Tekst: Kyra Delsing
Beeld: Twinkle/BBP Media

Voorkom p-hacking

De neiging om patronen in gegevens te ontdekken die statistisch significant lijken, terwijl er feitelijk geen effect is, heet p-hacking. Het fenomeen bestaat in verschillende vormen en maten.

P-hacking is een erkend probleem binnen de wetenschap. We zien het veelal terug als:

- Het continu monitoren van het experiment en dit stoppen zodra de waargenomen p-waarde onder een drempelwaarde (gewoonlijk 0,05) daalt.
- Het testen van vele hypothesen, maar alleen die testen rapporteren die significant zijn.
- Het uitsluiten van bezoekers of het aanpassen van data om toch de p-waarde onder de drempelwaarde te krijgen.

Uit recent onderzoek van Ron Berman blijkt dat p-hacking, ondanks de groei in volwassenheid van de *conversion rate optimization* (CRO)-markt, ook binnen de commerciële sector een veel voorkomend fenomeen is. Zo liet de analyse van 2.101 commerciële experimenten uitgevoerd op het A/B-testing platform Optimizely zien dat bij ongeveer 57 procent van de experimenten p-hacking plaatsvindt wanneer het experiment 90 procent zekerheid bereikt.

Vier methodes

Er zijn verschillende methodes waarmee je het risico op p-hacking tijdens je optimalisatiewerkzaamheden kunt verlagen. Hieronder lees je er vier.

1. WORDT ALLES GOED GEMETEN?

Allereerst controleren we altijd voorafgaand aan een experiment of nieuw traject zorgvuldig of de A/B-test-tool en de webanalytics-tool goed geïmplementeerd zijn. Op deze manier kun je naast de informatie of je A/B-testresultaten significant zijn (of niet) achterhalen of er andere inzichten zijn die je helpen het effect op gedrag te doorgronden.

2. BAYESIAANS HYPOTHESETESTEN

De analyse van onze A/B-testen doen we (in de meeste gevallen) aan de hand van Bayesiaanse statistiek. Binnen de Bayesiaanse methode liggen de resultaten een stuk genuanceerder: op basis van een testuitslag wordt bepaald hoe groot de kans is dat de variant beter presteert dan de huidige situatie. Een testuitslag heeft daardoor geen binaire uitslag (zoals bij frequentistische statistiek), maar een kans van 0 tot 100 procent.

De keuze voor Bayesiaanse statistiek betekent echter niet dat je geen risico meer loopt om bijvoorbeeld vroegtijdig een experiment te stoppen. De manier van rapporteren zorgt er alleen voor dat het advies op basis van je analyse minder stellig is en dus minder bepalend voor de mogelijke gevolgen van je experiment (wel of niet implementeren).

3. START MET EEN IMPACTANALYSE

Een impactanalyse vertelt je hoeveel A/B-testen per jaar je als bedrijf statistisch gezien kunt uitvoeren. Doordat je met deze berekening van tevoren vaststelt welk effect er per pagina behaald moet worden (je *power*) en hoe lang een test aan moet staan, voorkom je dat je hier tijdens het experiment, of na afloop, bij de analyse dingen gaat aanpassen en risico loopt op p-hacking.

4. GEWOON NIET SPIEKEN

Naast het feit dat wij op basis van de impactanalyse vaststellen hoe lang een A/B-test aan moet staan en hoe hoog het behaalde effect moet zijn, is spieken bij lopende testen bij ons op kantoor simpelweg verboden. •

Webanalisten.nl
Dit artikel is geschreven door Kyra Delsing, consumer psychology expert bij Online Dialogue, voor het online analyse- en optimalisatieplatform Webanalisten.nl. Het complete artikel is te lezen op http://www.webanalisten.nl/hoe-voorkom-je-p-hacking-bij-ab-testen.