Tom Alby

WEBANALYSE: Wie aus Daten Taten folgen, Teil III

Von den fünf Schritten der Datenanalyse (Problem definieren, Daten akquirieren, Daten analysieren, testen und präsentieren) wurden in Teil I und Teil II dieser Serie bereits die ersten drei behandelt. Bevor es zu nun zu dem nächsten Schritt geht, sollen noch weitere Aspekte der Datenakquise und -analyse beleuchtet werden. Abgerundet wird dieser Teil III mit Fallstricken in A/B- und multivariaten Tests, die es zu vermeiden gilt.

Nicht jeder Analyst hat die Zeit oder die Möglichkeiten für eine explorative Datenanalyse, wie sie im letzten Beitrag in Ausgabe 59 beschrieben wurde. Die durchgeführte Analyse war in dieser Form außerdem nur mit Rohdaten möglich, was in der Regel Programmierkenntnisse erfordert. Was für Anfänger zunächst abschreckend wirken mag, könnte in Zukunft der Standard für Webanalysten werden, da Analytics-Installationen zunehmend komplexer und Daten aus Webanalyse-Systemen mit anderen Datenquellen verknüpft werden. So steht eine Website selten isoliert da, sondern meistens in Verbindung mit anderen Systemen, sei es eine Marketingplattform, ein Customer-Relationship-Management-System oder ein Warenbestandssystem.

Nicht immer aber müssen Daten erst im Nachhinein verknüpft werden. Google Analytics & Co bieten viele Möglichkeiten, Daten aus anderen Systemen zu integrieren. Wird vorab hinreichend Gehirnschmalz investiert, können Daten so gesammelt werden, dass sie später ohne viel Programmieraufwand abgerufen werden können. Wichtig ist hierbei, dass keine personenbezogenen Daten in den Systemen der Anbieter gespeichert werden. Namen, E-Mail-Adressen etc. sind tabu.

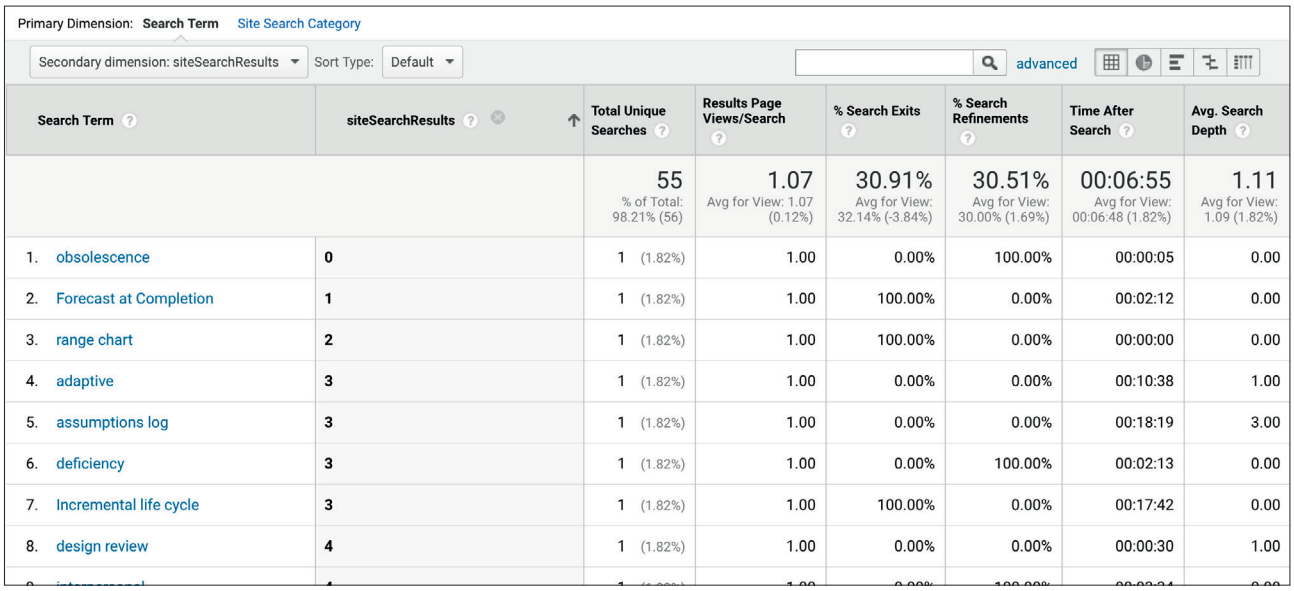
Benutzerdefinierte Dimensionen und Metriken: Die Freiheit nehm' ich mir!

