuff where it's easy to see.

Segment heatmaps

Segment heatmaps group website visitors based on attributes like their device type or operating system. This helps you understand how people behave on your website. By analyzing this data, you can optimize your website for your target audience. For example, you might need to make changes to your mobile site or adjust your content for users who come from search engines. Get the insights you need to improve your website with segment heatmaps!

What is A/B testing?

A/B testing is a powerful technique for improving your website's conversion rate and increasing your income. It's very simple: you change one thing on your website (let's call it version B), then compare it to the original version (version A) to see which works better.

By making small changes and comparing the results, you can learn a lot about what works best for your users. Maybe your call-to-action button needs to be in a different place, or perhaps changing the color of a T-shirt on your model will make a big difference. With A/B testing, you can easily optimize and tailor your website to meet the needs of your users.

And the best part? You don't need to be a technical genius to run an A/B test. There are plenty of tools out there that make it easy to set up and run experiments on your website. So why not give it a try and see how it can help you improve your website's performance?

What's the difference between A/B and multivariate testing?

It’s not uncommon for people to mix up A/B testing and multivariate testing. But there’s a significant difference between the two.

With A/B testing, you compare two different versions of your website to see which one performs better. This involves changing just one element at a time, like the color of a button or the wording of a headline.

Moreover, multivariate testing involves changing various elements at once. This approach can be useful if you want to test several different variations of your website to see which combination of changes works best.

Both A/B testing and multivariate testing can help make your website better, but it’s important to know how they differ from each other. That way, you can choose the right approach for your specific needs and goals.

So next time if you want to test different versions of your website, keep in mind the difference between A/B testing and multivariate testing. With this knowledge, you can make informed decisions about which testing strategy to use and get the most out of your experiments.

Why is a heatmap essential in A/B testing?

Heatmaps are helpful tools that show where users click, scroll, and hover on your website. By using heatmaps, you can quickly see which parts of your website work well and which ones need improvement. Heatmaps can even reveal user behavior patterns that other metrics might miss.

Use the insights from your heatmaps to plan future A/B tests and optimization strategies. This will help you create a great user experience that keeps people coming back to your website. Essential for any website owner or designer looking to improve engagement!

How to use heatmaps for A/B testing?

Are you looking to make data-driven decisions and optimize your website’s design for better user engagement? Here are some ways you can make effective use of heatmaps in A/B testing to achieve the best performance.

#1 Identify areas of interest

Heatmaps are a great tool for understanding user behavior on your website. They show you which parts of the page are being clicked on the most, highlighted in red for easy identification. For instance, in the figure below, version A has a more frequently clicked menu.

If you want to improve a specific call-to-action (CTA) button, you can easily make changes to its text or design. Ultimately, the choice of version depends on your goals and what you hope to achieve with your website. Take some time to think about your objectives and make an informed decision that works best for you.

#2 Analyze click patterns

Heatmaps can be a valuable tool for analyzing click patterns and identifying areas of your website that may need improvement. Compare heatmaps of different webpage versions side-by-side to see how users behave differently.

Use this information to make changes and continue testing until you reach full user engagement. Remember, optimization is an ongoing process, so don't hesitate to seek help from our team of experts.

#3 Check out page scrolling

It's a common issue that many website owners face – they tend to place important information in areas where users don't scroll down. If you happen to be a website owner, I'd suggest taking a look at the heatmap to ensure that the essential information is in the right place.

Remember that if users don't scroll down to a certain part of your website, it's better not to put any important information there. Otherwise, they might miss it. After all, you want to make sure your users have a great experience on your website and find what they're looking for easily.

#4 Test different page layouts

Different page layouts play a crucial role in creating unique user experiences. How you design the user experience on your website can have a big impact on your conversion rate and your revenue. It's important to consider your target audience when crafting your UX, as their preferences and needs may differ greatly.

Let's look at an example. If your target audience is young gamers, make sure that your website matches their personality and needs. Make it rich and dynamic. If your target audience is mainly older business people, focus on making your website easy to use and accessible for them. Have authority and reflect your brand personality.

#5 Optimize form completion

Did you know that 81% of people have abandoned a form after beginning to fill it out?

Making online forms easy to use is crucial for increasing conversion rates. By simplifying the form-filling process, you can encourage more people to complete the form.

A/B testing is a useful tool for analyzing user behavior and making informed decisions about your website's design. By using a click heatmap, you can see exactly how users are navigating your site and identify any areas where they may be getting stuck.

To get more people to complete your online forms, you can make changes based on how users interact with them. Paying attention to the user experience and making adjustments can help you get better results.

#6 Perfect your product images

When it comes to online shopping, product images are absolutely essential for driving sales. When shopping online, customers can't touch or see the products in person before buying them. That's why having high-quality images optimized for the web is important.

By testing different product images and optimizing them for the greatest impact, you can boost your sales and revenue. One effective way to do this is through A/B testing, which allows you to experiment with different images and see which ones perform best.

And when it comes to analyzing the results of your A/B tests, heatmaps are an invaluable tool. They give you a detailed look at how customers interact with your product images, so you can identify any areas where improvements can be made.

You can make your customers happy and earn more money by paying attention to your product images and using data to improve your online store.

#7 Find ideal product positioning

As we discussed in the section on page layouts, the placement of information is absolutely critical. In particular, the way you position your products can have a huge impact on whether customers choose to buy them.

It's crucial to give extra attention to your most profitable products since they make the most money. This way, you can make sure they get the attention they need. But, it's possible that the positioning of these products might not be ideal for your customers, which can hurt your revenue.

A/B testing and the usage of heatmaps are helpful tools to determine the best placement for your products. You can experiment with different placements and see which one your customers prefer. Data can help you find the perfect spot for your top products, so you can make more money and improve your customers' shopping experience.

Make your website simple and place your products in the right place to keep your customers coming back.

#8 Refine navigation and menus

Have you ever tried to buy something online but the website's menu is so complicated that you feel like giving up and leaving? I'm sure you understand the importance of having a user-friendly menu that helps shoppers easily find what they're looking for.

If your website's menu is confusing, potential customers may leave and you'll lose money. But, if you regularly check how customers are clicking and scrolling, you can find and fix problems, making it easier for them to buy from you.

#9 Streamline checkout flows

If your checkout process is complicated, it can have a negative impact on your earnings. Make it easy for your customers to shop and buy from you by understanding their needs and simplifying their experience.

If your checkout process is too complicated, customers might leave and you'll lose revenue. So, make sure to test and track it regularly, listen to your customer's feedback, and make changes to improve your sales.

You can make some simple changes to improve your sales and make your customers' shopping experience better.

#10 Boost sales with smart upselling

Discovering the ideal upsell product can feel like a refreshing oasis in the middle of a scorching desert. It's a game-changer that can significantly boost your revenue, but it often takes time and effort to find the right fit.

If you suggest a well-timed and relevant product at the end of the shopping process, it can really boost your sales. To identify the perfect upsell product, conducting A/B testing is a reliable way to see what resonates best with your customers.

Using click heatmap analysis and A/B testing can help you make your upsell offers better and get more sales. By planning carefully and paying attention to details, you can increase your cart value and make more money for your business.

#11 Guide users on their journey

It's really important to make sure your website is easy to use for your customers. If it's not clear what they should do or how to do it, they might leave without buying anything or signing up for your newsletter. This can be a nightmare for any business owner because it means fewer people will do what you want them to do, which can also hurt your revenue. So, make sure you have clear instructions and a simple design that helps your customers know what actions to take on your website.

On the other hand, when you provide a seamless and straightforward user experience, you can generate leads and build a loyal customer base. If you want to make your website easy to use for your customers, you need to know how they behave on your site. This means identifying any places where they might have trouble.

If you do your research and analyze your website's data, you can make better decisions that will improve your customers' experience. This will help you be more successful overall.

#12 Fine-tune font styles and sizes

The designers' community website, Dribbble.com estimates that 85% of fonts are sans-serif, with the rest being serif, monospaced, and everything else.

Picking the right font for your website is super important if you want it to look good and be trustworthy. It's best to use fonts that are easy to read because difficult ones can quickly turn potential customers away from your site.

The font you choose for your website can affect how people see your brand. If you pick the wrong one, it could make you look bad. As an example, imagine a criminal lawyer's website that uses the font, Comic Sans. It wouldn't seem professional and could hurt their reputation.

Thus, it's essential to select a font that aligns with your brand's personality and target audience. To ensure that your font choice resonates with your users, conduct thorough testing to see how they respond. Will they leave the site? Will they continue browsing? Or will they make a buy?

When you're selecting a font, keep these factors in mind to make sure it enhances your users' experience and reinforces your brand identity.

#13 Design engaging pop-ups

The key is to determine your goals and create pop-ups that align with them. For example, if your goal is to capture email addresses for your newsletter, a pop-up can be an effective tool to achieve this. A/B testing is also a great way to optimize your pop-ups and ensure that they are as effective as possible.

To make your pop-ups work better, try out different designs, messages, and timing until you find what works best for your business.

+1 Create a seamless onboarding experience

When new users visit your site, they may feel a bit disoriented, especially if your site has a more intricate or unusual interface. To keep people interested and using your product or service, it's important to make the first steps as easy as possible.

As we've discussed before, complicated navigation and unclear instructions can quickly deter customers. This is especially true for new users who may feel lost when trying to discover your site. Make things easier for your users by being helpful and considerate. Use pop-ups, information points, and clear instructions to guide them along the way.

A/B testing and heatmaps can also help you design and optimize these elements for the best possible onboarding experience.

How can you achieve your goals?

You read a lot about heatmaps, but are still not sure how to get started. Capturly's heatmap tool can help you to make your dreams happen. Boost your website's conversion rates and get the most out of your online presence

Capturly's heatmap helps you see how people use your website with heatmaps such as scroll, click, and segment heatmaps. With A/B testing, you can figure out what changes on your website make people more interested and likely to buy something.

Check out our Heatmap Guide to learn more about heatmaps and how they can improve your website. It has everything you need to know! We're committed to helping you achieve your goals with ease, so don't hesitate to reach out if you have any questions or need help.

Conclusion

To make a website that's successful and makes a lot of money, you need to regularly review and test everything on it. That means things like the design, layout, text, images, and font all need to be checked. Continuous A/B testing can make decision-making easier and streamline website optimization. By utilizing different heatmaps during A/B testing, you can easily identify areas where you may be losing or retaining users.

Heatmaps can show you important things you might not have noticed if you were only looking at the numbers. Remember, staying on top of these details can make a significant difference in the overall success of your website.

Source: https://splitmetrics.com/resources/minimum-detectable-effect-mde/

Minimum Detectable Effect (MDE)

1. What is the Minimum Detectable Effect (MDE)?

It's a minimum improvement over the conversion rate of the existing asset (baseline conversion rate) that you want the experiment to detect.

By setting MDE, you define the conversion rate increase sufficient for the system to declare the new asset a winner. The lower MDE you set, the slighter conversion changes will be detected by the system. Basically, MDE measures the experiment sensitivity.

Highly sensitive settings, or low MDE, come along with a big sample size. The lower the MDE is, the more traffic you need to detect minor changes, hence the more money you have to spend on driving that traffic.

So, by configuring MDE you are flexible about connecting the experiment design with the costs you

are ready to incur.

2. Why is setting the right MDE so important?

Minimum detectable effect is a crucial parameter for evaluating the cost and potential return on running A/B experiments. From the practical perspective of mobile app marketers, choosing the appropriate value of MDE means striking the balance between the cost of acquiring paid traffic for an experiment and achieving meaningful return on investment.

To put it in the most practical terms for SplitMetrics Optimize users: by setting a lower target MDE, you're instructing the system to collect more views of your app's product pages before marking the test as statistically significant and completing it.

3. What's the optimal MDE?

There's no such thing as an ideal MDE, so **SplitMetrics Optimize** can't recommend you the optimal value. This is a key custom parameter affecting your sample size and, by implication, the costs associated with the traffic. In other words, we suggest you define MDE by yourself, taking into consideration your individual risks – money and time.

This may sound like an overwhelming task, but don't worry – if this guide and other, freely available materials on our blog are just too much for you, we recommend turning to the SplitMetrics Agency. Using their experience they will recommend the optimal MDE for your experiments – among many other things to make your mobile app grow.

4. How does MDE affect my sample size?

MDE has a dramatic effect on the amount of traffic required to reach statistical significance. To know your maximum sample size, use the Evan Miller calculator for sequential A/B sampling. Make sure that you insert relative value for MDE rather than absolute.

By setting **smaller MDE**, you tell the system to detect slighter conversion rate changes, which requires more traffic and possibly time. On the other hand, the **larger MDE** you set, the less traffic (and possibly time) is required to finish the test.

For example, to reach the significance level of 5%, you'll require 2,922 total conversions with MDE = 10%. with MDE = 5% the sample size grows up to 11,141 total conversions.

Remember, that the sample size you see is only the maximum threshold required for a statistically significant result. Due to the nature of sequential A/B testing, the system will constantly check the difference between conversion rates of variations under testing. Once the difference is found, the test is finished and there's no need to score the entire sample size.

5. How to calculate the minimum detectable effect?

Although our system can count MDE for you, we strongly recommend setting it by yourself. This parameter depends on your own risks – *money* you're ready to allocate for the traffic acquisition and *time* you can wait for the experiment to run.

To get MDE that works for you, you have to understand:

- **Traffic acquisition costs**, or the money you invest in driving the required sample size;
- **Potential revenue** generated from ASO-acquired users, in other words, the money you can make from using the new asset with a higher conversion rate.

The best possible MDE implies that the potential revenue exceeds or compensates for the traffic acquisition costs.

The minimum detectable effect formula

Minimum detectable effect is calculated as a percent of the baseline conversion rate:

MDE = (desired conversion rate lift / baseline conversion rate) x 100%

Its practical application is discussed in the workflow below.

MDE calculation workflow

Step 1. Estimate the desired conversion rate lift

Let's say the conversion rate of your product page with the existing icon is 20% (baseline conversion rate). You assume that the new icon should have at least a 22% conversion rate for you to use it instead of the existing icon.

So, you have to configure an experiment in such a way that it declares the winner when the conversion rate difference is at least 22% – 20% = 2%. To set that up, you have to count your *estimated* MDE.

MDE is calculated as a percent of the baseline conversion rate:

MDE = desired conversion rate lift / baseline conversion rate x 100%

In this example, 2% of the 20% baseline conversion rate is 10% – this is your *estimated MDE* for the experiment.

Step 2. Calculate your sample size

Next step is to get your sample size, using the Evan Miller's calculator for sequential A/B testing.

- Sample size calculation for **experiments with two variations (A+B).** Here's what you should insert in the calculator:

1. Your *estimated* Minimum Detectable Effect: 10% (in this example). **Important!** Make sure that you use **relative** MDE.
2. Insert any value in the "Baseline conversion rate" field. As we use relative MDE, the baseline conversion rate is ignored in the sample size calculation;
3. Statistical power: 80% (default in SplitMetrics Optimize);
4. Significance level: 5% (default in SplitMetrics Optimize).

You will see the following:

Control wins if: 2,922 total conversions – this is the maximum sample size per two variations (A+B) needed to finish the experiment.

Treatment wins: 106 conversions ahead – means that the system will sequentially check the difference in conversions between variation A (control) and B, and may finish the experiment once the difference of 106 is found, even before reaching the maximum sample size.

- Sample size calculation for **experiments with multiple variations (A+B+C+…)**.

1. Here's what you should insert in the calculator:
1. *Estimated* MDE;
2. Any value as the baseline conversion rate;
3. Statistical power: 80%;
4. **Significance level:** when more than two variations are tested, you have to apply the Sidak correction to the significance level.

Why? Each pair of variations has its individual significance level. As the number of variations under testing grows, so does the overall significance level because those individual values accumulate. The Sidak correction balances out individual significance levels so that the overall significance level equals 5%.

To apply the Sidak correction, use the following significant level values:

Number of variations under testing (incl. control)	Significance level to set
3 (A+B+C)	3%

4 (A+B+C+D)	2%
5 (A+B+C+D+E)	1%

2. Get the total conversions.

The total conversions will appear after you insert all the above in the calculator.

For example, you want to run an experiment with 3 variations – A+B+C. Things you'll insert in the calculator will be:

1. *Estimated* MDE: 10% (can be another value, depending on the conversion rate lift you want the system to detect);
2. Statistical power: 80%
3. Significance level: 3%

Total conversions required: 3,472

3. Divide the total conversions by 2.

In the above example with 3 variations, you'll get:

total conversions / 2 = 3,472 / 2 = 1,736

4. Multiply the result received in step 3 by the number of variations, including control:

- For A+B+C: total conversions / 2 * 3
- For A+B+C+D: total conversions / 2 * 4
- For A+B+C+D+E: total conversions / 2 * 5

Back to the example, as you run an A+B+C experiment, 3 will be your multiplier:

total conversions / 2 * 3 = 5,208

5,208 is the rough estimation of the maximum sample size for an experiment with 3 variations (A+B+C).

Step 3. Calculate your traffic acquisition costs

In step 2, we've calculated the maximum required conversions for an experiment with two variations (A+B) – 2,922. Now that you know the maximum required sample size, you can calculate the possible **traffic acquisition costs.** Use this formula:

traffic acquisition costs = total conversions / baseline conversion rate * Cost per Click

Note: By dividing the total conversions by your baseline conversion rate you gauge your sample size **in visitors** (those who click on your ad banner).

Let's say your Cost per Click is $0.5 and the baseline conversion rate is 20% (convert it to the decimal form to use in the formula). Your traffic acquisition costs will be:

2,922 / 0.2 * $0.5 = $7,305

When you have SplitMetrics Optimize integrated with Facebook Pixel, you may configure “Complete Registration” as a conversion event. In such a case, the traffic acquisition costs will be calculated considering users who click on the “Get” button rather than those who click on an ad banner.

The formula for cost calculations in such cases will include Cost per Install (not Cost per Click):

traffic acquisition costs = total conversions x CPI

Note: As you can see, you don’t have to recalculate sample size in visitors. Just multiply CPI by the total conversions obtained in the Evan Miller calculator.

Let’s say your Cost per Install is $2.5 and the maximum sample size is 2,922 total conversions. Your traffic acquisition costs will be:

2,922 X $2.5 = $7,305

At this point, you have to make sure that these costs line up with the budget allocated for the traffic acquisition:

- **If no**t, set a bigger MDE, which will require a smaller sample size, hence smaller acquisition costs; however, if your variations are quite similar, a big value for MDE won’t be able to deliver a result.
- **If yes**, proceed with calculating the potential revenue generated from ASO-acquired users.

Step 4. Calculate the potential revenue

You may use different ways to calculate the potential revenue from the conversion rate lift, for example, based on the LTV of ASO-acquired app subscribers. In the above described example with two variations (A+B), you have to calculate how much money you will generate from a 2% conversion rate lift.

Once you have your Potential revenue ($Y) calculated, compare it with the Traffic acquisition costs ($X):

- if the potential revenue is greater than the traffic acquisition costs ($Y > $X), you can go with the *estimated* MDE in your experiment;
- if the traffic acquisition costs exceed the potential revenue ($Y < $X), estimate a bigger MDE and repeat all the steps above.

6. At what experiment stage should I set MDE?

MDE is configured after the experiment is created but before you start driving traffic. If you change your MDE after the traffic starts driving to the experiment, you will lose all the statistics.

7. Can I change my MDE during the experiment?

Don’t modify MDE after you start driving traffic to your experiment. Otherwise, all the statistics – visitors, conversions, improvement, etc. – will be reset.

8. MDE best practices

Evaluating the Minimum Detectable Effect (MDE) in mobile app marketing involves careful consideration of statistical and practical factors. Here are a couple of best practices to keep in mind:

Understand your business goals, baseline conversion rate and other metrics

Before determining the minimum detectable effect and configuring your experiments, have a clear understanding of your business goals and the key performance indicators (KPIs) that matter most to your mobile app's success. Without them, you won't be able to successfully establish the intersection between statistical significance and business relevance. Similarly, benchmarking helps you evaluate where you actually stand in terms of performance and whether further optimization is viable.

Balance statistical significance and practical impact

Strive for a balance between statistical significance and practical impact. A statistically significant result might not be practically meaningful if the observed effect is too small to matter to your business goals. On the other hand, setting the bar too high may make the experiment financially unviable. Aim to set an MDE that ensures changes are noticeable and provide a tangible benefit to users, but never lose focus on your ROI.

Consider context and differences in user behavior

Mobile app user behavior can vary significantly from other online platforms. Users on Google Play can display behavioral patterns much different to those on the App Store. Experiments for the same creatives can return different results. Consider the unique characteristics of mobile usage, such as attention spans and on-the-go interactions. Recognize that the MDE might be different for various user segments and app stores. This is well evidenced in our reports: the ASO Benchmarks & Mobile Trends Report and How Users Behave on the App Stores: the App Store vs. Google Play.

Use calculators or rely on SplitMetrics Optimize for sample size calculation

Leverage sample size calculators or statistical software to estimate the sample size needed to detect your desired minimum detectable effect with a specified level of confidence and power. These calculators take into account factors like baseline metric values, variability, desired significance level (alpha), and desired power. You may also rely on SplitMetrics Optimize to automate the process of setting the right value for MDE and calculating sample size. We highly recommend you read our guide on Calculating Sample Size for A/B Testing: Formulas, Examples & Errors.

9. How does SplitMetrics Optimize calculate my MDE?

To arrive at your best possible MDE, our algorithm will rely on your baseline conversion. Your ideal MDE will be the value which produces a sufficiently large sample size, yet comparable to that in classic A/B testing.

If SplitMetrics Optimize calculates MDE for you, be aware that the result won't appear straight away. The algorithm will gauge and display your MDE in the interface after your variations gain enough conversions.

Source: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

A normal Q–Q plot of randomly generated, independent standard exponential data, (X ~ Exp(1)). This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points follow a strongly nonlinear pattern, suggesting that the data are not distributed as a standard normal (X ~ N(0,1)). The offset between the line and the points suggests that the mean of the data is not 0. The median of the points can be determined to be near 0.7

A normal Q–Q plot comparing randomly generated, independent standard normal data on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points suggests that the data are normally distributed.

A Q–Q plot of a sample of data versus a Weibull distribution. The deciles of the distributions are shown in red. Three outliers are evident at the high end of the range. Otherwise, the data fit the Weibull(1,2) model well.

A Q–Q plot comparing the distributions of standardized daily maximum temperatures at 25 stations in the US state of Ohio in March and in July. The curved pattern suggests that the central quantiles are more closely spaced in July than in March, and that the July distribution is skewed to the left compared to the March distribution. The data cover the period 1893–2001.

In statistics, a Q–Q plot (quantile–quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally more diagnostic than comparing the samples' histograms, but is less widely known. Q–Q plots are commonly used to compare a data set to a theoretical model.[2][3] This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.[4] Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

The term "probability plot" sometimes refers specifically to a Q–Q plot, sometimes to a more general class of plots, and sometimes to the less commonly used P–P plot. The probability plot correlation coefficient plot (PPCC plot) is a quantity derived from the idea of Q–Q plots, which measures the agreement of a fitted distribution with observed data and which is sometimes used as a means of fitting a distribution to data.

Definition and construction

Q–Q plot for first opening/final closing dates of Washington State Route 20, versus a normal distribution.[5] Outliers are visible in the upper right corner.

A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two distributions being compared, some of the quantiles may not be defined, so an interpolated quantile may be plotted. If the Q–Q plot is based on data, there are multiple quantile estimators in use. Rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions.

A simple case is where one has two data sets of the same size. In that case, to make the Q–Q plot, one orders each set in increasing order, then pairs off and plots the corresponding values. A more complicated construction is the case where two data sets of different sizes are being compared. To construct the Q–Q plot in this case, it is necessary to use an interpolated quantile estimate so that quantiles corresponding to the same underlying probability can be constructed.

More abstractly,[4] given two cumulative probability distribution functions F and G, with associated quantile functions F−1 and G−1 (the inverse function of the CDF is the quantile function), the Q–Q plot draws the q-th quantile of F against the q-th quantile of G for a range of values of q. Thus, the Q–Q plot is a parametric curve indexed over [0,1] with values in the real plane R2.

Typically for an analysis of normality, the vertical axis shows the values of the variable of interest, say x with CDF F(x), and the horizontal axis represents N−1(F(x)), where N−1(.) represents the inverse cumulative normal distribution function.

Interpretation

The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q–Q plot follows the 45° line y = x. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line y = x. If the general trend of the Q–Q plot is flatter than the line y = x, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis. Conversely, if the general trend of the Q–Q plot is steeper than the line y = x, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. Q–Q plots are often arced, or S-shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

Although a Q–Q plot is based on quantiles, in a standard Q–Q plot it is not possible to determine which point in the Q–Q plot determines a given quantile. For example, it is not possible to determine the median of either of the two distributions being compared by inspecting the Q–Q plot. Some Q–Q plots indicate the deciles to make determinations such as this possible.

The intercept and slope of a linear regression between the quantiles gives a measure of the relative location and relative scale of the samples. If the median of the distribution plotted on the horizontal

axis is 0, the intercept of a regression line is a measure of location, and the slope is a measure of scale. The distance between medians is another measure of relative location reflected in a Q–Q plot. The "probability plot correlation coefficient" (PPCC plot) is the correlation coefficient between the paired sample quantiles. The closer the correlation coefficient is to one, the closer the distributions are to being shifted, scaled versions of each other. For distributions with a single shape parameter, the probability plot correlation coefficient plot provides a method for estimating the shape parameter – one simply computes the correlation coefficient for different values of the shape parameter, and uses the one with the best fit, just as if one were comparing distributions of different types.

Another common use of Q–Q plots is to compare the distribution of a sample to a theoretical distribution, such as the standard normal distribution N(0,1), as in a normal probability plot. As in the case when comparing two samples of data, one orders the data (formally, computes the order statistics), then plots them against certain quantiles of the theoretical distribution.[3]

Plotting positions

The choice of quantiles from a theoretical distribution can depend upon context and purpose. One choice, given a sample of size n, is k / n for k = 1, …, n, as these are the quantiles that the sampling distribution realizes. The last of these, n / n, corresponds to the 100th percentile – the maximum value of the theoretical distribution, which is sometimes infinite. Other choices are the use of (k − 0.5) / n, or instead to space the n points such that there is an equal distance between all of them and also between the two outermost points and the edges of the

[

0

,

1

]

{\displaystyle [0,1]} interval, using k / (n + 1).[6]

Many other choices have been suggested, both formal and heuristic, based on theory or simulations relevant in context. The following subsections discuss some of these. A narrower question is choosing a maximum (estimation of a population maximum), known as the German tank problem, for which similar "sample maximum, plus a gap" solutions exist, most simply m + m/n − 1. A more formal application of this uniformization of spacing occurs in maximum spacing estimation of parameters.

