

Aus der Forschung

Was ist P-Hacking?

P-Hacking heißt, Daten und Analysen anzupassen, bis ein Ergebnis signifikant ist, und diese Fassung zu berichten. Es kann p-Werte knapp unter 0,05 häufen.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic)	18. Oktober 2026	20 Min.	4.155
Führungskraft für Umsatzwachstum, Dozent und Doktorand			

Management Summary P-Hacking heißt, Daten oder statistische Analysen zu erheben oder auszuwählen, bis ein nicht signifikantes Ergebnis signifikant wird, und diese Fassung zu berichten. Dieser Konzeptleitfaden definiert es (Head et al.; Munafò et al.), grenzt es vom Schubladenproblem ab, nennt die Entscheidungen, über die es wirkt, erklärt, wie eine p-Kurve es über viele Studien erkennt und was sie übersieht, berichtet eine Text-Mining-Studie, die in gepoolten p-Werten vieler Disziplinen Belege dafür fand, seine Wirkung auf die Schlüsse von Meta-Analysen aber als wahrscheinlich gering einschätzte, zeigt, warum eine wegen ihrer Signifikanz gewählte Schätzung den Effekt eher überschätzt (Gelman und Carlin), und bietet ein Blatt für Analyseteams im Unternehmen. Die Quellen sind eine Text-Mining-Studie mit Nachanalyse von zwölf Meta-Analysen, ein Methodenartikel, ein analytisches Modell und eine Perspektive.

Inhalt	§ 1	Was ist P-Hacking?	3
	§ 2	Welche Analyseentscheidungen können zu P-Hacking werden?	5
	§ 3	Wie zeigt eine p-Kurve P-Hacking an, und was übersieht sie?	6
	§ 4	Wie verbreitet erschien P-Hacking in der veröffentlichten Forschung?	8
	Abbildung 1	Per Text-Mining gewonnene p-Werte aus Ergebnisteilen in den beiden Intervallen knapp unter 0,05	9
	§ 5	Ändert P-Hacking das Fazit einer Forschungsliteratur?	10
	§ 6	Warum überschätzt ein wegen seiner Signifikanz gewähltes Ergebnis den Effekt?	12
	§ 7	Wie erkennt ein Analyseteam P-Hacking in seinen eigenen Berichten?	14
	Abbildung 2	Ein Blatt für Analyseentscheidungen	15

§ 8	Was lassen diese Quellen für Business Analytics offen?	16
	Referenzen	17

Was ist P-Hacking?

§ 1 von 8

Gemeint ist, die Erhebung oder Auswertung von Daten so lange anzupassen, bis ein Ergebnis die Signifikanzschwelle unterschreitet, und dann genau diese Fassung zu berichten. Head et al. (2015)² fassen es in einen Satz: P-Hacking liegt vor, wenn Forschende Daten oder statistische Analysen so lange erheben oder auswählen, bis aus nicht signifikanten Ergebnissen signifikante werden („occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant“). Sie nennen es auch Inflationsbias oder selektives Berichten und beschreiben es als falsche Wiedergabe der wahren Effektstärken in veröffentlichten Studien („is the misreporting of true effect sizes in published studies“): Forschende probieren mehrere statistische Analysen oder Kriterien für die Datenauswahl aus („try out several statistical analyses and/or data eligibility specifications“) und berichten dann gezielt diejenigen, die signifikante Ergebnisse liefern („selectively report those that produce significant results“).

² Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106.

Munafò et al. (2017)⁴ beschreiben dasselbe aus Sicht der Auswertung. Modelle, Ausschlussregeln und Kovariaten können in Kenntnis der beobachteten Daten gewählt werden („with knowledge of the observed data“), und solche datenabhängigen Analyseentscheidungen sind das, was inzwischen P-Hacking heißt („Such data-contingent analysis decisions constitute what has become known as P-hacking“). Sie sprechen auch von Data Dredging („data dredging (also known as P-hacking“) und halten fest, dass jemand, der mehrere mögliche Zielgrößen erhoben hat, nach Vorliegen der Ergebnisse absichtlich oder unabsichtlich diejenigen zu den maßgeblichen erklären kann, die signifikant ausfallen („intentionally or unintentionally select the subset of outcomes that show statistically significant results as the outcomes of interest“).

⁴ Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021.

