Nutzen und Grenzen der Inferenzstatistik in der (Wirtschafts-)Psychologie

Alla Sawatzky, Verena Stumm

Hochschule Fresenius Köln, Pädagogische Hochschule Heidelberg
Hochschule Fresenius, Köln

ZUSAMMENFASSUNG

Inferenzstatistische Verfahren, wie *p*-Werte, Signifikanztests (NHST oder nach Neyman und Pearson), Konfidenzintervalle oder Bayes-Faktoren werden zum Teil seit rund 100 Jahren flächendeckend in psychologischer Forschung verwendet. Die Funktion der Inferenzstatistik besteht dabei darin, die (Un-)Sicherheit darüber zu quantifizieren, inwiefern sich das Ergebnis aus den beobachteten Studiendaten als Aussage über eine Population verallgemeinern lässt. Da sich jedoch 1) die meisten inferenzstatistischen Angaben unmittelbar aus den vorliegenden Daten ergeben und 2) in der Regel unbekannt ist, ob die vorliegenden Daten typisch oder untypisch für die Population sind, kann diese Funktion nicht erfüllt werden. Inferenzstatistische Angaben haben daher in den meisten Fragestellungen in der (Wirtschafts-)Psychologie keinen Informationsmehrwert. Da inferenzstatistische Angaben darüber hinaus häufig fehlinterpretiert werden, sollten sie nur in Ausnahmefällen, in denen eine geeignete Zielsetzung und Datensituation vorliegt, genutzt werden.

Schlüsselbegriffe: Inferenzstatistik, Nutzen, Fehlinterpretation, Induktionsproblem

1 Einleitung

Statistical inference enables bad science; statistical thinking enables good science (Tong, 2019, S. 257, Hervorhebung im Original)

And [the students] will realize that in many applications, a skilful and transparent descriptive data analysis is sufficient, and preferable to the application of statistical routines [like the Null Hypothesis Significance Test] chosen for their complexity and opacity. (Gigerenzer, 2004, S. 604)

Confirmation comes from repetition. Any attempt to avoid this statement leads at least to failure and more probably to destruction. (Tukey, 1969, S. 84)

[…] if we accept this convenient convention [of 5% as "significant"] […], we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us. (Fisher, 1935, S. 16, Hervorhebung im Original)

Der Einsatz von inferenzstatistischen Angaben erfreut sich seit einem knappen Jahrhundert konsistent großer Beliebtheit in der psychologischen Forschung (Hubbard & Ryan, 2000). Die dabei mit Abstand am häufigsten genutzten Verfahren sind *p*-Werte (*p-values*) und Signifikanzaussagen (Hubbard & Ryan, 2000; Fritz, Scherndl & Kühberger, 2012), deutlich seltener Konfidenzintervalle (Finch et al., 2004; Fritz et al., 2012), die zu den sogenannten frequentistischen Wahrscheinlichkeitsansätzen gehören, und besonders in jüngerer Zeit Bayesianische Statistiken, die einen epistemischen Wahrscheinlichkeitsbegriff nutzen (van de Schoot et al., 2017). Ebenso alt und überdauernd sind jedoch auch die Einwände gegen den Gebrauch der Inferenzstatistik in der Wissenschaft. Diese Einwände richten sich oft gegen spezifische inferenzstatistische Verfahren, meist das sogenannte Null Hypothesis Significance Testing (NHST, z. B. Bakan, 1966; Meehl, 1978; Cohen, 1994; Gigerenzer, 2004; Lambdin, 2012; Wasserstein, Schirm & Lazar, 2019), aber auch gegen den Gebrauch von Konfidenzintervallen (Hoekstra et al., 2014) oder Bayesianischen Statistiken (Fisher, 1934; Gigerenzer, 2004). Einzelne kritische Stimmen richten sich auch gegen den Einsatz

der Inferenzstatistik grundsätzlich (Boring, 1919; Tong, 2019). In dem vorliegenden Beitrag soll eine solche letztgenannte Position skizziert, begründet und gestärkt werden.