Expected value of the order statistic for a uniform distribution

The k / (n + 1) approach equals that of plotting the points according to the probability that the last of (n + 1) randomly drawn values will not exceed the k-th smallest of the first n randomly drawn values.[7][8]

Expected value of the order statistic for a standard normal distribution

In using a normal probability plot, the quantiles one uses are the rankits, the quantile of the expected value of the order statistic of a standard normal distribution.

More generally, Shapiro–Wilk test uses the expected values of the order statistics of the given distribution; the resulting plot and line yields the generalized least squares estimate for location and scale (from the intercept and slope of the fitted line).[9] Although this is not too important for the normal distribution (the location and scale are estimated by the mean and standard deviation, respectively), it can be useful for many other distributions.

However, this requires calculating the expected values of the order statistic, which may be difficult if the distribution is not normal.

Median of the order statistics

Alternatively, one may use estimates of the median of the order statistics, which one can compute based on estimates of the median of the order statistics of a uniform distribution and the quantile function of the distribution; this was suggested by Filliben (1975).[9]

This can be easily generated for any distribution for which the quantile function can be computed, but conversely the resulting estimates of location and scale are no longer precisely the least squares estimates, though these only differ significantly for n small.

Heuristics

Several different formulas have been used or proposed as affine symmetrical plotting positions. Such formulas have the form (k − a) / (n + 1 − 2a) for some value of a in the range from 0 to 1, which gives a range between k / (n + 1) and (k − 1) / (n − 1).

Expressions include:

k / (n + 1)

(k − 0.3) / (n + 0.4).[10]

(k − 0.3175) / (n + 0.365).[11][note 1]

(k − 0.326) / (n + 0.348).[12]

(k − ⅓) / (n + ⅓).[note 2]

(k − 0.375) / (n + 0.25).[note 3]

(k − 0.4) / (n + 0.2).[citation needed]

(k − 0.44) / (n + 0.12).[note 4]

(k − 0.5) / n.[14]

(k − 0.567) / (n − 0.134).[citation needed]

(k − 1) / (n − 1).[note 5]

For large sample size, n, there is little difference between these various expressions.

Filliben's estimate

The order statistic medians are the medians of the order statistics of the distribution. These can be expressed in terms of the quantile function and the order statistic medians for the continuous uniform distribution by:

N

(

i

)

=

G

(

U

(

i

)

)

{\displaystyle N(i)=G(U(i))}

where U(i) are the uniform order statistic medians and G is the quantile function for the desired distribution. The quantile function is the inverse of the cumulative distribution function (probability that X is less than or equal to some value). That is, given a probability, we want the corresponding quantile of the cumulative distribution function.

James J. Filliben uses the following estimates for the uniform order statistic medians:[15]

m

(

i

)

=

{

1

−

0.5

1

/

n

i

=

1

i

−

0.3175

n

+

0.365

i

=

2

,

3

,

…

,

n

−

1

0.5

1

/

n

i

=

n

.

{\displaystyle m(i)={\begin{cases}1-0.5^{1/n}&i=1\\\\{\dfrac {i-0.3175}{n+0.365}}&i=2,3,\ldots ,n-1\\\\0.5^{1/n}&i=n.\end{cases}}}

The reason for this estimate is that the order statistic medians do not have a simple form.

Software

The R programming language comes with functions to make Q–Q plots, namely qqnorm and qqplot from the stats package. The fastqq package implements faster plotting for large number of data points.

Source: https://www.optimizely.com/sample-size-calculator/#/?conversion=3&effect=20&significance=95

A/B test sample size calculator

Powered by Optimizely Experimentation's stats engine

Baseline Conversion Rate

Your control group's expected conversion rate: 3 % (put desired percent you want)

Minimum Detectable Effect

The minimum relative change in conversion rate you would like to be able to detect: 20% (put desired percent you want)

Statistical Significance

95% is an accepted standard for statistical significance, although Optimizely allows you to set your own threshold for significance based on your risk tolerance: 95 % (put desired percent you want)

Sample size per variation: 13,000

Sample size calculator

What is Experimentation stats engine?

Optimizely's sample size calculator is different from other statistical significance calculators. It is based on the formula used in Experimentation's stats engine. Stats engine calculates statistical significance using sequential testing and false discovery rate controls. When combined, these two techniques mean you no longer need to wait for a pre-set sample size to ensure the validity of your results. If Intelligence Cloud tells you that a result is 95% significant, you can make a decision with 95% confidence.

How many visitors do I need for my A/B test?

This statistical significance calculator allows you to calculate the sample size for each variation in your test you will need, on average, to measure the desired change in your conversion rate. In many cases, if Intelligence Cloud detects an effect larger than the one you are looking for, you will be able to end your test early.

Why is your calculator different from other sample size calculators?

Our A/B test sample size calculator is powered by the formula behind our new Stats Engine, which uses a two-tailed sequential likelihood ratio test with false discovery rate controls to calculate statistical significance.

With this methodology, you no longer need to use the sample size calculator to ensure the validity of your results. Instead, the A/B test calculator is best used as a tool for planning out your testing program to find out how long you may need to wait before Optimizely can determine whether your results are significant, depending on the effect you want to observe.

How do I determine the baseline conversion rate?

You can look at historical data on how this page has typically performed in the past, from a tool like Google Analytics or other website analytics you use.

What is minimum detectable effect (MDE)

In traditional hypothesis testing, the MDE is essentially the sensitivity of your test. In other words, it is the smallest relative change in conversion rate you are interested in detecting. For example, if your baseline conversion rate is 20%, and you set an MDE of 10%, your test would detect any changes that move your conversion rate outside the absolute range of 18% to 22% (a 10% relative effect is a 2% absolute change in conversion rate in this example).

How do I estimate a value for minimum detectable effect (MDE)?

Decide how willing you are to trade off sensitivity of your test versus how long you might need to run your test for. The smaller the MDE, the more sensitive you are asking your test to be, and the larger sample size you will need.

Keep in mind that statistical significance in Intelligence Cloud's stats engine shows you the chance that your results will ever be significant, while the experiment is running. If the effect that our Stats Engine observes is larger than the minimum detectable effect you are looking for, your test may declare a winner or loser up to twice as fast as if you had to wait for your pre-set sample size. Given more time, stats engine may also find a smaller MDE than the one you expect. Learn more

Where is statistical power in your sample size calculator?

Statistical power is essentially a measure of whether your test has adequate data to reach a conclusive result. Intelligence Cloud's stats engine runs tests that always achieve a power of one, meaning that the test always has adequate data to show you results that are valid at that moment, and will eventually detect a difference if there is one. This means that you can make a decision as soon as your results reach significance without worrying about power.

Source: https://medium.com/data-science/required-sample-size-for-a-b-testing-6f6608dd330a

This article describes some popular approaches for calculating minimum required sample size for A/B testing. There is nothing novel about the approaches described in this article. However, it will provide practical guidelines (with rigor) for data scientists who are new to the field. It will also help job candidates excel in data science interviews. Besides, the article reveals the math behind some popular online sample size calculators, too. The formulas used in this article are from *Fundamentals of Biostatistics (8th edition)*.

Why to calculate required sample size?

In A/B testing, we are often interested in testing if the treatment group is significantly different from the control group in a certain success metric (e.g., conversion rate). The null hypothesis is that there is no significant difference.

Type I error happens when we reject the null hypothesis when it should not be rejected. Type I error rate is the probability when Type I error happens, also known as significance level, or alpha. A common value for alpha is 0.05.

Type II error happens when we fail to reject the null hypothesis when it should be rejected. Type II error rate is also known as beta.

Statistical power is the probability that the test rejects the null hypothesis when it should be rejected. It is basically 1 minus beta. A common value for statistical power is 0.80 (so beta is 0.20).

In order to obtain meaningful results, we want our test to have sufficient statistical power. And, sample size influence statistical power. For example, when comparing two means, the follow formula can be used to calculate statistical power. As sample size increases, the statistical power increases. Therefore, for our test to have desirable statistical power (usually 0.80), we want to estimate the minimum sample size required.

Power for Comparing the Means of Two Normally Distributed Samples Using a Significance Level α

To test the hypothesis $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$ for the specific alternative $|\mu_1 - \mu_2| = \Delta$, with significance level α,

$$\text{Power} = \Phi\left(-z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right)$$

where (μ_1, σ_1^2), (μ_2, σ_2^2) are the means and variances of the two respective groups and n_1, n_2 are the sample sizes of the two groups.

(From Fundamentals of Biostatistics)

Next, we will understand the calculations of required sample size through a hypothetical example.

Example: Conversion Rate of an E-Commerce Website

Suppose an e-commerce website wants to test if implementing a new feature (e.g., layout or button) will significantly improve conversion rate (number of purchases divided by number of sessions/visits). We can randomly show the new webpage to 50% of the users. Then, we have a test group and a control group. Once we have enough data points, we can test if the conversion rate in

the treatment group is significantly higher (one side test) than that in the control group. The null hypothesis is that conversion rates are not significantly different in the two group.

Sample Size for Comparing Two Means

One way to perform the test is to calculate daily conversion rates for both the treatment and the control groups. Since the conversion rate in a group on a certain day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period.

The formula for estimate minimum sample size is as follows:

Sample Size Needed for Comparing the Means of Two Normally Distributed Samples of Equal Size Using a Two-Sided Test with Significance Level α and Power $1-\beta$

$$n = \frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\Delta^2} = \text{sample size for each group}$$

where $\Delta = |\mu_2 - \mu_1|$. The means and variances of the two respective groups are (μ_1, σ_1^2) and (μ_2, σ_2^2).

(From Fundamentals of Biostatistics)

For our example, let's assume that the mean daily conversion rate for the past 6 months is 0.15 and the sample standard deviation is 0.05. With the new feature, we expect to see a 3% absolute increase in conversion rate. Thus, for the conversion rate for the treatment group will be 0.18. We also assume taht the sample standard deviations are the same for the two group. Our parameters are as follows.

- mu1 = 0.15
- mu2 = 0.18
- sigma1 = sigma2 = 0.05

Assuming alpha = 0.05 and beta = 0.20 (power = 0.80), applying the formula, the required minimum sample size is 35 days. This is consistent with the result from this web calculator.

Sample Size for Comparing Two Proportions

The two-means approach considers each day+group as a data point. But what if we focus on individual users and visits? What if we want to know how many visits/sessions are required for the testing? In this case, the conversion rate for a group is basically all purchases divided by all sessions in that group. If each session is a Bernoulli trial (convert or not), each group follows a binomial distribution. To test the difference in conversion rate between the treatment and control groups, we need a test of two proportions. The formula for estimating the minimum required sample size is as follows.

Sample Size Needed to Compare Two Binomial Proportions Using a Two-Sided Test with Significance Level α and Power $1 - \beta$, Where One Sample (n_2) Is k Times as Large as the Other Sample (n_1) (Independent-Sample Case)

To test the hypothesis $H_0: p_1 = p_2$ vs. $H_1: p_1 \neq p_2$ for the specific alternative $|p_1 - p_2| = \Delta$, with a significance level α and power $1 - \beta$, the following sample size is required

$$n_1 = \left[\sqrt{\bar{p}\,\bar{q}\left(1+\frac{1}{k}\right)}\, z_{1-\alpha/2} + \sqrt{p_1 q_1 + \frac{p_2 q_2}{k}}\, z_{1-\beta}\right]^2 \Big/ \Delta^2$$

$$n_2 = k n_1$$

where p_1, p_2 = projected true probabilities of success in the two groups

$$q_1, q_2 = 1 - p_1, 1 - p_2$$

$$\Delta = |p_2 - p_1|$$

$$\bar{p} = \frac{p_1 + k p_2}{1 + k}$$

$$\bar{q} = 1 - \bar{p}$$

(From Fundamentals of Biostatistics)

Assuming 50–50 split, we have the following parameters:

- p1 = 0.15
- p2 = 0.18
- k = 1

Using alpha = 0.05 and beta = 0.20, applying the formula, the required sample size is 1,892 sessions per group. This is consistent with the results from this web calculator. This is also close to the result using Optimizely's sample size calculator for conversion rate.

Source: https://guessthetest.com/calculating-sample-size-in-a-b-testing-everything-you-need-to-know/

How to Correctly Calculate Sample Size in A/B Testing

By: Deborah O'Malley, M.Sc | Last updated December, 2023

What is sample size?

Very simply stated, in A/B testing, sample size describes the number of visitors you need to accurately run a valid test.

The group of users who take part in your experiment comprise the sample population.

When running an experiment, it's important to have a large enough sample so the portion of users accurately represents your entire audience.

If your sample size is too small, your test results will not be adequately powered to detect a meaningful effect. In other words, the results may appear exaggerated, inaccurate, and will not truly represent how your entire audience actually behaves.

How large does your sample size need to be to run a valid A/B test?

What's the minimum sample size needed to run an accurate A/B test?

It's such a simple question -- with a very difficult answer.

Ask 100 skilled testers their thoughts and they'll all tell you the same thing: it depends!

In short, the larger the sample size, the better.

And, as a result, the more certain you can be that your test findings are representative and truly reflect your overall population.

But that answer really doesn't really help at all. Does it?

The problem with trying to calculate sample size

The problem is, the population you're sampling needs to, in theory, be representative of the entire audience you're trying to target -- which is of course an abstract and impossible task.

You can't truly ever tap every single individual in your entire audience. Especially since, over time, your audience will presumably grow and change.

But what you can do is capture a broad, or large enough slice of your audience to get a reasonably accurate picture of how most users are likely to behave.

There will, of course, always be outliers, or users who behave entirely differently than anyone else.

But, again, with a large enough sample, these individual discrepancies become smoothed out and larger patterns become apparent -- like whether your audience responds better to version A or B.

So how big does my sample size need to be?

Keeping in mind, the actual answer is, "it depends."

A general rule of thumb is: for a highly reliable test, you need a minimum of 30,000 visitors and 3,000 conversions per variant.

If you follow this guideline, you'll generally achieve enough traffic and conversions to derive statistically significant results at a high level of confidence.

However, any testing purist will balk at this suggested guideline and tell you, you absolutely must calculate your sample size requirements.

How do you calculate sample size requirements?

A bit of good news.

If you're able to wrap your head around a few statistical terms, it's actually relatively easy to calculate your A/B test sample size requirements.

The fastest and most effective way to do so is using a sample size calculator.

Using a sample size calculator

There are many free, online A/B test sample size calculators available.

Some require more detailed information than others, but all will yield the same answer: the minimum number of people you need to yield accurate A/B test results.

My Favorite sample size calculator

My favorite, go-to sample size calculator is Evan Miller's calculator because it is so flexible and thorough.

You have the ability to tweak a number of different parameters to yield an accurate sample size calculation.

But, there are many calculators out there. Here are some other good ones:

- Optimizely
- CXL
- Unbounce
- Convertize

Every calculator will require you to plug-in different inputs or numbers.

This article specifically focuses on the inputs required in Evan Miller's calculator because it's one of the most complex.

If you can competently use his calculator, you can accurately use any.

And here's where it gets a bit tricky. . .

Sample size calculator terminology

In order to properly use this calculator, you need to understand every aspect of it. Let's explore what each term means:

Baseline conversion rate:

The conversion rate for your control or original version. (Usually labelled "version A" when setting up an A/B test).

You should be able to find this conversion rate within your analytics platform.

If you don't have or don't know the baseline conversion rate, make your best educated guess.

For an eCommerce site, the average conversion rate across desktop and mobile hovers between 2-5%. So, if you're at a loss, plug in a number between 2-5%.

If you want to be conservative in your guess, go with the lower end -- which will push your sample size requirements up, and vise versa.

Minimum Detectable Effect (MDE):

The MDE sounds complicated, but it's actually quite simple if you break the concept down into each of the three terms:

- *Minimum* = smallest
- *Effect* = conversion difference between the control and treatment
- Detectable = want you want to see from running the experiment

Therefore, the minimum detectable effect is the smallest conversion lift you're hoping to achieve.

Unfortunately, there's no magic number for your MDE. Again, it depends.

What does it depend on?

As this article explains, a few things:

- ***Historical data:*** observations you've made overtime that show, in general, most tests tend to achieve a certain lift, so this one should too.
- ***What's worth it:*** a number you choose, based on what you consider *worth it* to take the time and resources to run the experiment. For example, a testing agency may, by default, set the MDE at 10% for every experiment at because that's the minimum needed to declare a win for the client.
- ***Organizational maturity:*** a large, mature testing organization, with a lot of traffic, may set the MDE at 1-3% because, through ongoing optimization, getting gains any higher would be unrealistic.

As a very general rule of thumb, an MDE of 2-5% is reasonable.

Therefore, if you don't have enough data to historically inform your MDE, plug in a range between 2-5%.

If you don't have power to detect an MDE of 5%, the test results aren't trustworthy. The larger the organization and traffic, the sample the MDE will likely be.

In, and in that tune, the smaller the effect, the bigger your sample size needed.

An MDE can be expressed as an absolute or relative amount.

Absolute:

The actual raw number difference between the conversion rates of the control and variant.

For example, if the baseline conversion rate is 0.77% and you're expecting the variant to achieve a MDE of ±1%, the absolute difference is 0.23% (Variant: 1% - Control: 0.77% = 0.23%) OR 1.77% (Variant 1% + Control 0.7&% = 1.77%).

Relative:

The percentage difference between the baseline conversion rate and the MDE of the variant.

For example, if the baseline conversion rate is 0.77% and you're expecting the variant to achieve a MDE of ±1%, the relative difference between the percentages is 29.87% (increase from Control: 0.77% to Variant: 1% = 29.87% gain) or -23% (decrease from Control: 0.77% to Variant 1% =-23%).

In general, clients are used to seeing a relative percentage lift, so it's typically best way to use a relative percentage calculation and report results this way.

Statistical power 1−β:

Very simply stated, this concept is the probability of finding an "effect," or difference between the performance of the control and variant(s), assuming there is one.

A power of 0.80 is considered standard best practice. So, you can leave it as the default range on this calculator.

A power of 0.80 means there's an 80% chance that, if there is an effect, you'll accurately detect it without error. Meaning there's only a 20% chance you'd miss properly detecting the effect. A risk worth taking.

In testing, your aim is to ensure you have enough power to meaningfully detect a difference in conversion rates. Therefore, a higher power is always better. But the trade-off is, it requires a larger sample size.

In tune, the larger your sample size, the higher your power which is part of why a large sample size is required for accurate A/B testing.

Significance Level α:

Significance level α, also called "alpha (α)" is like a check point set before calling a test win or loss.

It acts as a tool to control how often we make incorrect conclusions.

As a very basic definition, significance level alpha represents the probability of committing a type I error -- which happens when there's a false positive, or think you've spotted a conversion difference that doesn't actually exist.

The closer to 0, the lower the probability of a type I error/false positive, but the higher the probability of a type II error/false negative.

Therefore, a happy middle ground, and commonly accepted level for α is 0.05.

This level mean accepting a (0.05) 5% chance of a false positive; which in turn, suggests there is a 95% chance the null hypothesis is correct and the treatment is no better than the control.

Calculating sample size ahead of running an A/B test

Once you've used the sample size calculator and crunched the numbers, you're ready to start running your A/B test.

Now, you might be thinking, wait a minute. . . Why do I need to calculate the sample size I need BEFORE I run my A/B test?

The answer is two-fold:

1. So you know you have a large enough sample size for your test to be adequately powered to accurately detect a meaningful effect.
2. So you don't prematurely stop the test and incorrectly declare a winner before one has truly emerged.

Remember: you need a large enough sample size, or amount of traffic, to adequately represent your entire audience.

What is peeking?

For most tests and websites, getting enough traffic takes time. But waiting is hard.

If you don't set the sample size ahead of time, you might be tempted to "peak" at the results prematurely and declare a winner when, in reality, it's far too early to tell.

Peeking at a test result is like baking a cake, opening the oven and deciding the cake is ready before it's fully finished.

Even though it's tempting to stop the test early and dig into that delicious win, the results are only half-baked.

Before pulling a cake out of the oven, a good baker pokes a tooth pick, or fork, into the cake batter to make sure it's truly ready. If the batter sticks to the fork, the cake needs to stay in the over longer longer.

The same goes for an A/B test.

In testing, calculating sample size ahead of time is like an experimenter's tuning *fork*.

It's a definitive, quantitative metric you can use to know when you've achieved enough traffic to stop the test and reliably declare a test version has won.

When your "fork" shows the targeted sample size hasn't yet been reached, you need to keep the test running longer.

(If you liked this analogy, are into baking, and love cake, you've got to A/B test the world's best chocolate icing recipe! Checkout the recipe post here).

An important caveat is you only need to calculate sample size ahead of time if you're running a traditional hypothesis-based test using the frequentist testing methodology.

If you're running a test with using the Bayesian methodology, you don't need to worry about calculating sample size ahead of time.

As well, occasionally, you may find your test variant is drastically losing.

You may, then, decide you want to stop the test early to mitigate losses. Doing so should be done with caution, however, and only after you've truly given the test variant a true chance.

In the first few days of running a test, especially on lower traffic sites, test results can shift radically.

So, as a rule of thumb, it's best to let a test run it's full course, for a minimum of at least two weeks.

If you don't have the risk tolerance to do so, remember, testing itself is a way to mitigate risk since only 50% of the variant is exposed to the treatment.

How long do you need to run your test to achieve a valid sample size?

Again, it depends.

What does it depend on? The short answer is: everything.

The longer answer is a myriad of factors including, but not limited to:

- The type of test you're running
- How many variants you're testing
- Seasonal factors
- Sales cycles
- And more. . .

However, the general A/B testing best practice is to let a test run for a minimum of 2-weeks but no longer than 6-8 weeks.

This time period is selected so that any trends observed over a one-week period, or less, can be confirmed and validated over again.

Summary

- To run a valid A/B test, the larger the sample size, the better.
- As a general guideline, test results are valid when you achieve at least, 30,000 visitors per variant with at least 3,000 conversions on that variant. However, generally even higher numbers are preferred.
- A sample size calculator, like this one, will help you determine how large your sample size needs to be to run a valid test that yields statistically significant results.
- You should always calculate your sample size needs AHEAD of actually running your A/B test and stop the test only when you've achieved a large enough sample size, as pre-determined by your sample size calculations.
- In situations with lower traffic, it's best to limit the number of variants tested, usually to 2 (A vs. B), so each version receives enough traffic to draw conclusive results in an adequate timeframe.
- The ideal testing time period is generally between 2-6 weeks. Anything shorter, results may not hold true overtime. Anything longer and other factors may start to confound and muddy test results.

Source: https://www.figpii.com/blog/a-b-test-segmentation-how-to-do-it-right/

A/B Test Segmentation: How To Do It Right

By Usman Adepoju January 17, 2024

Have you ever wondered how some websites seem to know exactly what you want? The secret often lies in A/B testing, which compares two web page versions to see which performs better.

But here's where it gets really interesting: segmentation. Imagine you're at a party. Everyone's having a good time, but only some people like the same music. What if you could change the tunes in each room to suit the tastes of the people there? That's segmentation in A/B testing.

It's about creating different experiences for different groups of website visitors, ensuring that each group finds what resonates best with them.

In this article, we're diving deep into the nuances of segmentation in A/B testing. We'll navigate the whys, the hows, and the best practices.

What is A/B Test Segmentation

A/B test segmentation is the strategic practice of dividing your audience into smaller, distinct groups to allow for more precise targeting in your tests. Instead of applying a one-size-fits-all method, segmentation acknowledges the diversity within your audience by recognizing and catering to the unique characteristics and preferences of different user groups.

The process involves identifying meaningful ways to categorize your audience – this could be based on the –

1. Demographics,
2. Online behavior,
3. Purchase history,
4. New or returning visitor
5. type of device they use to access your site.

For example, suppose you have an online store and need to test a new layout. In that case, segmentation allows you to test it specifically on returning customers or first-time visitors instead of testing a new layout on all visitors. This focused approach ensures that your collected feedback and data are relevant to each group.

The power of segmentation lies in its ability to deliver more accurate results. It enables you to understand how different groups interact with your website, providing insights that are not just generalized but deeply relevant to each segment.

This targeted testing leads to more effective optimizations, as you're not guessing what works for your entire audience but knowing what resonates with each specific group.

How do you segment A/B Tests?

Segmenting your A/B tests effectively is key to unlocking their full potential. Let's delve into how you can segment your tests to yield the most insightful and actionable results.

Understanding Your Audience

The first step in segmenting your A/B tests is to understand who's visiting your website thoroughly. It's about digging deep into your user data to uncover behaviors and preferences that define your audience.

This step is crucial because the more you know about your users, the better you can tailor your tests to their needs and interests. This analysis focuses on identifying key characteristics that differentiate your user groups. These characteristics range from age and location to browsing behavior and interaction patterns.

Understanding these nuances allows you to tailor your A/B tests more effectively. It's about knowing who your users are and how they uniquely interact with your site.

This knowledge is the foundation of creating meaningful segments that lead to more impactful A/B testing outcomes.

Setting Clear Objectives

Once you understand your audience, setting clear objectives is the next step in segmenting your A/B tests. This involves defining what you aim to achieve with your test.

- Are you looking to increase conversions, such as boosting product sales or sign-up rates?
- Or is your goal more about improving user engagement, like enhancing time spent on a page or interaction with a feature?

Determining which audience segments are most likely to impact these objectives is also important. For example, if increasing sign-ups is your goal, focus on segments that have shown interest but have yet to commit, like new visitors or users who abandoned the sign-up process.

Examples of Objectives To Set during A/B Testing Segmentation

1. **For Increasing Conversions**: Target users who reach the checkout page but abandon their carts. The objective could be to test different checkout processes or promotional offers to see which leads to a higher completion rate.
2. **For Improving User Engagement:** Focus on users who spend a lot of time on informational content but don't take action. Testing different calls-to-action or content layouts might reveal what drives these users to engage more actively.
3. **For Enhancing Navigation**: Look at segments that frequently use the search function or have high bounce rates from the homepage. Testing variations in menu layout or homepage design could lead to insights on improving site navigation.

