Michael Schiffinger

Senior Scientist,
Interdisziplinäres Institut für verhaltenswissenschaftlich orientiertes Management & Kompetenzzentrum für empirische Forschungsmethoden,
Wirtschaftsuniversität Wien

michael.schiffinger@wu.ac.at

A world of p(ain)

Wie signifikant ist „(statistisch) signifikant“?

Die verbreitete Praxis, aus inferenzstatistischen Signifikanztests eine Bestätigung oder Widerlegung von Annahmen abzuleiten, führt zwar nicht filmzitatgemäß in eine Welt des Schmerzes, aber potenziell durchaus in eine Welt von Fehlinterpretationen und -schlüssen. Dieser Artikel richtet sich primär an Managementforscher/innen, die Studien mit mehr oder minder „signifikanten“ Ergebnissen lesen oder (mit)verfassen, sich aber mit der eigentlichen Bedeutung (und den Limitationen) des nach wie vor oft publikationsentscheidenden Kürzels „$p < 0{,}05$“ noch nicht näher auseinandergesetzt haben. Aber auch Manager/innen sind in ihrer Arbeit durchaus mit „signifikanten“ Forschungsergebnissen konfrontiert, etwa im Rahmen von Marktforschungsstudien, Mitarbeiter/innenbefragungen, sonstiger Auftragsforschung oder gleichfalls bei der Lektüre von Fachartikeln. Das Ziel dieses Beitrags ist einerseits eine nachvollziehbare Darstellung der Grundlogik hinter dem üblichen Hypothesentesten, andererseits das Aufzeigen von Alternativen und Ergänzungen bei der Ergebnisdarstellung und -interpretation jenseits der binären Einordnung als „(nicht) signifikant“.

Der Signifikanztest als Hypothesentest

Etwas vereinfacht ausgedrückt lässt sich das übliche Vorgehen bei quantitativen Untersuchungen wie folgt skizzieren:

- Basierend auf einer Theorie wird eine Annahme (Hypothese) formuliert, etwa: “Es gibt einen positiven Zusammenhang zwischen organisationaler Wertschätzung und Unterstützung (*perceived organizational support*, POS) und freiwilligem Arbeitsengagement (*organizational citizenship behavior*, OCB)“. Die erwartete Effektgröße (Wie stark ist dieser Zusammenhang vermutlich?) wird dabei kaum je definiert. Es wird bloß unterstellt, dass der Effekt von null unterschiedlich ist, und zumeist wird wie im

obigen Beispiel auch die Richtung des Effekts spezifiziert (also ob ein positiver oder negativer Zusammenhang erwartet wird).

- Diese Annahme wird anhand einer Stichprobe und einem geeigneten Auswertungsverfahren empirisch geprüft, für das obige Beispiel etwa mit einer Korrelation. Das in der Forschungspraxis für die Auswertung und Ergebnisinterpretation meist entscheidende Element ist dabei die statistische Signifikanz des Ergebnisses, meist durch das Kürzel „$p < 0,05$“ bzw. „$p < 0,01$“ und/oder über Sternchen symbolisiert. In diesem Fall wird das Ergebnis als statistische Untermauerung („Bestätigung“) der Hypothese/des Zusammenhangs gewertet (als Umkehrschluss aus der „Widerlegung“ der zur Hypothese komplementären Nullhypothese). Liegt der p-Wert über 5%, wird das als Hinweis dafür interpretiert, dass der in der Hypothese angenommene Effekt nicht existiert.

So weit, so bekannt und scheinbar bewährt. Dieses Vorgehen ist indes in den letzten Jahren zunehmend in Misskredit geraten, wobei sich diese Diskussion eher in methodisch/statistisch orientierten Kreisen abspielt als in der angewandten Forschung.[2] Ein Anlass dafür ist die so genannte Replikationskrise: einige spektakuläre und prominente (und statistisch signifikante) Forschungsergebnisse insbesondere aus der Psychologie ließen sich nicht wiederholen (z.B. zur Wirkung von „machtvollen“ Körperhaltungen[3]), was Zweifel an der Gültigkeit dieser Forschungsarbeiten bzw. der dadurch scheinbar gestützten Theorien aufwirft.[4]

Der Ursprung des heute üblichen Hypothesentestens[1]

Das beschriebene Vorgehen entwickelte sich aus zwei Ansätzen mit Ursprung in den 1920ern:

Fisher: Für eine zu widerlegende Hypothese („nullify“, daher rührt auch der Name Nullhypothese) ist der ermittelte p-Wert als Maß der (Un-)Vereinbarkeit zwischen dieser Nullhypothese und den beobachteten Daten zu sehen. Das Verfahren ist v.a. dann anzuwenden, wenn man nur wenig über das Thema der Untersuchung weiß, und ein p-Wert unter 5% ist bloß ein Wink, weitere Studien durchzuführen.

Neyman/Pearson: Es werden zwei konkurrierende Hypothesen formuliert, samt einschlägiger (für jede Studie spezifischer) Entscheidungskriterien; anhand des p-Werts wird eine der beiden Hypothesen angenommen.

