Foto: Michael Blann / thinkstockphotos.de

Michael Witzenleiter

# A/B-TESTING: VON NULL AUF STATISTIKER

„Man sieht sich immer zweimal im Leben." Dieses Sprichwort kommt vielen Conversion-Optimierern bekannt vor, wenn sie es im Rahmen des A/B-Testings plötzlich wieder mit Konzepten und Begrifflichkeiten der Statistik wie Konfidenzlevel, Stichprobengröße oder dem sogenannten Alpha-Fehler zu tun bekommen.
Der Beitrag von Michael Witzenleiter erklärt kurz zusammengefasst das wichtigste Know-how, das Sie benötigen, um dabei nicht ins Schwitzen zu kommen, und frischt vielleicht bereits vorhandenes Wissen wieder auf. Und da, anders als in der Schule oder der Universität, diesmal echte praktische Anwendbarkeit dahinter steht, macht es unter dieser Perspektive nicht nur Sinn, sondern durchaus sogar Spaß, sich das nochmals anzueignen.

Mal Hand aufs Herz: War Statistik in der Schule oder im Studium Ihr Lieblingsthema? Die wenigsten werden diese Frage mit „Ja" beantworten. Die schlechte Nachricht ist: Sie werden, wenn Sie sinnvolle und aussagekräftige A/B-Tests durchführen möchten, nicht an Statistik vorbeikommen.

*„A/B-Tests ohne statistisches Know-how sind nicht besser, als auf sein Bauchgefühl zu hören!"*

Und wer möchte seine businessrelevanten Entscheidungen aus dem Bauch heraus treffen? A/B-Tests sind nichts anderes als Experimente oder Feldversuche, wie man sie in der Psychologie oder in der allgemeinen Statistik am laufenden Band durchführt. Als sogenannte „Hypothesen-Tests" (oder auch: t-Tests) geht es darum, vorab getroffene Annahmen zu beweisen (verifizieren) oder zu widerlegen (falsifizieren). Zum Beispiel, dass die Hervorhebung eines Gütesiegels oder die Anpassung von Call-to-Action-Elementen (wie z. B. Bestell-Buttons) zu einer höheren Bestellrate der Website-Besucher führt.

Die gute Nachricht ist: Sie brauchen sich dafür nicht vor umfangreiche Statistik-Lehrbücher setzen, sondern können sich das Wissen durch Lesen dieses Artikels aneignen. Sinn und Zweck ist es, Sie mit dem Rüstzeug auszustatten, um A/B-Tests fundiert konzipieren zu können und Ergebnisse und Kennzahlen sinnvoll nutz- und interpretierbar zu machen. Da die Begrifflichkeiten und Prinzipien dahinter bei gängigen A/B-Testing-Tools ähnlich sind, werden Sie sich damit in jedem Tool zurechtfinden.

DER AUTOR

**Michael Witzenleiter** ist deutscher Geschäftsführer des Testing-Anbieters Kameleoon und stellvertretender Geschäftsführer der Performance-Marketing-Agentur Kesselhaus. Er beschäftigt sich seit über zehn Jahren mit dem Thema Conversion-Optimierung und Testing und publiziert regelmäßig dazu im Blog *conversion-matters.de*.