Selecting Segmentation Criteria

After setting your objectives, the next step is choosing the proper criteria for segmentation. This choice is pivotal because it determines how effectively you can reach and impact your target audience.

The criteria should align closely with your objectives and be measurable to ensure meaningful test outcomes.

- **Demographic Segmentation**: This includes age, gender, and income level. For instance, if your objective is to target a product launch, you might focus on a specific age group or gender most likely to be interested in your product.

- **Geographic Segmentation**: Here, you segment by country, city, or even climate zone. This is particularly useful if you're testing market-specific strategies or your product or service has regional relevance.
- **Behavioral Segmentation**: This involves looking at purchase history, how users engage with your website, and how long they spend there. It's ideal for tailoring user experience based on past interactions with your site.

- **Technological Segmentation**: Segmenting by device type (mobile/desktop), browser, or operating system can be crucial, especially considering the varying user experiences on different devices.

Creating Hypotheses for Each Segment

With your audience segments defined, developing specific hypotheses for each segment is next. This involves predicting how different segments might react to various test variations.

Creating these hypotheses involves a thoughtful analysis of what might motivate or appeal to each segment. For example, a hypothesis could be that a younger audience segment will engage more with vibrant, interactive elements.

In contrast, a professional segment might respond better to a streamlined layout with detailed information.

Based on these hypotheses, you would tailor test elements like CTAs, images, and layout. The aim is to make each element resonate with the targeted segment through visual appeal, messaging tone, or the overall user experience.

This approach ensures that each variation in your A/B test is not just a random alteration but a strategic modification designed to elicit a specific response from a particular segment.

Testing and Iterating

Once you've defined your audience segments and crafted hypotheses, it's time to put your A/B tests into action. This phase involves a strategic approach to testing and continuous iteration, allowing you to refine your segmentation over time.

Begin by starting with broader segments and gradually narrowing them down as you gather more data. This approach ensures that your initial segmentation recognizes potentially significant segments, placing each group in the spotlight.

Moreover, be prepared to iterate on your segmentation strategies. As you conduct tests and gather results, you might discover new segments that deserve attention. Alternatively, specific initial segments can be further subdivided for more precise targeting.

Testing and iterating in A/B test segmentation are not one-time tasks but are continuous processes of refinement. This adaptability ultimately leads to more accurate and impactful outcomes in your testing endeavors.

Leveraging A/B Testing Tools

Effective A/B test segmentation goes hand in hand with the tools you use. Using A/B testing tools that offer advanced segmentation features for precise targeting ensures you can create and track experiments tailored to specific audience segments.

Moreover, it's essential that the testing tool can track and report results separately for each segment. This functionality is the key to evaluating the performance of different segments accurately.

When you have segment-specific data, you better understand how each group responds to your variations. This level of granularity ensures that your insights are not only precise but also highly actionable.

Choosing the right A/B testing tool with robust segmentation capabilities is a strategic move that can significantly enhance the effectiveness of your testing efforts.

Pre-segmentation in A/B Testing

Pre-segmentation is a crucial step in the A/B testing process, where you divide your audience into distinct groups before launching the actual test.

This approach is grounded in using existing data or hypotheses about your audience. By analyzing past behaviors, demographic information, or purchase history, you can identify segments within your audience likely to respond differently to your website's elements.

The role of pre-segmentation is to enable a more targeted and efficient A/B testing strategy. It's about using the insights you already have to anticipate how various segments might react to changes on your site.

For instance, you might use pre-segmentation to differentiate between new and returning visitors, hypothesizing that each group will interact differently with a new feature or design. This foresight allows you to tailor your A/B tests from the outset, increasing their relevance and effectiveness for each specific audience group.

This method sets the stage for more meaningful test results, helping you understand and cater to different user segments' unique needs and preferences.

Post-segmentation in A/B Testing

Post-segmentation comes into play after you've conducted your A/B test. It's about diving into the A/B Test results and examining how different segments of your audience reacted to the variations in the test.

This process is essential for understanding how specific changes impacted various groups within your audience.

In post-segmentation, you break down the overall results of your A/B test into smaller, segment-specific results.

This could mean analyzing how different age groups, geographic locations, or user behaviors influenced the test outcome. For example, you might discover that a new website feature was particularly effective with mobile users but didn't resonate as well with desktop users or that users from a specific region engaged more with a particular type of content.

The insights gained from post-segmentation can be incredibly valuable. They can reveal, for instance, that while the overall test showed a positive impact, specific segments had a negative response, or vice versa.

This level of detail helps fine-tune your website optimization strategies to cater to different audience segments' specific needs and preferences.

In summary, post-segmentation allows you to go beyond the surface of your A/B test results, providing a deeper understanding of how different segments interact with your website and respond to changes.

When you leverage the insights from post-segmentation, you can enhance user experience, increase engagement, and ultimately drive better conversion rates across all segments of your audience.[b]

Benefits of Segmentation in A/B Testing

1. **Enhanced Targeting and Personalization**

In A/B testing, segmentation enables the creation of content and offers that are finely tuned to the distinct needs and preferences of different user groups.

This targeted strategy elevates the relevance of each test for its specific segment, fostering more engaging and meaningful interactions.

Tailoring experiences to align with each group's unique characteristics deepens user connections and amplifies the impact of these interactions, enhancing overall user engagement and satisfaction.

2. **Improved Conversion Rates**

In A/B testing, targeting specific segments allows for a more focused approach to the factors driving those groups' conversions. This strategy helps identify changes that positively impact different segments, leading to an overall improvement in conversion rates.

Businesses can effectively enhance their strategies for better conversion outcomes by tailoring tests to each segment's unique characteristics and behaviors.

3. **Deeper Insights into Customer Behavior**

Segmentation in A/B testing offers a window into how different groups interact with your website, providing deeper insights into customer behavior.

This approach goes beyond surface-level analytics, delving into the 'why' behind user actions. By observing how distinct segments respond to various changes on your site, you better understand their preferences, motivations, and pain points.

This depth of knowledge is invaluable for making more informed decisions about website design, content, and features, ultimately leading to a more user-centric and effective online presence.

4. **Risk Mitigation**

Conducting tests on segmented groups is a key strategy for minimizing risk when introducing new features or changes.

It confines potential negative impacts to a smaller, controlled segment of your user base rather than risking a broader, possibly adverse effect. It enables safer experimentation with new ideas, allowing for the careful refinement of these concepts based on specific group feedback.

This method is essential in maintaining a stable and positive overall user experience while allowing for innovation and improvement.

5. **Long-Term Customer Value Optimization**

Long-Term Customer Value Optimization in A/B testing through segmentation focuses on enhancing customer relationships over time. This approach helps develop products and services that resonate more with specific customer groups, ensuring repeated business and a stronger connection.

Segmentation also allows for more effective communication, as different groups may prefer different messaging styles. Understanding these preferences can make your marketing more impactful.

Best Practices for Effective Segmentation is A/B Testing

1. **Start with Clear Objectives**

A clear game plan is essential when gearing up for A/B testing. Think about what you aim to achieve with your tests. Maybe you want to increase sign-ups, boost sales, or enhance user engagement on your website. Whatever your goal, understanding it is key.

2. **Use Data to Inform Segmentation**

Diving into A/B testing without data is like trying to hit a bullseye in the dark. You must base your segmentation on solid data and analytics to make it count. It's all about understanding who your users are and what they want.

Start by analyzing user behavior. This includes how they interact with your website, what pages they visit the most, and where they spend the most time.

Then, look at demographics like age, location, or even the device they use to browse. These details paint a clearer picture of your audience.

Remember, the more informed your segmentation is, the more relevant and practical your A/B tests will be.

3. **Keep Segments Manageable**

Regarding segmentation, it's easy to get carried away. But here's an essential tip: keep your segments manageable.

Creating too many segments can make your analysis more complex than it needs to be. Plus, it can dilute the clarity of your results.

Think of it this way: if you have a dozen segments, it's not only a challenge to track and analyze all of them, but it's also harder to draw precise, actionable insights.

You might have a lot of data but only a little useful information. So, focus on key segments that will give you valuable insights.

4. **Balance Broad and Narrow Segmentation**

Striking the right balance in segmentation is key in A/B testing. You need to blend broad and narrow segments to get a complete picture.

Broad segments give you a general overview. They're great for capturing overarching trends across a larger audience. Imagine you're looking at your entire customer base to see general patterns in

purchasing behavior. This broad approach helps identify widespread trends that apply to most of your users.

Conversely, narrow segments let you zoom in on specific behaviors or characteristics. You may be focusing on customers from a specific region or those who visited a particular page on your website. These narrow segments offer a detailed understanding of specific user experiences and preferences.

Both approaches have their place in A/B testing. Broad segments help you see the big picture, while narrow segments provide depth and detail.

5. **Ensure Statistical Significance**

When segmenting for A/B testing, it's crucial to base your segments on enough customer data to ensure statistical significance. This means each segment should be large enough to yield reliable results that accurately reflect the behaviors and preferences of that group.

Caution is key with smaller segments; they might not provide enough data to reach statistically significant conclusions and could lead to interpretations that don't accurately represent your broader customer base. Balancing the size of your segments with the quality of your data is essential for drawing meaningful and actionable insights from your A/B tests.

6. **Consider Segment Interaction**

When working with multiple segments in an A/B Test, it's important to understand how these segments interact and influence each other. It's not just about looking at each segment in isolation; it's about seeing the bigger picture of their relationships.

Imagine your segments as different circles in a Venn diagram. Some of these circles might overlap. This overlap is where you need to pay extra attention.

For example, some users fall into both categories if you have a segment based on age and another on geographic location. Understanding this overlap helps you avoid running tests that might conflict with each other or give you redundant information.

Over To You

Good segmentation in A/B testing validates business assumptions, uncovers potential new product areas, and offers a deeper understanding of your audience.

It's not just about what works in the test but what makes sense for sustainable business growth and customer engagement.

This approach ensures that each test contributes to a comprehensive strategy, blending clear results with strategic business insights.

Source: https://www.abtasty.com/blog/sequential-testing/

Sequential testing

In A/B tests where you can see the data coming in a continuous stream, it's tempting to stop the experiment before the planned end. It's so tempting that in fact a lot of practitioners don't even really know why one has to define a testing period beforehand.

Some platforms have even changed their statistical tools to take this into account and have switched to sequential testing which is designed to handle tests this way.

Sequential testing enables you to evaluate data as it's collected to determine if an early decision can be made, helping you cut down on A/B test duration as you can 'peak' at set points.

But, is this an efficient and beneficial type of testing? Spoiler: yes and no, depending on the way you use it.

Why do we need to wait for the predetermined end of the experiment?

Planning and respecting the data collection period of an experiment is crucial. Historical techniques use "fixed horizon testing" that establishes these guidelines for all to follow. If you do not respect this condition, then you don't have the guarantee provided by the statistical framework. This statistical framework guarantees that you only have a 5% error risk when using the common decision thresholds.

Sequential testing promises that when using the proper statistical formulas, you can stop an experiment as soon as the decision threshold is crossed and still have the 5% error risk guarantee. The test user here is the sequential Z-test, which is based on the classical Z-test with an added correction to take the sequential usage into account.

In the following sections, we will look at two objections that are often raised when it comes to sequential testing that may put it at odds with CRO practices.

Sequential testing objection 1: "*Each day has to be sampled the same*"

The first objection is that one should sample each day of the week the same way. This is basically to have a sampling that represents reality. This is the case in a classic A/B test. However, this rule may be broken if you use sequential testing since you can stop the test mid-week but this is not always applicable. Since in reality there are seven different days, your sampling unit should be by week and not by day to account for behavioral differences over the course of a week.

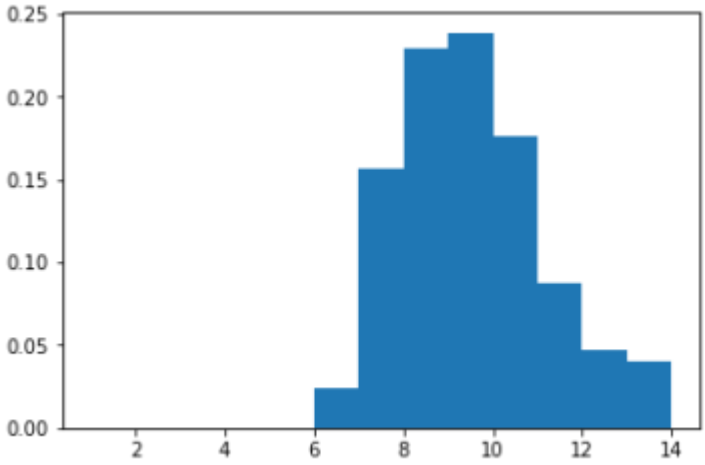
As experiments typically last 2-3 weeks, then the promise of sequential testing saving days isn't necessarily correct unless a winner appears very early in the process. However, it's more likely that the statistical test yielded significance during the last week. In this case, it's best to complete the data collection until each day is sampled evenly so that the full period is covered.

Let's consider the following simulation setting:

- One reference with a 5% conversion rate
- One variation with a 5.5% conversion rate (a 10% relative improvement)
- 5,000 visitors as daily traffic
- 14 days (2 weeks) of data collection

- We ran thousands of such experiments to get histograms for different decision index

In the following histogram, the horizontal axis is the day when the sequential testing crosses the significance threshold. The vertical axis is the ratio of experiments which stopped on this day.



In this setting, day 10 is the most likely day for the sequential testing to reach significance. This means that you will need to wait until the planned end of the test to respect the "same sampling each day" rule. And it's very unlikely that you will get a significant positive result in one week. Thus, in practice, determining the winner sooner with sequential testing doesn't apply in CRO.

Sequential testing objection 2: "*Yes, (effect) size does matter*"

In sequential testing, this is often a less obvious problem and may need some further clarification to be properly understood.

In CRO, we consider mainly two statistical indices for decision-making:

- The *pValue or any other confidence index,* which is linked to the fact that there exists (or not) a difference between the original and the variation. This index is used to validate or invalidate the test hypothesis. But a validated hypothesis is not necessarily a good business decision, so we need more information.
- The *Confidence Interval (CI)* around the estimated gain, which indicates the size of the effect. It's also central to business decisions. For instance, a variation can be a clear winner but with a very little margin that may not cover the implementation or operating costs such as coupon offerings that need to cover the coupon cost.

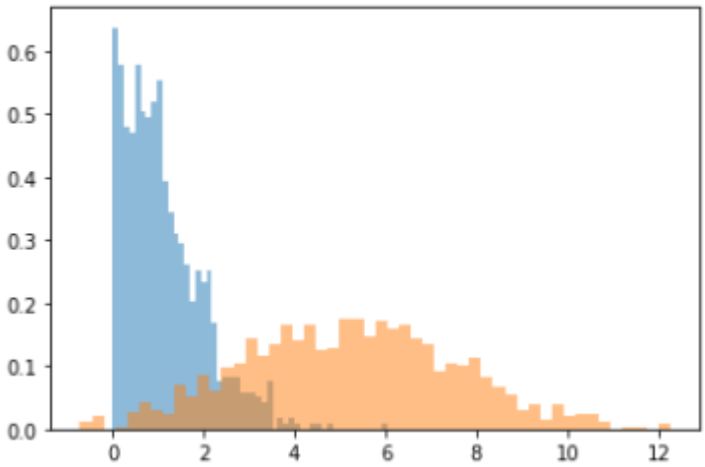
Confidence intervals can be seen as a best and worst case scenario. For example a CI = [1% ; 12%] means "in the worst case you will only get 1% relative uplift," which means going from 5% conversion rate to 5.05%.

If the variation has an implementation or operating cost, the results may not be satisfying. In that case, the solution would be to collect more data in order to have a narrower confidence interval, until you get a more satisfying lower bound, or you may find that the upper bound goes very low showing that the effect, even if it exists, is too low to be worth it from a business perspective.

Using the same scenario as above, the lower bound of the confidence interval can be plotted as follows:

- Horizontal axis – the percentage value of the lower bound

- Vertical axis – the proportion of experiments with this lower bound value
- Blue curve – sequential testing CI
- Orange curve – classical fixed horizon testing



We can see that sequential testing has a very low confidence interval for the lower bound. Most of the time, this is lower than 2% (in relative gain, which is very small). This means that you will get very poor information for business decisions.

Meanwhile, a classic fixed horizon testing (orange curve) will produce a lower bound >5% in half of the cases, which is a more comfortable margin. Therefore, you can continue the data collection until you have a useful result, which means waiting for more data. Even if by chance the sequential testing found a variant reaching significance in one week, you will still need to collect data for another week to do two things: have a useful estimation of the uplift and sample each day equally.

This makes sense in light of the purpose of sequential testing: quickly detect when a variation produces results that differ from the original, whether for the worse or better.

If done as soon as possible, it makes sense to stop the experiment as soon as the gain confidence interval lays mostly either on the positive or negative side. Then, for the positive side, the CI lower bound is close to 0, which doesn't allow for efficient business decisions. It's worth noting that for other applications other than CRO, this behaviour may be optimal and that's why sequential testing exists.

When does sequential testing in CRO make sense?

As we've seen, sequential testing should not be used to quickly determine a winning variation. However, it can be useful in CRO in order to detect losing variations as soon as possible (and prevent loss of conversions, revenue, …).

You may be wondering why it's acceptable to stop an experiment midway through when your variation is losing rather than when you have a winning variation. This is because of the following reasons:

- The most obvious one: To put it simply, you're losing conversions. This is acceptable in the context of searching for a better variation than the original. However, this makes little sense in cases where there is a notable loss, indicating that the variation has no more chances to

be a winner. **An alerting system** set at a low sensitivity level will help detect such impactful losses.

- The less obvious one: Sometimes when an experiment is only slightly "losing" for a good period of time, practitioners tend to let this kind of test run in the hopes that it may turn into a "winner". Thus, they accept this loss because the variation is only "slightly" losing but they often forget that another valuable component is lost in the process: traffic, which is essential for experimentation. For an optimal CRO strategy, one needs to take these factors into account and consider stopping this kind of useless experiment, doomed to have small effects. In such a scenario, an automated alert system will suggest stopping this kind of test and allocate this traffic to other experiments.

Therefore, sequential testing is, in fact, a valuable tool to alert and stop a losing variation.

However, one more objection could still be raised: by stopping the experiment midway, you are breaking the "sample each day the same" rule.

In this particular case, stopping a losing variation has very little chance to be a bad move. In order for the detected variation to become a winner, it first needs to gain enough conversions tobe comparable to the original version. Then it would need another set of conversions to be a "mild" winner and that still wouldn't be enough to be considered a business winner (and cover the implementation or exploitation costs of that winner). To be considered a winner for your business, the competing variation will need another high amount of conversions with a sufficient margin. This margin needs to be high enough to cover the cost of implementation, localization, and/or operating costs.

All the aforementioned events should happen in less than a week (ie. the number of days needed to complete the current week). This is very unlikely, which means it's safe and smart to stop such experiments.

Conclusion

It may be surprising or disappointing to see that there's no business value in stopping winning experiments early as others may believe. This is because a statistical winner is not a business winner. Stopping a test early is taking away the data you need to reach a significant effect size that would increase your chances of getting a winning variation.

With that in mind, the best way to use this type of testing is as an alert to help spot and stop tests that are either harmful to the business or not worth continuing.

Source: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

The Shapiro–Wilk test is a test of normality. It was published in 1965 by Samuel Sanford Shapiro and Martin Wilk.[1]

Theory

The Shapiro–Wilk test tests the null hypothesis that a sample x1, ..., xn came from a normally distributed population. The test statistic is

W

=

(

∑

i

=

1

n

a

i

x

(

i

)

)

2

∑

i

=

1

n

(

x

i

−

x

-

)

2

,

{\displaystyle W={\frac {{\left(\sum \limits _{i=1}^{n}a_{i}x_{(i)}\right)}^{2}}{\sum \limits _{i=1}^{n}{\left(x_{i}-{\overline {x}}\right)}^{2}}},}

where

x

(

i

)

{\displaystyle x_{(i)}} with parentheses enclosing the subscript index i is the ith order statistic, i.e., the ith-smallest number in the sample (not to be confused with

x

i

{\displaystyle x_{i}}).

x

-

=

(

x

1

+

...

+

x

n

)

/

n

{\displaystyle {\overline {x}}=\left(x_{1}+\cdots +x_{n}\right)/n} is the sample mean.

The coefficients

a

i

{\displaystyle a_{i}} are given by:[1]

(

a

1

,

…

,

a

n

)

=

m

T

V

−

1

C

,

{\displaystyle (a_{1},\dots ,a_{n})={m^{\mathsf {T}}V^{-1} \over C},}where C is a vector norm:[2]

C

=

‖

V

−

1

m

‖

=

(

m

T

V

−

1

V

−

1

m

)

1

/

2

{\displaystyle C=\left\|V^{-1}m\right\|={\left(m^{\mathsf {T}}V^{-1}V^{-1}m\right)}^{1/2}}and the vector m,

m

=

(

m

1

,

…

,

m

n

)

T

{\displaystyle m=(m_{1},\dots ,m_{n})^{\mathsf {T}}}is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally,

V

{\displaystyle V} is the covariance matrix of those normal order statistics.[3]

There is no name for the distribution of

W

{\displaystyle W}. The cutoff values for the statistics are calculated through Monte Carlo simulations.[2]

Interpretation

The null-hypothesis of this test is that the population is normally distributed. If the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed.[4]

Like most statistical significance tests, if the sample size is sufficiently large this test may detect even trivial departures from the null hypothesis (i.e., although there may be some statistically significant effect, it may be too small to be of any practical significance); thus, additional investigation of the effect size is typically advisable, e.g., a Q–Q plot in this case.[5]

Power analysis

Monte Carlo simulation has found that Shapiro–Wilk has the best power for a given significance, followed closely by Anderson–Darling when comparing the Shapiro–Wilk, Kolmogorov–Smirnov, and Lilliefors.[6][unreliable source?]

Approximation

Royston proposed an alternative method of calculating the coefficients vector by providing an algorithm for calculating values that extended the sample size from 50 to 2,000.[7] This technique is used in several software packages including GraphPad Prism, Stata,[8][9] SPSS and SAS.[10] Rahman and Govidarajulu extended the sample size further up to 5,000.[11]

Source: https://www.optimizely.com/optimization-glossary/statistical-significance/

What is statistical significance?

Statistical significance is a measure of how unusual your experiment results would be if there were actually no difference in performance between your variation and baseline and the discrepancy in lift was due to random chance alone.

It's become increasingly important for online businesses, marketers, and advertisers running A/B tests (such as testing conversion rates, ad copy, or email subject lines).

Achieving statistical significance helps ensure that conclusions drawn from experiments are reliable and not based on random fluctuations in data.

However, most experiments fail to reach a substantial significance level. Here's why:

- **Changes are too small:** Most changes to visitor experience aren't impactful and they fail to reach clinical significance due to sampling error.
- **Low baseline conversion rates:** Most data sets use metrics with low baseline as a proxy which often results in test results showing significant standard deviations.
- **Too many goals:** Often, teams don't focus on crucial metrics aligned with their hypothesis. This results in research findings falling short of the significance threshold.

Why is the concept of statistical significance important?

Statistical significance helps businesses make sound decisions based on data rather than random fluctuations. It relies on two key factors:

1. **Sample size:** The number of participants in your experiment. Larger samples generally provide more reliable results. For website tests, more traffic means quicker, more accurate results.
2. **Effect size:** The magnitude of difference between your test variations. It shows how much impact your changes have made.

Random sampling is crucial for bridging the statistically significant difference and getting accurate results. If you don't distribute your test variations randomly among your audience, you might introduce bias. For example: If all men see version A and all women see version B, you can't compare results fairly, even with a 50-50 split. Differences in behavior might be due to gender, not your test variations.

Example of real-world impact: In industries like pharmaceuticals, statistical significance in clinical trials can determine a drug's effectiveness. This can influence investor funding and the success or failure of a product.

Overall, statistical significance helps you distinguish between real improvements and random chance, guiding better business decisions.

Testing your hypothesis

Statistical significance is most practically used in hypothesis testing. For example, you want to know whether changing the color of a button on your website from red to green will result in more people clicking on it. If your button is currently red, that's called your "null hypothesis," which takes the

form of your experiment baseline. Turning your button green is known as your "alternative hypothesis."

To determine the observed difference in a statistical significance test, you will want to pay attention to two outputs: p-value and the confidence interval.

1. **P-value:** P-value is the likelihood of seeing evidence as strong or stronger in favor of a difference in performance between your variation and baseline, calculated assuming there actually is no difference between them and any lift observed is entirely owed to random chance.
2. **Confidence interval:** Confidence level is an estimated range of values that are likely, but not guaranteed, to include the unknown but exact value summarizing your target population if an experiment was replicated numerous times.

Get always valid results with Stats Engine

A strict set of guidelines is required to get valid results from experiments run with classical statistics: set a minimum detectable effect and sample size in advance, don't peek at results, and don't test too many goals or variations at the same time. These guidelines can be cumbersome and, if not followed carefully, can produce severely distorted and dubious test results for statisticians.

Fortunately, you can easily determine the practical significance of your experiments using Stats Engine, the advanced statistical model built-in to Optimizely. Here's how to calculate the estimated duration of your experiment:

- Total visitors needed = Sample size × Number of variations
- Estimated days to run = Total visitors needed ÷ Average daily visitors

Stats Engine operates by combining sequential testing and false discovery rate control to give you trustworthy results faster, regardless of sample size and type of data. Updating in real-time, this approach allows for:

- Real-time monitoring of results
- Adaptive testing that adjusts to true effect size
- Faster decision-making without sacrificing data integrity

With Stats Engine, statistical significance should generally increase over time as more evidence is collected. This evidence comes in two forms:

- Larger conversion rate differences
- Conversion rate differences that persist over more visitors

Check out the full stats engine report.

Best practices for reaching statistical significance

When running statistical tests, you might encounter challenges in reaching statistical significance. Here are some best practices you can follow:

- Run tests for at least one business cycle (7 days)
- Choose primary and secondary metrics carefully

- Design experiments with significant potential impact on user behavior

Source: https://pmc.ncbi.nlm.nih.gov/articles/PMC4877414/

Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

Keywords: Confidence intervals, Hypothesis testing, Null testing, *P* value, Power, Significance tests, Statistical testing

Introduction

Misinterpretation and abuse of statistical tests has been decried for decades, yet remains so rampant that some scientific journals discourage use of "statistical significance" (classifying results as "significant" or not based on a *P* value) [1]. One journal now bans all statistical tests and mathematically related procedures such as confidence intervals [2], which has led to considerable discussion and debate about the merits of such bans [3, 4].