Das heute übliche Prozedere kombiniert die 5%-Schwelle für eine Nullhypothese mit der binären Entscheidung für eine Hypothese.

Wofür der p-Wert (nicht) steht

Ein Teil der Irritation über die fehlende Replizierbarkeit rührt allerdings vielleicht nur daher, dass der p-Wert als Kriterium für statistische Signifikanz oft falsch interpretiert wird, etwa als Wahrscheinlichkeit, dass die (zu widerlegende) Nullhypothese zutrifft (bzw. 1-p als Wahrscheinlichkeit, dass die eigentliche Forschungshypothese zutrifft), oder als Wahrscheinlichkeit für ein reines Zufallsergebnis.[5]

Tatsächlich bezieht sich das eigentliche Forschungsinteresse meist auf die Frage: „Wie wahrscheinlich ist es (angesichts der Ergebnisse), dass ich mit meiner Hypothese/Theorie richtig liege?“ Dieser Erkenntniswunsch fördert (u.U. noch genährt durch die alternative Bezeichnung des p-Werts als „Irrtumswahrscheinlichkeit“) die verführerische Fehlinterpretation des p-Werts als direkter Indikator für die empirische Belastbarkeit einer Hypothese bzw. Theorie.[6]

Der p-Wert ist kein Maß für die (Un-)Wahrscheinlichkeit einer Hypothese, sondern ein Maß für die (Un-)Vereinbarkeit der Daten mit der Nullhypothese (meist: eines Nicht-Effekts).

Wie im nächsten Abschnitt genauer erläutert, gibt der p-Wert allerdings „nur“ die Wahrscheinlichkeit für den beobachteten Effekt (oder einen noch größeren) an, *falls die Nullhypothese zutrifft*. Der p-Wert sagt also höchstens indirekt etwas über die Wahrscheinlichkeit einer Hypothese angesichts der beobachteten Daten aus, sondern bloß über die (Un-)Wahrscheinlichkeit der beobachteten Daten angesichts einer Hypothese.

Ein („wahrer“) Populationseffekt, viele mögliche beobachtete Effekte

Wie lassen sich die Mechanismen hinter dem Hypothesentesten via Signifikanztest veranschaulichen? Die wohl bekannte Grundidee liegt in der Verallgemeinerung der beobachteten Stichprobenergebnisse auf eine (in der Praxis nicht immer präzise definierte) Population bzw. der diesbezügliche Abgleich mit den zuvor formulierten Hypothesen. Weniger offensichtlich bzw. präsent ist vielleicht, dass man nicht davon ausgehen sollte, mit den Stichprobenergebnissen den tatsächlichen Populationswert genau zu treffen, erst recht nicht bei kleineren Stichproben.

Abbildung 1[7] zeigt die Verteilung der zu erwartenden Korrelationen (etwa zwischen POS und OCB, um beim eingangs erwähnten Beispiel zu bleiben) für Stichproben mit n = 100[8]. Die rote Kurve links gibt diese Verteilung für die übliche Nullhypothese einer Nullkorrelation an; die blaue, gestrichelte Kurve für eine unterstellte „wahre“ Korrelation von 0,2 – eine typische Größenordnung für empirisch ermittelte Korrelationen im Bereich der Organisations- und Managementforschung.[9] Dabei wird in der Forschungshypothese „je höher POS, desto höher OCB“ von einem gerichteten (hier: positiven) Zusammenhang ausgegangen. Somit werden nur positive Korrelationen auf statistische Signifikanz geprüft („einseitige Testung“), negative Korrelationen gelten schon grundsätzlich als unvereinbar mit der Hypothese.

Angenommen, es gäbe tatsächlich keinen Zusammenhang zwischen POS und OCB (d.h., die wahre Korrelation in der Population ist 0), und man führt (viele) Studien dazu mit Stichproben von je 100 Personen durch. Dann ist zwar eine beobachtete Korrelation nahe bei 0 wahrscheinlicher (d.h., wird häufiger vorkommen) als eine merklich positive

oder negative Korrelation (rote Kurve), aber es gäbe doch einige Studien, in denen (trotz der Nullkorrelation in der Population) nennenswerte Zusammenhänge beobachtet würden: konkret in 5 % der Studien eine Korrelation von über 0,16 (rosa schattierter „α"-Bereich). Nachdem die oben genannte Hypothese (wie bei hinreichender theoretischer Begründung üblich) explizit einen positiven Zusammenhang zwischen POS und OCB unterstellt, ergibt sich das „erstrebte" $p < 0{,}05$ (bei $n = 100$) somit für alle Korrelationen über 0,16, die dann als statistisch signifikant bezeichnet werden – weil eben nur in 5 % der Studien mit $n = 100$ eine positive Korrelation von über 0,16 beobachtet werden würde, *falls es in der Population tatsächlich keine Korrelation gibt*.