Während Adobe Analytics mit seinen zu konfigurierenden eVars und Props schon vorab viel gedankliche Anstrengung erfordert, kommt Google Analytics mit einem Serviervorschlag als Standardinstallation daher, der dann allerdings nur von wenigen Anwendern angepasst wird. Das betrifft nicht nur die Art, wie die Daten gesammelt werden (siehe Teil I in Ausgabe 58), sondern auch, was zusätzlich zu der Standardinstallation erfasst wird. Dabei bietet Google Analytics in der kostenlosen Variante zurzeit je 20 benutzerdefinierte Dimensionen und Metriken, die spannende Anpassungen ermöglichen.

Ein Schritt zurück: Was sind überhaupt Dimensionen und Metriken? Je nach Webanalyse-System existieren unterschiedliche Definitionen, aber eine kleine Eselsbrücke in Form einer Alliteration hilft unabhängig vom System: Dimensionen sind deskriptiv, Metriken messen. Ein Beispiel aus einem anderen Kontext: Für die Sonntagsfrage („Welche Partei würden Sie wählen, wenn man nächsten Sonntag Bundestagswahlen wären?") wird nach der Partei gefragt (deskriptiv) und dann wird die Anzahl der Wähler für jede Partei gezählt (messen). Meistens

DER AUTOR

Tom Alby ist mehrfacher Buchautor und von der „brand eins" als „Datenfreak" bezeichneter Digitalexperte. Nach Stationen wie Google und Ask.com ist er seit 2018 CDTO bei Euler Hermes.



Primary Dimension: Search Term Site Search Category

Secondary dimension: siteSearchResults | Sort Type: Default | advanced

	Search Term	siteSearchResults	Total Unique Searches	Results Page Views/Search	% Search Exits	% Search Refinements	Time After Search	Avg. Search Depth
			55 % of Total: 98.21% (56)	1.07 Avg for View: 1.07 (0.12%)	30.91% Avg for View: 32.14% (-3.84%)	30.51% Avg for View: 30.00% (1.69%)	00:06:55 Avg for View: 00:06:48 (1.82%)	1.11 Avg for View: 1.09 (1.82%)
1.	obsolescence	0	1 (1.82%)	1.00	0.00%	100.00%	00:00:05	0.00
2.	Forecast at Completion	1	1 (1.82%)	1.00	100.00%	0.00%	00:02:12	0.00
3.	range chart	2	1 (1.82%)	1.00	100.00%	0.00%	00:00:00	0.00
4.	adaptive	3	1 (1.82%)	1.00	0.00%	0.00%	00:10:38	1.00
5.	assumptions log	3	1 (1.82%)	1.00	0.00%	0.00%	00:18:19	3.00
6.	deficiency	3	1 (1.82%)	1.00	0.00%	100.00%	00:02:13	0.00
7.	Incremental life cycle	3	1 (1.82%)	1.00	100.00%	0.00%	00:17:42	0.00
8.	design review	4	1 (1.82%)	1.00	0.00%	0.00%	00:00:30	1.00

Abb. 1: Der Site Search Report in Google Analytics mit einer benutzerdefinierten Dimension als sekundäre Dimension, die die Anzahl der Suchergebnisse anzeigt

werden noch weitere deskriptive Daten oder Merkmale des Wählers erhoben wie Alter, Einkommen, Geschlecht etc. Später wird dann ausgezählt, wie häufig welche Merkmalsausprägung in Kombination mit anderen vorgekommen ist. Ein Webanalyse-System ist nichts anderes: Es steht sozusagen vor dem Wahllokal beziehungsweise der Website und „fragt“, welchen Browser man nutzt, aus welchem Ort man kommt etc. Und hinterher wird gezählt, wie viele Nutzer aus Wanne-Eickel kommen und einen Internet Explorer nutzen, und es kann analysiert werden, ob sich diese Nutzer anders verhalten als Nutzer aus Gelsenkirchen, die mit einem Firefox auf der Website ankommen.

