konversionsKRAFT

MIT STATISTIK CHEAT SHEET

EBOOK

DER STATISTIK-SPICKZETTEL FÜR OPTIMIERER

MIT DIESEM KNOWHOW WIRST DU SCHNELL ZUM TESTING-PRO!

Dr. Julia Engelmann

Dr. Julia Engelmann

Head of Data Analytics

konversions**KRAFT**

Dr. Julia Engelmann ist Head of Data Analytics bei der Web Arts AG. Nach der Promotion im Bereich Makroökonomie war Julia Engelmann bei der Zalando SE in Berlin im Bereich Data Intelligence und Conversion Optimierung tätig. Bei der Web Arts AG kümmert sie sich um die Themen Webanalytics, Big Data und Personalisierung.

Der Statistik-Spickzettel für Optimierer

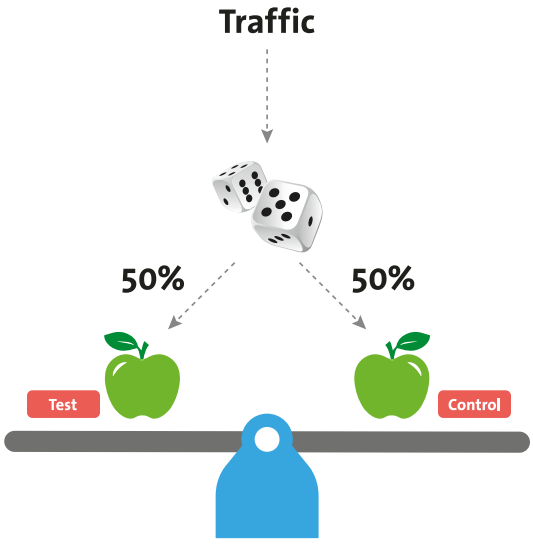
Wer sich mit A/B-Testing beschäftigt, stellt früher oder später fest, dass man dabei am Thema Statistik nicht vorbei kommt. In der Tat braucht man zumindest ein Grundverständnis der statistischen Begriffe und Zusammenhänge, damit man einen A/B-Test richtig plant und auswertet.

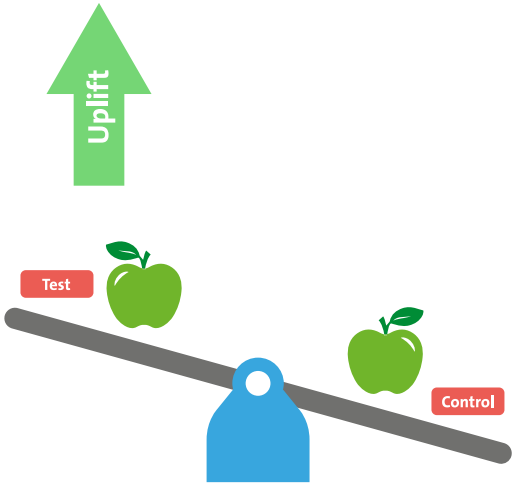
Wenn man nach den Begriffen „Statistik und A/B-Testing“ sucht, stößt man auf eine ganze Reihe von Artikeln, Büchern und Blogs. Allerdings sind auch viele der Inhalte entweder zu oberflächlich dargestellt oder viel zu komplex, um das große Bild und die Zusammenhänge zu verstehen.

Das Ziel dieses Guides ist es daher, die Statistik beim A/B-Testing kurz und knapp so zu erklären, dass man ohne Mathematik-Studium gut folgen kann und trotzdem in der Lage ist, die Ergebnisse eines A/B-Tests wirklich zu verstehen, zu interpretieren und die richtigen Ableitungen zu treffen.

Warum braucht man überhaupt Statistik beim A/B-Testing?

Damit man versteht, warum man sich überhaupt mit Begriffen wie Konfidenz, Signifikanz und Testgröße beschäftigen muss, ist es zunächst wichtig zu verstehen, was beim A/B-Testing eigentlich genau passiert: Ein A/B-Test ist aus statistischer Sicht nichts anderes als ein Experiment und daher untrennbar mit statistischen Methoden verbunden.

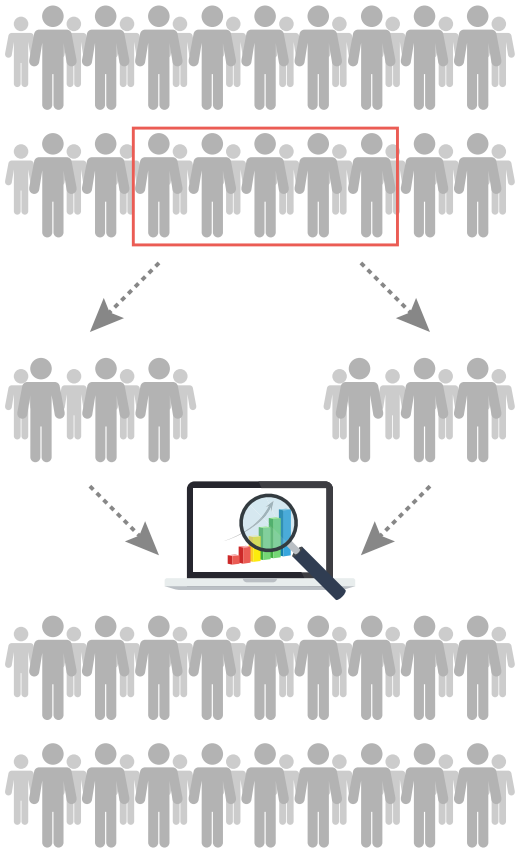




Beim A/B-Testing handelt es sich um einen sogenannten **statistischen Hypothesentest** und es geht darum, auf Basis einer **Stichprobe** herauszufinden, ob sich zwei Datensätze **signifikant** voneinander unterscheiden. Praktisch gesehen bedeutet das, dass man eine optimierte Variante einer Website B gegen die aktuelle Kontrollvariante A vergleicht und herausfinden möchte, ob die Besucher der optimierten Variante signifikant mehr kaufen als die Besucher der aktuellen Variante.