P-Hacking ist nicht dasselbe wie das Schubladenproblem, und die beiden Quellen ziehen diese Grenze mit unterschiedlichen Begriffen. Head et al.² unterscheiden zwei anerkannte Formen von Publikationsbias, die von Forschenden ausgehen („two widely recognized types of researcher-driven publication bias“): die

Selektion, den Schubladeneffekt, bei dem Studien mit nicht signifikanten Ergebnissen seltener veröffentlicht werden, und die Inflation, also das P-Hacking. Munafò et al.⁴ verwenden den Begriff Publikationsbias nur für die erste Form und beschreiben das Schubladenproblem damit, dass weit mehr Studien durchgeführt als veröffentlicht werden („the fact that many more studies are conducted than published“). In beiden Lesarten wirkt der Schubladeneffekt auf ganze Studien, P-Hacking innerhalb einer Studie, nämlich darauf, welche Analyse gezeigt wird.

In diesem Leitfaden heißt P-Hacking: Daten- oder Analyseentscheidungen werden getroffen, nachdem man gesehen hat, wie sie den p-Wert verändern, und berichtet wird die Fassung, die die Schwelle unterschritten hat, ohne die Fassungen, bei denen das nicht gelang. Dazu gehört, wann die Datenerhebung endet, welche Zielgröße berichtet wird, welche Beobachtungen und Gruppen bleiben und welche Kovariaten einfließen. Im Anschluss an Head et al.² gehört auch dazu, einen p-Wert über 0,05 als kleiner als 0,05 anzugeben; sie zählen das zum P-Hacking. P-Hacking unterscheidet sich vom Schubladenproblem, bei dem ganze Studien unveröffentlicht bleiben, von multiplem Testen, das vollständig berichtet und korrigiert wird und das der Leitfaden zur statistischen Schlussfolgerungsvalidität zusammen mit der Frage behandelt, was ein p-Wert tragen kann und was nicht, und von explorativer Analyse, die als explorativ gekennzeichnet ist. Die Absicht gehört nicht zur Definition: Nach Munafò et al.⁴ kann die Auswahl auch unabsichtlich geschehen.

Welche Analyseentscheidungen können zu P-Hacking werden?

§ 2 von 8

Als Beispiele nennen Head et al.²: Analysen mitten im Experiment, um zu entscheiden, ob weiter Daten erhoben werden; viele Zielvariablen zu erfassen und erst nach der Analyse zu entscheiden, welche berichtet werden; nachträglich zu entscheiden, ob Ausreißer bleiben oder entfernt werden; Versuchsgruppen nach der Analyse auszuschließen, zusammenzulegen oder aufzuteilen; Kovariaten nach der Analyse aufzunehmen oder wegzulassen; und die Suche abubrechen, sobald eine Analyse signifikant ausfällt. Ioannidis (2005)³ nennt für eine hypothetische genetische Assoziationsstudie den gerade passenden Ein- oder Ausschluss bestimmter Patienten oder Kontrollen und nachträgliche Subgruppenanalysen („serendipitous inclusion or exclusion of certain patients or controls, post hoc subgroup analyses“), Kontraste, die zu Beginn nicht festgelegt waren, geänderte Definitionen der Krankheit oder der Kontrollen sowie selektives oder verzerrtes Berichten. Bias definiert er als das Zusammenwirken von Faktoren in Design, Daten, Analyse und Darstellung, die Forschungsbefunde hervorbringen, wo keine entstehen sollten („the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced“), und er ergänzt, selektives oder verzerrtes Berichten sei eine typische Form davon („Selective or distorted reporting is a typical form of such bias.“).

³ Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Munafò et al.⁴ ergänzen den Wechsel der Zielgröße, im Englischen outcome switching: die Möglichkeit, die maßgeblichen Zielgrößen abhängig von den beobachteten Ergebnissen zu ändern („in the study depending on the observed results“). Außerdem verweisen sie auf analytische Flexibilität: In einem Datensatz mit vielen Variablen kann es Hunderte oder Tausende vertretbarer Wege geben, dieselben Daten auszuwerten („there may be hundreds or thousands of reasonable alternative approaches to analysing the same data“). In einer Studie, die sie anführen, zeigten nach fast 7.000 Auswertungswegen für einen einzigen fMRT-Datensatz über 90 Prozent der Voxel des Gehirns in mindestens einer Analyse eine signifikante Aktivierung („applying almost 7,000 analytical pipelines to a single fMRI dataset resulted in over 90% of brain voxels showing significant“). Ioannidis macht aus der Flexibilität eine der Folgerungen seines Modells: Je flexibler Designs, Definitionen, Zielgrößen und Analysen in einem Feld sind, desto

geringer ist die Wahrscheinlichkeit, dass seine Befunde zutreffen, weil Flexibilität mehr Gelegenheit bietet, aus negativen Ergebnissen positive zu machen. Diese Folgerung ist aus dem Modell abgeleitet, nicht in Daten gezählt.