Inferenzstatistik wird definiert als „Teil der Statistik mit dem Ziel, aus Stichprobenkennwerten auf die [entsprechenden] Populationswerte oder -verhältnisse zu schließen" (Inferenzstatistik, 2020) bzw. „über die erhobenen Daten hinaus allgemeinere Schlussfolgerungen für umfassendere Grundgesamtheiten zu ziehen" (Fahrmeir, Heumann, Künstler, Pigeot & Tutz, 2016, S. 12) oder solche Statistik, die „bestimmt […], mit welcher Sicherheit sich Ergebnisse, die an der untersuchten Stichprobe gewonnen wurden, auf die Grundgesamtheit verallgemeinern lassen" (Eid, Gollwitzer & Schmitt, 2017, S. 986). Damit wird die Inferenzstatistik von der deskriptiven Statistik abgegrenzt. Deskriptive (manchmal auch explorativ genannte, z. B. bei Fahrmeir et al., 2016) statistische Verfahren beschreiben Beziehungsmuster in tatsächlich vorliegenden Daten, etwa durch Mittelwertdifferenzen oder Korrelationen. Inferenzstatistische Angaben sollen demgegenüber die Unsicherheit quantifizieren, die bei einem Schluss von den Stichprobendaten auf die Populationverbunden ist: Mit welcher (Un-)Sicherheit kann man davon ausgehen, dass die in den Daten gefundene Mittelwertdifferenz oder Korrelation für die gesamte Population gilt? Im Folgenden soll der Unterschied zwischen einer deskriptiven Aussage und den fünf gängigsten inferenzstatistischen Aussagen umrissen werden. Für jedes inferenzstatistische Verfahren wird dabei die aus Sicht der Autorinnen zentrale Einschränkung der Aussagekraft – das Unwissen über die Typizität des Stichprobenkennwerts – aufgezeigt.

2 Deskriptive Analyse

Nehmen wir an, dass in einem experimentellen Zwei-Gruppen-Design die Wirkung einer Werbemaßnahme auf die Kaufbereitschaft untersucht werden soll. Die inhaltliche Hypothese lautet: „Die Werbemaßnahme führt zu einer höheren Kaufbereitschaft". Jede der beiden Gruppen (EG: Experimentalgruppe, KG: Kontrollgruppe) umfasst $n_{EG} = n_{KG} = 50$ Personen. Die deskriptive Datenauswertung ergibt eine Mittelwertdifferenz von $e = 1$ Skalenpunkt der Kaufbereitschaft auf einer 7-stufigen Ratingskala zugunsten der Werbemaßnahme (mit $M_{EG} = 4$ und $M_{KG} = 3$, $e = 4 - 3 = 1$, sogenannte *unstandardisierte Effektgröße*[1]). Dieses Ergebnis kann zweifelsohne als Bestätigung der inhaltlichen Hypothese interpretiert werden: Wie erwartet geben die Personen in der EG im Schnitt eine höhere Kaufbereitschaft an.

Selbstverständlich können (und sollten) weitere deskriptive und grafische Analysen angewendet werden, um die Wirkung der Werbemaßnahme auf die untersuchten Personen eingehender zu beleuchten. So kann analysiert werden, wie sich der Unterschied zwischen den Gruppen zu den Unterschieden innerhalb der Gruppen verhält. Die Betrachtung dieses Verhältnisses und zahlreicher anderer Beurteilungsmöglichkeiten der Größe des Unterschieds zwischen den Gruppen hat Folgen für eine differenziertere Einschätzung des Untersuchungsergebnisses. Haben wir in den beiden Gruppen etwa eine Streuung von einem Skalenpunkt festgestellt ($SD_{EG} = SD_{KG} = 1$)[2], so interpretieren wir den Unterschied in der Kaufbereitschaft und damit die Stärke der Werbewirkung womöglich als sehr viel deutlicher ($d = \frac{e}{SD_{gepoolt}} = \frac{1}{1} = 1$, sogenannte *standardisierte Effektgröße*) im Vergleich zu einem Szenario, in dem die Streuung in den Gruppen die Mittelwertdifferenz deutlich übertrifft (z. B. $SD_{EG} = SD_{KG} = 2$, $d = \frac{1}{2} = 0.5$). Zusätzlich können Medianvergleiche und die grafische Analyse der Datenverteilungen vorgenommen werden, um z. B. einflussreiche Einzelwerte oder Ausreißer zu identifizieren.