Despite such bans, we expect that the statistical methods at issue will be with us for many years to come. We thus think it imperative that basic teaching as well as general understanding of these methods be improved. Toward that end, we attempt to explain the meaning of significance tests, confidence intervals, and statistical power in a more general and critical way than is traditionally done, and then review 25 common misconceptions in light of our explanations. We also discuss a few more subtle but nonetheless pervasive problems, explaining why it is important to examine and synthesize all results relating to a scientific question, rather than focus on individual findings. We further explain why statistical tests should never constitute the sole input to inferences or decisions about associations or effects. Among the many reasons are that, in most scientific settings, the arbitrary classification of results into "significant" and "non-significant" is unnecessary for and often damaging to valid interpretation of data; and that estimation of the size of effects and the uncertainty surrounding our estimates will be far more important for scientific inference and sound judgment than any such classification.

More detailed discussion of the general issues can be found in many articles, chapters, and books on statistical methods and their interpretation [5–20]. Specific issues are covered at length in these

sources and in the many peer-reviewed articles that critique common misinterpretations of null-hypothesis testing and "statistical significance" [1, 12, 21–74].

Statistical tests, *P* values, and confidence intervals: a caustic primer

Statistical models, hypotheses, and tests

Every method of statistical inference depends on a complex web of assumptions about how data were collected and analyzed, and how the analysis results were selected for presentation. The full set of assumptions is embodied in a *statistical model* that underpins the method. This model is a mathematical representation of data variability, and thus ideally would capture accurately all sources of such variability. Many problems arise however because this statistical model often incorporates unrealistic or at best unjustified assumptions. This is true even for so-called "non-parametric" methods, which (like other methods) depend on assumptions of random sampling or randomization. These assumptions are often deceptively simple to write down mathematically, yet in practice are difficult to satisfy and verify, as they may depend on successful completion of a long sequence of actions (such as identifying, contacting, obtaining consent from, obtaining cooperation of, and following up subjects, as well as adherence to study protocols for treatment allocation, masking, and data analysis).

There is also a serious problem of defining the scope of a model, in that it should allow not only for a good representation of the observed data but also of hypothetical alternative data that might have been observed. The reference frame for data that "might have been observed" is often unclear, for example if multiple outcome measures or multiple predictive factors have been measured, and many decisions surrounding analysis choices have been made after the data were collected—as is invariably the case [33].

The difficulty of understanding and assessing underlying assumptions is exacerbated by the fact that the statistical model is usually presented in a highly compressed and abstract form—if presented at all. As a result, many assumptions go unremarked and are often unrecognized by users as well as consumers of statistics. Nonetheless, all statistical methods and interpretations are premised on the model assumptions; that is, on an assumption that the model provides a valid representation of the variation we would expect to see across data sets, faithfully reflecting the circumstances surrounding the study and phenomena occurring within it.

In most applications of statistical testing, one assumption in the model is a hypothesis that a particular effect has a specific size, and has been targeted for statistical analysis. (For simplicity, we use the word "effect" when "association or effect" would arguably be better in allowing for noncausal studies such as most surveys.) This targeted assumption is called the study hypothesis or *test hypothesis*, and the statistical methods used to evaluate it are called *statistical hypothesis tests*. Most often, the targeted effect size is a "null" value representing *zero* effect (e.g., that the study treatment makes no difference in average outcome), in which case the test hypothesis is called the *null hypothesis*. Nonetheless, it is also possible to test other effect sizes. We may also test hypotheses that the effect does or does not fall within a specific range; for example, we may test the hypothesis that the effect is no greater than a particular amount, in which case the hypothesis is said to be a *one-sided* or *dividing* hypothesis [7, 8].

Much statistical teaching and practice has developed a strong (and unhealthy) focus on the idea that the main aim of a study should be to test null hypotheses. In fact most descriptions of statistical testing focus *only* on testing null hypotheses, and the entire topic has been called "Null Hypothesis Significance Testing" (NHST). This exclusive focus on null hypotheses contributes to

misunderstanding of tests. Adding to the misunderstanding is that many authors (including R.A. Fisher) use "null hypothesis" to refer to any test hypothesis, even though this usage is at odds with other authors and with ordinary English definitions of "null"—as are statistical usages of "significance" and "confidence."

Uncertainty, probability, and statistical significance

A more refined goal of statistical analysis is to provide an evaluation of certainty or uncertainty regarding the size of an effect. It is natural to express such certainty in terms of "probabilities" of hypotheses. In conventional statistical methods, however, "probability" refers not to hypotheses, but to quantities that are hypothetical frequencies of data patterns under an assumed statistical model. These methods are thus called *frequentist* methods, and the hypothetical frequencies they predict are called "frequency probabilities." Despite considerable training to the contrary, many statistically educated scientists revert to the habit of misinterpreting these frequency probabilities as hypothesis probabilities. (Even more confusingly, the term "likelihood of a parameter value" is reserved by statisticians to refer to the probability of the observed data *given* the parameter value; it does not refer to a probability of the parameter taking on the given value.)

Nowhere are these problems more rampant than in applications of a hypothetical frequency called the *P* value, also known as the "observed significance level" for the test hypothesis. Statistical "significance tests" based on this concept have been a central part of statistical analyses for centuries [75]. The focus of traditional definitions of *P* values and statistical significance has been on null hypotheses, treating all other assumptions used to compute the *P* value as if they were known to be correct. Recognizing that these other assumptions are often questionable if not unwarranted, we will adopt a more general view of the *P* value as a statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (*all* the assumptions used to compute the *P* value) were correct.

Specifically, the distance between the data and the model prediction is measured using a *test statistic* (such as a t-statistic or a Chi squared statistic). The *P* value is then the probability that the chosen test statistic would have been *at least* as large as its observed value if *every* model assumption were correct, including the test hypothesis. This definition embodies a crucial point lost in traditional definitions: In logical terms, the *P* value tests *all* the assumptions about how the data were generated (the entire model), not just the targeted hypothesis it is supposed to test (such as a null hypothesis). Furthermore, these assumptions include far more than what are traditionally presented as modeling or probability assumptions—they include assumptions about the conduct of the analysis, for example that intermediate analysis results were not used to determine which analyses would be presented.

It is true that the smaller the *P* value, the more unusual the data would be *if* every single assumption were correct; but a very small *P* value does *not* tell us which assumption is incorrect. For example, the *P* value may be very small because the targeted hypothesis is false; but it may instead (or in addition) be very small because the study protocols were violated, or because it was selected for presentation based on its small size. Conversely, a large *P* value indicates only that the data are not unusual under the model, but does not imply that the model or any aspect of it (such as the targeted hypothesis) is correct; it may instead (or in addition) be large because (again) the study protocols were violated, or because it was selected for presentation based on its large size.

The general definition of a *P* value may help one to understand why statistical tests tell us much less than what many think they do: Not only does a *P* value *not* tell us whether the hypothesis

targeted for testing is true or not; it says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct—an assurance that is lacking in far too many studies.

Nonetheless, the *P* value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility, and in this sense may be viewed as measuring the fit of the model to the data. Too often, however, the *P* value is degraded into a dichotomy in which results are declared "statistically significant" if *P* falls on or below a cut-off (usually 0.05) and declared "nonsignificant" otherwise. The terms "significance level" and "alpha level" (α) are often used to refer to the cut-off; however, the term "significance level" invites confusion of the cut-off with the *P* value itself. Their difference is profound: the cut-off value α is supposed to be fixed in advance and is thus part of the study design, unchanged in light of the data. In contrast, the *P* value is a number computed from the data and thus an analysis result, unknown until it is computed.

Moving from tests to estimates

We can vary the test hypothesis while leaving other assumptions unchanged, to see how the *P* value differs across competing test hypotheses. Usually, these test hypotheses specify different sizes for a targeted effect; for example, we may test the hypothesis that the average difference between two treatment groups is zero (the null hypothesis), or that it is 20 or −10 or any size of interest. The effect size whose test produced *P* = 1 is the size most compatible with the data (in the sense of predicting what was in fact observed) *if* all the other assumptions used in the test (the statistical model) were correct, and provides a *point estimate* of the effect under those assumptions. The effect sizes whose test produced *P* > 0.05 will typically define a range of sizes (e.g., from 11.0 to 19.5) that would be considered more compatible with the data (in the sense of the observations being closer to what the model predicted) than sizes outside the range—again, *if* the statistical model were correct. This range corresponds to a 1 − 0.05 = 0.95 or 95 % *confidence interval*, and provides a convenient way of summarizing the results of hypothesis tests for many effect sizes. Confidence intervals are examples of *interval estimates*.

Neyman [76] proposed the construction of confidence intervals in this way because they have the following property: If one calculates, say, 95 % confidence intervals repeatedly *in valid applications*, 95 % of them, on average, will contain (i.e., include or cover) the true effect size. Hence, the specified confidence level is called the *coverage probability.* As Neyman stressed repeatedly, this coverage probability is a property of a long sequence of confidence intervals computed from valid models, rather than a property of any single confidence interval.

Many journals now require confidence intervals, but most textbooks and studies discuss *P* values only for the null hypothesis of no effect. This exclusive focus on null hypotheses in testing not only contributes to misunderstanding of tests and underappreciation of estimation, but also obscures the close relationship between *P* values and confidence intervals, as well as the weaknesses they share.

What *P* values, confidence intervals, and power calculations don't tell us

Much distortion arises from basic misunderstanding of what *P* values and their relatives (such as confidence intervals) do *not* tell us. Therefore, based on the articles in our reference list, we review prevalent *P* value misinterpretations as a way of moving toward defensible interpretations and presentations. We adopt the format of Goodman [40] in providing a list of misinterpretations that can be used to critically evaluate conclusions offered by research reports and reviews. Every one of

the bolded statements in our list has contributed to statistical distortion of the scientific literature, and we add the emphatic "No!" to underscore statements that are not only fallacious but also not "true enough for practical purposes."

Common misinterpretations of single *P* values

1. **The *P* value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave *P* = 0.01, the null hypothesis has only a 1 % chance of being true; if instead it gave *P* = 0.40, the null hypothesis has a 40 % chance of being true. No!**
   The *P* value *assumes* the test hypothesis is true—it is *not* a hypothesis probability and may be far from any reasonable probability for the test hypothesis. The *P* value simply indicates the degree to which the data conform to the pattern predicted by the test hypothesis and all the other assumptions used in the test (the underlying statistical model). Thus *P* = 0.01 would indicate that the data are not very close to what the statistical model (including the test hypothesis) predicted they should be, while *P* = 0.40 would indicate that the data are much closer to the model prediction, allowing for chance variation.

2. **The *P* value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the *P* value for the null hypothesis is 0.08, there is an 8 % probability that chance alone produced the association. No! This is a common variation of the first fallacy and it is just as false. To say that chance *alone* produced the observed association is logically equivalent to asserting that every assumption used to compute the *P* value is correct, *including the null hypothesis*. Thus to claim that the null *P* value is the probability that chance alone produced the observed association is completely backwards: The *P* value is a probability computed *assuming* chance was operating alone. The absurdity of the common backwards interpretation might be appreciated by pondering how the *P* value, which is a probability deduced *from* a set of assumptions (the statistical model), can possibly refer to the probability *of* those assumptions.**

Note: One often sees "alone" dropped from this description (becoming "the *P* value for the null hypothesis is the probability that chance produced the observed association"), so that the statement is more ambiguous, but just as wrong.

3. **A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected. No! A small *P* value simply flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; it may be small because there was a large random error or because some assumption other than the test hypothesis was violated (for example, the assumption that this *P* value was not selected for presentation because it was below 0.05). $P \leq 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large *or larger than* that observed no more than 5 % of the time if *only* chance were creating the discrepancy (as opposed to a violation of the test hypothesis or a mistaken assumption).**

4. **A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted. No! A large *P* value only suggests that the data are *not* unusual if all the assumptions used to compute the *P* value (including the test hypothesis) were correct. The same data would also not be unusual under many other hypotheses. Furthermore, even if the test hypothesis is wrong, the *P* value may be large because it was inflated by a large**

random error or because of some other erroneous assumption (for example, the assumption that this *P* value was not selected for presentation because it was above 0.05). *P* > 0.05 only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large *or largerthan* that observed more than 5 % of the time if *only* chance were creating the discrepancy.

5. **A large*P*value is evidence in favor of the test hypothesis. No! In fact, any *P* value less than 1 implies that the test hypothesis is *not* the hypothesis most compatible with the data, because any other hypothesis with a larger *P* value would be even more compatible with the data. A *P* value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller *P* values. Furthermore, a large *P* value often indicates only that the data are incapable of discriminating among many competing hypotheses (as would be seen immediately by examining the range of the confidence interval). For example, many authors will misinterpret *P* = 0.70 from a test of the null hypothesis as evidence for no effect, when in fact it indicates that, even though the null hypothesis is compatible with the data under the assumptions used to compute the *P* value, it is *not* the hypothesis most compatible with the data—that honor would belong to a hypothesis with *P* = 1. But even if *P* = 1, there will be many other hypotheses that are highly consistent with the data, so that a definitive conclusion of "no association" cannot be deduced from a *P* value, no matter how large.**

6. **A null-hypothesis*P*value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated. No! Observing *P* > 0.05 for the null hypothesis only means that the null is one among the many hypotheses that have *P* > 0.05. Thus, unless the point estimate (observed association) equals the null value exactly, it is a mistake to conclude from *P* > 0.05 that a study found "no association" or "no evidence" of an effect. If the null *P* value is less than 1 some association must be present in the data, and one must look at the point estimate to determine the effect size most compatible with the data under the assumed model.**

7. **Statistical significance indicates a scientifically or substantively important relation has been detected. No! Especially when a study is large, very minor effects or small assumption violations can lead to statistically significant tests of the null hypothesis. Again, a small null *P* value simply flags the data as being unusual if all the assumptions used to compute it (including the null hypothesis) were correct; but the way the data are unusual might be of no clinical interest. One must look at the confidence interval to determine which effect sizes of scientific or other substantive (e.g., clinical) importance are relatively compatible with the data, given the model.**

8. **Lack of statistical significance indicates that the effect size is small. No! Especially when a study is small, even large effects may be "drowned in noise" and thus fail to be detected as statistically significant by a statistical test. A large null *P* value simply flags the data as *not* being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; but the same data will also not be unusual under many other models and hypotheses besides the null. Again, one must look at the confidence interval to determine whether it includes effect sizes of importance.**

9. **The*P*value is the chance of our data occurring if the test hypothesis is true; for example,*P* = 0.05 means that the observed association would occur only 5 % of the time under the test hypothesis. No! The *P* value refers not only to what we observed, but also**

observations *more extreme* than what we observed (where "extremity" is measured in a particular way). And again, the *P* value refers to a data frequency when all the assumptions used to compute it are correct. In addition to the test hypothesis, these assumptions include randomness in sampling, treatment assignment, loss, and missingness, as well as an assumption that the *P* value was not selected for presentation based on its size or some other aspect of the results.

10. **If you reject the test hypothesis because $P \leq 0.05$, the chance you are in error (the chance your "significant finding" is a false positive) is 5 %. No! To see why this description is false, suppose the test hypothesis is in fact true. Then, if you reject it, the chance you are in error is 100 %, not 5 %. The 5 % refers only to how often you would reject it, and therefore be in error, over very many uses of the test across different studies when the test hypothesis and all other assumptions used for the test are true. It does not refer to your single use of the test, which may have been thrown off by assumption violations as well as random errors. This is yet another version of misinterpretation #1.**

11. **$P = 0.05$ and $P \leq 0.05$ mean the same thing. No! This is like saying reported height = 2 m and reported height ≤2 m are the same thing: "height = 2 m" would include few people and those people would be considered tall, whereas "height ≤2 m" would include most people including small children. Similarly, $P = 0.05$ would be considered a borderline result in terms of statistical significance, whereas $P \leq 0.05$ lumps borderline results together with results very incompatible with the model (e.g., $P = 0.0001$) thus rendering its meaning vague, for no good purpose.**

12. ***P*values are properly reported as inequalities (e.g., report "$P < 0.02$" when $P = 0.015$ or report "$P > 0.05$" when $P = 0.06$ or $P = 0.70$). No! This is bad practice because it makes it difficult or impossible for the reader to accurately interpret the statistical result. Only when the *P* value is very small (e.g., under 0.001) does an inequality become justifiable: There is little practical difference among very small *P* values when the assumptions used to compute *P* values are not known with enough certainty to justify such precision, and most methods for computing *P* values are not numerically accurate below a certain point.**

13. **Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance. No! This misinterpretation is promoted when researchers state that they have or have not found "evidence of" a statistically significant effect. The effect being tested either exists or does not exist. "Statistical significance" is a dichotomous description of a *P* value (that it is below the chosen cut-off) and thus is a property of a result of a statistical test; it is not a property of the effect or population being studied.**

14. **One should always use two-sided*P*values. No! Two-sided *P* values are designed to test hypotheses that the targeted effect measure equals a specific value (e.g., zero), and is neither above nor below this value. When, however, the test hypothesis of scientific or practical interest is a one-sided (dividing) hypothesis, a one-sided *P* value is appropriate. For example, consider the practical question of whether a new drug is *at least* as good as the standard drug for increasing survival time. This question is one-sided, so testing this hypothesis calls for a one-sided *P* value. Nonetheless, because two-sided *P* values are the usual default, it will be important to note when and why a one-sided *P* value is being used instead.**

There are other interpretations of *P* values that are controversial, in that whether a categorical "No!" is warranted depends on one's philosophy of statistics and the precise meaning given to the terms involved. The disputed claims deserve recognition if one wishes to avoid such controversy.

For example, it has been argued that *P* values overstate evidence against test hypotheses, based on directly comparing *P* values against certain quantities (likelihood ratios and Bayes factors) that play a central role as evidence measures in Bayesian analysis [37, 72, 77–83]. Nonetheless, many other statisticians do not accept these quantities as gold standards, and instead point out that *P* values summarize crucial evidence needed to gauge the error rates of decisions based on statistical tests (even though they are far from sufficient for making those decisions). Thus, from this frequentist perspective, *P* values do not overstate evidence and may even be considered as measuring one aspect of evidence [7, 8, 84–87], with 1 − *P* measuring evidence against the model used to compute the *P* value. See also Murtaugh [88] and its accompanying discussion.

Common misinterpretations of *P* value comparisons and predictions

Some of the most severe distortions of the scientific literature produced by statistical testing involve erroneous comparison and synthesis of results from different studies or study subgroups. Among the worst are:

- 15.

When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all*P* > 0.05), the overall evidence supports the hypothesis. No! This belief is often used to claim that a literature supports no effect when the opposite is case. It reflects a tendency of researchers to "overestimate the power of most research" [89]. In reality, every study could fail to reach statistical significance and yet when combined show a statistically significant association and persuasive evidence of an effect. For example, if there were five studies each with *P* = 0.10, none would be significant at 0.05 level; but when these *P* values are combined using the Fisher formula [9], the overall *P* value would be 0.01. There are many real examples of persuasive evidence for important effects when few studies or even no study reported "statistically significant" associations [90, 91]. Thus, lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

- 16.

When the same hypothesis is tested in two different populations and the resulting*P*values are on opposite sides of 0.05, the results are conflicting. No! Statistical tests are sensitive to many differences between study populations that are irrelevant to whether their results are in agreement, such as the sizes of compared groups in each population. As a consequence, two studies may provide very different *P* values for the same test hypothesis and yet be in perfect agreement (e.g., may show identical observed associations). For example, suppose we had two randomized trials A and B of a treatment, identical except that trial A had a known standard error of 2 for the mean difference between treatment groups whereas trial B had a known standard error of 1 for the difference. If both trials observed a difference between treatment groups of exactly 3, the usual normal test would produce *P* = 0.13 in A but *P* = 0.003 in B. Despite their difference in *P* values, the test of the hypothesis of no difference in effect across studies would have *P* = 1, reflecting the perfect agreement of the observed mean differences from the studies. Differences between results must be evaluated by directly, for example by estimating and testing those differences to produce a confidence interval and a *P* value comparing the results (often called analysis of heterogeneity, interaction, or modification).

- **17.**

When the same hypothesis is tested in two different populations and the same *P* values are obtained, the results are in agreement. No! Again, tests are sensitive to many differences between populations that are irrelevant to whether their results are in agreement. Two different studies may even exhibit identical *P* values for testing the same hypothesis yet also exhibit clearly different observed associations. For example, suppose randomized experiment A observed a mean difference between treatment groups of 3.00 with standard error 1.00, while B observed a mean difference of 12.00 with standard error 4.00. Then the standard normal test would produce $P = 0.003$ in both; yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$, reflecting the large difference (12.00 − 3.00 = 9.00) between the mean differences.

- **18.**

If one observes a small *P* value, there is a good chance that the next study will produce a *P* value at least as small for the same hypothesis. No! This is false even under the ideal condition that both studies are independent and all assumptions including the test hypothesis are correct in both studies. In that case, if (say) one observes $P = 0.03$, the chance that the new study will show $P \leq 0.03$ is only 3 %; thus the chance the new study will show a *P* value as small or smaller (the "replication probability") is exactly the observed *P* value! If on the other hand the small *P* value arose solely because the true effect exactly equaled its observed estimate, there would be a 50 % chance that a repeat experiment of identical design would have a larger *P* value [37]. In general, the size of the new *P* value will be extremely sensitive to the study size and the extent to which the test hypothesis or other assumptions are violated in the new study [86]; in particular, *P* may be very small or very large depending on whether the study and the violations are large or small.

Finally, although it is (we hope obviously) wrong to do so, one sometimes sees the null hypothesis compared with another (alternative) hypothesis using a two-sided *P* value for the null and a one-sided *P* value for the alternative. This comparison is biased in favor of the null in that the two-sided test will falsely reject the null only half as often as the one-sided test will falsely reject the alternative (again, under all the assumptions used for testing).

Common misinterpretations of confidence intervals

Most of the above misinterpretations translate into an analogous misinterpretation for confidence intervals. For example, another misinterpretation of $P > 0.05$ is that it means the test hypothesis has only a 5 % chance of being false, which in terms of a confidence interval becomes the common fallacy:

- **19.**

The specific 95 % confidence interval presented by a study has a 95 % chance of containing the true effect size. No! A reported confidence interval is a range between two numbers. The frequency with which an observed interval (e.g., 0.72–2.88) contains the true effect is either 100 % if the true effect is within the interval or 0 % if not; the 95 % refers only to how often 95 % confidence intervals computed from very many studies would contain the true size *if all the assumptions used to compute the intervals were correct*. It is possible to compute an interval that can be interpreted as having 95 % probability of containing the true value; nonetheless, such computations require not only the assumptions used to compute the confidence interval, but also further assumptions about the size of effects in the model. These further assumptions are summarized in what is called

a *prior distribution*, and the resulting intervals are usually called *Bayesian posterior (or credible) intervals* to distinguish them from confidence intervals [18].

Symmetrically, the misinterpretation of a small *P* value as disproving the test hypothesis could be translated into:

- **20.**

An effect size outside the 95 % confidence interval has been refuted (or excluded) by the data. No! As with the *P* value, the confidence interval is computed from many assumptions, the violation of which may have led to the results. Thus it is the combination of the data with the assumptions, along with the arbitrary 95 % criterion, that are needed to declare an effect size outside the interval is in some way incompatible with the observations. Even then, judgements as extreme as saying the effect size has been refuted or excluded will require even stronger conditions.

As with *P* values, naïve comparison of confidence intervals can be highly misleading:

- **21.**

If two confidence intervals overlap, the difference between two estimates or studies is not significant. No! The 95 % confidence intervals from two subgroups or studies may overlap substantially and yet the test for difference between them may still produce $P < 0.05$. Suppose for example, two 95 % confidence intervals for means from normal populations with known variances are (1.04, 4.96) and (4.16, 19.84); these intervals overlap, yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$. As with *P* values, comparison between groups requires statistics that directly test and estimate the differences across groups. It can, however, be noted that if the two 95 % confidence intervals fail to overlap, then when using the same assumptions used to compute the confidence intervals we will find $P < 0.05$ for the difference; and if one of the 95 % intervals contains the point estimate from the other group or study, we will find $P > 0.05$ for the difference.

Finally, as with *P* values, the replication properties of confidence intervals are usually misunderstood:

- **22.**

An observed 95 % confidence interval predicts that 95 % of the estimates from future studies will fall inside the observed interval. No! This statement is wrong in several ways. Most importantly, under the model, 95 % is the frequency with which *otherunobserved* intervals will contain the *true effect*, not how frequently the one interval being presented will contain future estimates. In fact, even under ideal conditions the chance that a future estimate will fall within the current interval will usually be much less than 95 %. For example, if two independent studies of the same quantity provide unbiased normal point estimates with the same standard errors, the chance that the 95 % confidence interval for the first study contains the point estimate from the second is 83 % (which is the chance that the difference between the two estimates is less than 1.96 standard errors). Again, an observed interval either does or does not contain the true effect; the 95 % refers only to how often 95 % confidence intervals computed from very many studies would contain the true effect if all the assumptions used to compute the intervals were correct.

- **23.**

If one 95 % confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one. No! When the model is correct, precision of

statistical estimation is measured directly by confidence interval *width* (measured on the appropriate scale). It is not a matter of inclusion or exclusion of the null or any other value. Consider two 95 % confidence intervals for a difference in means, one with limits of 5 and 40, the other with limits of −5 and 10. The first interval excludes the null value of 0, but is 30 units wide. The second includes the null value, but is half as wide and therefore much more precise.

In addition to the above misinterpretations, 95 % confidence intervals force the 0.05-level cutoff on the reader, lumping together all effect sizes with *P* > 0.05, and in this way are as bad as presenting *P* values as dichotomies. Nonetheless, many authors agree that confidence intervals are superior to tests and *P* values because they allow one to shift focus away from the null hypothesis, toward the full range of effect sizes compatible with the data—a shift recommended by many authors and a growing number of journals. Another way to bring attention to non-null hypotheses is to present their *P* values; for example, one could provide or demand *P* values for those effect sizes that are recognized as scientifically reasonable alternatives to the null.