In einem Analyseteam im Unternehmen können dieselben Entscheidungen anders heißen: eine Auswertung, die jede Woche angesehen und beendet wird, sobald sie gut aussieht; eine Kennzahl, die aus einem Dashboard mit vielen Kennzahlen für die Folie ausgewählt wird; ein Zeit- oder Attributionsfenster, das erst nachträglich festgelegt wird; ein großer Kunde, der als Ausreißer herausgenommen wird; Segmente, die so lange geschnitten werden, bis eines einen Unterschied zeigt; eine Kontrollvariable, die hinzukommt oder wegfällt. Jede dieser Entscheidungen ist legitim, wenn sie vorab festgelegt oder mit ihren Alternativen berichtet wird. Zu P-Hacking wird sie, wenn die Wahl dem p-Wert folgt und die Alternativen aus dem Bericht verschwinden. Der [Leitfaden zum A/B-Test](#) behandelt zwei davon für Online-Experimente: das vorzeitige Beenden und Segmentschnitte nach dem Test. Diese Zuordnung stammt vom Autor und ist nicht geprüft.

Wie zeigt eine p-Kurve P-Hacking an, und was übersieht sie?

§ 3 von 8

An ihrer Form nahe 0,05. Für Head et al.² ist eine p-Kurve die Verteilung der p-Werte einer Menge von Studien („A p-curve is the distribution of p-values for a set of studies.“). Gibt es keinen wahren Effekt, ist jeder p-Wert gleich wahrscheinlich („every p-value is equally likely to be observed“), und die Kurve verläuft flach; bei einem wahren Effekt ist die Verteilung exponentiell und rechtsschief („is exponential with a right skew“), kleine p-Werte sind also häufiger als p-Werte nahe 0,05. Machen Forschende aus nicht signifikanten Ergebnissen signifikante, verändert sich die Kurve nahe der Schwelle, und eine durch P-Hacking verzerrte p-Kurve zeigt einen Überschuss an p-Werten knapp unter 0,05 („a p-hacked p-curve will have an overabundance of p-values just below 0.05“). Darin unterscheidet sich P-Hacking vom Schubladeneffekt: Beide lassen einen Bruch der Kurve bei 0,05 erwarten, aber nur P-Hacking einen Überschuss knapp darunter („Both p-hacking and selective publication bias predict a discontinuity in the p-curve around 0.05, but only p-hacking predicts an overabundance of p-values just below 0.05“).

Ihr Test vergleicht zwei schmale Intervalle mit einer Breite von 0,005, das untere von 0,04 bis 0,045 und das obere von 0,045 bis 0,05. Ein einseitiger Vorzeichentest prüft, ob das obere Intervall mehr p-Werte

enthält als das untere. Ohne P-Hacking erwarten die Autoren dort eine Verteilung, die gleichförmig ist, wenn die wahren Effekte null sind, oder rechtsschief („uniform close to 0.05 (if the true effect sizes are zero), or right skewed“); das obere Intervall sollte dann nicht mehr enthalten. Die Grenzen nennen die Autoren selbst. Tests, die die Schiefe der Kurve über breitere Intervalle prüfen, können starkes P-Hacking erkennen, sind für mäßigeres und wohl realistischeres P-Hacking aber unempfindlich („can detect severe p-hacking, but are insensitive to more modest (and arguably more realistic) levels of p-hacking“), vor allem bei starken wahren Effekten, weil deren Rechtsschiefe die durch P-Hacking entstehende Linksschiefe überdeckt („as the right skew introduced to the p-curve will mask the left skew caused by p-hacking“). Ihr Test, der die beiden Intervalle knapp unter 0,05 vergleicht, ist der empfindlichere, und doch nennen sie ihn einen einfachen und konservativen Test („a simple, and conservative, test for p-hacking“). Die Form der Kurve hängt zudem vom wahren Effekt und von der Stärke des P-Hackings ab („on both the true effect (i.e., the p-curve before p-hacking) and the intensity of p-hacking“). Nach der Lesart dieses Leitfadens zeigt eine Kurve ohne den Überschuss deshalb nicht, dass niemand P-Hacking betrieben hat.