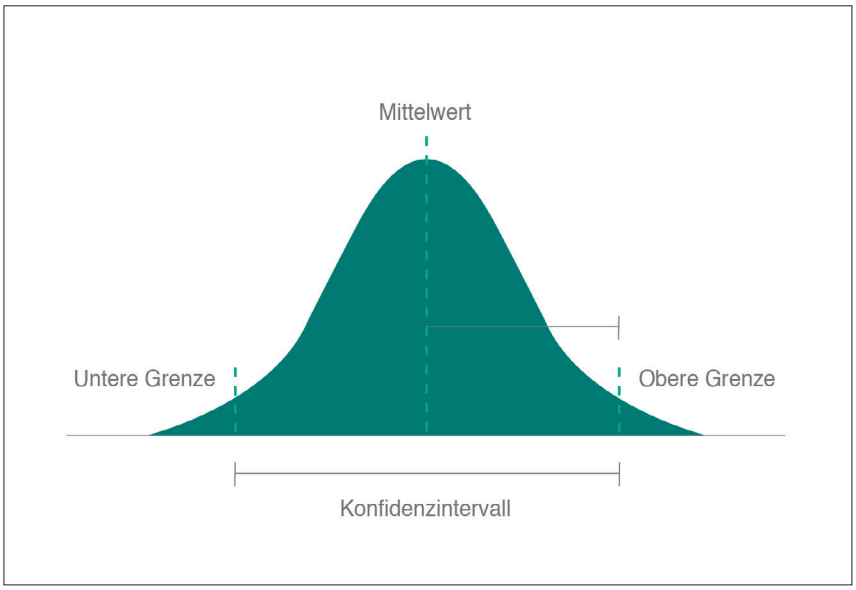



Abb. 1: Konfidenzintervall in Tests

## Was ist ein A/B-Test?

Bei einem A/B-Test bzw. Split-Test geht es darum, die derzeitige Version eines Systems oder in diesem Falle einer Webseite gegen eine veränderte Version zu testen, um Unterschiede in der Nutzerreaktion herauszufinden (siehe *de.wikipedia.org/wiki/A-B-Test*). Wie schon angedeutet, gilt es dabei, eine gute Hypothese pro Testvariante aufzustellen und mit dem Test zu überprüfen. Beispiele hierfür sind Layout-Änderungen oder zum Beispiel die Umbenennung von Navigationselementen. Dies kann man im einfachsten Fall mit zwei Varianten durchführen oder in Form eines A/B/n-Tests mit mehr als zwei Varianten. Grundsätzlich geht es immer darum, anhand einer Stichprobe im Rahmen eines Experiments zu einer allgemeinen Aussage über alle Nutzer zu gelangen. Hier beginnt die erste statistische Herausforderung: Wie wähle ich die ideale Stichprobe?

## Ideale Stichprobe und Varianz

Die Eingangsfrage jedes Experiments oder Tests in der Statistik ist die Definition der Stichprobengröße und -zusammensetzung, die am besten die Grundgesamtheit der Nutzer widerspiegelt. Anschauliches Beispiel: Bei einem Roulette-Spiel haben die Farben Rot und Schwarz dieselbe Wahrscheinlichkeit (48,6 %, es gibt noch die grüne 0). Das veranlasst viele, bei einer längeren Phase des Erscheinens von „Rot" krampfhaft auf „Schwarz" zu tippen, da beide gleich oft erscheinen müssten und nun „Schwarz" an der Reihe wäre. Das mag zwar bei rund 10.000 Runden Roulette in etwa stimmen, aber heißt noch lange nicht, dass die Verteilung auch bei zehn Runden 48,6 % ist. Dieser Irrtum über die „passende" Stichprobe dürfte schon einige ihrer Ersparnisse beraubt haben. Auf dieses Grundproblem wird bei der Testinterpretation noch einmal eingegangen.

Erschwert wird die Wahl der Stichprobe dadurch, dass der Traffic einer Webseite nie konstant die gleichen Eigenschaften aufweist. So konvertieren Besucher je nach Herkunft oder Endgerät besser oder schlechter als der „durchschnittliche Nutzer". Ebenso sind bestimmte Aktionen, die vielleicht während eines Tests laufen (zum Beispiel eine Sale-Aktion) oder schlicht und einfach die Tatsache, dass Nutzer je nach Wochentag besser oder schlechter konvertieren, potenzielle Störfeuer, die Ihre Testergebnisse verzerren können. Lange Test-Laufzeiten und große Stichprobengrößen helfen, solche Verzerrungen zu vermeiden. Die meisten Conversion-Optimierer empfehlen als grobe Faustregel mindestens 1.000 Conversions pro Variante bzw. eine Testlaufzeit von mindestens zwei, besser vier Wochen. Generell hängt die Stichprobengröße von folgenden vier Faktoren ab:

1. Die aktuelle Conversion-Rate des zu optimierenden Ziels (z. B. Bestellrate ist gleich 3,0 %)
2. Der generierte Uplift (die prozentuale Steigerung der Conversion, z. B. 5,0 %)
3. Das Konfidenzlevel (eine Aussagesicherheit von 95 % für Online-Tests wird empfohlen)
4. Die statistische Power (mehr im Abschnitt „Konfidenzlevel und Signifikanzlevel", die meisten Testing-Tools gehen von 80 % Power aus)

*Grundsätzlich gilt: Je größer der zu erwartende Uplift ist, desto weniger Fallzahlen benötigen Sie für einen aussagekräftigen Test.*

Für Seiten mit wenig Traffic kann es ein probates Mittel sein, bei Tests darauf zu achten, dass man die kritischsten und erfolgversprechendsten Punkte priorisiert angeht und sich nicht auf Feintunings mit geringen möglichen Uplifts konzentriert. Aber Vorsicht: Nur den zu erwartenden Uplift im Vorfeld hoch einzuschätzen, kann sich als Bärendienst erweisen, da Sie die benötigte Laufzeit Ihrer Tests unterschätzen könnten.