Die Frage danach, ob die Daten einen Beleg für die inhaltliche Hypothese darstellen ist jedoch im Prinzip durch den deskriptiven Vergleich der beiden Mittelwerte zumindest grob beantwortet: Wie erwartet hat die Werbemaßnahme zu einer höheren Kaufbereitschaft in unserer Stichprobe geführt.

[1] Zwar stammt der Begriff „Effektgröße" aus dem Hypothesentestansatz von Neyman und Pearson und hat dort eine spezifische Bedeutung, wie im weiteren Verlauf dargestellt wird, jedoch hat sich die Bezeichnung „Effekt" oder „Effektgröße" mittlerweile auch als Synonym für „Stärke der Beziehung" etabliert (siehe z. B. Cohen, 1988).

[2] Der Einfachheit halber wird auf Szenarien verzichtet, in denen sich die Streuungen in den beiden Gruppen deutlich unterscheiden (Varianzheterogenität oder Heteroskedastizität).

3 Typizität des Stichprobenergebnisses und das Induktionsproblem

Es stellt sich nun allerdings folgende Frage: Inwiefern können wir davon ausgehen, dass das Ergebnis aus unserer Untersuchung typisch ist? Können wir davon ausgehen, dass wir eine höhere Kaufbereitschaft nach der Werbemaßnahme auch in weiteren Stichproben finden? Mit anderen Worten, können wir davon ausgehen, dass die Werbemaßnahme nicht nur bei unseren 100 Versuchsteilnehmenden eine Wirkung zeigt, sondern bei allen Personen unserer Zielgruppe, d. h. in der Population?

Es lassen sich hierzu zwei Szenarien denken. Stellen wir uns vor, dass die Werbemaßnahme tatsächlich eine Wirkung auf die Kaufbereitschaft hat, dann sollten wir in unendlich vielen Untersuchungen häufiger einen Unterschied in der Kaufbereitschaft zugunsten der Experimentalgruppe finden. Keinen Unterschied in der Kaufbereitschaft oder sogar einen Unterschied zuungunsten der Experimentalgruppe sollten wir weniger häufig beobachten, auch wenn das zufallsbedingt vorkommen kann. Das typische Ergebnis wäre in einem solchen Szenario ein Unterschied in der Kaufbereitschaft zugunsten der Experimentalgruppe. Ein Ergebnis zuungunsten der Experimentalgruppe wäre zwar durchaus möglich, aber untypisch. Stellen wir uns für das zweite Szenario vor, dass die Werbemaßnahme tatsächlich keine Wirkung auf die Kaufbereitschaft hat. In diesem Fall sollten wir in unendlich vielen Untersuchungen häufiger keinen Unterschied in der Kaufbereitschaft zwischen der Experimental- und Kontrollgruppe finden. Einen Unterschied zwischen der Experimental- und Kontrollgruppe würden wir zufallsbedingt zwar ebenfalls erwarten zu beobachten, jedoch vergleichsweise selten. In diesem Szenario wäre das typische Ergebnis eins, das keinen Unterschied zwischen den beiden Versuchsgruppen anzeigt. Ein Ergebnis, das einen Unterschied zwischen den Versuchsgruppen anzeigt, wäre untypisch wenn auch möglich.