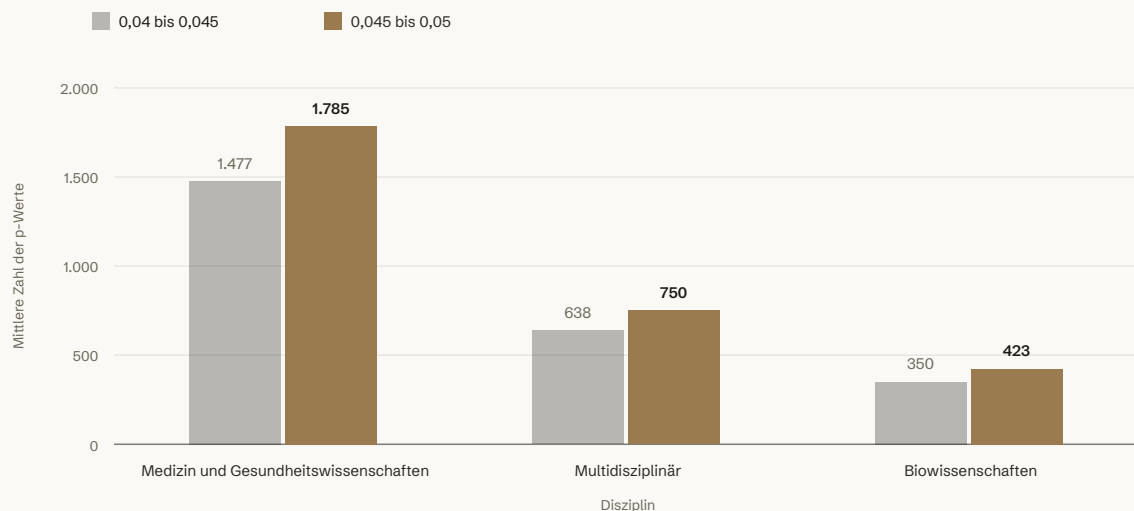
Wie verbreitet erschien P-Hacking in der veröffentlichten Forschung?

§ 4 von 8

Weit verbreitet, nach den Daten von Head et al.², auch wenn ihr Test nur einen Teil davon sehen kann. Sie durchsuchten per Text-Mining alle frei zugänglichen Artikel in der Datenbank PubMed nach p-Werten („used text-mining to search for p-values in all Open Access papers available in the PubMed database“) und bildeten Kurven je Disziplin, getrennt für p-Werte in Ergebnisteilen und in Abstracts. Über alle Disziplinen zusammengekommen waren beide Mengen stark rechtsschief, was die Autoren als Beleg lesen, dass Forschende überwiegend Effekte untersuchen, die nicht null sind. Die gepoolten Daten zeigten aber auch starke Belege für P-Hacking („strong evidence for p-hacking“): In den Ergebnisteilen lag der geschätzte Anteil der p-Werte im oberen Intervall über 14 Disziplinen bei 0,546 (untere Konfidenzgrenze 0,536), wo gleich viele p-Werte in beiden Intervallen 0,5 ergäben, in den Abstracts über 10 Disziplinen bei 0,537 (0,518). In den meisten Disziplinen lagen mehr p-Werte im oberen als im unteren Intervall („In most disciplines, there were more p-values in the upper than the lower bin.“). In den Ergebnisteilen war der Unterschied in jeder Disziplin signifikant, für die die Autoren eine gute Teststärke hatten; sie nennen Medizin und Gesundheitswissenschaften, Biowissenschaften und die Kategorie multidisziplinär, und Abbildung 1 zeigt diese drei. In den Abstracts waren nur die multidisziplinäre Kategorie und die Informations- und Computerwissenschaften signifikant.

Abbildung 1 Per Text-Mining gewonnene p-Werte aus Ergebnisteilen in den beiden Intervallen knapp unter 0,05

In jeder der drei Disziplinen mit den meisten p-Werten lagen im Intervall knapp unter 0,05 mehr p-Werte als im Intervall darunter, was die Autoren ohne P-Hacking nicht erwarten.



Der einseitige Binomialtest auf P-Hacking ergab $p < 0,001$ für Medizin und Gesundheitswissenschaften, $0,001$ für die Kategorie multidisziplinär und $0,005$ für die Biowissenschaften. Von den 14 Disziplinen in Tabelle 1 hatten neun mehr p-Werte im oberen als im unteren Intervall; signifikant auf dem 5-Prozent-Niveau war der Test außerdem für Psychologie und Kognitionswissenschaften (29 gegenüber 50, $p = 0,012$), die nicht dargestellt sind.

Quelle Head et al. (2015), S. 7, Tabelle 1: mittlere Zahl der p-Werte aus Ergebnisteilen frei zugänglicher Artikel in PubMed, auf Grundlage von 1.000 Bootstrap-Ziehungen von je einem p-Wert pro Ergebnisteil und auf ganze Zahlen gerundet, in den Intervallen 0,04 bis 0,045 und 0,045 bis 0,05, für die drei Disziplinen, für die die Autoren eine gute Teststärke angeben (S. 8). Einseitiger Binomialtest auf P-Hacking aus derselben Tabelle.