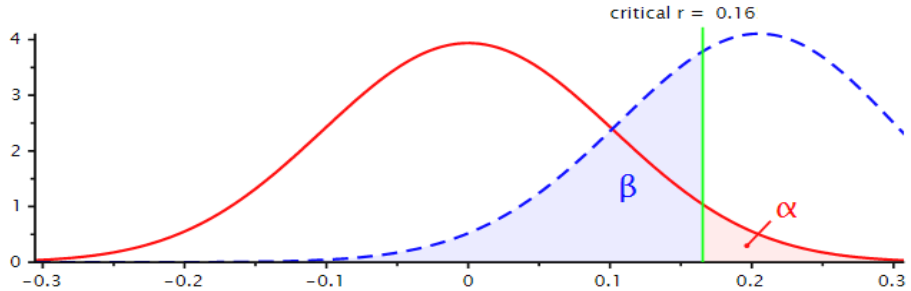

Abbildung 1: Verteilung der Stichprobenkorrelationen für n = 100, falls es in der Population keinen Zusammenhang gibt (rote Kurve links) bzw. für eine Populationskorrelation von 0,2 (blaue Kurve). Das *critical r* ist die Minimumkorrelation für ein statistisch signifikantes Ergebnis auf dem 5 %-Niveau (bei einseitiger Testung).

Ebenso wie in 5 % der Studien (bei tatsächlicher Nullkorrelation) eine Korrelation von über 0,16 beobachtet würde, ergäbe sich in weiteren 5 % der Studien eine Korrelation von unter -0,16. Eine Korrelation von ±0,16 wäre somit auch das *critical r* bei *zwei*seitiger Testung auf dem *10 %*-Niveau ($p < 0{,}10$, bisweilen als „marginal signifikant" bezeichnet), wenn also die Hypothese mangels hinreichender Theoriefundierung nur lautet, dass POS und OCB „irgendwie" zusammenhängen – entweder positiv oder negativ. Auf dem üblichen 5 %-Niveau läge das *critical r* (für $n = 100$) bei zweiseitiger Testung hingegen bei ±0,2, weil (bei Zutreffen der Nullhypothese einer Korrelation von 0 in der Population) nur 2,5 % der Studien eine Korrelation von 0,2 oder mehr ergäben und weitere 2,5 % eine Korrelation von -0,2 oder weniger. Ein „Umstieg" von zweiseitiger auf einseitige Testung bei gleichem Signifikanzniveau kann also gleichsam einen statistisch nicht signifikanten Effekt (hier z.B. eine Korrelation von 0,18) in ein statistisch signifikantes Ergebnis verwandeln. Gerichtete Hypothesen, die nicht vorab auf Basis einer Theorie, sondern erst auf Basis der beobachteten Ergebnisse abgeleitet werden, stellen indes eine zwar durchaus verbreitete aber dennoch höchst fragwürdige Forschungspraxis dar.[10]

Wenn nun die tatsächliche Korrelation zwischen POS und OCB „wie üblich" bei 0,2 liegt, dann würden die beobachteten Korrelationen in Studien mit n = 100 der gestrichelten blauen Kurve folgen: Korrelationen um 0,2 wären am wahrscheinlichsten/häufigsten, aber einige Studien würden sogar eine negative Korrelation berichten (und genauso viele respektive wenige – genauer gesagt jeweils 2,5 % – eine von 0,4 oder mehr). In diesem Fall wäre also nur bei knapp zwei Drittel der Studien die beobachtete Korrelation größer als 0,16 und somit statistisch signifikant, bei den anderen Studien (blau schattierter „β"-Bereich) würde (zu Unrecht) die Nullhypothese favorisiert.

Bei einer größeren Stichprobe bleibt das Prinzip gleich, aber die Schwankungsbereiche der zu erwartenden Korrelationen sind geringer, wie in der folgenden Abbildung für Stichproben mit n = 500 veranschaulicht. Hier würden bei einer tatsächlichen Korrelation von 0 weniger als 5 % der Studien eine Korrelation von über 0,1 oder unter -0,1 ergeben, und wenn nur positive Korrelationen relevant sind, wären bereits alle Korrelationen ab 0,07 statistisch signifikant – ein Zusammenhang, der ob seiner Geringfügigkeit praktisch irrelevant ist (und somit trotz statistischer Signifikanz auch nicht signifikant im üblichen Sinne von bedeutungsvoll, erheblich o.ä.). Wenn die tatsächliche Korrelation bei 0,2 liegt (blaue Kurve), ist es bei einem Stichprobenumfang von n = 500 beinahe unmöglich, eine „nicht signifikante" Korrelation von weniger als 0,07 zu beobachten.

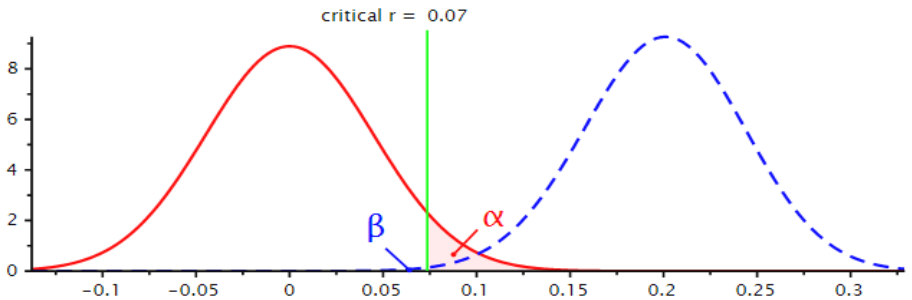

Abbildung 2: Verteilung der Stichprobenkorrelationen für n = 500, falls es in der Population keinen Zusammenhang gibt (rote Kurve links) bzw. für eine Populationskorrelation von 0,2 (blaue Kurve). Das *critical r* ist die Minimumkorrelation für ein statistisch signifikantes Ergebnis auf dem 5 %-Niveau (bei einseitiger Testung).