Und nun wieder ein Schritt tiefer hinein: Google Analytics & Co erlauben es dem Anwender, eigene Merkmale zu zählen. Und Dimensionen beschränken sich nicht nur auf Merkmale, die ein Nutzer mitbringt, sondern auch auf andere Objekte wie zum Beispiel Seiten, die ein Nutzer besucht. Zwar bietet Google Analytics zum Zeitpunkt dieser Ausgabe bereits 266 Dimensionen und 244 Metriken, aber diese sind nun mal generisch, also für jeden Anwender gleich.

Ein für fast alle Webseiten sinnvolles Beispiel einer benutzerdefinierten Dimension ist ein erweitertes Tracking der internen Suche. Jedes CMS bietet eine Suchfunktionalität und das Einrichten des Erfassens von Suchbegriffen ist mit jedem Webanalyse-Tool relativ einfach. Hieß es früher, dass Suchanfragen mit der Website-eigenen Suchmaschine auf Mängel in der Navigation schließen können, sind Nutzer heute eher so konditioniert, dass sie gleich die Suche nutzen, wenn sie einen Suchschlitz sehen. Was kann man also mit diesen Daten anfangen?

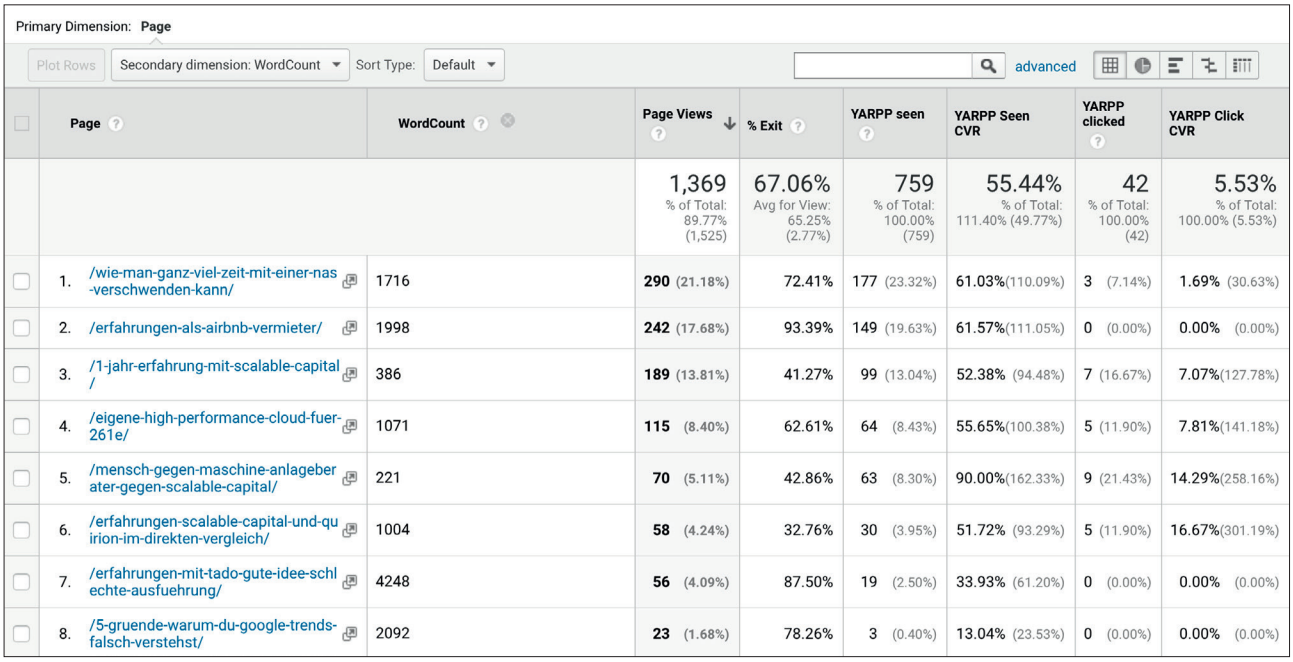
Sucht ein Nutzer mit einem Begriff und verlässt nach dem Durchführen der Suche die Webseite, ohne ein Ergebnis angeklickt zu haben, so stimmt etwas nicht mit der Suchergebnisseite. Allerdings ist noch nicht klar, was genau nicht stimmt. Es könnte sein, dass die Ergebnisse einfach schlecht sind. Oder aber, dass es gar keine Ergebnisse gibt. Je nach Variante führt dies zu unterschiedlichen Aktionen: Wurde kein Ergebnis angezeigt, so ist entweder die Suchanfrage irrelevant für die Seite oder die Suchanfrage deutet auf ein legitimes Interesse hin, das von der Website momentan noch nicht erfüllt wird. Wurden Ergebnisse angezeigt, aber kein Ergebnis angeklickt, so sollte die Relevanz der Suchergebnisse überprüft werden.