Ihre Lesart ist eindeutig: Das Text-Mining deutet darauf hin, dass P-Hacking weit verbreitet ist („Our text-mining suggests that p-hacking is widespread.“). Zwei Einschränkungen fügen sie hinzu. Alle untersuchten Kurven blieben stark rechtsschief („still a strong right skew in all the p-curves we examined“), was sie angesichts der Sorge, viele veröffentlichte Ergebnisse seien falsch, beruhigend finden. Und weil das Text-Mining viele verschiedene Fragestellungen zusammenfasst („many different types of questions“), bleibt offen, ob einzelne Felder oder Fragen innerhalb dieser Disziplinen Effekte untersuchen, die im Mittel null sind. Sie berichten außerdem, dass eine

frühere Text-Mining-Studie zu fünf medizinischen Zeitschriften keine Hinweise auf P-Hacking fand und dafür kritisiert wurde, p-Werte aus Abstracts zu verwenden, die eher nur die stärksten Ergebnisse enthalten.

Ändert P-Hacking das Fazit einer Forschungsliteratur?

§ 5 von 8

Wahrscheinlich nicht stark, jedenfalls in dem einen Gebiet, das Head et al.² prüften, auch wenn es die Schätzwerte aufblähen kann. Sie wiederholten den Test mit p-Werten aus Studien, die in zwölf Meta-Analysen von Evolutionsbiologen zur sexuellen Selektion eingingen („by evolutionary biologists studying sexual selection“). Neun der zwölf Kurven zeigten einen signifikanten Evidenzwert, also Belege dafür, dass der untersuchte Effekt nicht null ist. Wurden falsch berichtete p-Werte (16 Werte, die als kleiner als 0,05 angegeben, tatsächlich aber größer waren) dem oberen Intervall zugerechnet, hatten sieben der zwölf Kurven dort mehr p-Werte als im unteren, und signifikant war der Unterschied in einer, dem Datensatz mit der größten Stichprobe. Ohne die falsch berichteten p-Werte verschwanden die Hinweise auf P-Hacking („However, the evidence for p-hacking disappeared when we excluded misreported p-values from our analyses.“). Über alle zwölf lag der Anteil im oberen Intervall mit den falsch berichteten Werten bei 0,615 (untere Grenze 0,513, $p = 0,033$) und ohne sie bei 0,489 (0,375, $p = 0,443$). Die Autoren behalten diese Werte bei: Nicht signifikante Ergebnisse als kleiner als 0,05 zu berichten, sei ein Bestandteil von P-Hacking, der nicht übergangen werden sollte („is a component of p-hacking that should not be ignored“).

Ihr Urteil über die Folgen ist in beide Richtungen abgesichert. P-Hacking sei wahrscheinlich verbreitet, seine Wirkung erscheine aber gering im Vergleich zu den realen Effektstärken, die gemessen werden („its effect seems to be weak relative to the real effect sizes being measured“), und dieses Ergebnis deute darauf hin, dass P-Hacking den wissenschaftlichen Konsens aus Meta-Analysen wahrscheinlich nicht drastisch verändert („suggests that p-hacking probably does not drastically alter scientific consensus drawn from meta-analyses“). Die eine Meta-Analyse mit einem signifikanten Hinweis auf P-Hacking zeigte zugleich einen starken Evidenzwert, und ihre p-Werte zwischen 0,045 und 0,05 machten nur einen kleinen Teil der veröffentlichten signifikanten p-Werte aus („were only a small proportion of published significant p-values“). Deshalb, schreiben sie, werde P-Hacking die qualitativen Schlussfolgerungen dieser Meta-Analyse wohl nicht

ändern, könne aber die geschätzte mittlere Effektstärke aufgebläht haben („It is therefore unlikely that p-hacking would change the qualitative conclusions made in this meta-analysis, although p-hacking might have inflated the estimated mean effect size.“). Zwei Gründe nennen sie, warum eine Meta-Analyse dem standhalten kann: Unter sonst gleichen Bedingungen erhalten die kleinen Studien, die für P-Hacking am anfälligsten sind, weniger Gewicht, und zumindest in manchen Feldern verwenden Meta-Analysen häufig Daten, die mit der Hauptfrage der ursprünglichen Arbeit nicht direkt zusammenhängen; p-Werte zu Nebenfragen seien seltener von P-Hacking betroffen („The p-values associated with secondary questions are less likely to be p-hacked.“). Ihre Zusammenfassung hält beide Hälften fest: Die Belege sprechen für mittlere Effekte ungleich null, aber die geschätzte mittlere Effektstärke ist durch P-Hacking wahrscheinlich aufgebläht („estimated mean effect size has probably been inflated by p-hacking“).