Was bedeutet also ein statistisch signifikantes Ergebnis?

Was lässt sich punkto Aussagekraft der Hypothesenprüfung via Signifikanztest daraus ableiten? Kurz gesagt: bei kleinen Stichproben braucht es einen großen Effekt für ein statistisch signifikantes Ergebnis. Im obigen Beispiel mit n = 100 reicht (für ein einseitiges $p < 0{,}05$) eine Korrelation von 0,16. Bei n = 50 läge das entsprechende *critical r* bei 0,24, d.h. nur „überdurchschnittliche" Korrelationen (wenn man den genannten Durchschnittswert in der Managementforschung von 0,2 zugrunde legt) mit hypothesenkonformer Effektrichtung würden überhaupt veröffentlicht und diskutiert. Mehr als

die Hälfte der Studien mit diesem Stichprobenumfang (konkret rund 60%) verschwänden in der Schublade. Somit würden die veröffentlichten Studien (mit „Minimumkorrelation" von 0,24 und möglichen Korrelationen von bis zu ca. 0,5) den tatsächlichen Zusammenhang zwischen POS und OCB beträchtlich überschätzen. Denselben Effekt hätte potenziell eine bisweilen geforderte strengere Grenze für die statistische Signifikanz (etwa $p < 0{,}005$[11]): in diesem Fall wären im ersten Szenario (mit $n = 100$ und bei einseitiger Testung) nur Korrelationen > 0,25 statistisch signifikant. Wenn die Populationskorrelation bei 0,2 liegt, bedürfte es also auch hier schon einer überdurchschnittlichen Stichprobenkorrelation für ein statistisch signifikantes Ergebnis.[12]

Diese Verzerrung der berichteten im Vergleich zu den tatsächlichen Effekten firmiert in der Literatur unter der Bezeichnung Publikationsbias oder *file drawer problem* und ist insbesondere für Metaanalysen (statistische Zusammenfassung bestehender Forschungsergebnisse zu einem Thema) ein Risikofaktor für eine Überschätzung der tatsächlichen Effekte.[13] Eine solche Diskrepanz zwischen „signifikanz-selektiv" berichteten und (allen) beobachteten Effekten ist nicht nur für die akademische Forschung ein Thema, sondern auch für Praktiker/innen, etwa im Rahmen evidenzbasierten Managements und darauf basierender Entscheidungen.

Studien mit kleinen Stichproben bei geringen Populationseffekten bedeuten somit, dass *a)* ein statistisch signifikantes Ergebnis eher unwahrscheinlich ist und *b)* ein solches Ergebnis tendenziell ein „Ausreißer" wäre, der den tatsächlichen Effekt u.U. deutlich überschätzt bzw. zu Unrecht postuliert.[14] Bei einem großen Populationseffekt (der oft theoretisch trivial und empirisch bereits solide abgesichert ist) lassen sich zwar auch mit kleinen Stichproben Ergebnisse beobachten, die zur Nullhypothese eines Nicht-Effekts in klarem Widerspruch stehen, eine derartige „Nil-Nullhypothese" ist dann aber letztlich nur ein wertloser Strohmann und der Erkenntnisgewinn aus einer scheinbaren Widerlegung minimal.

Bei einer großen Stichprobe ist die statistische Signifikanz hingegen deshalb wenig aussagekräftig, weil selbst minimale und praktisch nicht relevante Effekte statistisch signifikant werden. Allerdings erhält man mit großen Stichproben (sofern für die Population hinreichend repräsentativ) ziemlich präzise Schätzungen des tatsächlichen Effekts. In der Forschungspraxis sind wirklich große Stichproben indes oft nur schwierig zu erreichen, vor allem wenn es um aggregierte Beobachtungseinheiten geht (etwa Teams, Abteilungen oder Organisationen anstatt Einzelpersonen). Zudem ist die Wahrscheinlichkeit für „spektakuläre" Ergebnisse geringer als bei kleinen

Ausschließliche Fokussierung auf p-Wert/statistische Signifikanz ist bei kleinen Stichproben potenziell frustrierend oder irreführend, bei großen Stichproben potenziell wertlos.

Stichproben, wo die beobachteten Effekte wie oben erläutert häufiger und weiter vom (oft eher bescheidenen) Populationseffekt abweichen können. Ironischerweise begünstigt die Anreizgestaltung im Wissenschaftsbetrieb tendenziell das Veröffentlichen dieser spektakulären aber mit hoher Wahrscheinlichkeit verzerrten und somit möglicherweise irreführenden Ergebnisse im Vergleich zu solideren, aber oft „langweiligeren" Resultaten.[15] Ähnliches gilt bei Auftragsforschung für die Praxis, wo scheinbar revolutionäre und disruptive Ergebnisse bei den Entscheidungsträger/innen vermutlich auf mehr Resonanz und Interesse stoßen als in Zahlen gegossene Unauffälligkeit, auch wenn letztere den Status quo besser abbilden mag.

p-Wert – what else?