Damit nicht alle Suchanfragen durchprobiert werden müssen, um zu sehen, wo es Ergebnisse gab und wo nicht, hilft eine benutzerdefinierte Dimension, die mit der Anzahl der vom Nutzer gesehenen Ergebnisse gefüllt wird. Das Ergebnis eines solchen Standardberichts, dem eine sekundäre Dimension hinzugefügt wurde, ist in Abbildung 1 zu sehen. In dem Beispiel sollten die Suchergebnisse für die Suchanfragen „Forecast at Completion“, „range chart“ und „incremental life cycle“ überprüft werden, da die Nutzer nach der Sichtung der Fundstellen die Seite verlassen haben.

Ein weiteres Beispiel, das die komplexe Implementierung des Falls aus dem letzten Teil der Ausgabe 59 zumindest teilweise ersetzt, ist in Abbildung 2 zu sehen. Hier wurden bei zwei Events benutzerdefinierte Metriken mitgesendet. Jedes Mal, wenn ein Benutzer auf einer Webseite so weit runterscrollt, dass das Ende des Textes erreicht ist und weitere Leseempfehlungen angezeigt werden (realisiert durch das WordPress-Plug-in Yet Another Related Posts Plugin, kurz YARPP), wird der Wert 1 (für einmal angezeigt) gesen-

TIPP

Benutzerdefinierte Dimensionen und Metriken sind keine Profi-Funktionalität, sondern gehören zu jeder Analytics-Implementierung.



Primary Dimension: Page

Plot Rows | Secondary dimension: WordCount | Sort Type: Default | advanced

	Page	WordCount	Page Views	% Exit	YARPP seen	YARPP Seen CVR	YARPP clicked	YARPP Click CVR
			1,369 % of Total: 89.77% (1,525)	67.06% Avg for View: 65.25% (2.77%)	759 % of Total: 100.00% (759)	55.44% % of Total: 111.40% (49.77%)	42 % of Total: 100.00% (42)	5.53% % of Total: 100.00% (5.53%)
1.	/wie-man-ganz-viel-zeit-mit-einer-nas-verschwenden-kann/	1716	290 (21.18%)	72.41%	177 (23.32%)	61.03%(110.09%)	3 (7.14%)	1.69% (30.63%)
2.	/erfahrungen-als-airbnb-vermieter/	1998	242 (17.68%)	93.39%	149 (19.63%)	61.57%(111.05%)	0 (0.00%)	0.00% (0.00%)
3.	/1-jahr-erfahrung-mit-scalable-capital/	386	189 (13.81%)	41.27%	99 (13.04%)	52.38% (94.48%)	7 (16.67%)	7.07%(127.78%)
4.	/eigene-high-performance-cloud-fuer-261e/	1071	115 (8.40%)	62.61%	64 (8.43%)	55.65%(100.38%)	5 (11.90%)	7.81%(141.18%)
5.	/mensch-gegen-maschine-anlageberater-gegen-scalable-capital/	221	70 (5.11%)	42.86%	63 (8.30%)	90.00%(162.33%)	9 (21.43%)	14.29%(258.16%)
6.	/erfahrungen-scalable-capital-und-quirion-im-direkten-vergleich/	1004	58 (4.24%)	32.76%	30 (3.95%)	51.72% (93.29%)	5 (11.90%)	16.67%(301.19%)
7.	/erfahrungen-mit-tado-gute-idee-schlechte-ausfuehrung/	4248	56 (4.09%)	87.50%	19 (2.50%)	33.93% (61.20%)	0 (0.00%)	0.00% (0.00%)
8.	/5-gruende-warum-du-google-trends-falsch-verstehst/	2092	23 (1.68%)	78.26%	3 (0.40%)	13.04% (23.53%)	0 (0.00%)	0.00% (0.00%)

Abb. 2: Benutzerdefinierter Report mit benutzerdefinierten Dimensionen, Metriken und berechneten Messwerten

det und in dem benutzerdefinierten Messwert „YARPP seen“ gespeichert. Klickt der Nutzer einen YARPP-Link an, so wird ein Wert 1 in einem weiteren benutzerdefinierten Messwert „YARPP clicked“ gespeichert. Aus diesen Messwerten werden zwei berechnete Messwerte erstellt, einmal „YARPP Seen CVR“, der Anteil der Pageviews, bei denen YARPP zu sehen war, zum andern „YARPP Click CVR“, der Anteil der Klicks auf YARPP, wenn YARPP sichtbar war. Zusätzlich wird hier eine benutzerdefinierte Dimension angezeigt, die die Anzahl der Wörter eines Artikels auflistet. All dies kann mit Bordmitteln mit WordPress, Google Tag-Manager und Google Analytics erstellt werden.