Ioannidis (2005)³ bildet denselben Mechanismus in einem Modell ab. Darin senkt Bias die Wahrscheinlichkeit, dass ein behaupteter Befund zutrifft („Thus, with increasing bias, the chances that a research finding is true diminish considerably.“). Das ist ein Ergebnis des Modells unter angenommenen Eingangswerten, keine Zählung falscher Befunde. Er hält außerdem fest, dass sich einem einzelnen Bericht nicht ansehen lässt, wie viel Testen ihm vorausging („usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding“).

Warum überschätzt ein wegen seiner Signifikanz gewähltes Ergebnis den Effekt?

§ 6 von 8

Weil bei einer Studie, die im Verhältnis zum wahren Effekt klein ist, nur große Schätzwerte die Signifikanzschwelle erreichen. Gelman und Carlin (2014)¹ fassen das in eine Kennzahl, den Übertreibungsfaktor (exaggeration ratio): den Erwartungswert des Betrags der Schätzung, geteilt durch die Effektstärke, sofern die Schätzung signifikant von null verschieden ist („the expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero“). Hinzu kommt bei ihnen die Typ-S-Fehlerrate, also die Wahrscheinlichkeit, dass eine signifikante Schätzung das falsche Vorzeichen hat. Für unverzerrte, normalverteilte Schätzungen sind beide bei 80 Prozent Teststärke klein, mit einem erwarteten Übertreibungsfaktor von 1,12 („an expected exaggeration factor of 1.12“). Für solche Schätzungen zeigen sich Probleme mit dem Übertreibungsfaktor ab einer Teststärke unter 0,5, Probleme mit den Typ-S-Fehlern ab einer Teststärke unter 0,1 („problems with the Type S error rate start to arise when power is less than 0.1“).

¹ Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Ihr erstes Beispiel ist ein veröffentlichter Befund aus einer Stichprobe von 2.972 Befragten einer Längsschnittbefragung, nach dem von den erstgeborenen Kindern der Eltern in der attraktivsten Kategorie 56 Prozent Mädchen waren, gegenüber 48 Prozent in den übrigen Gruppen („Of the first-born children of the parents in the most attractive category, 56% were girls, compared with 48% in the other groups“), ein Unterschied von 8 Prozentpunkten bei $p = 0,015$. Aus der Literatur zum Geschlechterverhältnis bei Geburten leiten sie als Hypothesen wahre Unterschiede von 0,1, 0,3 und 1,0 Prozentpunkten ab. Bei 0,1 Punkten hätte ein signifikantes Ergebnis mit 46 Prozent Wahrscheinlichkeit das falsche Vorzeichen („a 46% chance it will have the wrong sign“), und der geschätzte Effekt wäre im Erwartungswert 77-mal zu groß („in expectation the estimated effect will be 77 times too high“); bei 0,3 Punkten liegt die Wahrscheinlichkeit des falschen Vorzeichens bei etwa 40 Prozent und die Überschätzung beim Faktor 25; bei 1,0 Punkten hat ein signifikanter Befund mit 19 Prozent Wahrscheinlichkeit das falsche Vorzeichen, und der wahre Effekt würde im Erwartungswert um den Faktor 8 überschätzt („A statistically significant finding in this case has a 19% chance of appearing with the wrong sign, and the magnitude of the true effect would be overestimated by an expected factor of 8.“). In ihrem zweiten Beispiel,

einem berichteten Unterschied von 17 Prozentpunkten bei den Wahlpräferenzen von Frauen in verschiedenen Phasen des Menstruationszyklus, setzen sie den wahren Effekt auf 2 Punkte, was sie für die obere Grenze des Plausiblen halten; daraus ergeben sich eine Teststärke von 0,06, eine Wahrscheinlichkeit von 24 Prozent für das falsche Vorzeichen und ein erwarteter Übertreibungsfaktor von 9,7.

Ihre allgemeine Aussage lautet, dass statistische Signifikanz als Filter dazu führen kann, die Größe eines Effekts drastisch zu überschätzen („Using statistical significance as a screener can lead researchers to drastically overestimate the magnitude of an effect“). Im ersten Beispiel sehen sie von multiplen Vergleichen ab und nehmen die veröffentlichte Signifikanz beim Wort („ignore multiple comparisons issues and take the published claim of statistical significance at face value“), im zweiten nehmen sie den veröffentlichten Vergleich und seinen p-Wert beim Wort („the published comparison and p value at face value“); die berechnete Übertreibung stammt also allein aus dem Signifikanzfilter. Nach der Lesart dieses Leitfadens kommt P-Hacking noch hinzu, weil jede weitere Fassung der Analyse einer verrauschten Schätzung eine weitere Gelegenheit gibt, die Signifikanzschwelle zu erreichen; in eine ähnliche Richtung zielen Gelman und Carlin¹, wenn sie schreiben, multiple Vergleiche oder die Freiheitsgrade der Forschenden machten es leicht, große und signifikante Effekte zu finden, die allein aus Rauschen entstehen könnten („make it easy for researchers to find large and statistically significant effects that could arise from noise alone“).