Welche Optionen hat man nun als Forscher/in oder Manager/in, Studienergebnisse korrekt und umfassend darzustellen bzw. zu interpretieren, ohne in die „Signifikanzfalle" zu tappen, sprich: sich nur auf „$p < 0,05$" zu fokussieren und daraus potenziell falsche Schlüsse bezüglich der Belastbarkeit der untersuchten Hypothese(n) und Aussagekraft der Ergebnisse zu ziehen?

Abkehr vom dichotomen (und irrigen) „p < 0,05: Beleg für die Richtigkeit der Hypothese, p > 0,05: kein Ergebnis"

Der p-Wert gibt wie gesagt an, wie (in)kompatibel die beobachteten Ergebnisse mit der Annahme eines Nicht-Effekts sind (bzw. bei einseitiger Testung wie im oben geschilderten Beispiel zu POS und OCB auch mit allen Annahmen zu negativen Zusammenhängen). Anstatt *p < 0,05, p < 0,01, n.s.* oder ähnlichen „Standardfloskeln" (mit entsprechender Sternchendekoration) ist es demnach besser und informativer, einfach den ermittelten p-Wert direkt anzugeben (auch für „nicht signifikante" Ergebnisse).[18]

Was besagt ein Konfidenzintervall?

Ähnlich wie der p-Wert werden auch Konfidenzintervalle (KI) oft falsch interpretiert.[16] Die intuitive (und bevorzugte) Deutung eines 95%-KI als *„Bereich, in dem mit 95% Wahrscheinlichkeit der ‚wahre' Populationswert (bzw. –effekt) liegt"*, ist letztlich nicht korrekt (taugt allerdings bedingt als grobe Annäherung[17]). Die tatsächliche Interpretation ist eher umständlich und wenig intuitiv: bei vielen Studien (mit der gleichen Stichprobengröße) werden 95% der ermittelten KI den Populationswert enthalten. KI und statistische Signifikanz hängen zudem direkt zusammen: bei einem statistisch signifikanten Ergebnis schließt das entsprechende KI den Wert 0 nicht ein (bzw. umgekehrt: wenn das KI den Wert 0 einschließt, ist das Ergebnis statistisch nicht signifikant).

Effektschätzung statt Hypothesen"bestätigung"

Die Ergebnisse inferenzstatistischer Auswertungen beinhalten neben dem p-Wert vor allem eine Schätzung der Effektgröße anhand der beobachteten Daten. Diese Effektschätzung ist für die Interpretation und potenzielle Bedeutsamkeit der Ergebnisse oft weitaus bedeutsamer als die reine Antwort auf *„(statistisch) signifikant ja oder nein?"*. Um beim obigen Beispiel zu bleiben: angenommen, die beobachtete Korrelation

zwischen POS und OCB beträgt 0,5 (was im Übrigen bei einseitiger Testung auf dem 5%-Niveau schon ab n = 12 statistisch signifikant wäre). Dass dieser Zusammenhang – gemessen an den üblichen Korrelationen in der Managementforschung – weit überdurchschnittlich ist und wie sich das erklären ließe, ist für die Ergebnisdiskussion vermutlich deutlich ergiebiger als die bloße Feststellung: „Das Ergebnis stützt die Hypothese eines Zusammenhangs zwischen POS und OCB".

Nicht nur den geschätzten Effekt, sondern auch das Konfidenzintervall betrachten

Neben dem beobachteten Effekt hilft auch das so genannte Konfidenzintervall bei der Einordnung der Ergebnisse. Wie zuvor beschrieben und aus den beiden obigen Abbildungen ersichtlich, bedeutet eine größere Stichprobe eine präzisere Schätzung und somit ein engeres Konfidenzintervall. Für n = 500 und r = 0,3 wäre das 95%-KI [0,22; 0,38], für n = 100 und dieselbe Korrelation [0,11; 0,47].[19] Anders als die reine Punktschätzung („die beobachtete Korrelation beträgt 0,3") gibt das Konfidenzintervall somit ein Stück weit Aufschluss über die (Un)Genauigkeit der Schätzung.

Untersuchungsergebnisse mit früheren Resultaten vergleichen und einordnen

Bei der theoretischen Herleitung und Aufarbeitung des Forschungsstands oder auch der Ableitung von evidenzbasierten Handlungsempfehlungen stellen sich Forscher/innen und Manager/innen bereitwillig „auf die Schultern von Riesen", aber die Beobachtungen im Rahmen empirischer Studien werden meist isoliert berichtet und diskutiert, ohne explizit auf die Größenordnung und Verlässlichkeit beobachteter Effekte aus früheren vergleichbaren Studien einzugehen.