Aus diesem Bericht lassen sich mehrere Handlungen ableiten (wenngleich dies noch nicht viele Datenpunkte sind):

» Es sieht so aus, als stiege, je kürzer ein Text ist, die Wahrscheinlichkeit, dass er bis zum Ende gelesen wird (negative moderate bis starke Korrelation mit einem Koeffizienten von -0.66, wenn man es nachrechnet). Eine Ausnahme bildet der Text an Platz 3. Dieser Text sollte also überarbeitet werden.
» Ebenso scheint der Text über die Missverständnisse von Google Trends nicht gerne zu Ende gelesen zu werden; hier sollte auch nachgearbeitet werden.
» Die YARPP-Klickraten variieren stark. Hier sollte die Relevanz der Empfehlungen überprüft und gegebenenfalls nachjustiert werden.

Als kleines Geschenk zur Jubiläumsausgabe der Website Boosting ist eine komplette und detaillierte Anleitung für einen solchen Report, inklusive Implementierung im Google Tag-Manager, unter *https://alby.link/websiteboosting60* zu finden.

Zusammengefasst: Die Möglichkeiten für benutzerdefinierte Dimensionen und Metriken sind je nach Anwendungszweck fast endlos, aber genau so individuell wie jede Website. Dennoch wird nur in sehr wenigen Analytics-Installationen davon Gebrauch gemacht, obwohl gerade durch diese Anpassungen handlungsrelevantere Berichte erstellt werden können.

Geht das noch besser? Wie eine gute Hypothese aufgebaut wird

In dem vorherigen Abschnitt wurden Optimierungsmöglichkeiten aufgezeigt. Aber woher wissen wir, was wirklich eine Verbesserung bringt? Nehmen wir das obige Beispiel der YARPP-Klicks. Die enttäuschende Klickrate von unter 10 % könnte daher kommen, dass die Empfehlungen zum Teil schlecht sind. Oder dass bei manchen Texten das Informationsbedürfnis bereits gestillt ist, bei anderen Texten aber weitere Infos hilfreich sind. Sie könnte aber auch daher kommen, dass sich die Box mit den Empfehlungen nicht besonders gut abhebt (unwahrscheinlich, denn bei manchen Texten wird ja trotzdem geklickt). Oder die Überschrift „Mehr lesen“ wirkt nicht besonders attraktiv und sollte in etwas wie „Was Du unbedingt noch wissen musst“ umformuliert werden. Hier sind schon alle wichtigen Komponenten einer Hypothese vorhanden:

» Es gibt eine mit Daten unterfütterte Beobachtung,
» eine erläuterte Vermutung, was eine positive Veränderung bewirken könnte,
» und einen KPI, auf den sich diese Veränderung auswirkt, mit einer Schätzung, wie hoch diese Veränderung sein wird (dazu unten gleich mehr).

Die Webanalyse bietet mit ihren Daten die Ausgangslage für das Testen

potenzieller Optimierungen und, sofern es sich um ein integriertes System handelt, gleichzeitig die Begleitung des Tests durch Messung seiner Performance. Das kann gut sein, aber auch schlecht. Wenn eine Seite zum Beispiel mehrere Besuche bis zur Conversion benötigt, das Testsystem aber auf den Sessions der Webanalyse-Plattform beruht, so ist etwas mehr Aufwand notwendig, um über die Session hinauszugehen.

Warum ein nicht signifikantes Testergebnis doch gut sein kann

Wie schon bei den Webanalyse-Systemen existieren auch bei den Test-Plattformen häufig grundlegende Missverständnisse, die zu falschen Schlussfolgerungen führen können. Das beste Beispiel ist das der statistischen Signifikanz. Ein statistisch signifikantes Ergebnis bedeutet nicht, dass eine Hypothese („die rote Überschrift klickt besser") wahr oder falsch wäre. Ein Konfidenzniveau sagt auch nichts darüber aus, wie stark ein Effekt ist. Natürlich wünschen sich Marketingverantwortliche Sicherheit in den Aussagen eines Testergebnisses, denn häufig geht es um Geld, wenn etwas verändert werden soll. Genau diese Sicherheit gibt es nicht (oder wenn, dann nur sehr selten). Aber dieser Umstand lässt sich auch zum Vorteil verwenden.