Als Abhilfe schlagen sie eine Designanalyse vor, die auf einer Effektstärke aus Literaturübersichten oder anderen Informationen außerhalb der vorliegenden Daten beruht („determined from literature review or other information external to the data at hand“). Frühere Schätzwerte sind dafür ein schwacher Anhaltspunkt, weil sie selbst zur Überschätzung neigen („will themselves tend to be overestimates“), soweit die veröffentlichte Literatur nach Signifikanz auswählt („selects on statistical significance“); eine veröffentlichte Punktschätzung reicht nicht aus („using a published point estimate is not enough“). Und eine Designberechnung braucht eine angenommene Effektstärke und fügt einer vorhandenen Analyse nichts hinzu, wenn diese Effektstärke aus denselben Daten geschätzt wird („A design calculation requires an assumed effect size and adds nothing to an existing data analysis if the postulated effect size is estimated from the very same data.“). Dass die Eingabe ein Urteil bleibt, sagen sie offen („We recognize that any assumption of effect sizes is just that, an assumption.“).

Wie erkennt ein Analyseteam P-Hacking in seinen eigenen Berichten?

§ 7 von 8

Nicht an den p-Werten. Eine p-Kurve braucht viele unabhängige Ergebnisse, und ein Team mit einer Handvoll Analysen zu einer Frage kann keine zeichnen; geprüft wird deshalb die Dokumentation der Entscheidungen, nicht die Form einer Verteilung. Head et al.² empfehlen Forschenden, ihre Forschung klar als vorab festgelegt oder explorativ zu kennzeichnen („Clearly label research as prespecified“), damit Lesende die Ergebnisse mit angemessener Vorsicht behandeln können („so that readers can treat results with appropriate caution“), und Daten wo immer möglich verblindet auszuwerten („Perform data analysis blind wherever possible.“). Den Zeitschriften empfehlen sie, in ihren Richtlinien von Autorinnen und Autoren zu verlangen, den gesamten Analyseprozess offenzulegen und nicht nur die letzten Tests, aus denen die berichteten p-Werte stammen („about the entire analysis process (not just the final tests used to generate reported p-values“). Munafò et al.⁴ beschreiben eine Form der Verblindung, bei der während der Aufbereitung und Bereinigung der Daten die Bezeichnungen der Bedingungen oder Variablen so maskiert werden, dass sich die Ausgabe nicht im Sinne der Forschungshypothese deuten lässt („can be masked so that the output is not interpretable in terms of the research hypothesis“), und sie argumentieren, dass Analyseentscheidungen, die vor dem Blick auf die Daten festgehalten werden („prior to observing the data“), datenunabhängig bleiben; der Leitfaden zur Reproduzierbarkeit behandelt die Präregistrierung. Abbildung 2 macht daraus einen Eintrag je berichtetem Ergebnis und ergänzt die Prüfung von Gelman und Carlin¹ auf die plausible Größe des Effekts.

Abbildung 2 Ein Blatt für Analyseentscheidungen

Ergebnis, Entscheidungen, Zeitpunkt, alle gerechneten Fassungen, plausiblen Effekt und Kennzeichnung festhalten, bevor das Ergebnis in einen Bericht geht.

Ergebnis	Entscheidungen	Festgelegt	Fassungen	Externe Größe	Kennzeichnung
Welche Schätzung kommt in den Bericht?	Abbruch, Zielgröße, Ausschlüsse, Gruppen, Kovariaten?	Vor oder nach Blick auf Ergebnisse?	Wie viele gerechnet, alle gezeigt?	Plausibler Effekt aus anderen Daten?	Vorab festgelegt oder explorativ?

Fiel eine Entscheidung nach Blick auf die Ergebnisse, alle gerechneten Fassungen zeigen und das Ergebnis als explorativ kennzeichnen.

Eine Zeile je berichtetem Ergebnis. Die externe Größe stammt aus anderen Daten, nicht aus der Schätzung selbst. Liegt eine Schätzung aus einer kleinen Analyse weit darüber, dürfte sie den Effekt überschätzen.