Der methodisch höchstentwickelte Ansatz in diese Richtung ist die so genannte bayesianische Statistik, bei der die beobachteten Ergebnisse formell mit Vorwissen/-annahmen über den untersuchten Effekt kombiniert werden und so zu einer „aktualisierten" Schätzung führen, wie wahrscheinlich z.B. bestimmte Korrelationen zwischen POS und OCB sind. Mit diesem Vorgehen ist es anders als mit Signifikanztests möglich, Aussagen über die Wahrscheinlichkeit von Hypothesen angesichts der Daten zu treffen, und nicht nur Aussagen über die Wahrscheinlichkeit der Daten unter einer Hypothese. Vereinfacht gesagt bietet die bayesianische Statistik das, was in der üblichen („frequentistischen") Statistik mit p-Wert/statistischer Signifikanz auf die oben genannte Fehlinterpretation hinausläuft (Wahrscheinlichkeit für die Gültigkeit einer bestimmten Hypothese).

Der Haken bei der Sache? Neben der insgesamt komplizierteren Berechnung beruhen die Ergebnisse bzw. Schlussfolgerungen bei diesem Ansatz wie gesagt teilweise auf mehr oder weniger fundierten oder auch rein subjektiven Vorannahmen. Ohne hier näher auf die bisweilen nachgerade tribalistisch geführte Diskussion um den „besseren" statistischen Zugang einzugehen, lässt sich zusammenfassend festhalten, dass *a)* bei korrekter Verwendung beide Verfahren meist zu sehr ähnlichen Ergebnissen und

Schlussfolgerungen führen[20], *b)* bayesianische Analysen in den Sozial- und Wirtschaftswissenschaften vor allem verglichen mit anderen Disziplinen immer noch eher Exotenstatus haben.[21] Eine für Forschung wie Praxis wichtige Erkenntnis aus dem bayesianischen Ansatz ist indes: ein fundiertes Urteil über die Gültigkeit einer Hypothese lässt sich nie anhand einer einzelnen Studie fällen, sondern nur unter Einbeziehung früheren Wissens oder zumindest theoretisch fundierter Überlegungen, wie plausibel der beobachtete Effekt (in diesem Ausmaß) tatsächlich ist.[22]

Wie lässt sich all das in Handlungsempfehlungen gießen, wenn es um die Darstellung eigener Forschungsergebnisse geht? Die folgende Tabelle formuliert dazu abschließend einige Leitlinien in Abhängigkeit von der Größe der zugrunde liegenden Stichprobe und der beobachteten Effekte.

	Geringer bis "üblicher" beobachteter Effekt	**Überdurchschnittlich großer beobachteter Effekt**
Kleine Stichprobe	• „Dumm gelaufen": womöglich (zu) großer p-Wert • Zwecks „Ergebnisrettung" u.U. von Hypothesentestung auf explorativ Fazit umschwenken oder ggf. sogar Signifikanzniveau anpassen, falls begründbar (z.B. 10% statt der üblichen 5%) • Plausibilität des Ergebnisses betonen und Unsicherheit der Schätzung wegen kleiner Stichprobe einräumen	• Großen Effekt (und vermutlich geringen p-Wert) anmerken • Bei Diskussion abwägen zwischen möglichen substanziellen Gründen für die Beobachtung und der Möglichkeit, dass der beobachtete den tatsächlichen Effekt (deutlich?) überschätzt • Ergebnis nicht als belastbare Theoriebestätigung darstellen • Unsicherheit ggf. per Konfidenzintervall veranschaulichen
Große Stichprobe	• Effekt in Beziehung zu bestehenden Ergebnissen setzen (wo reiht sich Beobachtung ein?) • Plausibilität und Stichhaltigkeit hervorheben • Kann u.U. die Glaubwürdigkeit anderer, „spektakulärerer" Ergebnisse in derselben Studie erhöhen • Auch ein „(Fast-)Nicht-Effekt" mit entsprechend engem Konfidenzintervall ist potenziell ein inhaltlich interessantes Ergebnis	• Plausibilität der Ergebnisse (kritisch) beurteilen (mögliche Verzerrung der Ergebnisse durch Methodenartefakte oder eine spezielle Stichprobe?) • Bei Ergebnisdiskussion die potenziellen Implikationen betonen (Was ergibt sich aus der „ungewöhnlich deutlichen" Beobachtung?) • Fokus auf statistische Signifikanz hier besonders wertlos

Tabelle 1: Anregungen zur Ergebnisdarstellung je nach Stichproben- und Effektgröße

Diese Empfehlungen sind zugegeben nicht unmittelbar evidenzbasiert und subjektiv, anders als die bereits zuvor genannten Anregungen, die sich auch in einigen Richtlinien von Fachzeitschriften finden:[23] exakte p-Werte angeben, Effektgrößen und Konfidenzintervalle berichten und letztere auch bei der Ergebnisinterpretation und -diskussion betonen anstatt sich auf „statistisch signifikant oder nicht?" zu konzentrieren. Ihre Umsetzbarkeit bzw. potenzielle Nützlichkeit im Einzelnen hängt zudem auch von der persönlichen Positionierung im Spannungsfeld zwischen *methodological rigor*, *„publish or perish"* und begrenzten Ressourcen (sowie ggf. den Rückmeldungen seitens der Gutachter/innen) ab.