Was bedeutet eigentlich statistische Signifikanz? Wird zum Beispiel die oben formulierte Hypothese getestet, so wird in der Welt der Statistik eine Art Gegenhypothese aufgestellt, die sogenannte Nullhypothese. Sie besagt, dass die Optimierung keinen Einfluss auf die Klickrate hat. Ob das Ergebnis statistisch signifikant ist, wird an dem berühmt-berüchtigten p-Wert abgelesen, welcher die Wahrscheinlichkeit ist, dass man die gemessenen Werte erhält, wenn die Nullhypothese nicht abgelehnt wird.

An einem einfachen Beispiel erklärt: Beim Würfeln erhält einer der Spieler mit seinem Würfel dreimal eine 6. Ist das unmöglich? Nein. Ist es wahrscheinlich? Nein, aber es kann Zufall sein. Wir gehen zunächst einmal davon aus, dass der Spieler ein redlicher Mensch ist und keinen gezinkten Würfel nutzt. Wie oft hintereinander aber muss der Spieler noch eine 6 würfeln, bis wir unsere Gutmütigkeit aufgeben und davon ausgehen, dass unsere Unschuldsvermutung nicht zutrifft? Natürlich kann man auch zehnmal hintereinander eine 6 würfeln. Die Wahrscheinlichkeit ist gering (0,000001653817 %), aber wenn wir nun Fraktur mit dem Spieler reden, so kann es immer noch sein, dass wir uns irren und er unschuldig ist (auch Fehler 1. Art genannt).

Genau so funktioniert die Statistik in einem A/B-Test, vereinfacht dargestellt. Der einzige Unterschied: Es wird vorher festgelegt, ab welcher Grenze die Unschuldsvermutung beziehungsweise die Nullhypothese abgelehnt wird, und nicht erst während des Spiels beziehungsweise Experiments. Üblicherweise wird ein Wert von 5 % als Signifikanzniveau gewählt, das heißt, wenn der p-Wert unter 5 % (häufig als 0.05 notiert) geht, dann wird die Nullhypothese abgelehnt und das Ergebnis ist statistisch signifikant.

INFO

Nicht einmal Wissenschaftler können den p-Wert gut erklären, aber dennoch hält man sich starr daran. Es lohnt sich, die Mechanik zu verstehen, um Testergebnisse richtig einordnen zu können.

Woher kommt eigentlich diese Grenze? Es existiert kein wissenschaftlicher Grund für die 5 %, allerdings nutzte einer der ersten wissenschaftlichen Artikel zu diesem Thema diese Zahl als Beispiel (eine Wahrscheinlichkeit von 1:20 ist schließlich einfach zu erklären), und kaum jemand macht sich die Mühe, einen anderen Wert zu wählen. Signifikanzniveaus von 1 % oder 10 % sind eher selten. Anwender der Test-Tools verändern diese Werte sowieso selten, denn meistens sind sie fest eingebaut (siehe zum Beispiel den Konfidenzrechner von Konversionskraft). Es lohnt sich aber schon, darüber nachzudenken, ob der Wert wirklich sinnvoll ist. Wenn wir jedem Spieler die Freundschaft kündigen, der eine Serie von Würfen hat, deren Wahrscheinlichkeit unter dem Signifikanzniveau von 5 % liegt, dann wären wir schnell einsam. Die Wahrscheinlichkeit, zweimal hintereinander eine 6 zu würfeln, beträgt 2,78 %, also deutlich unter 5 %.