Nun stellt sich für das Ergebnis aus unserer Einzeluntersuchung die Frage, ob die beobachtete Mittelwertdifferenz zugunsten der Experimentalgruppe eine typische oder untypische ist. Handelt es sich um ein typisches Ergebnis, dann hätte die Werbemaßnahme wohl tatsächlich eine steigernde Wirkung auf die Kaufbereitschaft. Ist das Ergebnis jedoch untypisch, dann hätte die Werbemaßnahme vermutlich keine steigernde Wirkung auf die Kaufbereitschaft. Die entscheidende Frage, um die Hypothese „Die Werbemaßnahme führt zu einer höheren Kaufbereitschaft" zu beurteilen ist also die nach der Typizität des Untersuchungsergebnisses.

Eine Antwort auf diese Frage ist hoch relevant für zwei unterschiedliche Ziele, die wir mit der Untersuchung verfolgen könnten. Ein Ziel könnte darin bestehen, die inhaltliche Hypothese hinsichtlich ihres Zutreffens zu beurteilen: Führt die Werbemaßnahme zu einer höheren Kaufbereitschaft in unserer Zielgruppe? Ein anderes Ziel könnte darin bestehen, darüber zu entscheiden, ob es sich lohnt, die Werbemaßnahme zur Steigerung der Kaufbereitschaft großflächig einzusetzen. Für beide Ziele ist es höchst erstrebenswert zu wissen, wie repräsentativ oder typisch das Ergebnis aus unserer Untersuchung für die Population ist.

Die Frage nach der Repräsentativität oder Typizität einer oder mehrerer Beobachtungen (Untersuchungen) für die Grundgesamtheit aller möglichen Beobachtungen ist bekannt unter dem Schlagwort des Induktionsproblems (*problem of induction*) oder des *Principle of Uniformity of Nature* (Henderson, 2020; Hume, 1748, 1739-40). Die Schwierigkeit, die sich daraus ergibt, wenn wir anhand von Beobachtetem auf (noch) nicht Beobachtetes schließen wollen, wird besonders gut in der folgenden Veranschaulichung von Bertrand Russell deutlich:

We have a firm belief that [the sun] will rise [tomorrow], because it has risen in the past [due to the laws of motion]. The only reason for believing that the laws of motion will remain in operation [tomorrow] is that they have operated hitherto, so far as our knowledge of the past enables us to judge. […] Domestic animals expect food when they see the person who feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken. […] The mere fact that something has happened a certain number of times causes animals and men to expect that it will happen again. Thus our

instincts certainly cause us to believe the sun will rise to-morrow, but we may be in no better a position than the chicken which unexpectedly has its neck wrung. (Russell, 1912, S. 94-98, Hervorhebungen im Original)

Inwiefern bereits erfolgte Beobachtungen also typisch oder repräsentativ sind für die gesamte Population aller möglichen Beobachtungen, kann – zumindest nach heutigem Wissensstand – grundsätzlich nicht beurteilt werden: Wir haben bisher in 100 % der Beobachtungen festgestellt, dass die Sonne aufgegangen ist. Dies ist jedoch keine Garantie dafür, dass die Sonne auch morgen wieder aufgehen wird. Genauso wenig ist eine 100 % Quote von Fütterungen in der Vergangenheit eine Garantie für das Huhn, auch morgen wieder gefüttert zu werden. Dasselbe gilt natürlich auch für Untersuchungen zur Werbewirkung. Auch wenn wir sehr viele Untersuchungen durchführen, die allesamt ergeben, dass die Werbemaßnahme die Kaufbereitschaft um einen Skalenpunkt steigert, könnte die nächste Untersuchung prinzipiell ein ganz anderes Ergebnis zeigen. Diesen Umstand berücksichtigt Popper (1935, 2002) in seiner Empfehlung, eine Hypothese[3] hinsichtlich ihrer Bewährung *bisher* zu beurteilen. Wir müssen uns damit zufriedengeben zu sagen: „Bisher sieht es nicht so aus, dass die Hypothese definitiv falsch ist" oder bestenfalls: „Bisher hat noch keine Untersuchung gegen die Hypothese gesprochen". Im besten Fall können wir laut Popper (1935, 2002) also wissen, dass sich eine Hypothese im Sinne einer vorläufigen Annahme bisher gut bewährt hat. Auch wenn wir sehr viele Untersuchungen zu der Wirkung der Werbemaßnahme durchgeführt haben, die alle eine im Schnitt höhere Kaufbereitschaft für die Experimentalgruppe zeigen, können wir lediglich sagen, dass es *vorläufig* danach aussieht, dass die Werbung zu einer höheren Kaufbereitschaft führt. Eine einzelne Untersuchung, die höhere Kaufbereitschaft für die Werbemaßnahme ergibt, ist damit nur ein erster Schritt bei der Beurteilung des Bewährungsgrads der Hypothese bzw. der Beurteilung der Typizität des Ergebnisses.