Was punkto Stichprobenumfang als groß oder klein gelten kann, ist dabei je nach Forschungsfrage und -design bzw. der relevanten Auswertung unterschiedlich. Bei einer Korrelationsanalyse wie im wiederholt genannten Beispiel kann eine Stichprobe von n = 500 durchaus als groß bezeichnet werden. Bei Mehrebenenmodellen für geclusterte Daten, einem mehrfaktoriellen Design mit jeweils mehreren Gruppenkategorien, einer logistischen Regression bei sehr asymmetrischer Gruppenzugehörigkeit, der Suche nach Interaktions- anstatt Haupteffekten etc. kann diese Stichprobengröße hingegen auch unzureichend sein und nur bei erwartbar großen Effekten (oder einem „Glückstreffer") ein statistisch signifikantes Ergebnis erbringen, ähnlich wie eine Korrelationsanalyse bei einem n < 50. Auch bei der Einordnung von Effektgrößen sollte der präzise Forschungskontext berücksichtigt werden, anstatt sich undifferenziert und universell auf gängige „Klein – mittel – groß"-Kategorisierungen zu berufen.[24]

Während sich diese Empfehlungen und Überlegungen zur Ergebnisdarstellung vornehmlich an Forscher/innen richten, gilt gleichermaßen für Manager/innen als abschließender Praxistipp: ein gerüttelt Maß an Skepsis gegenüber allzu enthusiastischen Schlussfolgerungen aufgrund von Einzelergebnissen, deren Belastbarkeit und Erkenntniswert primär auf dem vermeintlichen Gütesiegel der statistischen Signifikanz beruhen. Dies gilt umso mehr für Studien, die auf Basis einer (möglicherweise aus nachvollziehbaren Sachzwängen) kleinen Stichprobe mit Verweis auf „signifikante" p-Werte einen (ungewöhnlich großen?) Effekt als substanzielles Forschungsergebnis berichten und diskutieren.[25]

Literatur

1 Gigerenzer, Gerd (2004). Mindless statistics. Journal of Socio-Economics, 33(5), 587-606. DOI 10.1016/j.socec.2004.09.033

2 Wasserstein, Ronald L. & Lazar, Nicole A. (2016). The ASA statement on p-values: context, process, and purpose. The American Statistician, 70(2), 129-133. DOI 10.1080/00031305.2016.1154108

3 Simmons, Joseph P. & Simonsohn, Uri (2017). Power posing: p-curving the evidence. Psychological Science, 28(5), 687-693. DOI 10.1177/0956797616658563

[4] Yong, Ed (2012). Replication studies: Bad copy. Nature, 485(7398): 298-300. DOI 10.1038/485298a | Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349(6251), 1-8. DOI 10.1126/science.aac4716

[5] Nickerson, Raymond S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods, 5(2), 241-301. DOI 10.1037/1082-989x.5.2.241 | Greenland, Sander, Senn, Stephen J., Rothman, Kenneth J., Carlin, John B., Poole, Charles, Goodman, Steven N. & Altman, Douglas G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology, 32, 337-350. DOI 10.1007/s10654-016-0149-3

[6] Cohen, Jacob (1994). The earth is round (p < .05). American Psychologist, 49(12), 997-1003. DOI 10.1037/0003-066X.49.12.997

[7] Die Grafiken wurden mit dem Programm G*Power erstellt (Faul, Franz, Erdfelder, Edgar, Lang, Albert-Georg & Buchner, Axel (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175-191. DOI 10.3758/BF03193146).

[8] Zur Einordnung: in den Rohdaten einer Metaanalyse zu Einflussfaktoren von Karriereerfolg lag die mittlere Stichprobengröße (Median) mit ca. 300 genau zwischen diesen beiden Szenarien, wobei allerdings weniger als 10 % der Primärstudien nur ein n von 100 oder weniger aufweisen, aber rund 30 % ein n von über 500 (Heslin, Peter A., Mayrhofer, Wolfgang, Schiffinger, Michael, Eggenhofer-Rehart, Petra, Latzke, Markus, Reichel, Astrid, Steyrer, Johannes & Zellhofer, Dominik (2019). Still relevant? An updated meta-analysis of classic career success predictors. Academy of Management Annual Meeting Proceedings, 2019(1), 11541. DOI 10.5465/AMBPP.2019.11541abstract).

[9] Paterson, Ted A., Harms, P.D., Steel, Piers & Credé, Marcus (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. Journal of Leadership & Organizational Studies, 23(1), 66-81. DOI: 10.1177/1548051815614321 | Bosco, Frank A., Aguinis, Herman, Singh, Kulraj, Field, James G. & Pierce, Charles A. (2015). Correlational effect size benchmarks. Journal of Applied Psychology, 100(2), 431-449. DOI: 10.1037/a0038047

[10] Kerr, Norbert L. (1998). HARKing: Hypothesizing After the Results are Known. Personality and Social Psychology Review, 2(3), 196-217. DOI 10.1207/s15327957pspr0203_4

[11] Benjamin, Daniel J., et al. (2018). Redefine statistical significance. Nature Human Behaviour 2(1), 6-10. DOI 10.1038/s41562-017-0189-z