As with *P* values, further cautions are needed to avoid misinterpreting confidence intervals as providing sharp answers when none are warranted. The hypothesis which says the point estimate is the correct effect will have the largest *P* value (*P* = 1 in most cases), and hypotheses inside a confidence interval will have higher *P* values than hypotheses outside the interval. The *P* values will vary greatly, however, among hypotheses inside the interval, as well as among hypotheses on the outside. Also, two hypotheses may have nearly equal *P* values even though one of the hypotheses is inside the interval and the other is outside. Thus, if we use *P* values to measure compatibility of hypotheses with data and wish to compare hypotheses with this measure, we need to examine their *P* values directly, not simply ask whether the hypotheses are inside or outside the interval. This need is particularly acute when (as usual) one of the hypotheses under scrutiny is a null hypothesis.

Common misinterpretations of power

The *power* of a test to detect a correct alternative hypothesis is the pre-study probability that the test will reject the test hypothesis (e.g., the probability that *P* will not exceed a pre-specified cut-off such as 0.05). (The corresponding pre-study probability of failing to reject the test hypothesis when the alternative is correct is one minus the power, also known as the Type-II or beta error rate) [84] As with *P* values and confidence intervals, this probability is defined over repetitions of the same study design and so is a frequency probability. One source of reasonable alternative hypotheses are the effect sizes that were used to compute power in the study proposal. Pre-study power calculations do not, however, measure the compatibility of these alternatives with the data actually observed, while power calculated from the observed data is a direct (if obscure) transformation of the null *P* value and so provides no test of the alternatives. Thus, presentation of power does not obviate the need to provide interval estimates and direct tests of the alternatives.

For these reasons, many authors have condemned use of power to interpret estimates and statistical tests [42, 92–97], arguing that (in contrast to confidence intervals) it distracts attention from direct comparisons of hypotheses and introduces new misinterpretations, such as:

- **24.**

If you accept the null hypothesis because the null *P* value exceeds 0.05 and the power of your test is 90 %, the chance you are in error (the chance that your finding is a false negative) is 10 %. No! If the null hypothesis is false and you accept it, the chance you are in error is 100 %, not 10 %. Conversely, if the null hypothesis is true and you accept it, the chance you are in error is 0 %. The

10 % refers only to how often you would be in error over very many uses of the test across different studies when the particular alternative used to compute power is correct *and* all other assumptions used for the test are correct in all the studies. It does not refer to your single use of the test or your error rate under any alternative effect size other than the one used to compute power.

It can be especially misleading to compare results for two hypotheses by presenting a test or *P* value for one and power for the other. For example, testing the null by seeing whether $P \leq 0.05$ with a power less than $1 - 0.05 = 0.95$ for the alternative (as done routinely) will bias the comparison in favor of the null because it entails a lower probability of incorrectly rejecting the null (0.05) than of incorrectly accepting the null when the alternative is correct. Thus, claims about relative support or evidence need to be based on direct and comparable measures of support or evidence for both hypotheses, otherwise mistakes like the following will occur:

- **25.**

If the null*P*value exceeds 0.05 and the power of this test is 90 % at an alternative, the results support the null over the alternative. This claim seems intuitive to many, but counterexamples are easy to construct in which the null *P* value is between 0.05 and 0.10, and yet there are alternatives whose own *P* value exceeds 0.10 and for which the power is 0.90. Parallel results ensue for other accepted measures of compatibility, evidence, and support, indicating that the data show lower compatibility with and more evidence against the null than the alternative, despite the fact that the null *P* value is "not significant" at the 0.05 alpha level and the power against the alternative is "very high" [42].

Despite its shortcomings for interpreting current data, power can be useful for designing studies and for understanding why replication of "statistical significance" will often fail even under ideal conditions. Studies are often designed or claimed to have 80 % power against a key alternative when using a 0.05 significance level, although in execution often have less power due to unanticipated problems such as low subject recruitment. Thus, if the alternative is correct and the actual power of two studies is 80 %, the chance that the studies will both show $P \leq 0.05$ will at best be only $0.80(0.80) = 64$ %; furthermore, the chance that one study shows $P \leq 0.05$ and the other does not (and thus will be misinterpreted as showing conflicting results) is $2(0.80)0.20 = 32$ % or about 1 chance in 3. Similar calculations taking account of typical problems suggest that one could anticipate a "replication crisis" even if there were no publication or reporting bias, simply because current design and testing conventions treat individual study results as dichotomous outputs of "significant"/"nonsignificant" or "reject"/"accept."

A statistical model is much more than an equation with Greek letters

The above list could be expanded by reviewing the research literature. We will however now turn to direct discussion of an issue that has been receiving more attention of late, yet is still widely overlooked or interpreted too narrowly in statistical teaching and presentations: That the statistical model used to obtain the results is correct.

Too often, the full statistical model is treated as a simple regression or structural equation in which effects are represented by parameters denoted by Greek letters. "Model checking" is then limited to tests of fit or testing additional terms for the model. Yet these tests of fit themselves make further assumptions that should be seen as part of the full model. For example, all common tests and confidence intervals depend on assumptions of random selection for observation or treatment and random loss or missingness within levels of controlled covariates. These assumptions have

gradually come under scrutiny via sensitivity and bias analysis [98], but such methods remain far removed from the basic statistical training given to most researchers.

Less often stated is the even more crucial assumption that the analyses themselves were not guided toward finding nonsignificance or significance (analysis bias), and that the analysis results were not reported based on their nonsignificance or significance (reporting bias and publication bias). Selective reporting renders false even the limited ideal meanings of statistical significance, *P* values, and confidence intervals. Because author decisions to report and editorial decisions to publish results often depend on whether the *P* value is above or below 0.05, selective reporting has been identified as a major problem in large segments of the scientific literature [99–101].

Although this selection problem has also been subject to sensitivity analysis, there has been a bias in studies of reporting and publication bias: It is usually assumed that these biases favor significance. This assumption is of course correct when (as is often the case) researchers select results for presentation when $P \leq 0.05$, a practice that tends to exaggerate associations [101–105]. Nonetheless, bias in favor of reporting $P \leq 0.05$ is not always plausible let alone supported by evidence or common sense. For example, one might expect selection for $P > 0.05$ in publications funded by those with stakes in acceptance of the null hypothesis (a practice which tends to understate associations); in accord with that expectation, some empirical studies have observed smaller estimates and "nonsignificance" more often in such publications than in other studies [101, 106, 107].

Addressing such problems would require far more political will and effort than addressing misinterpretation of statistics, such as enforcing registration of trials, along with open data and analysis code from all completed studies (as in the AllTrials initiative, http://www.alltrials.net/). In the meantime, readers are advised to consider the entire context in which research reports are produced and appear when interpreting the statistics and conclusions offered by the reports.

Conclusions

Upon realizing that statistical tests are usually misinterpreted, one may wonder what if anything these tests do for science. They were originally intended to account for random variability as a source of error, thereby sounding a note of caution against overinterpretation of observed associations as true effects or as stronger evidence against null hypotheses than was warranted. But before long that use was turned on its head to provide fallacious support for null hypotheses in the form of "failure to achieve" or "failure to attain" statistical significance.

We have no doubt that the founders of modern statistical testing would be horrified by common treatments of their invention. In their first paper describing their binary approach to statistical testing, Neyman and Pearson [108] wrote that "it is doubtful whether the knowledge that [a *P* value] was really 0.03 (or 0.06), rather than 0.05…would in fact ever modify our judgment" and that "The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision." Pearson [109] later added, "No doubt we could more aptly have said, 'his final or provisional decision.'" Fisher [110] went further, saying "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." Yet fallacious and ritualistic use of tests continued to spread, including beliefs that whether *P* was above or below 0.05 was a universal arbiter of discovery. Thus by 1965, Hill [111] lamented that "too often we weaken our capacity to interpret data and to take reasonable

decisions whatever the value of *P*. And far too often we deduce 'no difference' from 'no significant difference.'"

In response, it has been argued that some misinterpretations are harmless in tightly controlled experiments on well-understood systems, where the test hypothesis may have special support from established theories (e.g., Mendelian genetics) and in which every other assumption (such as random allocation) is forced to hold by careful design and execution of the study. But it has long been asserted that the harms of statistical testing in more uncontrollable and amorphous research settings (such as social-science, health, and medical fields) have far outweighed its benefits, leading to calls for banning such tests in research reports—again with one journal banning *P* values as well as confidence intervals [2].

Given, however, the deep entrenchment of statistical testing, as well as the absence of generally accepted alternative methods, there have been many attempts to salvage *P* values by detaching them from their use in significance tests. One approach is to focus on *P* values as continuous measures of compatibility, as described earlier. Although this approach has its own limitations (as described in points 1, 2, 5, 9, 15, 18, 19), it avoids comparison of *P* values with arbitrary cutoffs such as 0.05, (as described in 3, 4, 6–8, 10–13, 15, 16, 21 and 23–25). Another approach is to teach and use correct relations of *P* values to hypothesis probabilities. For example, under common statistical models, one-sided *P* values can provide lower bounds on probabilities for hypotheses about effect directions [45, 46, 112, 113]. Whether such reinterpretations can eventually replace common misinterpretations to good effect remains to be seen.

A shift in emphasis from hypothesis testing to estimation has been promoted as a simple and relatively safe way to improve practice [5, 61, 63, 114, 115] resulting in increasing use of confidence intervals and editorial demands for them; nonetheless, this shift has brought to the fore misinterpretations of intervals such as 19–23 above [116]. Other approaches combine tests of the null with further calculations involving both null and alternative hypotheses [117, 118]; such calculations may, however, may bring with them further misinterpretations similar to those described above for power, as well as greater complexity.

Meanwhile, in the hopes of minimizing harms of current practice, we can offer several guidelines for users and readers of statistics, and re-emphasize some key warnings from our list of misinterpretations:

a. **Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise *P* values (not just whether *P* values are above or below 0.05 or some other threshold).**

b. **Careful interpretation also demands critical examination of the assumptions and conventions used for the statistical analysis—not just the usual statistical assumptions, but also the hidden assumptions about how results were generated and chosen for presentation.**

c. **It is simply false to claim that statistically nonsignificant results support a test hypothesis, because the same results may be even more compatible with alternative hypotheses—even if the power of the test is high for those alternatives.**

d. **Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes, or whether statistical results have been misrepresented as supporting one hypothesis when those results are better explained by**

other hypotheses (see points 4–6). We caution however that confidence intervals are often only a first step in these tasks. To compare hypotheses in light of the data and the statistical model it may be necessary to calculate the *P* value (or relative likelihood) of each hypothesis. We further caution that confidence intervals provide only a best-case measure of the uncertainty or ambiguity left by the data, insofar as they depend on an uncertain statistical model.

e. **Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases [68, 119–125]. Even when this is done, however, all the earlier cautions apply. Furthermore, the outcome of any statistical procedure is but one of many considerations that must be evaluated when examining the totality of evidence. In particular, statistical significance is neither necessary nor sufficient for determining the scientific or practical significance of a set of observations. This view was affirmed unanimously by the U.S. Supreme Court, (Matrixx Initiatives, Inc., et al. v. Siracusano et al. No. 09–1156. Argued January 10, 2011, Decided March 22, 2011), and can be seen in our earlier quotes from Neyman and Pearson.**

f. **Any opinion offered about the *probability*, *likelihood*, *certainty*, or similar property for a hypothesis cannot be derived from statistical methods alone. In particular, significance tests and confidence intervals do not by themselves provide a logically sound basis for concluding an effect is present or absent with certainty or a given probability. This point should be borne in mind whenever one sees a conclusion framed as a statement of probability, likelihood, or certainty about a hypothesis. Information about the hypothesis beyond that contained in the analyzed data and in conventional statistical models (which give only data probabilities) must be used to reach such a conclusion; that information should be explicitly acknowledged and described by those offering the conclusion. Bayesian statistics offers methods that attempt to incorporate the needed information directly into the statistical model; they have not, however, achieved the popularity of *P* values and confidence intervals, in part because of philosophical objections and in part because no conventions have become established for their use.**

g. **All statistical methods (whether frequentist or Bayesian, or for testing or estimation, or for inference or decision) make extensive assumptions about the sequence of events that led to the results presented—not only in the data generation, but in the analysis choices. Thus, to allow critical evaluation, research reports (including meta-analyses) should describe in detail the full sequence of events that led to the statistics presented, including the motivation for the study, its design, the original analysis plan, the criteria used to include and exclude subjects (or studies) and data, and a thorough description of all the analyses that were conducted.**

In closing, we note that no statistical method is immune to misinterpretation and misuse, but prudent users of statistics will avoid approaches especially prone to serious abuse. In this regard, we join others in singling out the degradation of *P* values into "significant" and "nonsignificant" as an especially pernicious statistical practice [126].

Source: https://www.jmp.com/en/statistics-knowledge-portal/t-test/two-sample-t-test

The Two-Sample *t*-Test

What is the two-sample *t*-test?

The two-sample *t*-test (also known as the independent samples *t*-test) is a method used to test whether the unknown population means of two groups are equal or not.

Is this the same as an A/B test?

Yes, a two-sample *t*-test is used to analyze the results from A/B tests.

When can I use the test?

You can use the test when your data values are independent, are randomly sampled from two normal populations and the two independent groups have equal variances.

What if I have more than two groups?

Use a multiple comparison method. Analysis of variance (ANOVA) is one such method. Other multiple comparison methods include the Tukey-Kramer test of all pairwise differences, analysis of means (ANOM) to compare group means to the overall mean or Dunnett's test to compare each group mean to a control mean.

What if the variances for my two groups are not equal?

You can still use the two-sample *t*- test. You use a different estimate of the standard deviation.

What if my data isn't nearly normally distributed?

If your sample sizes are very small, you might not be able to test for normality. You might need to rely on your understanding of the data. When you cannot safely assume normality, you can perform a *nonparametric* test that doesn't assume normality.

Using the two-sample *t*-test

The sections below discuss what is needed to perform the test, checking our data, how to perform the test and statistical details.

What do we need?

For the two-sample *t*-test, we need two variables. One variable defines the two groups. The second variable is the measurement of interest.

We also have an idea, or hypothesis, that the means of the underlying populations for the two groups are different. Here are a couple of examples:

- We have students who speak English as their first language and students who do not. All students take a reading test. Our two groups are the native English speakers and the non-native speakers. Our measurements are the test scores. Our idea is that the mean test scores for the underlying populations of native and non-native English speakers are not the same. We want to know if the mean score for the population of native English speakers is different from the people who learned English as a second language.
- We measure the grams of protein in two different brands of energy bars. Our two groups are the two brands. Our measurement is the grams of protein for each energy bar. Our idea is

that the mean grams of protein for the underlying populations for the two brands may be different. We want to know if we have evidence that the mean grams of protein for the two brands of energy bars is different or not.

See how to perform a two-sample *t*-test using statistical software

- Download JMP to follow along using the sample data included with the software.
- To see more JMP tutorials, visit the JMP Learning Library.

Two-sample *t*-test assumptions

To conduct a valid test:

- Data values must be independent. Measurements for one observation do not affect measurements for any other observation.
- Data in each group must be obtained via a random sample from the population.
- Data in each group are normally distributed.
- Data values are continuous.
- The variances for the two independent groups are equal.

For very small groups of data, it can be hard to test these requirements. Below, we'll discuss how to check the requirements using software and what to do when a requirement isn't met.

Two-sample *t*-test example

One way to measure a person's fitness is to measure their body fat percentage. Average body fat percentages vary by age, but according to some guidelines, the normal range for men is 15-20% body fat, and the normal range for women is 20-25% body fat.

Our sample data is from a group of men and women who did workouts at a gym three times a week for a year. Then, their trainer measured the body fat. The table below shows the data.

Table 1: Body fat percentage data grouped by gender

Group	Body Fat Percentages				
Men	13.3	6.0	20.0	8.0	14.0
19.0	18.0	25.0	16.0	24.0	
15.0	1.0	15.0			
Women	22.0	16.0	21.7	21.0	30.0
26.0	12.0	23.2	28.0	23.0	

You can clearly see some overlap in the body fat measurements for the men and women in our sample, but also some differences. Just by looking at the data, it's hard to draw any solid conclusions about whether the underlying populations of men and women at the gym have the same mean body fat. That is the value of statistical tests – they provide a common, statistically valid way to make decisions, so that everyone makes the same decision on the same set of data values.

Checking the data

Let's start by answering: Is the two-sample *t*-test an appropriate method to evaluate the difference in body fat between men and women?

- The data values are independent. The body fat for any one person does not depend on the body fat for another person.
- We assume the people measured represent a simple random sample from the population of members of the gym.
- We assume the data are normally distributed, and we can check this assumption.
- The data values are body fat measurements. The measurements are continuous.
- We assume the variances for men and women are equal, and we can check this assumption.

Before jumping into analysis, we should always take a quick look at the data. The figure below shows histograms and summary statistics for the men and women.

Figure 1: Histogram and summary statistics for the body fat data

The two histograms are on the same scale. From a quick look, we can see that there are no very unusual points, or *outliers*. The data look roughly bell-shaped, so our initial idea of a normal distribution seems reasonable.

Examining the summary statistics, we see that the standard deviations are similar. This supports the idea of equal variances. We can also check this using a test for variances.

Based on these observations, the two-sample *t*-test appears to be an appropriate method to test for a difference in means.

How to perform the two-sample *t*-test

For each group, we need the average, standard deviation and sample size. These are shown in the table below.

Table 2: Average, standard deviation and sample size statistics grouped by gender

up	Sample Size (n)	Average (X-bar)	Standard deviation (s)
Women	10	22.29	5.32
Men	13	14.95	6.84

Without doing any testing, we can see that the averages for men and women in our samples are not the same. But how different are they? Are the averages “close enough” for us to conclude that mean body fat is the same for the larger population of men and women at the gym? Or are the averages too different for us to make this conclusion?

We'll further explain the principles underlying the two sample *t*-test in the statistical details section below, but let's first proceed through the steps from beginning to end. We start by calculating our test statistic. This calculation begins with finding the difference between the two averages:

22.29−14.95=7.3422.29−14.95=7.34

This difference in our samples estimates the difference between the population means for the two groups.

Next, we calculate the pooled standard deviation. This builds a combined estimate of the overall standard deviation. The estimate adjusts for different group sizes. First, we calculate the pooled variance:

s2p=((n1−1)s21)+((n2−1)s22)n1+n2−2sp2=((n1−1)s12)+((n2−1)s22)n1+n2−2

s2p=((10−1)5.322)+((13−1)6.842)(10+13−2)sp2=((10−1)5.322)+((13−1)6.842)(10+13−2)

=(9×28.30)+(12×46.82)21=(9×28.30)+(12×46.82)21

=(254.7+561.85)21=(254.7+561.85)21

=816.5521=38.88=816.5521=38.88

Next, we take the square root of the pooled variance to get the pooled standard deviation. This is:

√38.88=6.2438.88=6.24

We now have all the pieces for our test statistic. We have the difference of the averages, the pooled standard deviation and the sample sizes. We calculate our test statistic as follows:

t=difference of group averagesstandard error of difference=7.34(6.24×√(1/10+1/13))=7.342.62=2.80t=difference of group averagesstandard error of difference=7.34(6.24×(1/10+1/13))=7.342.62=2.80

To evaluate the difference between the means in order to make a decision about our gym programs, we compare the test statistic to a theoretical value from the *t*- distribution. This activity involves four steps:

1. We decide on the risk we are willing to take for declaring a significant difference. For the body fat data, we decide that we are willing to take a 5% risk of saying that the unknown population means for men and women are not equal when they really are. In statistics-

speak, the significance level, denoted by α, is set to 0.05. It is a good practice to make this decision before collecting the data and before calculating test statistics.

2. We calculate a test statistic. Our test statistic is 2.80.
3. We find the theoretical value from the *t*- distribution based on our null hypothesis which states that the means for men and women are equal. Most statistics books have look-up tables for the *t*- distribution. You can also find tables online. The most likely situation is that you will use software and will not use printed tables.

   To find this value, we need the significance level (α = 0.05) and the *degrees of freedom*. The degrees of freedom (*df*) are based on the sample sizes of the two groups. For the body fat data, this is:

   df=n1+n2−2=10+13−2=21df=n1+n2−2=10+13−2=21

   The *t* value with α = 0.05 and 21 degrees of freedom is 2.080.
4. We compare the value of our statistic (2.80) to the *t* value. Since 2.80 > 2.080, we reject the null hypothesis that the mean body fat for men and women are equal, and conclude that we have evidence body fat in the population is different between men and women.

Statistical details

Let's look at the body fat data and the two-sample *t*-test using statistical terms.

Our null hypothesis is that the underlying population means are the same. The null hypothesis is written as:

Ho:μ1=μ2Ho:μ1=μ2

The alternative hypothesis is that the means are not equal. This is written as:

Ho:μ1≠μ2Ho:μ1≠μ2

We calculate the average for each group, and then calculate the difference between the two averages. This is written as:

¯¯¯¯¯x1−¯¯¯¯¯x2x1¯−x2¯

We calculate the pooled standard deviation. This assumes that the underlying population variances are equal. The pooled variance formula is written as:

s2p=((n1−1)s21)+((n2−1)s22)n1+n2−2sp2=((n1−1)s12)+((n2−1)s22)n1+n2−2

The formula shows the sample size for the first group as n_1 and the second group as n_2. The standard deviations for the two groups are s_1 and s_2. This estimate allows the two groups to have different numbers of observations. The pooled standard deviation is the square root of the variance and is written as s_p.

What if your sample sizes for the two groups are the same? In this situation, the pooled estimate of variance is simply the average of the variances for the two groups:

s2p=(s21+s22)2sp2=(s12+s22)2

The test statistic is calculated as:

t=(¯¯¯¯¯x1−¯¯¯¯¯x2)spV1/n1+1/n2t=(x1¯−x2¯)sp1/n1+1/n2

The numerator of the test statistic is the difference between the two group averages. It estimates the difference between the two unknown population means. The denominator is an estimate of the standard error of the difference between the two unknown population means.

Technical Detail: For a single mean, the standard error is s/√ns/n . The formula above extends this idea to two groups that use a pooled estimate for *s* (standard deviation), and that can have different group sizes.

We then compare the test statistic to a *t* value with our chosen alpha value and the degrees of freedom for our data. Using the body fat data as an example, we set α = 0.05. The degrees of freedom (*df*) are based on the group sizes and are calculated as:

df=n1+n2−2=10+13−2=21df=n1+n2−2=10+13−2=21

The formula shows the sample size for the first group as n_1 and the second group as n_2. Statisticians write the *t* value with α = 0.05 and 21 degrees of freedom as:

t0.05,21t0.05,21

The *t* value with α = 0.05 and 21 degrees of freedom is 2.080. There are two possible results from our comparison:

- The test statistic is lower than the *t* value. You fail to reject the hypothesis of equal means. You conclude that the data support the assumption that the men and women have the same average body fat.
- The test statistic is higher than the *t* value. You reject the hypothesis of equal means. You do not conclude that men and women have the same average body fat.

t-Test with unequal variances

When the variances for the two groups are not equal, we cannot use the pooled estimate of standard deviation. Instead, we take the standard error for each group separately. The test statistic is:

t=(¯¯¯¯¯x1−¯¯¯¯¯x2)√s21/n1+s22/n2t=(x1¯−x2¯)s12/n1+s22/n2

The numerator of the test statistic is the same. It is the difference between the averages of the two groups. The denominator is an estimate of the overall standard error of the difference between means. It is based on the separate standard error for each group.

The degrees of freedom calculation for the *t* value is more complex with unequal variances than equal variances and is usually left up to statistical software packages. The key point to remember is that if you cannot use the pooled estimate of standard deviation, then you cannot use the simple formula for the degrees of freedom.

Testing for normality

The normality assumption is more important when the two groups have small sample sizes than for larger sample sizes.

Normal distributions are symmetric, which means they are “even” on both sides of the center. Normal distributions do not have extreme values, or outliers. You can check these two features of a

normal distribution with graphs. Earlier, we decided that the body fat data was “close enough” to normal to go ahead with the assumption of normality. The figure below shows a normal quantile plot for men and women, and supports our decision.

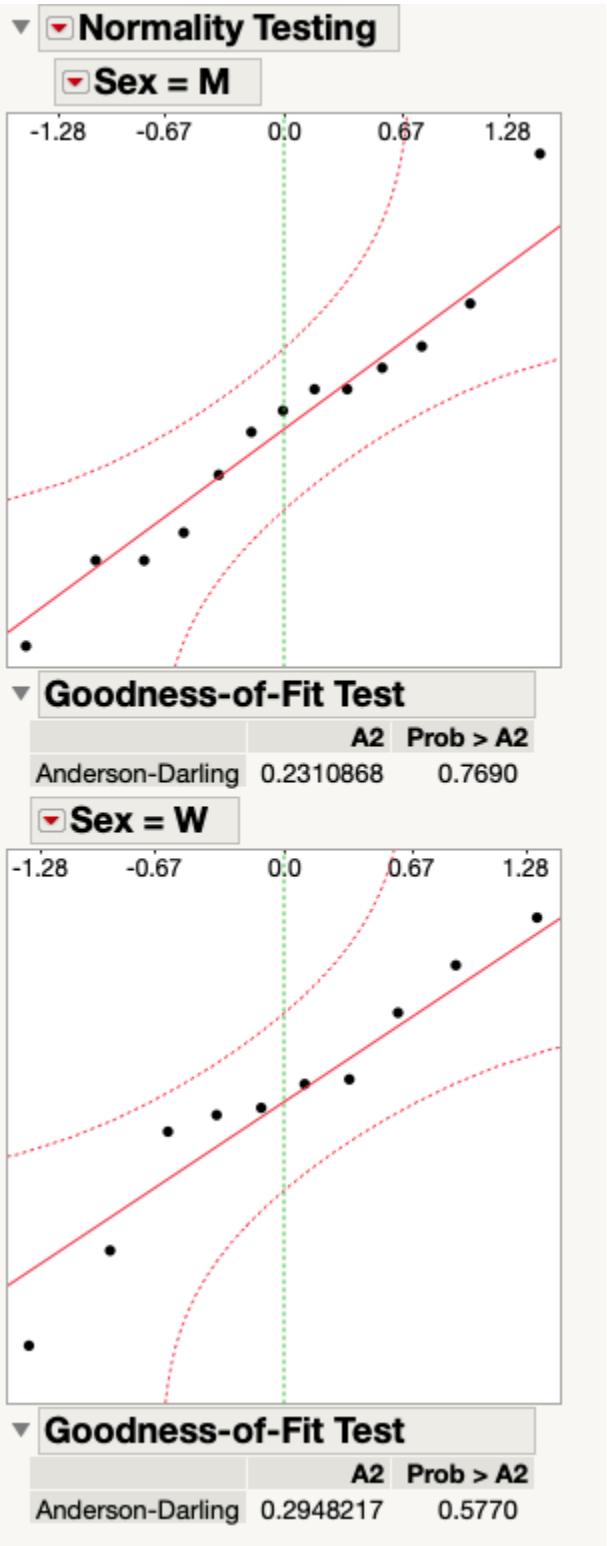


Figure 2: Normal quantile plot of the body fat measurements for men and women

You can also perform a formal test for normality using software. The figure above shows results of testing for normality with JMP software. We test each group separately. Both the test for men and

the test for women show that we cannot reject the hypothesis of a normal distribution. We can go ahead with the assumption that the body fat data for men and for women are normally distributed.