4 Inferenzstatistische Analyse

Was kann die Inferenzstatistik nun zu dem Prozess der Beurteilung der Hypothese beitragen? Mit anderen Worten: Welchen Nutzen hat die Inferenzstatistik für die Beurteilung der Typizität eines Stichprobenergebnisses? Inferenzstatistische Verfahren lassen sich grob nach zwei Wahrscheinlichkeitsschulen unterteilen, der frequentistischen (oder klassischen) und der Bayesianischen (Romeijn, 2017).

Innerhalb der frequentistischen Schule wird Wahrscheinlichkeit als Auftretenshäufigkeit auf lange Sicht aufgefasst. Damit geben frequentistische Ansätze an, wie wahrscheinlich, d. h. typisch ein bestimmtes Stichprobenergebnis wäre im Sinne einer zu erwartenden Auftretenshäufigkeit *in einer bestimmten hypothetischen Population*. Innerhalb der Bayesianischen Schule wird Wahrscheinlichkeit als Ausmaß der Zuversicht oder des Glaubens verstanden. Bayesianische Ansätze können damit angeben, wie groß die Wahrscheinlichkeit, d. h. unsere Zuversicht in eine Hypothese sein sollte *auf der Basis von bisherigen Beobachtungen* und des vergangenen Glaubens an die Hypothese.

Bereits an dieser Stelle wird deutlich, dass weder frequentistische noch Bayesianische inferenzstatistische Ansätze eine Lösung für das Induktionsproblem darstellen. Sie geben zwar an, wie wahrscheinlich das Ergebnis in verschiedenen hypothetischen Populationen wäre bzw. für wie wahrscheinlich wir die Hypothese halten sollten angesichts bisheriger Informationen. Sie sind jedoch nicht in der Lage, eine Auskunft darüber zu geben, wie typisch das Stichprobenergebnis per se ist.

Welche Aussagen die fünf gängigsten inferenzstatistischen Verfahren – p-Werte, der NHST, Neyman-Pearson Entscheidungstheorie, Konfidenzintervalle und der Bayes-Faktor – genau treffen und weshalb sie keine Aussage über die Typizität eines empirischen Ergebnisses erlauben, wird im Folgenden kurz erläutert. Für einen besseren Lesefluss verzichten wir auf die Darstellung mathematischer Details im Text, Erläuterungen zu den Verfahren befinden sich aber im Anhang des Artikels.

[3] Popper setzt den Bewährungsgrad der Hypothese bzw. Theorie sehr vereinfacht gesagt mit dem Ausmaß der Typizität von Beobachtungen gleich (Popper, 1935, 2002).

4.1 p-Wert

Zunächst soll das Verfahren und die Bedeutung einer frequentistischen Wahrscheinlichkeit, des p-Werts (im Datenanalyseprogramm IBM SPSS als „Sig." oder „Signifikanz" bezeichnet), umrissen werden. Der p-Wert gibt an, wie wahrscheinlich es wäre, den in einer Untersuchung beobachteten Stichprobenkennwert (oder solche, die noch kleiner oder größer sind) in einer bestimmten hypothetischen Population zu erhalten.