Die ideale Größe der Stichprobe bzw. die benötigte Testdauer, die sich ergibt, lässt sich mit vielen Online-Tools berechnen, was unbedingt vor einem Test gemacht werden sollte, zum Beispiel mit: *www.evanmiller.org/ab-testing/sample-size.html*.

Es gilt, die Stichprobe möglichst realistisch abzustecken. Bei „schwankenden" Stichproben haben Sie sonst eine sehr hohe Variabilität (auch: Varianz). Das bedeutet, dass die Streuung der betrachteten Variable (z. B. Conversion-Rate) sehr groß ist (siehe

> „Die größte Fehlerquelle eines Tests ist Ihre Ungeduld!“

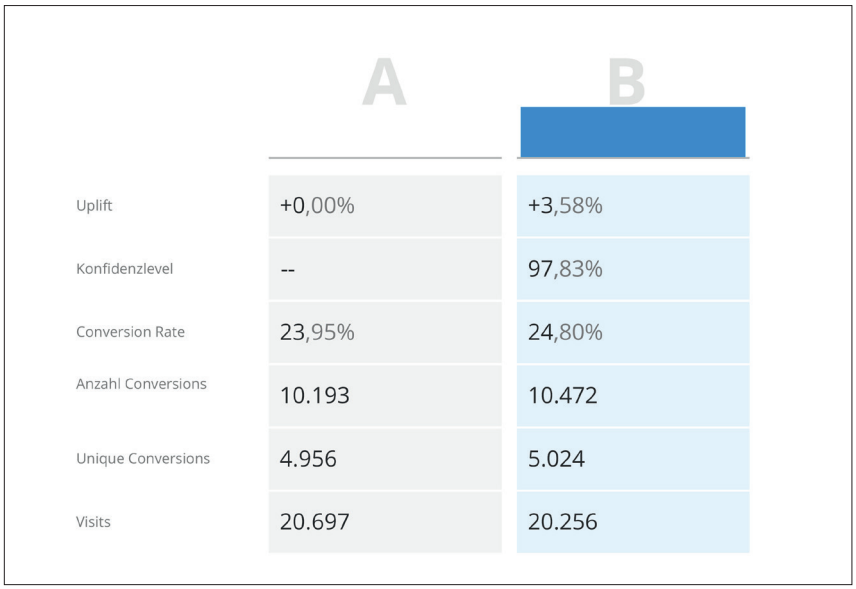
Abb. 1). Zum Beispiel liegt der Uplift (also die Verbesserung) der Variante im Schnitt (Mittelwert) in Ihrem Test bei 3,3 %. Wenn Sie sich die Stichprobe Ihres Tests genauer ansehen, wird es wie im Roulette-Beispiel Schwankungen geben. Mal wird die Conversion-Rate auf 100 Besucher bei 1,0 % liegen, mal sogar bei 5,6 %. Diese Abweichung um 2,3 % vom Mittelwert nennt man den statistischen Fehler. Damit lässt sich das Konfidenzintervall berechnen – die Spannweite der Werte von unten nach oben. Zum Beispiel lässt sich damit sagen, dass mit einer 95-prozentigen Wahrscheinlichkeit (Konfidenzlevel) der Uplift zwischen 1,0 % und 5,6 % liegt. Je schmaler die Spannweite, desto besser ist der wahre Wert des Uplifts eingegrenzt. Da viele A/B-Testing-Tools die Interpretation ihrer Daten möglichst einfach und verständlich gestalten möchten, wird meist nur der Mittelwert des Uplifts ausgewiesen. Daher lohnt sich ein genauer Blick darauf, wie sich die Varianz des Tests verhält bzw. wie hoch das Konfidenzintervall des Tests ist. Sonst gehen Sie ähnlich wie am Roulette-Tisch davon aus, dass Sie sofort 3,3 % mehr Conversions erzielen, und wundern sich, dass im dauerhaften Betrieb dann andere Werte (1 %–5,6 %) auftreten. Auch hier gilt das sogenannte Prinzip der „regression to the mean“, das besagt, dass eine hohe Fallzahl dazu führt, dass die Varianz geringer wird und somit eine „Regression“ (Angleichung) der Werte an den Mittelwert erfolgt.