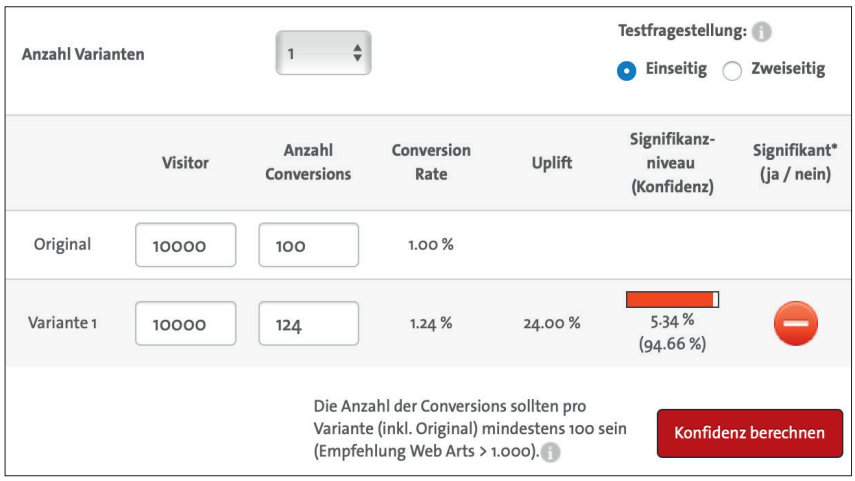


Abb. 3: Es fehlt eine einzige Conversion für ein statistisch signifikantes Ergebnis, allerdings ist die Sample-Größe zu klein *(Quelle: Konversionskraft)*

Oft übersehen: Sample-Größe zu klein

Zurück zu unserem Beispiel: Wie wahrscheinlich ist ein Uplift der YARPP-Klickrate von 24 %, wenn die andere nun getestete Überschrift nichts bringt (Unschuldsvermutung)? In dem Beispiel in Abbildung 3 beträgt die Wahrscheinlichkeit mit den angegebenen (zur Vereinfachung der Rechnung ausgedachten) Parametern 5,34 %. Die Nullhypothese wird also nicht abgelehnt. Aber nur eine Conversion mehr, und der p-Wert läge bei 0,0469 und das Testergebnis wäre statistisch signifikant. Eine einzige Conversion. Ein Statistiker würde nun nicht nachträglich das Signifikanzniveau ändern, damit das Testergebnis doch noch statistisch signifikant wird. Aber man könnte auch einfach eine andere Frage stellen: Was

Version	Include	Trials	Successes	Apprx probability of being best	95% chance conversion rate between
A	☑	15000	150	2%	0.8% and 1.2%
B	☑	15000	186	98%	1% and 1.5%
C	☐	0	0		
D	☐	0	0		

Abb. 4: A/B-Testrechner basierend auf Bayes *(Quelle https://alby.link/bayescalculator)*

würde es kosten, wenn tatsächlich ein Irrtum vorliegt und der Uplift von 24 % im Test einfach nur durch Zufall entstanden wäre, die Änderung aber doch implementiert wird? Es wird auch keine Freundschaft gekündigt, nur weil ein Freund zwei- oder sogar dreimal hintereinander eine 6 würfelt. Natürlich muss dennoch weiterhin der gesunde Menschenverstand eingesetzt werden, insbesondere wenn wenige Daten vorliegen.

Genau das ist in diesem Beispiel ein Problem und davor warnt so gut wie kein Online-Rechner: Die Sample-Größe für diesen Test ist zu klein! Tatsächlich werden mehr als 23.000 Besucher pro Variante benötigt, und da eine andere Überschrift auch schlechter sein könnte, wäre dies eigentlich ein zwei-

seitiger Test, der noch mehr Besucher pro Variante erfordern würde.

Noch einmal zurück zu dem Würfelspiel: Nach drei Würfen einer 6 wären wir uns noch unsicher, ob ein Würfel gezinkt ist. Je mehr Würfe einer 6 wir aber gesehen haben, desto sicherer sind wir uns, dass dies kein Zufall ist, zum Beispiel, wenn ein Spieler fünfmal hintereinander eine 6 würfelt. Was aber, wenn wir den Spieler nicht gleich vom Tisch jagen und dieser dann fünfmal hintereinander keine 6 würfelt? Das ist bei einer Webseite nicht anders: Steigt die Conversion-Rate zu Beginn eines Tests auf über 50 % (unwahrscheinlich, aber kann passieren!), so kann das purer Zufall sein. Wir können also nicht einfach warten, bis der p-Wert unter das Signifikanzniveau fällt, und den Test dann beenden. Wie viele Würfe oder Conversions müssen wir aber gesehen haben, bis wir das tun dürfen?