Für unser Werbewirkung-Beispiel würden wir z. B. einen p-Wert von .007 (0.7 %) für eine hypothetische Population erhalten, in der die Werbemaßnahme keine Wirkung hat (für Details siehe Anhang A). Die hypothetische Population, in der es keine Wirkung der Werbung gibt, machen wir an der Zahl 0 fest: Das typische Ergebnis, d. h. die typische Mittelwertdifferenz ist in dieser Population ungefähr null (*Erwartungswert*). Das Ergebnis, das wir beobachtet haben, ist nun nicht null, sondern eins. Wie (un-)typisch ist dieses Ergebnis für eine Population, in der null das typische Ergebnis, d. h. der Erwartungswert ist? Dies wird in der Wahrscheinlichkeit $p = .007$ ausgedrückt: „Wenn wir auf lange Sicht null Skalenpunkte Unterschied in der Kaufbereitschaft zwischen der Experimental- und der Kontrollgruppe erwarten, dann würden wir in nur 0.7 % von unendlich vielen Untersuchungen einen Unterschied von einem Skalenpunkt oder mehr feststellen" (für Details siehe Anhang A). Bei unserem Ergebnis von einem Skalenpunkt Unterschied handelt es sich also in einer Population mit dem tatsächlichen – wahren oder durchschnittlichen – Unterschied von null Skalenpunkten um ein untypisches Ergebnis.

Ein kleiner p-Wert wird gelegentlich als Beleg dafür interpretiert, dass der wahre Unterschied der Kaufbereitschaft in der Population nicht null beträgt: Da die Wahrscheinlichkeit sehr klein ist, einen Unterschied von einem Skalenpunkt zu erhalten, wenn der wahre Unterschied null beträgt, halten wir es für plausibel, dass der tatsächliche Unterschied nicht null ist. Diese Schlussfolgerung ist jedoch nicht zulässig. Der p-Wert sagt uns lediglich etwas darüber, wie typisch unser Stichprobenergebnis *in einer hypothetischen Population* ist, jedoch nicht, wie typisch das Stichprobenergebnis selbst ist. Hat unsere Untersuchung eine typische Mittelwertdifferenz ergeben, d. h., könnten wir damit rechnen, dass wir auch in Zukunft Unterschiede rund um einen Skalenpunkt erhalten, dann könnten wir tatsächlich zuversichtlich(er) sein, dass der wahre Unterschied nicht null Skalenpunkte beträgt. Würden also Replikationsversuche immer wieder eine höhere Kaufbereitschaft für Personen zeigen, die der Werbemaßnahme unterzogen wurden, wäre es plausibel, davon auszugehen, dass die Werbemaßnahme die Kaufbereitschaft nicht nur bei den untersuchten Personen steigert, sondern generell eine Wirkung zeigt. Haben wir in unserer Untersuchung jedoch eine untypische Mittelwertdifferenz erhalten, d. h., ist genau dieser seltene Fall von 0.7 % eingetreten und wir könnten damit rechnen, dass wir zukünftig wesentlich geringere Mittelwertdifferenzen rund um null Skalenpunkte beobachten würden, wäre es wohl nicht zu rechtfertigen von einer Werbewirkung der Maßnahme auszugehen.