Dabei betrachtet man aber nur einen Ausschnitt aus der eigenen Nutzerbasis der Website. Es werden nie alle jemals dagewesenen Besucher einer Website im Test erfasst, sondern nur eine Stichprobe aus dieser Nutzerbasis, die man auch als **statistische Grundgesamtheit** bezeichnet.

Auf Basis dieser Stichprobe wird nun ein Test durchgeführt. Man möchte mit Hilfe der Stichprobe herausfinden, ob die Testvariante systematisch besser funktioniert und ob man den gemessen Uplift in der Conversion Rate verallgemeinern, also auf die gesamte Nutzerbasis übertragen, kann. Diese Methode bezeichnet man in der Statistik als Inferenzstatistik oder Hypothesentest:



Die Ergebnisse einer Stichprobe sollen dahingehend überprüft werden, ob sie allgemein gültige Aussagen für alle Website-Nutzer liefern. Man möchte am Ende eines A/B-Tests also sagen können: „ Die Testvariante führt mit hoher Wahrscheinlichkeit zu einer höheren Conversion Rate, deshalb sollten wir sie live stellen.“

Das Problem dabei ist allerdings, und hier kommt die Statistik ins Spiel, die auftretende Unsicherheit. Gerade weil man beim A/B-Test nur auf eine Stichprobe schaut, ist das Ergebnis eines Tests immer mit einer gewissen Unsicherheit verbunden. Man kann beim Test nie eine 100%ige Sicherheit erzielen und genau wissen, wie sich die nächsten 100.000 Besucher der Website verhalten werden. Diese Information haben wir zum heutigen Zeitpunkt schlichtweg nicht. Was man aber tun kann, und genau das passiert beim A/B Testing, ist, die ersten 10.000 Besucher zu beobachten und die Inferenzstatistik zu benutzen, um zu prognostizieren, wie sich die nächsten 90.000 Besucher vermutlich verhalten werden.

Die Ergebnisse eines A/B-Tests werden aber immer ein Stück weit schwanken, auch wenn man den selben Test unter gleichen Bedingungen wiederholt laufen lassen würde. Der Grund dafür sind **einfach zufällige Schwankungen**, die man bei jedem statistischen Experiment beobachtet, zum Beispiel allein deswegen, weil von Test zu Test unterschiedliche Besucher und Verhalten in der Stichprobe betrachtet werden.

Zusammengefasst geht es beim A/B-Testing also darum, auf Basis der gesammelten Informationen zum Verhalten der Besucher im konkreten Test möglichst gesicherte Aussagen über das Verhalten der gesamten Nutzerbasis einer Website zu machen, ohne deren konkretes Verhalten genau zu kennen.

Das Problem mit der statistischen Stichprobe

Das klingt doch soweit erstmal ganz logisch oder? Warum genau wird es dann so kompliziert?

Das ganze Verfahren funktioniert nur, wenn die betrachteten Stichproben, also die Test- und Kontrollgruppe, eine **repräsentative Teilmenge** der Grundgesamtheit sind.

Die Begriffe Stichprobe und Grundgesamtheit wurden bereits mehrfach erwähnt und bedürfen einer kurzen Erläuterung:

Grundgesamtheit = Gruppe, über die man allgemein gültige Aussagen treffen möchte, ohne dass man in der Lage ist, deren Verhalten genau zu beobachten. Beim A/B-Testing möchte man verlässliche Aussagen zum allgemeinen Verhalten aller Website Besucher treffen, ohne dass deren Verhalten im Test beobachtet werden kann.

Stichprobe = Ausschnitt aus der Grundgesamtheit. Beim A/B-Testing sind das die Test- und Kontrollgruppe, deren Verhalten man analysiert und miteinander vergleicht.

In einer perfekten Welt wäre die Stichprobe ein exaktes Abbild der Grundgesamtheit. Das heißt, das Verhalten der Besucher in der Stichprobe entspricht dem Verhalten der Besucher in der Grundgesamtheit. Ist die Stichprobe dann zusätzlich noch ausreichend groß, spricht man in der Statistik von der **Repräsentativität.**

Ein Beispiel

Angenommen, 10.000 von 100.000 Besuchern würden kaufen. Daraus ergäbe sich eine „wahre“ Conversion Rate von 10% für die Grundgesamtheit. In der perfekten Welt des A/B-Testers wäre die durchschnittliche Conversion Rate jeder gezogenen Stichprobe identisch mit dieser Conversion Rate der Grundgesamtheit. Anders gesagt: bei einer Stichprobe von 10 Besuchern würde 1 von ihnen kaufen und bei einer Stichprobe von 100 Besuchern würden 10 kaufen.

Diese perfekte Welt gibt es leider so in der Praxis nicht. In der Praxis ist jede Stichprobe und deren Ergebnis mit einer gewissen Schwankung verbunden. Es kann sein, dass in einem Fall nur 2 der 100 Besucher kaufen, oder 20 oder sogar gar keiner.

Diese **Variabilität** der Stichprobe wird in der Statistik auch als **Varianz** bezeichnet. Diese misst, inwiefern eine zufällig ausgewählte Stichprobe von dem wahren **Mittelwert** abweicht. Je höher die Varianz ist, desto stärker schwankt der Mittelwert über die einzelnen Stichproben hinweg. Desto höher ist also die Unsicherheit, die mit einem Testergebnis verbunden ist.

Regression zur Mitte

Glücklicherweise gibt es ein statistisches Verfahren, mit dem man diese Varianz kontrollieren kann. Dieses bezeichnet man als **„regression to the mean“ (Regression zur Mitte)**. Dieses Phänomenen beschreibt, dass eine Variable mit geringer Anzahl von Beobachtungen extreme Werte annehmen kann, sie sich jedoch langfristig mit einer großen Fallzahl immer mehr an den wahren Mittelwert annähert.

Ganz konkret bedeutet das für den Fall des A/B-Testings, dass die Variabilität mit steigender **Testgröße (sample size)** abnimmt. Das Ergebnis wird immer stabiler und nähert sich dem wahren, aber unbekannten Wert der Grundgesamtheit immer mehr an.

Aber was genau ist in diesem Zusammenhang mit **stabil** gemeint und wie misst man, ob ein Testergebnis stabil ist oder nicht?