Damit kommen wir zu einem weiteren, oft ignorierten Parameter, der sogenannten Teststärke oder „Power" eines Tests. Damit ist gemeint, wie sensitiv der Test für einen Unterschied ist, wenn dieser tatsächlich existiert. In Bezug zum p-Wert können wir auch sagen,

» dass der p-Wert möglichst gering sein soll, um die Wahrscheinlichkeit zu minimieren, dass wir einen Spieler des Tricksens bezichtigen, obwohl er unschuldig ist (Fehler 1. Art),
» und die Teststärke gleichzeitig möglichst hoch sein soll, damit die Wahrscheinlichkeit, dass es einen Effekt gibt und wir ihn nicht sehen, auch gering ist (Fehler 2. Art).

In der Regel wird eine Trennschärfe von 80 % gewählt, also eine Wahrscheinlichkeit von 80 %, dass ein Effekt, wenn es ihn gibt, auch festgestellt wird.

Um nun die notwendige Sample-Größe berechnen zu können, fehlt nur noch eine Einschätzung der vermuteten Verbesserung. Das klingt zunächst kontraintuitiv, schließlich will man eigentlich mit dem Test herausfinden, wie groß die Verbesserung ist. Aber offensichtlich ist ein größerer Effekt früher sichtbar als ein kleinerer, und damit auch ein kleinerer Effekt gemessen werden kann, ist diese Angabe wichtig. Ist ein Würfel so gezinkt, dass immer die 6 gewürfelt wird, so wird der Falschspieler schneller entlarvt, als wenn der Würfel zwischendurch auch auf andere Seiten fällt. Ein guter Rechner für die Sample-Größe findet sich unter *https://abtestguide.com/abtestsize/*. Eine kleine Warnung vorab: Die Anzahl der notwendigen Nutzer pro Variante ist immer weit höher, als man es sich wünscht.

Ein Test, zwei verschiedene Ergebnisse

Viele Test-Tools funktionieren nach dem bisher beschriebenen System, zum Beispiel Adobe Target, aber auch viele Signifikanzrechner im Netz. Google Optimize dagegen basiert auf einem anderen Ansatz, dessen Gebrauch in der Statistik noch vor wenigen Jahrzehnten immense Diskussionen ausgelöst hat (und zum Teil immer noch tut): die Bayes'sche Inferenz. Während der oben beschriebene Ansatz vereinfacht ausgedrückt so funktioniert, dass man ein Modell hat und dann schaut, wie die Daten dazu passen, wird mit der Bayes'schen Inferenz geschaut, wie ein Modell zu den Daten passt, denn jeder neue Datenpunkt wird in die Berechnung einbezogen. Das klingt logischer, aber der Bayes-Ansatz birgt eine gewisse Subjektivität in sich, da er nur die Daten nimmt, die beobachtet wurden. Nur weil in einem Würfel-Spiel noch nicht dreimal eine 6 gewürfelt wurde, heißt das nicht, dass das nicht möglich wäre. Ein Vorteil dagegen ist, dass in diesem Verfahren unter Umständen eine etwas geringere Sample-Größe benötigt wird.

Für unser Beispiel hat ein Testrechner auf Basis der Bayes-Statistik ein Ergebnis berechnet, das etwas anders aussieht als das des Signifikanztests (siehe Abbildung 4). Es wurde eine ausreichende Sample-Größe gewählt (hier werden nur 15.000 Nutzer pro Variante benötigt, ein kleiner Vorteil gegenüber dem anderen Testverfahren). Die 98%ige Wahrscheinlichkeit, dass Variante B besser ist, darf nicht verwechselt werden mit einem umgekehrten p-Wert, dafür kann sie so interpretiert werden, wie ein p-Wert oft missinterpretiert wird. Spannend ist dabei die letzte Spalte, in der ein Intervall angegeben wird für die Wahrscheinlichkeit, dass die Conversion-Rate zwischen zwei Werten liegt. Offensichtlich gibt es eine Schnittmenge zwischen Variante A und B. Es könnte sogar sein, dass Variante A besser ist als Variante B. Sicherheit gibt es also auch in diesem Modell nicht, aber wahrscheinlich würde sich ein Anwender mit diesem Test-Ergebnis wohler fühlen als mit dem des anderen Verfahrens.

Wer testet (und jeder sollte testen!), kann dies nicht gewissenhaft tun, ohne sich ein Basis-Wissen der Test-Tools und ihrer Mechaniken anzueignen. Denn wie bei den Webanalyse-Systemen gilt auch hier, dass das beste System kein Gehirn ersetzt.

Im nächsten und letzten Teil wird es dann darum gehen, die Ergebnisse der Webanalyse richtig und wirkungsvoll darzustellen. ¶