Testing for unequal variances

Testing for unequal variances is complex. We won't show the calculations in detail, but will show the results from JMP software. The figure below shows results of a test for unequal variances for the body fat data.

Tests that the Variances are Equal

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.6073	1	21	0.4445
Brown-Forsythe	0.5042	1	21	0.4855
Levene	0.5175	1	21	0.4798
Bartlett	0.6009	1	.	0.4382
F Test 2-sided	1.6545	12	9	0.4561

Figure 3: Test for unequal variances for the body fat data

Without diving into details of the different types of tests for unequal variances, we will use the *F* test. Before testing, we decide to accept a 10% risk of concluding the variances are equal when they are not. This means we have set $\alpha = 0.10$.

Like most statistical software, JMP shows the *p*-value for a test. This is the likelihood of finding a more extreme value for the test statistic than the one observed. It's difficult to calculate by hand. For the figure above, with the *F* test statistic of 1.654, the *p*- value is 0.4561. This is larger than our α value: $0.4561 > 0.10$. We fail to reject the hypothesis of equal variances. In practical terms, we can go ahead with the two-sample *t*-test with the assumption of equal variances for the two groups.

Understanding p-values

Using a visual, you can check to see if your test statistic is a more extreme value in the distribution. The figure below shows a *t*- distribution with 21 degrees of freedom.

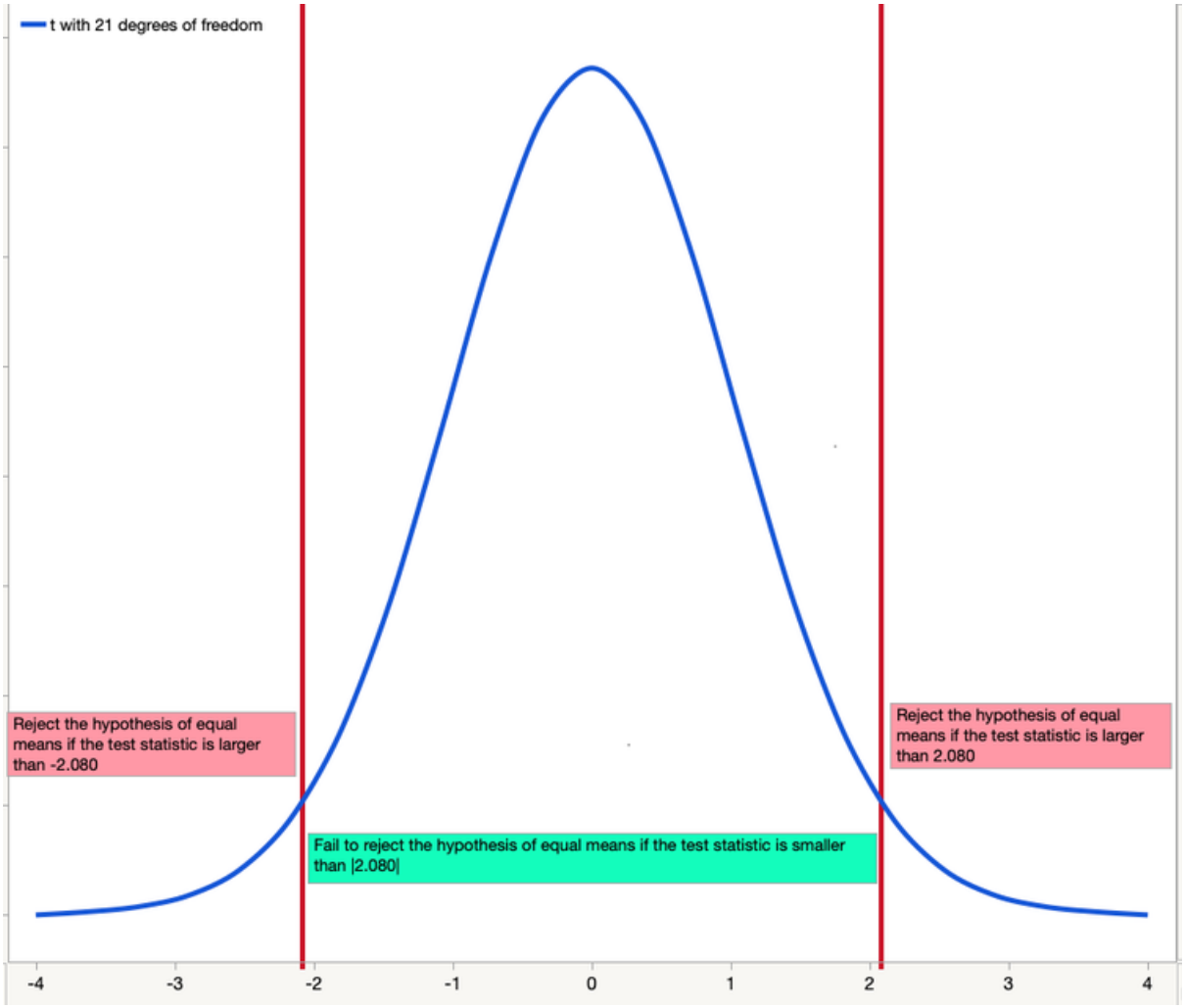


Figure 4: t-distribution with 21 degrees of freedom and α = .05

Since our test is two-sided and we have set α = .05, the figure shows that the value of 2.080 "cuts off" 2.5% of the data in each of the two tails. Only 5% of the data overall is further out in the tails than 2.080. Because our test statistic of 2.80 is beyond the cut-off point, we reject the null hypothesis of equal means.

Putting it all together with software

The figure below shows results for the two-sample *t*-test for the body fat data from JMP software.

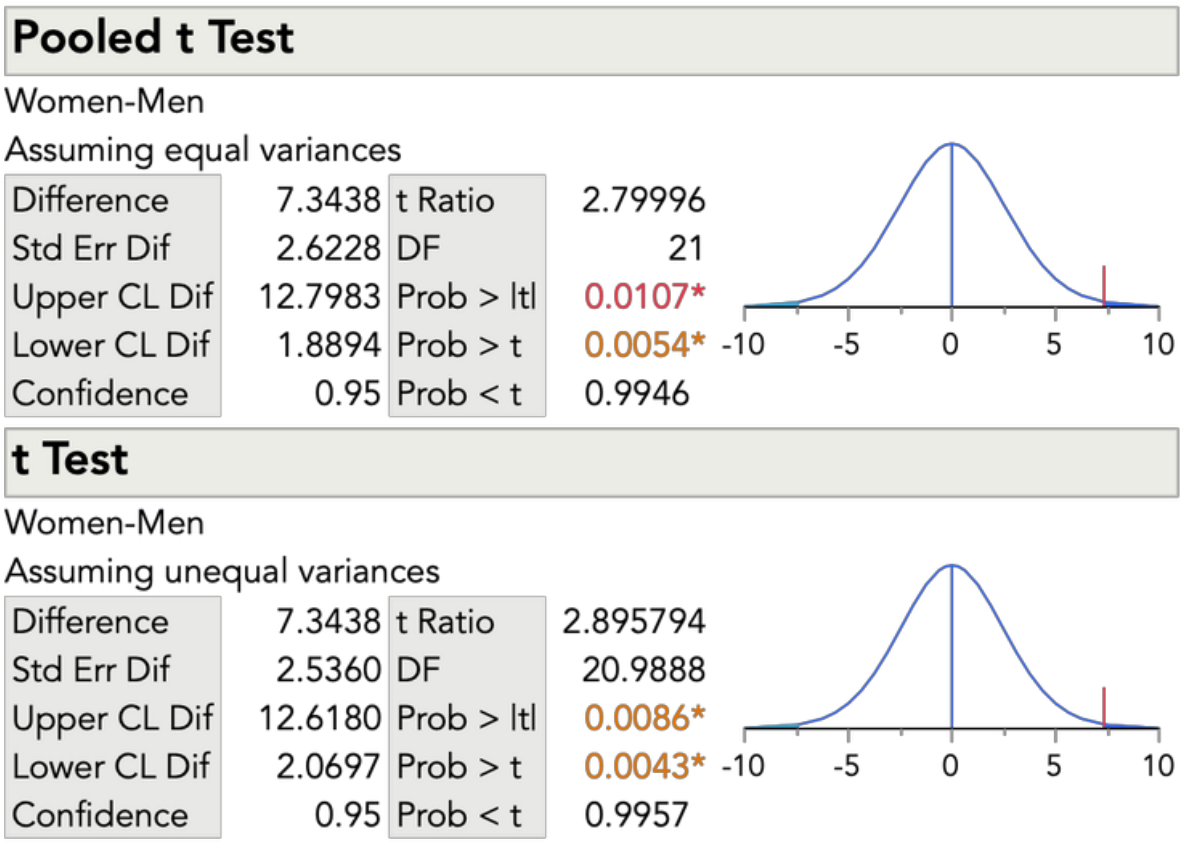


Figure 5: Results for the two-sample t-test from JMP software

The results for the two-sample *t*-test that assumes equal variances are the same as our calculations earlier. The test statistic is 2.79996. The software shows results for a two-sided test and for one-sided tests. The two-sided test is what we want (Prob > |t|). Our null hypothesis is that the mean body fat for men and women is equal. Our alternative hypothesis is that the mean body fat is not equal. The one-sided tests are for one-sided alternative hypotheses – for example, for a null hypothesis that mean body fat for men is less than that for women.

We can reject the hypothesis of equal mean body fat for the two groups and conclude that we have evidence body fat differs in the population between men and women. The software shows a *p*-value of 0.0107. We decided on a 5% risk of concluding the mean body fat for men and women are different, when they are not. It is important to make this decision before doing the statistical test.

The figure also shows the results for the *t*- test that does not assume equal variances. This test does not use the pooled estimate of the standard deviation. As was mentioned above, this test also has a complex formula for degrees of freedom. You can see that the degrees of freedom are 20.9888. The software shows a *p*- value of 0.0086. Again, with our decision of a 5% risk, we can reject the null hypothesis of equal mean body fat for men and women.

Other topics

What if I have more than two groups?

If you have more than two independent groups, you cannot use the two-sample *t*- test. You should use a multiple comparison method. ANOVA, or analysis of variance, is one such method. Other multiple comparison methods include the Tukey-Kramer test of all pairwise differences, analysis of

means (ANOM) to compare group means to the overall mean or Dunnett's test to compare each group mean to a control mean.

What if my data are not from normal distributions?

If your sample size is very small, it might be hard to test for normality. In this situation, you might need to use your understanding of the measurements. For example, for the body fat data, the trainer knows that the underlying distribution of body fat is normally distributed. Even for a very small sample, the trainer would likely go ahead with the *t*-test and assume normality.

What if you know the underlying measurements are not normally distributed? Or what if your sample size is large and the test for normality is rejected? In this situation, you can use nonparametric analyses. These types of analyses do not depend on an assumption that the data values are from a specific distribution. For the two-sample *t* -test, the Wilcoxon rank sum test is a nonparametric test that could be used.

Source: https://www.statsig.com/perspectives/interpret-pvalues-confidence-intervals

Ever wondered how scientists determine if a new drug works better than the old one, or how marketers know if a campaign truly made an impact?

It's all about statistics, and one of the go-to tools in this realm is the **t-test**.

In this blog, we're diving into the world of t-tests, p-values, and confidence intervals. Whether you're crunching numbers for a project or just curious about statistical testing, we've got you covered. So grab a coffee, and let's get started!

Related reading: **T-test fundamentals: Building blocks of experiment analysis**.

Understanding t-tests and their applications

T-tests are handy statistical tools used to compare means between groups. They help us figure out if observed differences are **statistically significant** or just due to chance. Basically, they're essential for hypothesis testing, especially when dealing with small sample sizes.

There are different types of t-tests, each suited for specific situations:

- **One-sample t-test**: This compares a sample mean to a known population mean. For example, if you're studying patients with **Everley's syndrome** and want to compare their mean blood sodium concentration to a standard value, you'd use this test. It's perfect when you have a single sample and a reference value.
- **Independent two-sample t-test**: Use this when comparing means between two separate groups. It tests if the two samples could come from the same population. Say you're comparing transit times through the alimentary canal with two different treatments—this test has got you covered. It's ideal for two independent groups.
- **Paired t-test**: This one compares means from the same group under different conditions. It accounts for variability between pairs, giving you a more sensitive analysis. If you have matched subjects or repeated measures on the same individuals, this is the test to use.

When conducting t-tests, it's important to consider assumptions like **normality** and **equal variances**. If variances aren't equal, **Welch's t-test** can handle the situation. And interpreting p-values correctly is crucial—a low p-value suggests significant differences, while a high p-value indicates we don't have enough evidence to reject the null hypothesis. **Confidence intervals** complement p-values by quantifying the precision of our estimates.

Interpreting p-values in t-tests

P-values are a big deal in **hypothesis testing**. In the context of **t-tests**, they indicate the likelihood of observing a difference between means as extreme as the one found in your sample, assuming the null hypothesis is true.

Here's how to interpret them:

- **If the p-value is less than your significance level (usually 0.05)**: You reject the null hypothesis, suggesting there's a statistically significant difference between the means.
- **If the p-value is greater than your significance level**: You fail to reject the null hypothesis, indicating insufficient evidence to conclude a significant difference.

But remember, a small p-value doesn't necessarily mean the difference is large or practically meaningful. That's where **effect size** and **confidence intervals** come into play, offering additional context about the magnitude and precision of the difference. Likewise, a non-significant p-value doesn't prove the null hypothesis—it just suggests a lack of strong evidence against it.

When working with p-values, be mindful of factors like sample size, variability, and potential confounding variables. These can all influence your results. Sometimes, **visualizing the distribution of p-values** helps identify patterns or issues in your data, guiding further analysis and decision-making.

The role of confidence intervals in t-tests

Confidence intervals are crucial in t-tests because they quantify the uncertainty around the estimated mean difference. They provide a range of plausible values for the true population mean difference, considering sample variability and size.

To calculate a confidence interval for a mean difference in a t-test, you use the sample means, standard errors, and the appropriate t-distribution critical value. Interpreting them is straightforward:

- **If the interval doesn't contain zero**: There's a statistically significant difference between the means at your chosen confidence level.
- **If the interval includes zero**: You can't conclude a significant difference between the means.

This aligns with the p-value approach—a confidence interval excluding zero corresponds to a p-value less than the significance level (e.g., 0.05). But confidence intervals offer more—they show the range of plausible values for the true mean difference, not just whether a difference exists.

Keep in mind, the width of the confidence interval depends on sample size and variability. **Larger samples and lower variability lead to narrower intervals**, indicating greater precision in your estimate. So, when reporting t-test results, it's best practice to include both the p-value and the confidence interval for a comprehensive view.

Practical considerations and best practices

Sample size plays a significant role in the reliability of t-test results. **Larger sample sizes yield more precise estimates and narrower confidence intervals**, increasing the likelihood of detecting true differences. If your sample sizes are small or variances are unequal, **Welch's t-test** can be a better choice.

To ensure accurate interpretation of t-test results, here are some tips:

- **Avoid common pitfalls**: Don't confuse statistical significance with practical significance. A significant p-value doesn't always imply a meaningful difference in real-world terms.
- **Be cautious with multiple t-tests**: Conducting many tests increases the risk of Type I errors (false positives). Adjust your significance level accordingly or consider alternative methods.
- **Interpret p-value histograms wisely**: When looking at **p-value histograms**, patterns may reveal issues with your data or tests. Unusual patterns might warrant consulting a statistician.

Remember, t-tests are just one tool in your statistical toolkit. **Consider the context and limitations of your data**, and use t-tests alongside other methods like confidence intervals and effect sizes for a

comprehensive understanding. Platforms like **Statsig** can help streamline this process, offering robust tools for statistical analysis and experimentation.

Closing thoughts

Grasping t-tests, p-values, and confidence intervals is key to making sense of statistical analyses. These tools help determine whether differences in data are meaningful or just happenstance. By understanding and applying these concepts, you empower yourself to make informed, data-driven decisions.

If you're eager to learn more or need tools to assist with your analysis, platforms like **Statsig** offer great resources to deepen your understanding and streamline your work.

Source: https://www.investopedia.com/terms/t/t-test.asp

What Is a T-Test?

A t-test is an inferential statistical test used to determine if there is a significant difference between the means of two groups and how they are related.

T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.

The t-test assists in hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

Key Takeaways

- A t-test can shed light on a statistically significant difference between the means of two data sets.
- It is used for hypothesis testing in statistics.
- Calculating a t-test requires the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.
- T-tests can be dependent or independent.

Understanding the T-Test

A t-test compares the mean values of two samples to determine a statistically significant difference.

For example, the grades of students from a physics class and those of a different group of students from a writing class would not likely have the same mean and standard deviation.

Similarly, samples taken from the placebo-fed control group of a drug test and those taken from the drug-prescribed group should have a slightly different mean and standard deviation.

Four assumptions are made while using a t-test:

1. The data collected must follow a continuous or ordinal scale, such as the scores for an IQ test.
2. The data is collected from a randomly selected portion of the total population
3. The data will result in a normal distribution of a bell-shaped curve.
4. Equal or homogenous variance exists when the standard variations are equal.

Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement. It assumes a null hypothesis, which means that it assumes the two means are equal.

Using the t-test formulas, values are calculated and compared against the standard values. This comparison helps to determine the effect of chance on the difference, and whether the difference is outside that chance range.

The t-test questions whether the difference between the groups represents a true difference in the study or merely a random difference.

Based on the results, the assumed null hypothesis is accepted or rejected. If the null hypothesis is rejected, it indicates that data readings are strong and are probably not due to chance.

- **Null hypothesis rejected**: Differences are statistically significant
- **Null hypothesis accepted**: Differences are not statistically significant

The t-test is just one of many tests used for this purpose. Others may be more appropriate depending on the number of variables or the size of the sample.

For example, statisticians use a z-test for data sets with a large sample size. Other testing options include the chi-square test and the f-test.

Example of When a T-Test Would Be Useful

Imagine that a drug manufacturer tests a new medicine. Following standard procedure, the drug is given to one group of patients, and a placebo is given to another group called the control group.

The placebo is a substance with no therapeutic value and serves as a benchmark to measure how the other group, administered the actual drug, responds.

After the drug trial, the members of the control group reported an increase in average life expectancy of three years. Members of the group that was prescribed the new drug reported an increase in average life expectancy of four years.

Initial observation indicates that the drug is working. However, it is also possible that the observation may be due to chance.

A t-test could be used to determine if the results are significant and applicable to the entire population, or whether they are random and not due to the drug intervention.

Using the T-Test

Calculating a t-test requires three fundamental data values:

1. The difference between the mean values from each data set, also known as the mean difference
2. The standard deviation of each group
3. The number of data values of each group

The t-test produces two values as its output: t-value and degrees of freedom. The t-value, or t-score, is a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets.

The numerator is the difference between the mean of the two sample sets. The denominator is the variation that exists within the sample sets and is a measurement of the dispersion or variability.

This calculated t-value is then compared against a value obtained from a critical value table called the T-distribution table.

Higher values of the t-score indicate that a large difference exists between the two sample sets. The smaller the t-value, the more similarity exists between the two sample sets.

Degrees of freedom refers to the values in a study that have the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis.

Computation of these values usually depends upon the number of data records available in the sample set.

Important

A large t-score, or t-value, indicates that the groups are different while a small t-score indicates that the groups are similar.

Types of T-Tests

Paired Sample T-Test Formula

The paired t-test, or correlated t-test, is a dependent type of test and is performed when the samples consist of matched pairs of similar units, or when there are cases of repeated measures.

For example, there may be instances where the same patients are repeatedly tested before and after receiving a particular treatment. Each patient is being used as a control sample against themselves.

This method also applies to cases where the samples are related or have matching characteristics, like a comparative analysis involving children, parents, or siblings.

The formula to compute the t-value and degrees of freedom for a paired t-test is:

T=mean1−mean2s(diff)(n)where:mean1 and mean2=The average values of each of the sample setss(diff)=The standard deviation of the differences of the paired data valuesn=The sample size (the number of paired differences)n−1=The degrees of freedom$T = \frac{(mean1 - mean2}{s(diff)}\sqrt{(n)}$
where:*mean*1 and *mean*2=The average values of each of the sample sets*s*(diff)=The standard deviation of the differences of the paired data values*n*=The sample size (the number of paired differences)*n*−1=The degrees of freedom

Equal Variance or Pooled T-Test Formula

The equal variance t-test is an independent t-test and is used when the number of samples in each group is the same, or the variance of the two data sets is similar.

The formula to calculate t-value and degrees of freedom for equal variance t-test is:

T-value=mean1−mean2(n1−1)×var12+(n2−1)×var22n1+n2−2×1n1+1n2where:mean1 and mean2=Average values of eachof the sample setsvar1 and var2=Variance of each of the sample setsn1 and n2=Number of records in each sample setT-value$= \frac{mean1 - mean2}{\sqrt{\frac{(n1-1)\times var12 + (n2-1)\times var22}{n1+n2-2}\times\frac{1}{n1}+\frac{1}{n2}}}$
where:*mean*1 and *mean*2=Average values of eachof the sample sets*var*1 and *var*2=Variance of each of the sample sets*n*1 and *n*2=Number of records in each sample set

and,

Degrees of Freedom=n1+n2−2where:n1 and n2=Number of records in each sample set
Degrees of Freedom=*n*1+*n*2−2**where:***n*1 and *n*2=Number of records in each sample set

Unequal Variance T-Test Formula

The unequal variance t-test is an independent t-test and is used when the number of samples in each group is different, and the variance of the two data sets is also different. This test is also called Welch's t-test.

The formula to calculate t-value and degrees of freedom for an unequal variance t-test is:

T-
value=mean1−mean2(var1n1+var2n2)where:mean1 and mean2=Average values of eachof the sampl e setsvar1 and var2=Variance of each of the sample setsn1 and n2=Number of records in each sampl e setT-value=(*n*1*var*1+*n*2*var*2)*mean*1−*mean*2
where:*mean*1 and *mean*2=Average values of eachof the sample sets*var*1 and *var*2=Variance of each of the sample sets*n*1 and *n*2=Number of records in each sample set

and,

Degrees of Freedom=(var12n1+var22n2)2(var12n1)2n1−1+(var22n2)2n2−1where:var1 and var2=Vari ance of each of the sample setsn1 and n2=Number of records in each sample set
Degrees of Freedom=*n*1−1(*n*1*var*12)2+*n*2−1(*n*2*var*22)2(*n*1*var*12+*n*2*var*22)2
where:*var*1 and *var*2=Variance of each of the sample sets*n*1 and *n*2=Number of records in each sam ple set

Which T-Test to Use

The following flowchart can determine which t-test to use based on the characteristics of the sample sets. The key items to consider include:

- The similarity of the sample records
- The number of data records in each sample set
- The variance of each sample set

Example of an Unequal Variance T-Test

Assume that the diagonal measurement of paintings received in an art gallery is taken. One group of samples includes 10 paintings, while the other includes 20 paintings. The data sets, with the corresponding mean and variance values, are as follows:

Set 1	Set 2
19.7	28.3
20.4	26.7
19.6	20.1
17.8	23.3
18.5	25.2

	Set 1	Set 2
	18.9	22.1
	18.3	17.7
	18.9	27.6
	19.5	20.6
	21.95	13.7
		23.2
		17.5
		20.6
		18
		23.9
		21.6
		24.3
		20.4
		23.9
		13.3
Mean	19.4	21.6

	Set 1	Set 2
Variance	1.4	17.1

Is the difference from 19.4 to 21.6 due to chance alone, or do differences exist in the overall populations of all the paintings received in the art gallery?

We establish the problem by assuming the null hypothesis that the mean is the same between the two sample sets and conduct a t-test to test if the hypothesis is plausible.

Since the number of data records is different (n1 = 10 and n2 = 20) and the variance is also different, the t-value and degrees of freedom are computed for the above data set using the formula for the Unequal Variance T-Test.

The t-value is -2.24787. Since the minus sign can be ignored when comparing the two t-values, the computed value is 2.24787.

The degrees of freedom value is 24.38 and is reduced to 24 (the formula definition requires rounding down the value to the least possible integer value).

One can specify a level of probability (alpha level, level of significance, *p*) as a criterion for acceptance. In most cases, a 5% value can be assumed.

Using the degree of freedom value as 24 and a 5% level of significance, a look at the t-value distribution table gives a value of 2.064.

Comparing this value against the computed value of 2.247 indicates that the calculated t-value is greater than the table value at a significance level of 5%.

Therefore, it is safe to reject the null hypothesis that there is no difference between means.

Rejecting the null hypothesis means the population set has intrinsic differences, and they are not by chance.

How Is the T-Distribution Table Used?

The T-Distribution Table is available in one-tail and two-tails formats. The one-tail format is used for assessing cases that have a fixed value or range with a clear direction, either positive or negative. For instance, what is the probability of the output value remaining below -3, or getting more than seven when rolling a pair of dice? The two-tails format is used for range-bound analysis, such as asking if the coordinates fall between -2 and +2.

What Is an Independent T-Test?

The samples of independent t-tests are selected independent of each other where the data sets in the two groups don't refer to the same values. They may include a group of 100 randomly unrelated patients split into two groups of 50 patients each. One of the groups becomes the control group and is administered a placebo, while the other group receives a prescribed treatment. This constitutes two independent sample groups that are unpaired and unrelated to each other.

What Does a T-Test Explain and How Is It Used?

A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment has an effect on the population of interest, or whether two groups are different from one another.

The Bottom Line

A t-test is used to determine if there is a statistically significant difference between the means of two population samples. It is used in statistics for hypothesis testing and can indicate whether differences between two populations are meaningful or random.

The t-test calculation uses three data: the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.

There are different variations of the t-test formula. Which one to use depends on different factors. However, each variation is used to investigate the same statistical question.