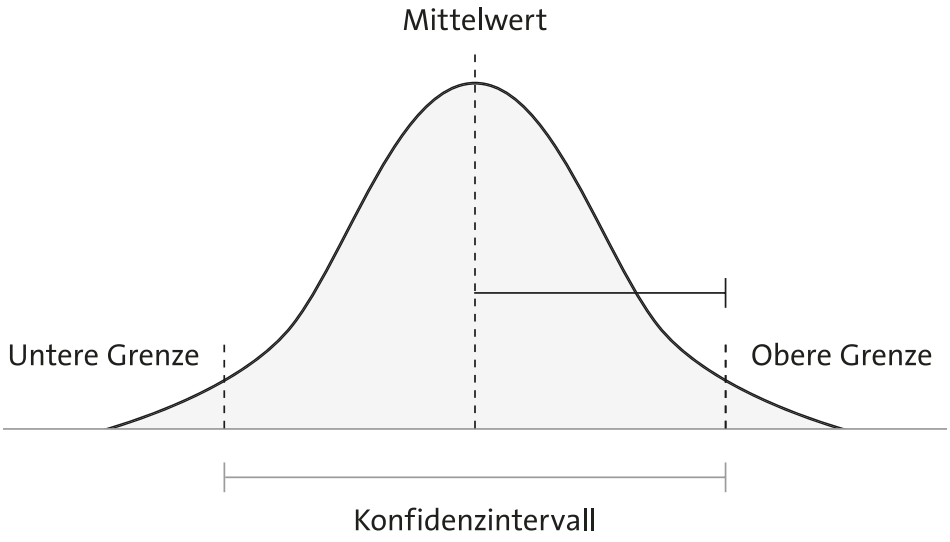
Konfidenz und statistischer Fehler

Den Begriff Konfidenz sollte man kennen, damit man die Ergebnisse eines A/B-Tests richtig interpretieren kann.

	Besucher	Conversions	Conversion Rate in %	Uplift	Untere Grenze 95% Konfidenzintervall des Uplifts	Obere Grenze 95% Konfidenzintervall des Uplifts
Kontrollvariante	80.000	3.400	4,25%			
Testvariante	84.000	3.900	4,64%	9,24%	2,52%	15,96%

Angenommen, in der Testvariante waren 84.000 Besucher, von denen 3.900 gekauft haben. Daraus ergibt sich eine Conversion Rate von 4,64 %. Die Kontrollvariante erzielt eine Conversion Rate von 4,25 %. Daraus ergibt sich dann ein Uplift von 9,24 %. Dieser Uplift ist allerdings nicht der wahre, tatsächliche Uplift, sondern nur eine Schätzung auf Basis einer Stichprobe. Wie schon beschrieben, ist die Schätzung mit einer gewissen Unsicherheit verbunden und unterliegt Schwankungen. Diese Schwankungen werden in Form eines Intervalls abgebildet, welches man auch als **Konfidenzintervall** bezeichnet. Innerhalb dieses Intervalls wird sich der wahre Uplift mit einer gewissen Wahrscheinlichkeit bewegen.

Im obigen Beispiel bedeutet dies für das konkrete Testergebnis, dass der wahre Uplift mit einer Wahrscheinlichkeit von 95 % in einem Intervall von 2,52 % bis 15,96 % liegt. Die Abweichungen von der Conversion Rate 9,24 % betragen nach oben und nach unten jeweils 6,72 %. Diese Abweichung wird als **statistischer Fehler (margin of error)** bezeichnet. Dieser gibt an, wie stark der tatsächliche Uplift vom beobachteten Wert der Stichprobe nach oben bzw. unten abweichen kann. Der statistische Fehler spiegelt also im Prinzip die Unsicherheit wider, mit dem ein Testergebnis verbunden ist und ergibt sich in Abhängigkeit von der beobachteten Varianz.



Folgende Aussagen kann man also zum Konfidenzintervall treffen:

„Wir sind zu 95 % sicher, dass der wahre Effekt auf die Conversion Rate zwischen 2,52 % und 15,96 % liegt.“

oder

„Wenn man den Test 20 Mal wiederholen würde, würde der beobachtete Effekt in 19 Fällen zwischen 2,52 % und 15,96 % liegen.“

Das Konfidenzintervall ist also ein beobachtetes Intervall, innerhalb dessen ein bestimmter Prozentsatz bei ständiger Wiederholung des Tests liegen wird. Wie hoch dieser Prozentsatz sein soll, wird vorab anhand des **Konfidenzlevels** bestimmt.

Das Konfidenzlevel gibt an, wie sicher man über das Testergebnis sein möchte. In der Regel wird hier ein Level von 95 % gewählt. Das bedeutet konkret:

„Die Wahrscheinlichkeit, dass sich der wahre Effekt zwischen 2,52 % und 15,96 % bewegt, liegt bei 95 %.“

oder auch

„Wir sind zu 95 % sicher, dass die Testvariante zu einem Uplift führt. Dieser Uplift liegt dann höchstwahrscheinlich zwischen 2,52 % und 15,96 %.“

Auf Basis des Konfidenzlevels, dem Mittelwert und dem statistischen Fehler (margin of error) kann dann das Konfidenzintervall berechnet werden. Die obere Intervallgrenze ergibt sich, indem der statistische Fehler zum beobachteten Uplift addiert wird. Die untere Grenze, indem er vom Mittelwert (der Conversion Rate) abgezogen wird.

Das Konfidenzlevel kann beliebig gewählt werden. Man kann auch festlegen, dass man zu 99 % sicher über das Testergebnis sein möchte. In diesem Fall würde man einen laufenden Test erst dann abschalten, wenn er dauerhaft mindestens eine Konfidenz von 99% im Testing Tool aufweisen würde.

In den meisten Fällen ist eine Konfidenz von 95 % ausreichend. Dieser Wert hat allerdings keine statistisch-methodisch begründete Berechtigung. Man hat allerdings festgestellt, dass ein Konfidenzlevel von 95 % eine hinreichende Sicherheit für CRO Zwecke bietet. Bei medizinischen Studien, wie z.B. der Wirksamkeit von lebensrettenden Medikamenten, wird hingegen oft ein höheres Niveau von 99 % gewählt, da man sich hier noch weniger Fehler erlauben darf.