## Warum ein A/A-Test dabei sinnvoll sein kann

Ein weiterer Tipp aus der Praxis ist die Durchführung eines sogenannten A/A-Tests vor dem eigentlichen geplanten Test. Ein A/A-Test ist kurz gesagt ein A/B-Test, bei dem die Variante keinerlei Veränderung enthält, also deckungsgleich mit der Original-Version ist. Man führt diese Art Test aus zwei Hauptgründen durch: Zum einen, um den sogenannten „Noise“ zu ermitteln, also die oben genannten Störfaktoren und Schwankungen im Uplift, um später diese Schwankungen mit den Ausschlägen des Tests zu vergleichen. Der zweite Grund ist, das Conversion-Tracking zu überprüfen, was meist eine gängige Fehlerquelle bei einem A/B-Test sein kann. Selbstredend führt auch hier die höhere Varianz zu Beginn des Tests für einige Überraschungen bei Neueinsteigern ins Thema Conversion-Optimierung, doch sofern die Aussteuerung des Tests und die eingerichteten Ziele korrekt sind, gleichen sich die Conversion-Rates der Varianten an. Da mit steigender Fallzahl der Uplift immer geringer wird, empfiehlt es sich, hier nicht auf ein Konfidenzlevel von 95 % zu warten, sondern vornweg eine Stopp-Bedingung in Form eines Enddatums oder einer maximalen Traffic-Menge zu definieren. Es ist wichtig zu erwähnen, dass ein A/A-Test tatsächlich einen signifikanten Unterschied zwischen den Varianten belegen kann, der dann jedoch vom Zufall abhängig ist; dazu mehr beim Thema „Alpha-Fehler.“

## Nullhypothese und Alternativhypothese

Zwei wichtige Konzepte im Rahmen von A/B-Tests sind die Konzepte der Nullhypothese sowie der Alternativhypothese. Bei der Nullhypothese geht es darum, dass man die Annahme, dass zwischen Original und Variante kein Zusammenhang besteht, also eine Variante keine bessere Conversion-Rate bringt, zu widerlegen bzw. zu bestätigen versucht. Das Konzept der Alternativhypothese basiert auf der Annahme, dass es eine Abhängigkeit gibt und man genau diese Hypothese bestätigen bzw. widerlegen möchte. Bei einem A/B-Test werden beide Hypothesen gegeneinander getestet. Sie sind also quasi wie Yin und Yang miteinander verbunden, wobei man in der Regel bei einem A/B-Test das Ziel verfolgt, die Nullhypothese abzulehnen und die Alternativhypothese anzunehmen. Also kurz gesagt möchte man beweisen, dass die Variante mit hoher Wahrscheinlichkeit mehr Conversions liefert. Eine Ausnahme hierbei kann der A/A-Test bilden. Die Krux bei diesen Verifizierungen bzw. Falsifizierungen ist nun die folgende: Sofern das Testergebnis deckungsgleich mit dem Ergebnis in der Grundgesamtheit ist, gibt es keine möglichen Fehlerquellen. Das bedeutet, sofern der Test bestätigt, dass die Variante keine Steigerung (Uplift) bringt und dies in der Realität der Grundgesamtheit ebenfalls der Fall ist, dann ist alles gut. Gleiches gilt für den Nachweis eines Uplifts in der Stichprobe und in der Grundgesamtheit. Das Problem ist nun jedoch, dass das in der Praxis so gut wie nie auftritt, also eine Variante zwar im Test besser abschneidet als das Original, was aber im laufenden Betrieb in der Grundgesamtheit nicht der Fall ist. In dem Fall spricht man vom sogenannten Fehler 1. Art oder auch Alpha-Fehler, während man im Falle der Identifikation keiner Gewinnervariante im Test, die jedoch in der Grundgesamtheit vorhanden ist, vom sogenannten Fehler 2. Ordnung oder auch Beta-Fehler spricht.



| | A | B |
|---|---|---|
| Uplift | +0,00% | +3,58% |
| Konfidenzlevel | -- | 97,83% |
| Conversion Rate | 23,95% | 24,80% |
| Anzahl Conversions | 10.193 | 10.472 |
| Unique Conversions | 4.956 | 5.024 |
| Visits | 20.697 | 20.256 |