Quelle Entscheidungsblatt des Autors. Das Feld für die Entscheidungen folgt den Praktiken, die Head et al. (2015), S. 1, und Munafò et al. (2017), S. 3, nennen; das Feld für den Zeitpunkt folgt Munafò et al. (2017), S. 3; die Felder für Fassungen und Kennzeichnung folgen Head et al. (2015), S. 12; das Feld für die externe Größe folgt Gelman und Carlin (2014), S. 642 und 648. Die Felder sind Eingaben der Lesenden; es werden keine Unternehmensdaten geliefert.

Ein Beispiel ohne Unternehmensdaten: Ein Quartalsreview fragt, ob sich die Abschlussquoten nach einer Änderung der Rabattrichtlinie unterscheiden. Der geplante Vergleich, die Abschlussquote über alle Deals vor und nach der Änderung, zeigt keinen signifikanten Unterschied. Daraufhin werden Deals unter einer Größenschwelle weggelassen, der Rest wird nach Segmenten aufgeteilt, und die Zielgröße wechselt von der Abschlussquote zum durchschnittlichen Auftragsvolumen; eine von zwölf Fassungen, das Auftragsvolumen im Mid-Market-Segment, ergibt $p = 0,03$. Der Eintrag zeigt das geplante Ergebnis mit seinem p-Wert, nennt die drei Entscheidungen, die nach dem ersten Ergebnis fielen, zählt die zwölf Fassungen und führt sie alle auf, nennt als externe Größe die Veränderung des Auftragsvolumens nach einer früheren Richtlinienänderung, sofern das Team eine kennt, die nicht selbst wegen ihrer Signifikanz ausgewählt wurde, und kennzeichnet die Mid-Market-Zahl als explorativ: als Frage für einen

geplanten Vergleich, nicht als Befund. Der Eintrag macht einen Vorher-nachher-Vergleich nicht kausal; er zeigt, wie weit die berichtete Zahl ausgewählt wurde. Das sind Beispiele des Autors, sie sind nicht geprüft.

Was lassen diese Quellen für Business Analytics offen? § 8 von 8

Wie oft es dort vorkommt und ob ein Eintrag wie in Abbildung 2 es seltener macht. Die Belege von Head et al. stammen aus p-Werten in frei zugänglichen, in PubMed erfassten Artikeln und aus zwölf Meta-Analysen in einem Gebiet der Evolutionsbiologie; ihr Test ist konservativ und fasst viele Fragestellungen zusammen; ob sich in den Meta-Analysen überhaupt P-Hacking zeigt, hängt davon ab, ob falsch berichtete p-Werte mitzählen, und ihr Urteil, seine Folgen seien gering, stützt sich auf diese zwölf. Die Zahlen von Gelman und Carlin¹ sind Designberechnungen unter angenommenen wahren Effekten für zwei veröffentlichte Studien, deren veröffentlichte Signifikanz sie beim Wort nehmen. Der Essay von Ioannidis ist ein Modell mit angenommenen Eingangswerten, keine Messung. Munafò et al.⁴ schlagen Maßnahmen vor und schreiben, es gebe einige Belege aus Simulationen und empirischen Studien für deren wahrscheinliche Wirksamkeit („There is some evidence from both simulations and empirical studies supporting the likely effectiveness of these measures“), ihre breite Einführung werde aber eine fortlaufende Bewertung und Verbesserung erfordern („will require iterative evaluation and improvement“). Keine der vier Arbeiten hat P-Hacking in der Datenanalyse von Unternehmen gemessen.

Nach der Lesart dieses Leitfadens lassen sich zwei Dinge mit einiger Zuversicht auf Business Analytics übertragen: Ist eine Analyse im Verhältnis zum plausiblen Effekt klein, überschätzt eine Schätzung, die behalten wurde, weil sie die Signifikanzschwelle erreichte, den Effekt wahrscheinlich; und ein einzelner Bericht lässt sich nicht an seinen eigenen p-Werten auf P-Hacking prüfen, sodass die Dokumentation der Entscheidungen die Prüfung ist. Ob eine solche Dokumentation verändert, was Analyseteams berichten, prüfen die Quellen nicht.

 Ende des Beitrags

Referenzen

Zitate stehen in der Sprache des Originals; die deutschen Wiedergaben im Fließtext sind sinngemäß und stammen vom Autor.

- 1** Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- 2** Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- 3** Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- 4** Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 0021. <https://doi.org/10.1038/s41562-016-0021>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist P-Hacking? Sinan Isoglu.
<https://isoglu.com/de/insights/journal/was-ist-p-hacking/>
Online-Fassung: <https://isoglu.com/de/insights/journal/was-ist-p-hacking/>

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand
© 2026 Sinan Isoglu · isoglu.com · Alle Rechte vorbehalten