Jedoch hat das gewählte Konfidenzlevel auch einen Einfluss auf die benötigte Testlaufzeit: Je höher man das Level wählt, desto länger muss der Test laufen (Mehr dazu unter dem Punkt: Testlaufzeit bestimmen).

Signifikanz und statistische Fehlertypen

Ein A/B-Test ist statistisch gesehen nichts anderes als ein Experiment, um eine vorab definierte Hypothese zu be- oder widerlegen. Praktisch gesehen bedeutet das, dass man beim Testen in den allermeisten Fällen die Hypothese hat, dass die optimierte Testvariante zu einer höheren Conversion Rate führt. Diese möchte man durch das Sammeln und Beobachten einer ausreichend großen Stichprobe belegen. Aufgrund der Unsicherheit mit der ein Testergebnis in einem solchen Fall verbunden ist, können beim Testen unterschiedliche Szenarien auftreten, die man folgendermaßen klassifizieren kann:

1. Laut Testergebnis ist die Testvariante besser als die Kontrollvariante, was in Wirklichkeit auch stimmt.

2. Laut Testergebnis ist die Testvariante besser als die Kontrollvariante, was in Wirklichkeit aber nicht stimmt. Diesen Fehler bezeichnet man als Fehler 1. Art, Irrtumswahrscheinlichkeit oder Alpha-Fehler.

3. Laut Test ist die Testvariante nicht besser als die Kontrollvariante, was in Wirklichkeit aber nicht stimmt. Das ist der Fehler 2. Art, auch Beta-Fehler genannt.

4. Laut Testergebnis ist die Testvariante nicht besser als die Kontrollvariante, was in Wirklichkeit auch stimmt.

Beim A/B-Testing können also zwei verschiedene Arten von Fehlern auftreten, welche dazu führen können, dass man auf Basis eines Testergebnisses eine falsche Schlussfolgerung zieht. Um zu verstehen, was genau die einzelnen Fehlertypen bedeuten, ist es wichtig die Begriffe der **Nullhypothese und Alternativhypothese** zu kennen:

Die Nullhypothese besagt, dass es keinen Zusammenhang zwischen zwei Datensätzen gibt. Auf Basis eines statistischen Tests kann diese Hypothese entweder abgelehnt oder nicht abgelehnt, also angenommen werden.

Das Konzept ist ähnlich, wie das eines Angeklagten vor Gericht: Dieser gilt als unschuldig, solange nichts anderes bewiesen werden kann.

Die Alternativhypothese besagt genau das Gegenteil, nämlich dass es sehr wohl einen Zusammenhang zwischen zwei Datensätzen gibt.

	H0 ist falsch	**H0 ist richtig**
Test lehnt H0 ab	richtige Entscheidung	Fehler 1. Art (Alpha-Fehler)
Test lehnt H0 nicht ab	Fehler 2. Art (Beta-Fehler)	richtige Entscheidung

Diese beiden Hypothesen werden bei einem statistischen Hypothesentest gegeneinander getestet. Bezogen auf einen A/B-Test bedeutet dies, dass die Nullhypothese besagt, dass die Testvariante zu keiner signifikanten Erhöhung der Conversion Rate führt. Die Alternativhypothese hingegen besagt, dass die Testvariante sehr wohl zu einer Erhöhung führt (bei einer einseitigen Testbetrachtung).

Ziel des A/B-Tests ist es in der Regel, die Nullhypothese abzulehnen und die Alternativhypothese anzunehmen. Entweder, wir können auf Basis der Testergebnisse die Nullhypothese ablehnen und so zeigen, dass die Testvariante einen signifikanten Uplift erzeugt. Oder wir können sie nicht ablehnen und akzeptieren, dass es sehr wahrscheinlich keinen signifikanten Uplift gibt.

Apropos Signifikanz ...

Der Fehler 1. Art und statistische Signifikanz

Der Fehler 1. Art tritt auf, wenn die Nullhypothese fälschlicherweise abgelehnt wird. Bei einem A/B-Test bedeutet das, dass man auf Basis der Ergebnisse entscheiden würde, dass die Testvariante besser ist als die Kontrollvariante, obwohl das in Wirklichkeit nicht stimmt. Man erklärt also eine Variante für einen Testsieger, obwohl diese in Wahrheit keiner ist.

Dieser Fehler 1. Art kann schwerwiegende Folgen haben, denn man trifft hier womöglich eine falsche Business-Entscheidung, indem man die angebliche Gewinnervariante live stellt, obwohl sie gar keine Verbesserung bringt. Das Ziel eines Tests sollte es also auch sein, diesen Fehler 1. Art möglichst gering zu halten. Aber wie geht das?

Die Höhe des Fehlers 1. Art wird beim A/B-Test mit Hilfe des Signifikanzniveaus kontrolliert: Ein Testergebnis gilt als statistisch signifikant, wenn der **p-Wert** kleiner ist als das **Signifikanzlevel**. Gleich zwei neue Begriffe in einem Satz, die einer Erklärung bedürfen. man folgendermaßen klassifizieren kann:

Signifikanzlevel

Das Signifikanzlevel wird vor dem Test vom Nutzer festgelegt und ergibt sich direkt aus dem oben beschriebenen Konfidenzlevel durch folgende Formel:

Signifikanzlevel in % = 100 % - Konfidenzlevel in %

Bei den meisten Tests wird mit einem Konfidenzlevel von 95 % gearbeitet. Das bedeutet, dass der Nutzer vor dem Test festgelegt hat, dass er zu 95 % sicher über das Testergebnis sein möchte. Andersherum heißt das, dass die Irrtumswahrscheinlichkeit 5 % beträgt.