Source: https://en.wikipedia.org/wiki/Type_I_and_type_II_errors

Type I error, or a false positive, is the erroneous rejection of a true null hypothesis in statistical hypothesis testing. A type II error, or a false negative, is the erroneous failure in bringing about appropriate rejection of a false null hypothesis.[1]

Type I errors can be thought of as errors of commission, in which the status quo is erroneously rejected in favour of new, misleading information. Type II errors can be thought of as errors of omission, in which a misleading status quo is allowed to remain due to failures in identifying it as such. For example, if the assumption that people are innocent until proven guilty were taken as a null hypothesis, then proving an innocent person as guilty would constitute a Type I error, while failing to prove a guilty person as guilty would constitute a Type II error. If the null hypothesis were inverted, such that people were by default presumed to be guilty until proven innocent, then proving a guilty person's innocence would constitute a Type I error, while failing to prove an innocent person's innocence would constitute a Type II error. The manner in which a null hypothesis frames contextually default expectations influences the specific ways in which type I errors and type II errors manifest, and this varies by context and application.

Knowledge of type I errors and type II errors is applied widely in fields of in medical science, biometrics and computer science. Minimising these errors is an object of study within statistical theory, though complete elimination of either is impossible when relevant outcomes are not determined by known, observable, causal processes.

Definition

Statistical background

In statistical test theory, the notion of a statistical error is an integral part of hypothesis testing. The test goes about choosing about two competing propositions called null hypothesis, denoted by

H

0

{\textstyle H_{0}} and alternative hypothesis, denoted by

H

1

{\textstyle H_{1}}. This is conceptually similar to the judgement in a court trial. The null hypothesis corresponds to the position of the defendant: just as he is presumed to be innocent until proven guilty, so is the null hypothesis presumed to be true until the data provide convincing evidence against it. The alternative hypothesis corresponds to the position against the defendant. Specifically, the null hypothesis also involves the absence of a difference or the absence of an association. Thus, the null hypothesis can never be that there is a difference or an association.

If the result of the test corresponds with reality, then a correct decision has been made. However, if the result of the test does not correspond with reality, then an error has occurred. There are two situations in which the decision is wrong. The null hypothesis may be true, whereas we reject

H

0

{\textstyle H_{0}}. On the other hand, the alternative hypothesis

H

1

{\textstyle H_{1}} may be true, whereas we do not reject

H

0

{\textstyle H_{0}}. Two types of error are distinguished: type I error and type II error.[2]

Type I error

The first kind of error is the mistaken rejection of a null hypothesis as the result of a test procedure. This kind of error is called a type I error (false positive) and is sometimes called an error of the first kind. In terms of the courtroom example, a type I error corresponds to convicting an innocent defendant.

Type II error

The second kind of error is the mistaken failure to reject the null hypothesis as the result of a test procedure. This sort of error is called a type II error (false negative) and is also referred to as an error of the second kind. In terms of the courtroom example, a type II error corresponds to acquitting a criminal.[2]

Crossover error rate

The crossover error rate (CER) is the point at which type I errors and type II errors are equal. A system with a lower CER value provides more accuracy than a system with a higher CER value.

False positive and false negative

Further information: False positives and false negatives

In terms of false positives and false negatives, a positive result corresponds to rejecting the null hypothesis, while a negative result corresponds to failing to reject the null hypothesis; "false" means the conclusion drawn is incorrect. Thus, a type I error is equivalent to a false positive, and a type II error is equivalent to a false negative.

Table of error types

Tabulated relations between truth/falseness of the null hypothesis and outcomes of the test:[3]

Table of error types

Null hypothesis (

H

0

{\textstyle {\boldsymbol {H_{0}}}}) is

True False

Decision

about null

hypothesis (

H

0

{\textstyle {\boldsymbol {H_{0}}}})

Not reject

Correct inference

(true negative)

(probability =

1

−

α

{\textstyle 1-\alpha })

Type II error

(false negative)

(probability =

β

{\textstyle \beta })

Reject Type I error

(false positive)

(probability =

α

{\textstyle \alpha })

Correct inference

(true positive)

(probability =

1

−

β

{\textstyle 1-\beta })

Error rate

See also: Sensitivity and specificity and False positive rate § Comparison with other error rates

The results obtained from negative sample (left curve) overlap with the results obtained from positive samples (right curve). By moving the result cutoff value (vertical bar), the rate of false positives (FP) can be decreased, at the cost of raising the number of false negatives (FN), or vice versa (TP = True Positives, TPR = True Positive Rate, FPR = False Positive Rate, TN = True Negatives).

A perfect test would have zero false positives and zero false negatives. However, statistical methods are probabilistic, and it cannot be known for certain whether statistical conclusions are correct. Whenever there is uncertainty, there is the possibility of making an error. Considering this, all statistical hypothesis tests have a probability of making type I and type II errors.[4]

The type I error rate is the probability of rejecting the null hypothesis given that it is true. The test is designed to keep the type I error rate below a prespecified bound called the significance level, usually denoted by the Greek letter α (alpha) and is also called the alpha level.[5] Usually, the significance level is set to 0.05 (5%), implying that it is acceptable to have a 5% probability of incorrectly rejecting the true null hypothesis.[6]

The rate of the type II error is denoted by the Greek letter β (beta) and related to the power of a test, which equals 1−β.[citation needed]

These two types of error rates are traded off against each other: for any given sample set, the effort to reduce one type of error generally results in increasing the other type of error.[citation needed]

The quality of hypothesis test

The same idea can be expressed in terms of the rate of correct results and therefore used to minimize error rates and improve the quality of hypothesis test. To reduce the probability of committing a type I error, making the alpha value more stringent is both simple and efficient. For example, setting the alpha value at 0.01, instead of 0.05. To decrease the probability of committing a type II error, which is closely associated with analyses' power, either increasing the test's sample size or relaxing the alpha level, ex. setting the alpha level to 0.1 instead of 0.05, could increase the analyses' power.[citation needed] A test statistic is robust if the type I error rate is controlled.

Varying different threshold (cut-off) values could also be used to make the test either more specific or more sensitive, which in turn elevates the test quality. For example, imagine a medical test, in which an experimenter might measure the concentration of a certain protein in the blood sample. The experimenter could adjust the threshold (black vertical line in the figure) and people would be diagnosed as having diseases if any number is detected above this certain threshold. According to the image, changing the threshold would result in changes in false positives and false negatives, corresponding to movement on the curve.[citation needed]

Example

Since in a real experiment it is impossible to avoid all type I and type II errors, it is important to consider the amount of risk one is willing to take to falsely reject H0 or accept H0. The solution to this question would be to report the p-value or significance level α of the statistic. For example, if the p-value of a test statistic result is 0.0596, then there is a probability of 5.96% that we falsely reject H0 given it is true. Or, if we say, the statistic is performed at level α, like 0.05, then we allow to falsely reject H0 at 5%. A significance level α of 0.05 is relatively common, but there is no general rule that fits all scenarios.

Vehicle speed measuring

The speed limit of a freeway in the United States is 120 kilometers per hour (75 mph). A device is set to measure the speed of passing vehicles. Suppose that the device will conduct three measurements of the speed of a passing vehicle, recording as a random sample X1, X2, X3. The traffic police will or will not fine the drivers depending on the average speed

X

-

{\displaystyle {\bar {X}}}. That is to say, the test statistic

T

=

X

1

+

X

2

+

X

3

3

=

X

-

{\displaystyle T={\frac {X_{1}+X_{2}+X_{3}}{3}}={\bar {X}}}

In addition, we suppose that the measurements X1, X2, X3 are modeled as normal distribution N(μ,2). Then, T should follow N(μ,2/

3

{\displaystyle {\sqrt {3}}}) and the parameter μ represents the true speed of passing vehicle. In this experiment, the null hypothesis H0 and the alternative hypothesis H1 should be

H0: μ=120 against H1: μ>120.

If we perform the statistic level at α=0.05, then a critical value c should be calculated to solve

P

(

Z

⩾

c

−

120

2

3

)

=

0.05

{\displaystyle P\left(Z\geqslant {\frac {c-120}{\frac {2}{\sqrt {3}}}}\right)=0.05}

According to change-of-units rule for the normal distribution. Referring to Z-table, we can get

c

−

120

2

3

=

1.645

⇒

c

=

121.9

{\displaystyle {\frac {c-120}{\frac {2}{\sqrt {3}}}}=1.645\Rightarrow c=121.9}

Here, the critical region. That is to say, if the recorded speed of a vehicle is greater than critical value 121.9, the driver will be fined. However, there are still 5% of the drivers are falsely fined since the recorded average speed is greater than 121.9 but the true speed does not pass 120, which we say, a type I error.

The type II error corresponds to the case that the true speed of a vehicle is over 120 kilometers per hour but the driver is not fined. For example, if the true speed of a vehicle μ=125, the probability that the driver is not fined can be calculated as

P

=

(

T

<

121.9

|

μ

=

125

)

=

P

(

T

−

125

2

3

<

121.9

−

125

2

3

)

=

φ

(

−

2.68

)

=

0.0036

{\displaystyle P=(T<121.9|\mu =125)=P\left({\frac {T-125}{\frac {2}{\sqrt {3}}}}<{\frac {121.9-125}{\frac {2}{\sqrt {3}}}}\right)=\phi (-2.68)=0.0036}

which means, if the true speed of a vehicle is 125, the driver has the probability of 0.36% to avoid the fine when the statistic is performed at level α=0.05, since the recorded average speed is lower than 121.9. If the true speed is closer to 121.9 than 125, then the probability of avoiding the fine will also be higher.

The tradeoffs between type I error and type II error should also be considered. That is, in this case, if the traffic police do not want to falsely fine innocent drivers, the level α can be set to a smaller value, like 0.01. However, if that is the case, more drivers whose true speed is over 120 kilometers per hour, like 125, would be more likely to avoid the fine.

Etymology

In 1928, Jerzy Neyman (1894–1981) and Egon Pearson (1895–1980), both eminent statisticians, discussed the problems associated with "deciding whether or not a particular sample may be judged as likely to have been randomly drawn from a certain population":[7] and, as Florence Nightingale David remarked, "it is necessary to remember the adjective 'random' [in the term 'random sample'] should apply to the method of drawing the sample and not to the sample itself".[8]

They identified "two sources of error", namely:

the error of rejecting a hypothesis that should have not been rejected, and

the error of failing to reject a hypothesis that should have been rejected.

In 1930, they elaborated on these two sources of error, remarking that

in testing hypotheses two considerations must be kept in view, we must be able to reduce the chance of rejecting a true hypothesis to as low a value as desired; the test must be so devised that it will reject the hypothesis tested when it is likely to be false.

In 1933, they observed that these "problems are rarely presented in such a form that we can discriminate with certainty between the true and false hypothesis". They also noted that, in deciding whether to fail to reject, or reject a particular hypothesis amongst a "set of alternative hypotheses", H1, H2..., it was easy to make an error,

[and] these errors will be of two kinds:

we reject H0 [i.e., the hypothesis to be tested] when it is true,[9]

we fail to reject H0 when some alternative hypothesis HA or H1 is true. (There are various notations for the alternative).

In all of the papers co-written by Neyman and Pearson the expression H0 always signifies "the hypothesis to be tested".

In the same paper they call these two sources of error, errors of type I and errors of type II respectively.[10]

Related terms

See also: Coverage probability

Null hypothesis

Main article: Null hypothesis

It is standard practice for statisticians to conduct tests in order to determine whether or not a "speculative hypothesis" concerning the observed phenomena of the world (or its inhabitants) can be supported. The results of such testing determine whether a particular set of results agrees reasonably (or does not agree) with the speculated hypothesis.

On the basis that it is always assumed, by statistical convention, that the speculated hypothesis is wrong, and the so-called "null hypothesis" that the observed phenomena simply occur by chance (and that, as a consequence, the speculated agent has no effect) – the test will determine whether this hypothesis is right or wrong. This is why the hypothesis under test is often called the null hypothesis (most likely, coined by Fisher (1935, p. 19)), because it is this hypothesis that is to be either nullified or not nullified by the test. When the null hypothesis is nullified, it is possible to conclude that data support the "alternative hypothesis" (which is the original speculated one).

The consistent application by statisticians of Neyman and Pearson's convention of representing "the hypothesis to be tested" (or "the hypothesis to be nullified") with the expression H0 has led to circumstances where many understand the term "the null hypothesis" as meaning "the nil hypothesis" – a statement that the results in question have arisen through chance. This is not necessarily the case – the key restriction, as per Fisher (1966), is that "the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the 'problem of distribution', of which the test of significance is the solution."[11] As a consequence of this, in experimental science the null hypothesis is generally a statement that a particular treatment has no effect; in observational science, it is that there is no difference between the value of a particular measured variable, and that of an experimental prediction.[citation needed]

Statistical significance

If the probability of obtaining a result as extreme as the one obtained, supposing that the null hypothesis were true, is lower than a pre-specified cut-off probability (for example, 5%), then the result is said to be statistically significant and the null hypothesis is rejected.

British statistician Sir Ronald Aylmer Fisher (1890–1962) stressed that the null hypothesis

is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

— Fisher, 1935, p.19

Application domains

Medicine

In the practice of medicine, the differences between the applications of screening and testing are considerable.

Medical screening

Screening involves relatively cheap tests that are given to large populations, none of whom manifest any clinical indication of disease (e.g., Pap smears).

Testing involves far more expensive, often invasive, procedures that are given only to those who manifest some clinical indication of disease, and are most often applied to confirm a suspected diagnosis.

For example, most states in the US require newborns to be screened for phenylketonuria and hypothyroidism, among other congenital disorders.

Hypothesis: "The newborns have phenylketonuria and hypothyroidism".

Null hypothesis (H0): "The newborns do not have phenylketonuria and hypothyroidism".

Type I error (false positive): The true fact is that the newborns do not have phenylketonuria and hypothyroidism but we consider they have the disorders according to the data.

Type II error (false negative): The true fact is that the newborns have phenylketonuria and hypothyroidism but we consider they do not have the disorders according to the data.

Although they display a high rate of false positives, the screening tests are considered valuable because they greatly increase the likelihood of detecting these disorders at a far earlier stage.

The simple blood tests used to screen possible blood donors for HIV and hepatitis have a significant rate of false positives; however, physicians use much more expensive and far more precise tests to determine whether a person is actually infected with either of these viruses.

Perhaps the most widely discussed false positives in medical screening come from the breast cancer screening procedure mammography. The US rate of false positive mammograms is up to 15%, the highest in world. One consequence of the high false positive rate in the US is that, in any 10-year period, half of the American women screened receive a false positive mammogram. False positive mammograms are costly, with over $100 million spent annually in the U.S. on follow-up testing and treatment. They also cause women unneeded anxiety. As a result of the high false positive rate in the US, as many as 90–95% of women who get a positive mammogram do not have the condition. The lowest rate in the world is in the Netherlands, 1%. The lowest rates are generally in Northern Europe where mammography films are read twice and a high threshold for additional testing is set (the high threshold decreases the power of the test).

The ideal population screening test would be cheap, easy to administer, and produce zero false negatives, if possible. Such tests usually produce more false positives, which can subsequently be sorted out by more sophisticated (and expensive) testing.

Medical testing

False negatives and false positives are significant issues in medical testing.

Hypothesis: "The patients have the specific disease".

Null hypothesis (H0): "The patients do not have the specific disease".

Type I error (false positive): The true fact is that the patients do not have a specific disease but the physician judges the patient is ill according to the test reports.

Type II error (false negative): The true fact is that the disease is actually present but the test reports provide a falsely reassuring message to patients and physicians that the disease is absent.

False positives can also produce serious and counter-intuitive problems when the condition being searched for is rare, as in screening. If a test has a false positive rate of one in ten thousand, but only

one in a million samples (or people) is a true positive, most of the positives detected by that test will be false. The probability that an observed positive result is a false positive may be calculated using Bayes' theorem.

False negatives produce serious and counter-intuitive problems, especially when the condition being searched for is common. If a test with a false negative rate of only 10% is used to test a population with a true occurrence rate of 70%, many of the negatives detected by the test will be false.

This sometimes leads to inappropriate or inadequate treatment of both the patient and their disease. A common example is relying on cardiac stress tests to detect coronary atherosclerosis, even though cardiac stress tests are known to only detect limitations of coronary artery blood flow due to advanced stenosis.

Biometrics

Biometric matching, such as for fingerprint recognition, facial recognition or iris recognition, is susceptible to type I and type II errors.

Hypothesis: "The input does not identify someone in the searched list of people".

Null hypothesis: "The input does identify someone in the searched list of people".

Type I error (false reject rate): The true fact is that the person is someone in the searched list but the system concludes that the person is not according to the data.

Type II error (false match rate): The true fact is that the person is not someone in the searched list but the system concludes that the person is someone whom we are looking for according to the data.

The probability of type I errors is called the "false reject rate" (FRR) or false non-match rate (FNMR), while the probability of type II errors is called the "false accept rate" (FAR) or false match rate (FMR).

If the system is designed to rarely match suspects then the probability of type II errors can be called the "false alarm rate". On the other hand, if the system is used for validation (and acceptance is the norm) then the FAR is a measure of system security, while the FRR measures user inconvenience level.

Security screening

Main articles: Explosive detection and Metal detector

False positives are routinely found every day in airport security screening, which are ultimately visual inspection systems. The installed security alarms are intended to prevent weapons being brought onto aircraft; yet they are often set to such high sensitivity that they alarm many times a day for minor items, such as keys, belt buckles, loose change, mobile phones, and tacks in shoes.

Hypothesis: "The item is a weapon".

Null hypothesis: "The item is not a weapon".

Type I error (false positive): The true fact is that the item is not a weapon but the system still sounds an alarm.

Type II error (false negative) The true fact is that the item is a weapon but the system keeps silent at this time.

The ratio of false positives (identifying an innocent traveler as a terrorist) to true positives (detecting a would-be terrorist) is, therefore, very high; and because almost every alarm is a false positive, the positive predictive value of these screening tests is very low.

The relative cost of false results determines the likelihood that test creators allow these events to occur. As the cost of a false negative in this scenario is extremely high (not detecting a bomb being brought onto a plane could result in hundreds of deaths) whilst the cost of a false positive is relatively low (a reasonably simple further inspection) the most appropriate test is one with a low statistical specificity but high statistical sensitivity (one that allows a high rate of false positives in return for minimal false negatives).

Computers

The notions of false positives and false negatives have a wide currency in the realm of computers and computer applications, including computer security, spam filtering, malware, optical character recognition, and many others.

For example, in the case of spam filtering:

Hypothesis: "The message is spam".

Null hypothesis: "The message is not spam".

Type I error (false positive): Spam filtering or spam blocking techniques wrongly classify a legitimate email message as spam and, as a result, interfere with its delivery.

Type II error (false negative): Spam email is not detected as spam, but is classified as non-spam.

While most anti-spam tactics can block or filter a high percentage of unwanted emails, doing so without creating significant false-positive results is a much more demanding task. A low number of false negatives is an indicator of the efficiency of spam filtering.

See also

icon Mathematics portal

Binary classification – Dividing things between two categories

Detection theory – Means to measure signal processing ability

Ethics in mathematics – Emerging field of applied ethics

False discovery rate – Statistical method for handling multiple comparisons

False positive paradox – Logic error due to ignoring the base rate

Family-wise error rate – Probability of making type I errors when performing multiple hypotheses tests

Information retrieval performance measures – Obtaining information resources relevant to an information need

Lemma (mathematics) – Theorem for proving more complex theorems

Jerzy Neyman – Polish American mathematician

Neyman–Pearson lemma – Theorem about the power of the likelihood ratio test

Null hypothesis – Position that there is no relationship between two phenomena

Probability of a hypothesis for Bayesian inference – Method of statistical inference

Egon Pearson – British statistician (1895–1980)

Precision and recall – Pattern-recognition performance metrics

Prosecutor's fallacy – Logic error due to ignoring the base rate

Prozone phenomenon – Immunologic phenomenon occurring in high antigen or antibody levels

Receiver operating characteristic – Diagnostic plot of binary classifier ability

Sensitivity and specificity – Statistical measure of a binary classification

Statisticians' and engineers' cross-reference of statistical terms

Testing hypotheses suggested by the data – Problem of circular reasoning in statistics

Type III error – Term in statistical hypothesis testing

STATISTICS

Type I (α) and Type II (β) Errors and Power (1-β)

Type I Error (False Positive)

• Alpha (α) is the probability that the test will lead to the rejection of the hypothesis tested when that hypothesis is true.

Hypothesis: The medical device results in an improved outcome. α=0.05 means that there is only a 5% probability that this is wrong; i.e., low chance of a false positive. In other words, we are 95% sure that the new device is better than the control.

There could be dire consequences if we are wrong because the new device would be used even though it is no better than the control. Therefore, we have to be confident that we are right (i.e., a 95% probability that the new device is better than the control).

Type II Error (False Negative)

• Beta (β) is the probability that the test will reject the hypothesis tested when a specific alternative hypothesis is true; 1-β is the “power.”

Hypothesis: The medical device results in no improvement in outcome. β=0.2 means that there is only a 20% probability that the new device is shown by the study to be the same as the control, when it is actually better; i.e., a 20% chance of a false negative.

Not as much a problem if we are wrong in saying that the new device is not better than control, because the new device would just not be used. Therefore we can accept only an 80% probability (i.e., power) that there is no difference between the device and control.

In meeting the criterion of $\alpha \leq 0.05$ it is shown that there is a high probability (95%) that the new device is better than the control. The study should have a sufficient number of patients in each group (i.e., a sufficient power; 80% is enough) such that if there were no difference it would have been detected.

See: http://calculators.stat.ucla.edu/

Source: https://www.statsig.com/blog/why-the-uplift-in-a-b-tests-often-differs-from-real-world

Why the uplift in A/B tests often differs from real-world results

As someone who has worked with numerous clients over the years, I've frequently encountered the same frustration: A/B tests show promising results, but once the feature is launched, the anticipated uplift just doesn't materialize.

This disconnect can be puzzling and disappointing, especially when decisions and expectations are built around these tests. Understanding why the uplift seen in A/B tests often differs from real-world outcomes is essential for product managers, data scientists, and stakeholders.

Today, we'll explore some of the key reasons behind this discrepancy and offer insights on how to manage expectations effectively.

Human bias in analysis and interpretation

Human bias plays a significant role in the disconnect between A/B test results and real-world performance. The natural inclination to "win" a test can introduce bias into both the analysis and interpretation of results. Confirmation bias, where analysts favor information that supports their preconceptions, can lead to selective reporting of positive outcomes while overlooking negative or neutral results.

A common example I've encountered with clients involves tests that yield inconclusive (non-significant) results. Instead of ending the test, they often decide to let it run longer to "give the treatment an opportunity to win." If the results later turn positive (significant), they immediately close the test, interpreting the delayed success as a true positive effect. This approach, however, is highly susceptible to bias and can lead to misleading conclusions.

Blind analysis and peer review are two strategies that can help counteract this bias. Blind analysis involves analyzing data without knowing which group is the control or treatment, reducing the chance of biased interpretations. Peer review adds an additional layer of scrutiny, where other experts review the methodology and findings, helping to catch any biases or errors that might have been overlooked.

False positives

False positives occur when an A/B test incorrectly indicates that a change has a significant effect when it doesn't. This error can mislead stakeholders into believing that a feature will perform better than it actually will post-launch.

Consider this example: Suppose only 10% of your A/B test ideas actually have a positive effect. This is a situation reported by several large companies. If you run tests with 80% statistical power and a 5% significance level (for one-tailed tests), 8% of your A/B tests will yield true positives (10% * 80%), while 4.5% will be false positives (90% * 5%). This means that more than a third (36% = 4.5% / (4.5% + 8%)) of the positive significant results are falsely significant! That's a significant proportion of misleading results.

While reducing the significance level can decrease the number of false positives, it would also require longer test durations, which may not always be feasible.

Sequential testing and overstated effect sizes

Sequential testing, where data is analyzed at multiple points during the experiment, is a common practice in A/B testing. It allows teams to monitor results and potentially stop a test early if results seem favorable. However, research has shown that even when done properly, sequential testing can introduce bias and overstate effect sizes.

This overestimation occurs because stopping a test at the first sign of positive results might capture a momentary peak rather than a true, long-term effect. While there are methods to correct this bias and create an unbiased estimator, we won't delve into those techniques in this post. It's crucial, however, to be aware that sequential testing can lead to inflated expectations, which may not hold up in the real world.

Novelty effect and user behavior

Another key reason for the discrepancy is the novelty effect. When users are exposed to something new, especially in a controlled testing environment, they often react positively simply because it's different or exciting. This effect is particularly pronounced among existing users who are already familiar with the product.

For example, imagine a new user interface feature that initially boosts engagement during the A/B test phase. Users, intrigued by the fresh design, may interact with it more frequently. However, as time passes and the novelty wears off, their behavior tends to revert to their usual patterns. The initial uplift observed in the test diminishes, leading to less impressive results post-launch.

External validity and real-world factors

Another critical factor to consider is external validity, which refers to how well the results of an A/B test generalize to real-world settings. A/B tests are typically conducted in controlled environments where variables can be carefully managed. However, the real world is far more complex, with numerous external factors influencing user behavior.

For instance, seasonality, marketing efforts, or competitive actions can significantly impact the performance of a feature post-launch. A feature that performed well during a test in a quiet period might not do as well during a busy season, or vice versa. This variability can lead to a significant difference between test results and real-world outcomes.

Limited exposure in testing

The limited exposure of some A/B tests can also lead to discrepancies between test results and real-world performance. Tests are often conducted in specific parts of a product or funnel, meaning that not all users are exposed to the treatment. This can limit the generalizability of the test results.

For example, suppose there are two pages where a product can be purchased, and 50% of purchases are made on each page. If you test only one page and see a significant positive lift in purchases, the real effect on overall purchases, assuming independence between the pages, will be halved. This underlines the importance of considering the full user experience when interpreting A/B test results.

Strategies for mitigating discrepancies

To better align test results with real-world performance, teams can consider several strategies:

1. **Repeated Tests**: Running repeated tests involves conducting the same test multiple times to verify the results. This approach is excellent for mitigating human bias, false positives, novelty effects, and external validity issues. However, it requires more time and operational resources, which may not always be feasible.