Diese entscheidende Frage nach der Typizität des beobachteten Stichprobenergebnisses kann der p-Wert jedoch nicht beantworten. Die einzige Möglichkeit, um die Typizität eines empirischen Ergebnisses zu beurteilen, besteht in Wiederholungen der Untersuchung. Stellen wir uns vor, dass wir 20 Untersuchungen zur Wirkung der Werbemaßnahme durchführen. Wir berechnen nun für jedes der 20 Experimente den p-Wert und erhalten für alle 20 Studien kleine Zahlen, z. B. p = .001 bis p = .06 mit einem durchschnittlichen p-Wert von .009 oder 0.9 %. In allen 20 Studien haben wir also Unterschiede in der Kaufbereitschaft zugunsten der Werbemaßnahme beobachtet – 20 Unterschiede, von denen jeder ziemlich untypisch ist für eine Population mit einem wahren Unterschied von null. Das typische Stichprobenergebnis scheint also bisher eins zu sein, das zugunsten der Werbemaßnahme ausfällt. Dies würden wir vermutlich als Beleg für die Werbewirkungshypothese betrachten. Hätten wir in den 20 Studien im Schnitt jedoch einen p-Wert von z. B. 27 % ermittelt, würden wir wahrscheinlich zögern, die Maßnahme als generell wirksam anzusehen. Offenbar haben wir in den 20 Studien Unterschiede in der Kaufbereitschaft beobachtet, die ziemlich typisch sind für eine Population mit einem wahren Unterschied von null. Das typische Stichprobenergebnis scheint also bisher eins zu sein, das nicht (konsistent) zugunsten der Werbemaßnahme ausfällt.

Zum selben Schluss kämen wir nun aber auch, wenn wir statt der p-Werte direkt die ermittelten Mittelwertdifferenzen aus den 20 Studien heranziehen würden. Mehr noch, die Inspektion der Mittelwertdifferenzen würde uns eine wesentlich differenziertere Aussage über die Wirkung der Werbemaßnahme ermöglichen: Wie groß waren die unstandardisierten Mittelwertdifferenzen im Schnitt, wie einheitlich? Wurden z. B. größere oder kleinere Unterschiede in der Kaufbereitschaft für verschiedene Konsumentengruppen beobachtet? Inwiefern wich die Verteilung der standardisierten Mittelwertdifferenzen von der Verteilung der unstandardisierten Mittelwertdifferenzen ab? Wurden z. B. systematisch größere Streuungen für verschiedene Konsumentengruppen beobachtet? Hing die Größe der Mittelwertdifferenzen mit der jeweils genutzten Stichprobengröße oder anderen Merkmalen der Untersuchungen zusammen? Die Antworten auf solche Fragen erlauben nuancierte Aussagen über die Wirkung der Maßnahme in unterschiedlichen Stichproben.

Insgesamt muss festgehalten werden, dass ein p-Wert aus einer einzelnen Untersuchung lediglich angibt, wie wahrscheinlich, d. h. typisch ein Stichprobenergebnis ist in einer *hypothetischen* Population. Der p-Wert trifft jedoch keine Aussage über die Typizität eines Stichprobenergebnisses, also darüber, wie wahrscheinlich oder typisch das Stichprobenergebnis ist in der *tatsächlichen* Population. Damit kann die durch den p-Wert angegebene Wahrscheinlichkeit aus einer einzelnen Untersuchung auch nicht als (Un-)Sicherheit darüber interpretiert werden, inwiefern die Population einen Erwartungswert von 0 hat (z. B. „Die Wahrscheinlichkeit ist mit 0.7 % sehr klein, dass es keine Werbewirkung auf die Kaufabsicht gibt") oder inwiefern das Stichprobenergebnis reliabel ist (z. B. „der Vorteil der Werbewirkung in der Untersuchung ist mit 0.7 % überzufällig"). Für eine Einzeluntersuchung hat die Angabe des p-Werts daher keinen Erkenntnismehrwert. Hat man mehrere Untersuchungen (zur gleichen inhaltlichen Hypothese) vorliegen, erscheint die Analyse der p-Werte ebenfalls überflüssig[4].