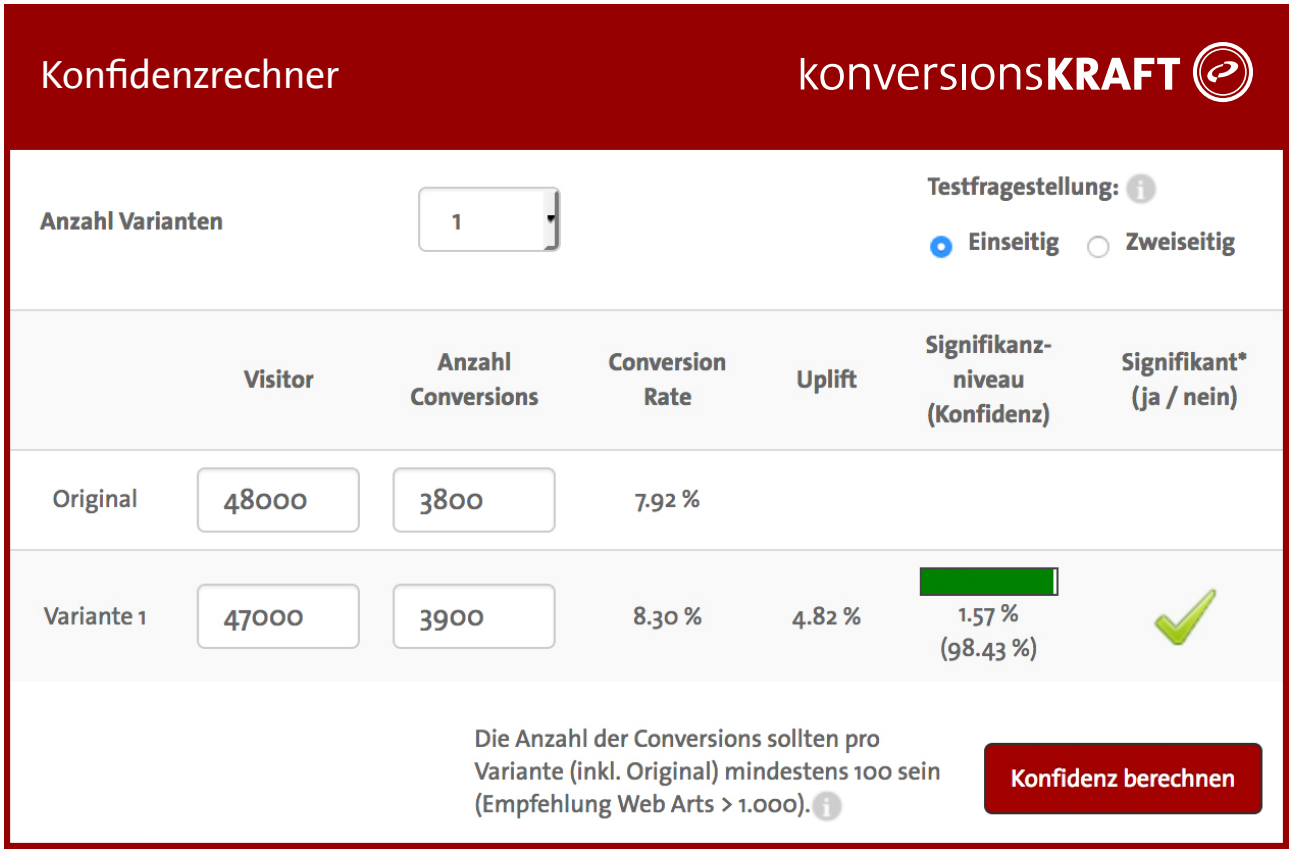
Das Signifikanzlevel liegt in diesem Fall also bei 5 %. Man ist bereit, maximal eine Wahrscheinlichkeit von 5 % zu akzeptieren, eine Testvariante fälschlicherweise als Sieger zu identifizieren, ob wohl sie gar keiner ist.

Der p-Wert

Der p-Wert ist letztendlich nichts anderes, als das konkrete Ergebnis für das Signifikanzlevel in einem durchgeführten A/B-Test. Er gibt an, wie hoch die Wahrscheinlichkeit in dem konkreten Testszenario ist, dass man eine Testvariante fälschlicherweise als Sieger identifiziert. Der Wert ergibt sich, indem man die Konfidenz, die im konkreten Testreport angezeigt wird zugrunde legt und sie von 100% abzieht.

Ein Beispiel:

In dem konkreten Test liegt die Konfidenz bei 98,43 %. Der p-Wert beträgt in diesem Fall also 1,57 %. Das heißt, hier beträgt die Wahrscheinlichkeit, dass die Testvariante, die hier als Gewinner angezeigt wird, in Wirklichkeit aber kein Gewinner ist, 1,57 % beträgt.
Anders gesagt: Die Wahrscheinlichkeit, dass das Testergebnis nur rein zufällig zustande gekommen ist, liegt bei 1,57 %.



Und wann bezeichnet man ein Testergebnis als signifikant?

Auf Basis des p-Wertes eines konkreten Tests wird entschieden, ob eine Testvariante zu einer signifikant höheren Conversion Rate führt oder nicht.

Legt man also vor einem Test fest, dass man maximal bereit ist, eine Irrtumswahrscheinlichkeit von 5 % zu akzeptieren, ist ein Testergebnis also signifikant, wenn der beobachtete p-Wert kleiner oder maximal gleich dem Signifikanzlevel, also den 5 %, ist.

Die meisten Testing Reports zeigen aber nicht direkt den konkreten p-Wert, sondern das Konfidenzlevel an. Da aber wie besprochen beide Werte durch folgende Gleichung zusammenhängen (p-Wert = 100 % - Beobachtete Konfidenz eines Tests), bedeutet das für die Interpretation der Ergebnisse eines konkreten Tests:

Ein Ergebnis gilt als signifikant, wenn das beobachtete Konfidenzlevel bei mindestens 95% liegt.

Fehler 2. Art und statistische Power

Ein weiterer Fehler, der beim A/B-Testing auftreten kann, ist der sogenannte Fehler 2. Art oder auch Beta-Fehler genannt.

Der Fehler 2. Art tritt ein, wenn die Nullhypothese in Wahrheit falsch ist, der Test sie aber nicht ablehnt. Der Test sagt also, es gibt keinen signifikanten Unterschied oder Zusammenhang, obwohl es in Wahrheit einen gibt.