2. **Using Smaller Significance Levels**: Reducing the significance level (e.g., from 0.05 to 0.01) directly decreases the likelihood of false positives. This is an easy strategy to implement, but it will extend the test duration, which could be a drawback in time-sensitive situations.
3. **Employing Holdout Groups**: Holdout groups involve keeping a segment of users who are not exposed to the test, serving as a control group. While setting this up internally can be challenging, many A/B testing platforms like Statsig and Eppo offer this as a feature. This method has a similar effect to repeated tests, providing more reliable results.
4. **Maintaining a Healthy Skepticism About Test Results**: Always approach test results with a critical eye, especially when the outcomes are unexpectedly positive. This mindset is particularly important in avoiding human bias and ensuring objective analysis.
5. **Conducting Blind Analyses**: Analyzing data without knowing which group is the control or treatment helps reduce bias. This technique ensures that conclusions are drawn based on the data alone, without any preconceived notions influencing the outcome.
6. **Involving Peer Reviews**: Peer review adds an additional layer of scrutiny to your analysis, helping to identify potential biases or errors. This collaborative approach can significantly improve the reliability of your conclusions.
7. **Checking Effect Sizes Over Time:** Monitoring effect sizes over time can reveal trends that might not be apparent in a single snapshot. This approach helps identify whether observed effects are stable or if they diminish as the novelty wears off or other factors come into play.
8. **Correcting Biases in Sequential Testing:** Being aware of and correcting for biases introduced by sequential testing can help ensure that the effect sizes reported are accurate and reliable.
9. **Calculating the Overall Effect When There Is Limited Exposure in Testing:** When tests are conducted in limited areas of a product or funnel, it's crucial to calculate the overall effect across all relevant areas. This approach ensures that the impact on the entire user base is accurately represented.

By incorporating these strategies, teams can set more realistic expectations and improve the accuracy of their predictions, leading to better decision-making and ultimately more successful product launches.

Takeaways

Understanding the reasons behind the discrepancy between A/B test results and real-world outcomes is crucial for anyone involved in product development and decision-making. By being aware of factors like human bias, false positives, sequential testing, novelty effects, and external validity, and by implementing strategies to mitigate these issues, you can better manage expectations and achieve more reliable results. Ultimately, these practices lead to more informed decisions and successful product launches.

Source: https://towardsdatascience.com/figuring-out-the-most-unusual-segments-in-data-af5fbeacb2b2/

Wise pizza

Figuring out the most unusual segments in data

How to find segments to focus on using common sense and machine learning

Analysts often have tasks of finding the "interesting" segments – the segments where we could focus our efforts to get the maximum potential impact. For example, it may be interesting to determine what customer segments have the most significant effect on churn. Or you could try to understand what types of orders affect customer support workload and the company's revenue.

Of course, we could look at graphs to find such outstanding features. But it may be time-consuming because we usually track dozens or even hundreds of customers' characteristics. More than that, we need to look at combinations of different factors so that it may lead to a combinatorial explosion. With such tasks, a framework would be really helpful because it could save you hours of analysis.

In this article, I would like to share with you two approaches for finding the most outstanding slices of data:

- based on common sense and basic maths,
- based on machine learning – our data science team at Wise has open-sourced a library ***Wise Pizza*** that gives you answers in three lines of code.

Example: Churn for bank customers

You can find the complete code for this example on GitHub.

We will be using data for bank customers' churn as an example. This dataset can be found on Kaggle under CC0: Public Domain license.

We will try to find the segments with the most significant impact on churn using different approaches: graphs, common sense and machine learning. But let's start with data preprocessing.

The dataset lists customers and their characteristics: credit score, country of residency, age & gender, how much money customers have on balance etc. Also, for each customer, we know whether they churned or not – parameter exited.

Our main goal is to find the customer segments with the highest impact on the number of churned customers. After that, we could try to understand the problems specific to these user groups. If we focus on fixing issues for these segments, we will have the most significant effect on the number of churned customers.

To simplify calculations and interpretations, we will define segments as sets of filters, for example, gender = Male or gender = Male, country = United Kingdom.

We will be working with discrete characteristics, so we have to transform continuous metrics, such as age or balance. For this, we could look at distributions and define suitable buckets. For example, let's look at age.

Code example for bucketing continuous characteristic

```
def get_age_group(a):
    if a &lt; 25:
        return '18 - 25'
    if a &lt; 35:
        return '25 - 34'
    if a &lt; 45:
        return '35 - 44'
    if a &lt; 55:
        return '45 - 54'
    if a &lt; 65:
        return '55 - 64'
    return '65+'


raw_df['age_group'] = raw_df.age.map(get_age_group)
```

The most straightforward way to find intriguing segments in data is to look at visualisations. We can look at churn rates split by one or two dimensions using bar charts or heat maps.

Let's look at the correlation between age and churn. Churn rates are low for customers under 35 years – less than 10%. While for customers between 45 and 64 years, retention is the worst – almost half of customers have churned.

Let's add one more parameter (gender) to try to find more complex relations. Barchart won't be able to show us two-dimensional relationships, so let's switch to a heatmap.

Churn rates for females are higher for all age groups, so gender is an influential factor.

Such visualisations can be pretty insightful, but there are a couple of problems with this approach:

- we don't take into account the size of segments,
- it may be time-consuming to look at all possible combinations of characteristics you have,
- it's challenging to visualize more than two dimensions in one graph.

So let's move on to more structured approaches that will help us to get a prioritized list of interesting segments with estimated effects.

Common sense approach

Assumptions

How could we calculate the potential impact of fixing problems for a specific segment? We can compare it to the "ideal" scenario with a lower churn rate.

You may wonder how we could estimate the benchmark for churn rate. There are several ways to do it:

- **benchmarks from the market:** you can try to search for typical churn rates levels for products in your domain,
- **high-performing segments in your product:** usually, you have a bit better-performing segments (for example, you can split by country or platform) and you can use them as a benchmark,
- **average value:** the most conservative approach is looking at the global mean value and estimating the potential effect of reaching the average churn rates for all segments.

Let's play safe and use the average churn rate from our dataset as a benchmark – 20.37%.

Listing all possible segments

The next step is to build all possible segments. Our dataset has ten dimensions with 3–6 unique values for each. The total number of combinations is around 1.2M. It looks computationally costly even though we have just a few dimensions and different values for them. In actual tasks, you usually have dozens of characteristics and unique values.

We definitely need to think about some performance optimizations. Otherwise, we may have to spend hours waiting for results. Here are a couple of tips on reducing computations:

- First of all, we don't need to build all possible combinations. It will be reasonable to limit the depth to 4–6. The possibility that your product team should focus on a user segment defined by 42 different filters is pretty low.
- Secondly, we may define the size of the effect we are interested in. Let's say we would like to increase the retention rate by at least 1% point. It means we are not interested in segments with a size of less than 1% of all users. Then we can stop splitting a segment further if its size is below this threshold – it will reduce the number of operations.
- Last but not least, you can significantly reduce the data size and resources spent on calculations in real-life datasets. For that, you can group all small characteristics for each dimension into an other group. For example, there are hundreds of countries, and each country's users' share usually follows Zipf's law as with many other real data relations. So you will have many countries with a size of less than 1% of all users. As we discussed earlier, we are not interested in such small user groups, and we can just group them all into one segment country = other to make calculations easier.

We will be using recursion to build all combinations of filters up to max_depth. I like this concept of computer science because, in many cases, it allows you to solve complex problems elegantly. Unfortunately, data analysts rarely face the need to write recursive code – I can remember three tasks through 10 years of data analysis experience.

The idea of recursion is pretty straightforward – it's when your function calls itself during the execution. It's handy when you are working with hierarchies or graphs. If you would like to learn more about recursion in Python, read this article.

The high-level concept in our case is the following:

- We start with the entire dataset and no filters.

- Then we try to add one more filter (if the segment size is big enough and we haven't reached maximum depth) and apply our function to it.
- Repeat the previous step until conditions are valid.

```
num_metric = 'exited'
denom_metric = 'total'
max_depth = 4

def convert_filters_to_str(f):
    lst = []
    for k in sorted(f.keys()):
        lst.append(str(k) + ' = ' + str(f[k]))

    if len(lst) != 0:
        return ', '.join(lst)
    return ''

def raw_deep_dive_segments(tmp_df, filters):
    # return segment
    yield {
        'filters': filters,
        'numerator': tmp_df[num_metric].sum(),
        'denominator': tmp_df[denom_metric].sum()
    }

    # if we haven't reached max_depth then we can dive deeper
    if len(filters) &lt; max_depth:
        for dim in dimensions:
            # check if this dimensions has already been used
            if dim in filters:
                continue
```

```
        # deduplication of possible combinations
        if (filters != {}) and (dim &lt; max(filters.keys())):
            continue

        for val in tmp_df[dim].unique():
            next_tmp_df = tmp_df[tmp_df[dim] == val]

            # checking if segment size is big enough
            if next_tmp_df[denom_metric].sum() &lt; min_segment_size:
                continue

            next_filters = filters.copy()
            next_filters[dim] = val

            # executing function for subsequent segment
            for rec in raw_deep_dive_segments(next_tmp_df, next_filters):
                yield rec

# aggregating all segments for dataframe
segments_df = pd.DataFrame(list(raw_deep_dive_segments(df, {})))
```

As a result, we got around 10K segments. Now we can calculate the estimated effects for each, filter segments with negative effects and look at the user groups with the highest potential impact.

```
baseline_churn = 0.2037
segments_df['churn_share'] = segments_df.churn/segments_df.total
segments_df['churn_est_reduction'] = (segments_df.churn_share - baseline_churn)
    *segments_df.total
segments_df['churn_est_reduction'] = segments_df['churn_est_reduction']
  .map(lambda x: int(round(x)))

filt_segments_df = segments_df[segments_df.churn_est_reduction &gt; 0]
```

```
    .sort_values('churn_est_reduction', ascending = False).set_index('segment')
```

It should be a Holly Graal that gives all the answers. But wait, there are too many duplicates and segments subsequent to one another. Could we reduce duplication and keep only the most informative user groups?

Grooming

Let's look at a couple of examples.

The churn rate for the child segment age_group = 45–54, gender = Male is lower than age_group = 45–54. Adding a gender = Male filter doesn't bring us closer to the specific problem. So we can eliminate such cases.

The example below shows the opposite situation: the churn rate for the child segment is significantly higher, and, more than that, the child segment includes 80% of churned customers from the parent node. In this case, it's reasonable to eliminate a credit_score_group = poor, tenure_group = 8+ segment because the main problem is within a is_active_member = 0 group.

Let's filter all those not-so-interesting segments.

```
import statsmodels.stats.proportion


# getting all parent - child pairs

def get_all_ancestors_recursive(filt):
    if len(filt) &gt; 1:
        for dim in filt:
            cfilt = filt.copy()
            cfilt.pop(dim)
            yield cfilt
            for f in get_all_ancestors_recursive(cfilt):
                yield f


def get_all_ancestors(filt):
    tmp_data = []
    for f in get_all_ancestors_recursive(filt):
        tmp_data.append(convert_filters_to_str(f))
    return list(set(tmp_data))


tmp_data = []
```

```
for f in tqdm.tqdm(filt_segments_df['filters']):
    parent_segment = convert_filters_to_str(f)
    for af in get_all_ancestors(f):
        tmp_data.append(
            {
                'parent_segment': af,
                'ancestor_segment': parent_segment
            }
        )

full_ancestors_df = pd.DataFrame(tmp_data)

# filter child nodes where churn rate is lower

filt_child_segments = []

for parent_segment in tqdm.tqdm(filt_segments_df.index):
    for child_segment in full_ancestors_df[full_ancestors_df.parent_segment ==
parent_segment].ancestor_segment:
        if child_segment in filt_child_segments:
            continue

        churn_diff_ci = statsmodels.stats.proportion.confint_proportions_2indep(
            filt_segments_df.loc[parent_segment][num_metric],
            filt_segments_df.loc[parent_segment][denom_metric],
            filt_segments_df.loc[child_segment][num_metric],
            filt_segments_df.loc[child_segment][denom_metric]
        )

        if churn_diff_ci[0] &gt; -0.00:
```

```
        filt_child_segments.append(
          {
            'parent_segment': parent_segment,
            'child_segment': child_segment
          }
        )

filt_child_segments_df = pd.DataFrame(filt_child_segments)
filt_segments_df =
filt_segments_df[~filt_segments_df.index.isin(filt_child_segments_df.child_segment.values)]

# filter parent nodes where churn rate is lower

filt_parent_segments = []

for child_segment in tqdm.tqdm(filt_segments_df.index):
   for parent_segment in full_ancestors_df[full_ancestors_df.ancestor_segment ==
child_segment].parent_segment:
      if parent_segment not in filt_segments_df.index:
        continue

      churn_diff_ci = statsmodels.stats.proportion.confint_proportions_2indep(
        filt_segments_df.loc[parent_segment][num_metric],
        filt_segments_df.loc[parent_segment][denom_metric],
        filt_segments_df.loc[child_segment][num_metric],
        filt_segments_df.loc[child_segment][denom_metric]
      )
      child_coverage =
filt_segments_df.loc[child_segment][num_metric]/filt_segments_df.loc[parent_segment][num_metr
ic]

      if (churn_diff_ci[1] &lt; 0.00) and (child_coverage &gt;= 0.8):
```

```
        filt_parent_segments.append(
          {
            'parent_segment': parent_segment,
            'child_segment': child_segment
          }
        )


filt_parent_segments_df = pd.DataFrame(filt_parent_segments)

filt_segments_df =
filt_segments_df[~filt_segments_df.index.isin(filt_parent_segments_df.parent_segment.values)]
```

Now we have around 4K interesting segments. With this toy dataset, we see little difference after this grooming for the top ones. However, with real-life data, these efforts often pay out.

Root causes

The last thing we can do to leave the most meaningful slices is to keep only the root nodes of our segments. These segments are the root causes, and others are included in them. If you would like to dig deeper into one of the root causes, look at child nodes.

To get only the root causes, we need to eliminate all segments for which we have a parent node in our final list of interesting ones.

```
root_segments_df = filt_segments_df[~filt_segments_df.index.isin(
  full_ancestors_df[full_ancestors_df.parent_segment.isin(
    filt_segments_df.index)].ancestor_segment
  )
]
```

So here it is, now we have a list of user groups to focus on. We got only one-dimensional segments at the top since there are few complex relations in data where a couple of characteristics explain the full effect.

It’s crucial to discuss how we could interpret the results. We got a list of customer segments with the estimated impact. Our estimation is based on the hypothesis that we could decrease the churn rate for the whole segment to reach the benchmark level (in our example – the average value). So we estimated the impact of fixing the problems for each user group.

You must keep in mind that this approach only gives you a high-level view of what user groups to focus on. It doesn’t take into account whether it’s possible to fix these problems entirely or not.

We’ve written quite a lot of code to get results. Maybe there’s another approach to solving this task using data science and machine learning that won’t require so much effort.

Pizza time

Actually, there is another way. Our data science team at Wise has developed a library ***Wise Pizza*** that could find the most intriguing segments in a blink of an eye. It's open-sourced under Apache 2.0 license, so you also could use it for your tasks.

If you are interested to learn more about **Wise Pizza** library, don't miss Egor's presentation on Data Science Festival.

Applying Wise Pizza

The library is easy to use. You need to write just a couple of lines and specify the dimensions and number of segments you want in a result.

```
# pip install wise_pizza - for installation

import wise_pizza


# building a model

sf = wise_pizza.explain_levels(

    df=df,

    dims=dimensions,

    total_name="exited",

    size_name="total",

    max_depth=4,

    min_segments=15,

    solver="lasso"

)


# making a plot

sf.plot(width=700, height=100, plot_is_static=False)
```

As a result, we also got a list of the most interesting segments and their potential impact on our product churn. Segments are similar to the ones we've obtained using the previous approach. However, the impact estimations differ a lot. To interpret Wise Pizza results correctly and understand the differences, we need to discuss how it works in more detail.

How it works

The library is based on Lasso and LP solvers. If we simplify it, the library does something similar to one-hot-encoding, adding flags for segments (the same ones we've calculated before) and then uses Lasso regression with churn rate as a target variable.

As you may remember from machine learning, the Lasso regression tends to have many zero coefficients, selecting a few significant factors. ***Wise Pizza*** finds the appropriate alpha coefficient for Lasso regression so that you will get a specified number of segments as a result.

For revising Lasso (L1) and Ridge (L2) regularisations, you could consult the article.

How to interpret results

Impact is estimated as the result of multiplication of coefficient and segment size.

So as you could see, it's completely different to what we've estimated before. The common sense approach estimates the impact of completely fixing the problems for user groups, while Wise Pizza's impact shows incremental effects to other selected segments.

The advantage of this approach is that you can sum up different effects. However, you need to be accurate during the results' interpretations because the impact for each segment depends on other selected segments since they may be correlated. For example, in our case, we have three correlated segments:

- age_group = 45-54
- num_of_products = 1, age_group = 44–54
- is_active_member = 1, age_group = 44–54.

The impact for age_group = 45–54 grasps potential effects for the whole age group, while others estimate additional impact from specific subgroups. Such dependencies may lead to significant results differences depending on min_segments parameter, because you will have different sets of final segments and correlations between them.

It's crucial to pay attention to the whole picture and interpret ***Wise Pizza*** results correctly. Otherwise, you may jump to the wrong conclusions.

I appreciate this library as an invaluable tool for getting quick insights from data and the first segment candidates to dive deeper. However, suppose I need to do opportunity sizing and more robust analysis to share the potential impact of our focus with my product team. In that case, I still use a common sense approach with a reasonable benchmark because it's much easier to interpret.

TL;DR

1. Finding interesting slices in your data is a common task for analysts (especially at the discovery stage). Luckily, you don't need to make dozens of graphs to solve such questions. There are frameworks which are more comprehensive and easy-to-use.
2. You can use the ***Wise Pizza*** ML library to get quick insights on the segments with the most significant impact on average (*it also allows you to look at the difference between two datasets*). I usually use it to get the first list of meaningful dimensions and segments.
3. ML approach can give you a high-level view and prioritization in a blink of an eye. However, I recommend you to pay attention to results interpretation and make sure you and you stakeholders fully understand it. However, if you need to do a robust estimation of potential effect on KPIs of fixing problems for the whole user group, it's worth using a good old common sense approach based on arithmetics.

Source: https://en.wikipedia.org/wiki/Two-proportion_Z-test

Two-proportion Z-test

The Two-proportion Z-test (or, Two-sample proportion Z-test) is a statistical method used to determine whether the difference between the proportions of two groups, coming from a binomial distribution is statistically significant.[1] This approach relies on the assumption that the sample proportions follow a normal distribution under the Central Limit Theorem, allowing the construction of a z-test for hypothesis testing and confidence interval estimation. It is used in various fields to compare success rates, response rates, or other proportions across different groups.

Hypothesis test

The z-test for comparing two proportions is a Statistical hypothesis test for evaluating whether the proportion of a certain characteristic differs significantly between two independent samples. This test leverages the property that the sample proportions (which is the average of observations coming from a Bernoulli distribution) are asymptotically normal under the Central Limit Theorem, enabling the construction of a z-test.

The test involves two competing hypotheses:

Null hypothesis (H0): The proportions in the two populations are equal, i.e.,

p

1

=

p

2

{\displaystyle p_{1}=p_{2}}.

Alternative hypothesis (H1): The proportions in the two populations are not equal, i.e.,

p

1

≠

p

2

{\displaystyle p_{1}\neq p_{2}} (two-tailed) or

p

1

>

p

2

{\displaystyle p_{1}>p_{2}} /

p

1

<

p

2

{\displaystyle p_{1}<p_{2}} (one-tailed).

The z-statistic for comparing two proportions is computed using:[2]

z

=

p

^

1

−

p

^

2

p

^

(

1

−

p

^

)

(

1

n

1

+

1

n

2

)

{\displaystyle z={\frac {{\hat {p}}_{1}-{\hat {p}}_{2}}{\sqrt {{\hat {p}}(1-{\hat {p}})\left({\frac {1}{n_{1}}}+{\frac {1}{n_{2}}}\right)}}}}

Where:

p

^

1

{\displaystyle {\hat {p}}_{1}} = sample proportion in the first sample

p

^

2

{\displaystyle {\hat {p}}_{2}} = sample proportion in the second sample

n

1

{\displaystyle n_{1}} = size of the first sample

n

2

{\displaystyle n_{2}} = size of the second sample

p

^

{\displaystyle {\hat {p}}} = pooled proportion, calculated as

p

^

=

x

1

+

x

2

n

1

+

n

2

{\displaystyle {\hat {p}}={\frac {x_{1}+x_{2}}{n_{1}+n_{2}}}}, where

x

1

{\displaystyle x_{1}} and

x

2

{\displaystyle x_{2}} are the counts of successes in the two samples.

The pooled proportion is used to estimate the shared probability of success under the null hypothesis, and the standard error accounts for variability across the two samples.

The z-test determines statistical significance by comparing the calculated z-statistic to a critical value. E.g., for a significance level of

α

=

0.05

{\displaystyle \alpha =0.05} we reject the null hypothesis if

|

z

|

>

1.96

{\displaystyle |z|>1.96} (for a two-tailed test). Or, alternatively, by computing the p-value and rejecting the null hypothesis if

p

<

α

{\displaystyle p<\alpha }.

Confidence interval

The confidence interval for the difference between two proportions, based on the definitions above, is:

(

p

^

1

–

p

^

2

)

±

z

α

/

2

p

^

1

(

1

–

p

^

1

)

n

1

+

p

^

2

(

1

–

p

^

2

)

n

2

{\displaystyle ({\hat {p}}_{1}-{\hat {p}}_{2})\pm z_{\alpha /2}{\sqrt {{\frac {{\hat {p}}_{1}(1-{\hat {p}}_{1})}{n_{1}}}+{\frac {{\hat {p}}_{2}(1-{\hat {p}}_{2})}{n_{2}}}}}}

Where:

z

α

/

2

{\displaystyle z_{\alpha /2}} is the critical value of the standard normal distribution (e.g., 1.96 for a 95% confidence level).

This interval provides a range of plausible values for the true difference between population proportions.

Using the z-test confidence intervals for hypothesis testing would give the same results as the chi-squared test for a two-by-two contingency table.[3]: 216–7 [4]: 875 Fisher's exact test is more suitable for when the sample sizes are small.

Notice how the variance estimation is different between the hypothesis testing and the confidence intervals. The first uses a pooled variance (based on the null hypothesis), while the second has to estimate the variance using each sample separately (so as to allow for the confidence interval to accommodate a range of differences in proportions). This difference may lead to slightly different results if using the confidence interval as an alternative to the hypothesis testing method.

Minimum detectable effect (MDE)

The minimum detectable effect (MDE) is the smallest difference between two proportions (

p

1

{\displaystyle p_{1}} and

p

2

{\displaystyle p_{2}}) that a statistical test can detect for a chosen Type I error level (

α

{\displaystyle \alpha }), statistical power (

1

−

β

{\displaystyle 1-\beta }), and sample sizes (

n

1

{\displaystyle n_{1}} and

n

2

{\displaystyle n_{2}}). It is commonly used in study design to determine whether the sample sizes allows for a test with sufficient sensitivity to detect meaningful differences.

The MDE for when using the (two-sided) z-test formula for comparing two proportions, incorporating critical values for

α

{\displaystyle \alpha } and

1

−

β

{\displaystyle 1-\beta }, and the standard errors of the proportions:[5][6]

MDE

=

|

p

1

−

p

2

|

=

z

1

–

α

/

2

p

0

(

1

–

p

0

)

(

1

n

1

+

1

n

2

)

+

z

1

–

β

p

1

(

1

−

p

1

)

n

1

+

p

2

(

1

−

p

2

)

n

2

{\displaystyle {\text{MDE}}=|p_{1}-p_{2}|=z_{1-\alpha /2}{\sqrt {p_{0}(1-p_{0})\left({\frac {1}{n_{1}}}+{\frac {1}{n_{2}}}\right)}}+z_{1-\beta }{\sqrt {{\frac {p_{1}(1-p_{1})}{n_{1}}}+{\frac {p_{2}(1-p_{2})}{n_{2}}}}}}

Where:

z

1

−

α

/

2

{\displaystyle z_{1-\alpha /2}}: Critical value for the significance level.

z

1

–

β

{\displaystyle z_{1-\beta }}: Quantile for the desired power.

p

0

=

p

1

=

p

2

{\displaystyle p_{0}=p_{1}=p_{2}}: When assuming the null is correct.

The MDE depends on the sample sizes, baseline proportions (

p

1

,

p

2

{\displaystyle p_{1},p_{2}}), and test parameters. When the baseline proportions are not known, they need to be assumed or roughly estimated from a small study. Larger samples or smaller power requirements leads to a smaller MDE, making the test more sensitive to smaller differences. Researchers may use the MDE to assess the feasibility of detecting meaningful differences before conducting a study.

[Proof]

Assumptions and conditions

To ensure valid results, the following assumptions must be met:

Independent random samples: The samples must be drawn independently from the populations of interest.

Large sample sizes: Typically,

n

1

{\displaystyle n_{1}} and

n

2

{\displaystyle n_{2}} should exceed 30. [citation needed]

Success/failure condition: [citation needed]

n

1

p

^

1

>

10

{\displaystyle n_{1}{\hat {p}}_{1}>10} and

n

1

(

1

−

p

^

1

)

>

10

{\displaystyle n_{1}(1-{\hat {p}}_{1})>10}

n

2

p

^

2

>

10

{\displaystyle n_{2}{\hat {p}}_{2}>10} and

n

2

(

1

−

p

^

2

)

>

10

{\displaystyle n_{2}(1-{\hat {p}}_{2})>10}

The z-test is most reliable when sample sizes are large, and all assumptions are satisfied.

Software implementation

R

Use prop.test() with continuity correction disabled:

prop.test(x = c(120, 150), n = c(1000, 1000), correct = FALSE)

Output includes z-test equivalent results: chi-squared statistic, p-value, and confidence interval:

2-sample test for equality of proportions without continuity correction

data: c(120, 150) out of c(1000, 1000)

X-squared = 3.8536, df = 1, p-value = 0.04964

alternative hypothesis: two.sided

95 percent confidence interval:

-5.992397e-02 -7.602882e-05

sample estimates:

prop 1 prop 2

0.12 0.15

Python

Use proportions_ztest from statsmodels:

from statsmodels.stats.proportion import proportions_ztest

```
z, p = proportions_ztest([120, 150], [1000, 1000], 0)

# For CI: from statsmodels.stats.proportion import proportions_diff_confint_indep
```