Abb. 2: Beispiel-Report eines A/B-Tests

## Konfidenzlevel und Signifikanzlevel

Um den Alpha- und Beta-Fehler zu verstehen, sollte man sich mit den Themen Konfidenzlevel und Signifikanzlevel beschäftigen, die in der Praxis häufig verwechselt werden. Das Konfidenzlevel wurde bereits angesprochen und ist die sogenannte „Aussagewahrscheinlichkeit". Ein Konfidenzlevel von 95 %, das empfohlen wird, bedeutet, dass mit 95-prozentiger Wahrscheinlichkeit eine Aussage in der Grundgesamtheit zutrifft, also z. B. Variante A besser konvertiert als das Original. Das Signifikanzlevel besagt, dass man in dem Fall eine Irrtumswahrscheinlichkeit (100 % minus Konfidenzlevel) von 5 % akzeptiert. Der sogenannte p-Wert ist nun das Signifikanzlevel im tatsächlichen Test, also sozusagen der gemessene Zufall im Experiment. Damit lässt sich das Signifikanzniveau überprüfen. Einfach ausgedrückt ist der Test signifikant, sobald der p-Wert kleiner als das Signifikanzlevel, also im konkreten Beispiel 5 % ist.

Die sogenannte statistische Power eines Tests lässt sich als 100 % minus Wahrscheinlichkeit des Beta-Fehlers definieren. Hierbei setzen die meisten A/B-Testing-Tools wie angesprochen einen Wert von 80 % an.

## Welche Kennzahlen sollte ein Report beinhalten?

Üblicherweise dürften Sie bei der Auswertung Ihres Tests über ein A/B-Testing-Tool einen ähnlichen Report wie Abbildung 2 erhalten. Die wichtigsten Kennzahlen eines Tests sind in der Regel die folgenden: der erzielte Uplift (also die Steigerung der Conversion-Rate), die Conversion-Rate der Referenz und der Varianten, das Konfidenzlevel (also die Aussagekraft eines Tests) sowie die Anzahl der Conversions und des Traffics im Gesamten, der Ihnen einen Anhaltspunkt bezüglich der Aussagekraft des Tests liefert.

Sofern Sie Split-Tests über eigene Systeme durchführen und danach die Performancewerte der Varianten messen, müssen Sie vermutlich das Konfidenzlevel selbst berechnen. Auch hier sei wieder auf die Online-Tools aus dem oberen Abschnitt verwiesen.

## Die Hauptquelle für Testfehler

Sobald der erste Test gestartet ist, steigt naturgemäß die Spannungskurve und regelmäßig wird der Report des Tests aufgerufen und bewertet. Das stellt die größte Gefahr für die Bewertung Ihrer Testergebnisse dar. Denn so wie der Börsenguru André Kostolany empfiehlt, Aktien zu kaufen und möglichst lange liegen zu lassen, gilt auch bei A/B-Tests die Empfehlung, Geduld mitzubringen. Denn selbst wenn der Uplift zwischen zwei Versionen zunächst sehr groß erscheint, kann er sich im Laufe eines Tests sehr schnell nach oben und unten verändern. Ein sehr plastisches Beispiel dazu lässt sich in folgendem Video bei Vimeao entdecken: *www.vimeo.com/140802384*. Das einzige Mittel dagegen ist, regelmäßig die Entwicklung Ihres Konfidenzlevels pro Variante zu beobachten und Tests erst dann abzustellen, wenn auch über einen längeren Zeitverlauf keine größeren Sprünge oder Schwankungen mehr auftreten.

## Fazit

Das Statistik-Wissen, das Sie für A/B-Tests brauchen, ist sehr überschaubar und schnell gelernt. Sofern Ihnen das Konzept von Stichproben und Aussagewahrscheinlichkeiten (Stichwort: Konfidenzlevel und Signifikanzlevel) bekannt ist, dürften Sie im laufenden Betrieb keine Probleme mit unterschiedlichen Tools und Benennungen von Kennzahlen haben. Sofern doch einmal Begriffsunklarheiten auftauchen, lassen sich online unzählige Spickzettel für Ihre nächsten Statistikgespräche finden (zum Beispiel: *goo.gl/Rbi7CS*). Wobei die größte Gefahr für Ihre Tests in Ihrer möglichen Ungeduld lauert. Aber auch die lässt sich durch Erwartungsmanagement vorweg und intensives Training in den Griff bekommen. ¶