[12] Betensky, Rebecca A. (2019). The p-value requires context, not a threshold. The American Statistician, 73(sup1), 115-117. DOI 10.1080/00031305.2018.1529624

[13] Franco, Annie, Malhotra, Neil & Simonovits, Gabor (2014). Publication bias in the social sciences: Unlocking the file drawer. Science, 345(6203), 1502-1505. DOI 10.1126/science.1255484 | Rosenthal, Robert (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86(3), 638-641. DOI 10.1037/0033-2909.86.3.638 | Rothstein, Hannah R., Sutton, Alexander J. & Borenstein, Michael (Hrsg.) (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. New York: Wiley. DOI 10.1002/0470870168

[14] Ioannidis, John P.A. (2005). Why most published research findings are false. PLoS Med 2(8), e124. DOI 10.1371/journal.pmed.0020124

[15] Campbell, Harlan & Gustafson, Paul (2019). The world of research has gone berserk: Modeling the consequences of requiring "greater statistical stringency" for scientific publication. The American Statistician, 73(sup1), 358-373. DOI 10.1080/00031305.2018.1555101 | Higginson, Andrew D. & Munafò, Marcus R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. PLOS Biology 14(11), e2000995. DOI 10.1371/journal.pbio.2000995

[16] Hoekstra, Rink, Morey, Richard D., Rouder, Jeffrey N. & Wagenmakers, Eric-Jan (2014). Robust misinterpretation of confidence intervals. Psychonomic Bulletin & Review, 21, 1157-1164. DOI 10.3758/s13423-013-0572-3

[17] https://handbook-5-1.cochrane.org/chapter_12/12_4_1_confidence_intervals.htm

[18] Wilkinson, Leland & Task Force on Statistical Inference, American Psychological Association, Science Directorate (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594-604. DOI 10.1037/0003-066X.54.8.594 (S. 599)

[19] Unter https://www.wu.ac.at/fileadmin/wu/d/i/ivm/corr_ci.xlsx ist ein Excel-Sheet verfügbar, mit dem je nach Korrelationskoeffizient und Stichprobengröße das 99%-, 95%- und 90%-Konfidenzintervall berechnet wird. Alternativ gibt es z.B. https://www2.ccrb.cuhk.edu.hk/stat/confidence%20interval/CI%20for%20correlation.htm.

[20] Lakens, Daniel (2019). The practical alternative to the p-value is the correctly used p-value. PsyArXiv, April 9. DOI 10.31234/osf.io/shm8v

[21] van de Schoot, Rens, Winter, Sonja D., Ryan, Oisin, Zonder van-Zwijnenburg, Marielle & Depaoli, Sarah (2017). A systematic review of Bayesian articles in psychology: The last 25 years. Psychological Methods, 22(2), 217-239. DOI 10.1037/met0000100

[22] Nuzzo, Regina (2014). Scientific method: statistical errors. Nature, 506(7487), 150-152. DOI 10.1038/506150a

[23] Hahn, Eugene D. & Ang, Siah Hwee (2017). From the editors: New directions in the reporting of statistical results in the Journal of World Business. Journal of World Business, 52(2), 125-126. DOI 10.1016/j.jwb.2016.12.003 | Bettis, Richard A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. Strategic Management Journal, 33(1), 108-113. DOI 10.1002/smj.975 | Köhler, Tine, Landis, Ronald S. & Cortina, José M. (2017). From the editors: Establishing methodological rigor in quantitative management learning and education research: the role of design, statistical methods, and reporting standards. Academy of Management Learning & Education, 16(2), 173-192. DOI 10.5465/amle.2017.0079 (S. 184f.)

[24] Cortina, José M. & Landis, Ronald S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In: Lance, Charles E. & Vandenberg, Robert J. (Hrsg.): *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences*. New York, NY [u.a.]: Routledge, 287-308

[25] Tversky, Amos & Kahneman, Daniel (1971). Belief in the law of small numbers. Psychological Bulletin, 76(2), 105-110. DOI 10.1037/h0031322. Für ein konkretes Beispiel aus der Managementforschung: Peterson, Randall S., Smith, D. Brent, Martorana, Paul V. & Owens, Pamela D. (2003). The impact of chief executive officer personality on top management team dynamics: One mechanism by which leadership affects organizational performance. Journal of Applied Psychology, 88(5), 795-808. DOI 10.1037/0021-9010.88.5.795 | Hollenbeck, John R., DeRue, D. Scott & Mannor, Michael (2006). Statistical power and parameter stability when subjects are few and tests are many: Comment on Peterson, Smith, Martorana, and Owens (2003). Journal of Applied Psychology, 91(1), 1-5. DOI 10.1037/0021-9010.91.1.1 | Peterson, Randall S., Smith, D. Brent & Martorana, Paul V. (2006). Choosing between a rock and a hard place when data are scarce and the questions important: Reply to Hollenbeck, DeRue, and Mannor (2006). Journal of Applied Psychology, 91(1), 6-8. DOI 10.1037/0021-9010.91.1.6

Information zum Autor

PD Dr. Michael Schiffinger ist seit 2000 an der Wirtschaftsuniversität Wien tätig.