In der A/B-Testing-Welt heißt das: Auf Basis der Testergebnisse entscheiden wir, dass die Testvariante nicht besser ist als die Kontrollvariante, obwohl sie in Wahrheit doch besser ist. Der Test erkennt in diesem Fall also nicht, dass tatsächlich ein Gewinner vorliegt und zeigt als Ergebnis, dass es keinen signifikanten Unterschied zwischen den Varianten gibt.

So wie der Fehler 1. Art eng mit dem Begriff der statistischen Signifikanz verbunden ist, ist der Fehler 2. Art eng mit dem Begriff der **statistischen Power** verknüpft.

Die statistische Power eines Tests ergibt sich durch:

Power = 100 % - Wahrscheinlichkeit für den Fehler 2. Art

Die Power gibt an, wie hoch die Wahrscheinlichkeit ist, dass man mit dem Test einen Effekt nachweisen kann, wenn es denn tatsächlich auch einen Effekt gibt. Die Power wird daher auch als Teststärke bezeichnet. Diese gibt an, wie gut ein Test in der Lage ist, einen Unterschied signifikant nachzuweisen.

Ab einer Power von 80 % (entspricht einer Wahrscheinlichkeit für den Fehler 2. Art von <= 20%) spricht man von einer guten Teststärke.

Wie hoch die Power eines konkreten Tests ist, wird nicht direkt im Testing Report angezeigt. Die Power wird aber beim A/B-Test direkt über die Dauer der Testlaufzeit und somit der Stichprobengröße bestimmt.

Damit sind wir schon beim nächsten wichtigen Thema. Der Frage wie lange ein Test eigentlich laufen muss.

Einfluss der Signifikanz, Konfidenz und Power auf die Testlaufzeit

Jeder, der schonmal einen A/B-Test gemacht hat, kennt das: Man startet den Test, wartet ein paar Tage ab und meint, eine erste Tendenz in den Ergebnissen erkennen zu können. Nach einer weiteren Woche sehen die Ergebnisse aber ganz anders aus. Wann sollte man den Test denn jetzt abschalten?

Die Begriffe Signifikanz, Konfidenz, Power, Fehler 1. und 2. Art sowie das Thema Testlaufzeit hängen sehr eng zusammen. Jeder Test benötigt eine Mindestgröße an Testteilnehmern, um herauszufinden, ob es einen systematischen Unterschied im Verhalten gibt, wenn man den Testteilnehmern verschiedene Varianten einer Website zeigt. In der Statistik bezeichnet man diese Mindestgröße als **notwendige Stichprobe (minimum sample size).**

A/B Test Mindestlaufzeit / Mindestconversion-Anzahl

Kundenname

Im Sheet "Guideline" findest du mehr Informationen dazu, welche Informationen du hier eingeben musst und woher du diese bekommst.

Dein Input		Dein Output	
Aktuelle Conversion Rate der Testseite:	3,00%	Benötigte Anzahl Visitors pro Variante	161.226
Erwarteter Uplift in %:	5%	Benötigte Anzahl Visitors gesamt	322.453
Anzahl aller Varianten (inkl. Control):	2	Benötigte Anzahl Conversions pro Variante	4.837
Conversions/Monat:	8.000	Benötigte Anzahl Conversions gesamt	9.674

Test-Settings:

Power	80%
Konfidenz	95%

Mindestlaufzeit deines Tests	Max. Anzahl Tests p.a.
36 Tage	10

Wie hoch die Power eines konkreten Tests ist, wird nicht direkt im Testing Report angezeigt. Die Power wird aber beim A/B-Test direkt über die Dauer der Testlaufzeit und somit der Stichprobengröße bestimmt.

Wie oben schon beschrieben, sollte die gezogene Stichprobe ein repräsentatives Abbild der Grundgesamtheit darstellen. Aufgrund der Schwankungen (Varianz) wird es jedoch immer eine gewisse Fehlerquote geben. Eine weitere Bedingung für Repräsentativität ist neben der Bedingung der Zufälligkeit, mit der die Stichprobe gezogen werden muss, eine ausreichende Größe der Stichprobe.

Beim A/B-Testing sind es die folgenden Faktoren, die einen Einfluss darauf haben wie groß die Stichprobe sein muss, also wie viele Besucher man im Test benötigt:

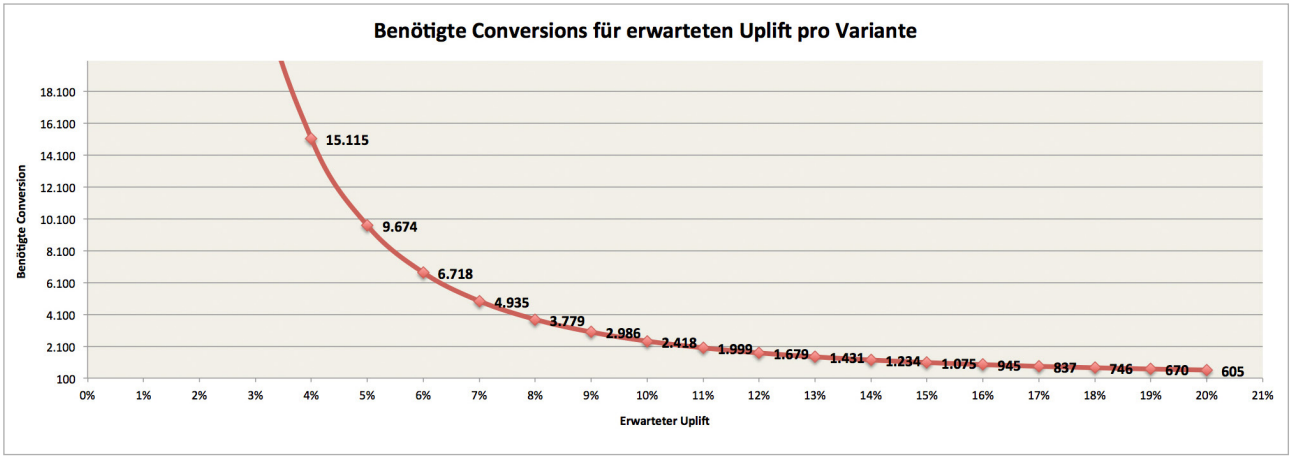
1. Der zu erwartende Uplift in der Conversion Rate,
2. die Konfidenz, die man erreichen möchte und
3. die Power, die man erreichen möchte.

Tipp

In diesem Artikel befindet sich am Ende ein Excel Sheet zum kostenlosen Download. Dort kannst Du unter Angabe der Conversions und Besucherzahlen berechnen, wie lange der Test laufen muss, um signifikante Ergebnisse zu erzielen.

1. Zu erwartender Uplift

Der Zusammenhang zwischen dem zu erwartenden Uplift und der Testdauer ist in der unteren Grafik dargestellt. Die Grafik zeigt, dass es einen negativen Zusammenhang zwischen dem erwarteten Uplift und der Testdauer (angegeben in benötigten Conversions) gibt.



Das heißt: Je niedriger der zu erwartende Uplift, desto länger die berechnete Testlaufzeit. Ist der zu erwartende Effekt hingegen sehr groß, ist die Testlaufzeit kurz. Dieser Zusammenhang lässt sich folgendermaßen übersetzen: Ist der Uplift sehr groß, ist er auch einfacher und schneller zu finden als ein kleiner Uplift nach dem ich lange suchen muss.

Wenn man einen A/B-Test macht, der einen relativ geringen Effekt auf die Conversion Rate hat, muss man also von vorne herein eine längere Testlaufzeit einplanen, da man eine größere Stichprobe benötigt. Aber woher soll man denn vorher wissen wie hoch der Uplift ist, genau das möchte man ja mit dem Test herausfinden?

Tatsächlich ist die Frage nach der Höhe des zu erwartenden Uplifts kaum bis gar nicht zu beantworten, bevor man es nicht getestet hat. Dennoch ist es extrem wichtig, zu versuchen, eine realistische Schätzung vorzunehmen.

Tipps zur Einschätzung des zu erwartenden Uplifts:

- Testing-Historie: Ergebnisse aus vergleichbaren Tests.

- Business Relevanz: Ab wann ist ein Uplift überhaupt relevant, dass man auf Grundlage davon eine wichtige Business-Entscheidung trifft und zum Beispiel ein neues Feature implementieren würde? Wie hoch müsste der Uplift in der Conversion Rate mindestens sein, damit die Kosten der Implementierung mindestens gedeckt werden?
  Aus diesen Kosten-Nutzen-Überlegungen lässt sich vorab prüfen ob der benötigte Uplift überhaupt realistisch ist und wie lange der Test mindestens laufen muss, um diesen Uplift zu erreichen.

- Kontrast: Handelt es sich eher um eine kleinere Änderung auf der Seite oder ist sie dem Nutzer direkt sichtbar? Tendenziell führen kontrastreichere Konzepte auch zu größeren Effekten.

- Conversion Pfad: Wie nahe ist die Änderung an der finalen Conversion? Testet man nur auf der Startseite, so ist der Effekt in der Regel kleiner, da sich die Änderung ganz am Anfang der Customer Journey befindet. Auf der Produktseite hingegen fällt der Kunde die Kaufentscheidung, so dass die Effekte hier in der Regel größer sind.

- Erfahrung: Es lässt sich nicht vermeiden, dass man bei der Einschätzungen des Uplifts auch mal falsch liegt. Je mehr Tests man aber durchführt, desto mehr Erkenntnisse gewinnt man auch. Diese helfen dabei, den Uplift treffgenauer einzuschätzen.

Generell empfiehlt es sich, mit einer konservativen Schätzung des Uplifts zu arbeiten. Denn wenn man den Uplift unterschätzt, dann überschätzt man die benötigte Testlaufzeit, aber stellt so zumindest sicher, dass die Stichprobe auf jeden Fall ausreichend groß ist. Überschätzt man den Uplift, so führt das zu einer Testlaufzeit, die nicht ausreicht ist, um die minimal notwendige Stichprobe zu erreichen. Folglich stellt man den Test zu früh ab. Das kann zu Fehlentscheidungen führen.

In dem Excel Sheet zur Testlaufzeitberechnung kann man mit unterschiedlichen Uplift-Werten „spielen“ um einen Eindruck zu gewinnen, welchen Einfluss dies auf deine benötigte Stichprobengröße hat. Die Schätzung des zu erwartenden Uplifts vor dem Test ist extrem schwierig und man wird vermutlich nicht den richtigen Wert treffen. Dieses Vorgehen hilft jedoch dabei herauszufinden, ob bei den gegebenen Traffic- und Conversion-Zahlen ein Test überhaupt sinnvoll wäre (im Sinne von benötigter Testlaufzeit), wenn man einen bestimmten Uplift-Wert zugrunde legt.

2. Konfidenzlevel

Die Höhe des gewählten Konfidenzlevels hat einen direkten Einfluss darauf, wie lange ein Test laufen muss.

Zur Erinnerung: Die Konfidenz gibt an, wie sicher man über das Testergebnis sein möchte. Dabei gilt folgender Zusammenhang: Je sicherer man sein möchte, dass der gefundene Effekt kein Zufall ist, desto höher musst man das Konfidenzlevel wählen und desto länger muss der Test laufen.

Tipp

In unserem Excel Sheet zur Berechnung der Testdauer ist ein Konfidenzlevel von 95 % vorausgewählt. Da dieser Wert Standard für die meisten A/B-Tests ist, kann man damit für fast alle Tests auch arbeiten und muss hier nichts verändern.

3. Power

Ganz ähnlich verhält es sich mit der Power eines Tests. Je sicherer man vorab sein möchte, dass ein Uplift auch wirklich gefunden wird, desto höher muss man das Level der Power im Test wählen. Je höher man also das Power Level setzt, desto wahrscheinlicher ist es, auch wirklich einen Uplift zu finden. Ist dieser sehr groß, reicht bereits eine geringe Testlaufzeit für eine gute Testpower. Ist der Effekt jedoch gering, so benötigt es entsprechend (viel) mehr Zeit.

Tipp

Ähnlich wie beim Konfidenzlevel arbeiten wir mit einem Standard-Wert für unsere A/B-Tests. Dieser beträgt 80 % und ist ausreichend für die meisten Fälle. In dem Excel Sheet wird dieser Wert deshalb auch standardmäßig für die Berechnungen verwendet.

Zum Abschluss der theoretischen Grundlagen zur Testlaufzeit hier noch ein paar **praktische Tipps**, wenn es um die Frage geht, ob man einen laufenden Test anhalten sollte oder nicht:

- Ein Test sollte mindestens 2-4 Wochen durchgehend laufen.
- Pro Variante sollten mindestens 1.500 - 2.000 Conversions erreicht werden.
- Beachten der regulären Zeit zur Kaufentscheidung (Business Cycle Betrachtung).
- Regelmäßiges Monitoring ist wichtig, um zu verhindern, dass Business-Ziele bedroht werden. Aber Testergebnisse nach zwei Tagen Laufzeit haben keine Aussage!
- Nicht den Test stoppen, sobald die gemessene Konfidenz das festgelegte Konfidenzlevel erreicht hat, sondern ein paar Tage abwarten, um sicher zu stellen, dass die Konfidenz stabil ist.
- Entwicklung der Konfidenz während der Testlaufzeit analysieren und sicherstellen, dass sich diese auf einem stabil hohen Niveau einpendelt.

Statistik Cheat Sheet

Wir hoffen, dass wir mit diesem Guide dabei helfen können, das Thema Statistik beim A/B-Testing transparenter und verständlicher zu machen. Mit diesem Guide hast Du ein solides Framework an der Hand, welches Dir dabei hilft, Deine Tests gut zu planen, zu analysieren und zu interpretieren.

Im folgenden Cheat Sheet werden alle Begriffe noch einmal kurz zusammen gefasst, wenn du etwas schnell nachschauen möchtest.

Stichprobe
Ein Datensatz, der aus einer Grundgesamtheit ausgewählt wird, z.B. bestimmte Anzahl von Besuchern einer Website. Die Stichprobe muss zufällig ausgewählt und ausreichend groß sein, um ein repräsentatives Abbild der Grundgesamtheit zu sein.

Grundgesamtheit
Die Menge aller vorhandenen Elemente, über die man eine Aussage treffen möchte, z.B. alle Besucher, die jemals eine bestimmte Website besucht haben.

Hypothesentest/Inferenzstatistik
Hier wird auf Basis einer Stichprobe aus einer Grundgesamtheit eine bestimmte Hypothese überprüft. Beim A/B-Testing wird auf Basis einer Test- und Kontrollgruppe überprüft, ob es signifikante Unterschiede im Verhalten der beiden Gruppen gibt.

p-Wert
Gibt an, wie hoch die Wahrscheinlichkeit in dem konkreten Testszenario dafür ist, dass man eine Testvariante fälschlicherweise als Sieger identifiziert, obwohl dies in Wirklichkeit gar nicht der Fall ist.

Signifikanzlevel
Wird vom Nutzer vorab festgelegt und gibt die Wahrscheinlichkeit an, die man bereit ist zu akzeptieren, dass eine Testvariante fälschlicherweise als Gewinner identifiziert wird. Das Level wird beim A/B-Testing in der Regel auf 5 % festgelegt.

Konfidenzlevel
Gibt an, wie sicher man sein möchte, dass der beobachtete Uplift in einem Test nicht rein zufällig entstanden ist sondern dass es einen systematischen Unterschied zwischen der Test- und Kontrollvariante gibt. Die Konfidenz wird in der Regel auf 95 % festgelegt. % bewegt.

Mittelwert
Der Mittelwert einer Stichprobe ist die zentrale statistische Kennzahl. Beim A/B-Testing ist das in der Regel die Conversion Rate in einem bestimmten Datensatz von Website-Besuchern.

Varianz
Statistische Maßzahl für die Variabilität oder Schwankung einer Metrik innerhalb einer Stichprobe. Statistisch gesehen ist die Varianz die quadrierte Summe der Abweichungen vom Mittelwert.
Beim A/B Testing ist jede Stichprobe diesen Schwankungen unterworfen, deswegen braucht man das Konzept der statistischen Signifikanzprüfung.

Statistischer Fehler
Statistische Maßzahl, die angibt, wie stark die Unsicherheit aufgrund zufälliger Abweichungen innerhalb einer Stichprobe ist. Der Fehler ist abhängig von der Standardabweichung, die aus der Varianz resultiert.
Der statistische Fehler wird vom Mittelwert abgezogen bzw. addiert um die Grenzen des Konfidenzintervalls zu berechnen.

Nullhypothese
Annahme, dass es keinen Zusammenhang zwischen zwei Datensätzen gibt. Bei einem statistischen Hypothesentest wird diese Hypothese auf Basis der Stichproben-Ergebnisse entweder angenommen oder abgelehnt.
Beim A/B-Test besagt die Nullhypothese, dass die Testvariante keinen signifikanten Effekt auf die Conversion Rate hat. Man hat das Ziel, diese Hypothese mit Hilfe eines Tests zu widerlegen.

Fehler 1.Art
Wahrscheinlichkeit, dass man die Nullhypothese ablehnt, obwohl sie richtig ist.
Beim A/B Testing bedeutet dies, dass man auf Basis der Testergebnisse darauf schließt, dass die Testvariante zu einem signifikanten Uplift in der Conversion Rate führt, obwohl dies gar nicht der Fall ist.

Fehler 2.Art
Wahrscheinlichkeit, dass man die Nullhypothese annimmt, obwohl sie in Wahrheit falsch ist.
Beim A/B-Testing bedeutet dies, dass man auf Basis der Testergebnisse darauf schließt, dass die Testvariante zu keinem signifikanten Uplift in der Conversion Rate führt, obwohl sie das in Wahrheit doch tut.

Statistische Power
Die Power gibt an, wie hoch die Wahrscheinlichkeit ist, dass man die Nullhypothese ablehnt, wenn sie auch tatsächlich falsch ist. Das bedeutet eigentlich nichts anderes, als dass man mit dem Test einen Effekt nachweisen kann, wenn es denn tatsächlich auch einen Effekt gibt.

konversionsKRAFT