4.2 NHST

Das mit Abstand am weitesten verbreitete frequentistische Verfahren, der Nullhypothesensignifikanztest (NHST) zieht als Grundlage den p-Wert heran. Hierbei wird der p-Wert in der Regel mit den Cut-Off-Werten von 5 %, 1 % und 0.1 % (sogenannte *Signifikanzniveaus*) verglichen. Liegt der p-Wert unter diesen Cut-Off-Werten, so wird das Ergebnis des NHST als signifikant bezeichnet, manchmal auch weiter unterteilt in *signifikant* ($p < .05$, markiert mit einem Stern *), *sehr signifikant* ($p < .01$, markiert mit zwei Sternen **) und *höchst signifikant* ($p < .001$, markiert mit drei Sternen ***). Der Begriff *Signifikanz* innerhalb des NHST bezeichnet also den Umstand, dass der p-Wert aus einer Untersuchung kleiner ist als 5 %, 1 % oder 0.1 %. Als Konsequenz einer Signifikanz, d. h. eines signifikanten Ergebnisses folgt die Ablehnung der sogenannten Nullhypothese, die einen Erwartungswert von 0 postuliert. In der Regel wird die Ablehnung der Nullhypothese dabei der Annahme der inhaltlichen Hypothese gleichgesetzt. Für das Werbewirkungsbeispiel erhalten wir z. B. ein sehr signifikantes Ergebnis, da der p-Wert von 0.7 % kleiner ist als 1 % und größer als 0.1 %. In der Folge lehnen wir die Nullhypothese, dass die Werbung tatsächlich keinen Effekt auf die Kaufabsicht hat ab. Damit nehmen wir die Forschungshypothese, dass es tatsächlich einen Effekt der Werbung gibt an.

Zuweilen wird hierbei das Signifikanzniveau als Irrtumswahrscheinlichkeit interpretiert. Auch wenn das Signifikanzniveau in der Forschungspraxis auf viele verschiedene (nicht korrekte) Weisen gedeutet wird (siehe z. B. Badenes-Ribera, 2016; Haller und Kraus; 2002, Lyu Peng & Hu, 2018; Oakes, 1986), soll nur eine der Fehlinterpretationen exemplarisch herausgegriffen werden. Diese lautet: „Wenn ich mich dafür entscheide, dass die Werbemaßnahme tatsächlich wirkt, irre ich mich, d. h., entscheide ich mich fälschlicherweise gegen die Nullhypothese mit einer Wahrscheinlichkeit von 1 % (Signifikanzniveau)". Eine solche Interpretation ist jedoch nur dann plausibel (wenn auch logisch oder mathematisch nicht haltbar), wenn man davon ausgehen kann, dass das beobachtete Stichprobenergebnis ein typisches ist. Beobachtet man sehr viele Stichprobenergebnisse, die alle einen p-Wert kleiner als 1 % aufweisen, d. h. sehr signifikant sind, besteht zwar die Möglichkeit, dass diese Ergebnisse auch bei Gültigkeit der Nullhypothese (Werbung hat keinen Effekt)

[4] Hier ist die Analyse der p-Werte zum Zweck der Beantwortung von inhaltlichen Hypothesen gemeint. Selbstverständlich ist es von großem Interesse und hoher Relevanz, p-Werte und deren Eigenschaften innerhalb von Forschungsbemühungen zur Statistik zu analysieren.

auftreten[5]. Jedoch wäre es in diesem Fall naheliegend, anzunehmen, dass die Werbemaßnahme tatsächlich wirkt und dabei das Risiko dafür, dass diese Annahme falsch ist, als relativ gering anzusehen. Handelt es sich bei dem signifikanten Ergebnis jedoch um eine Ausnahme und würde man in Replikationsversuchen ausnahmslos nicht-signifikante Ergebnisse erhalten, so ist die Wahrscheinlichkeit, dass man sich mit der Annahme der Werbewirkung irrt, natürlich sehr viel höher als 1 %. Abgesehen von den zahlreichen sonstigen Problemen des NHST (z. B. Cohen, 1994; Gigerenzer, 2004, 2018; Sonderausgabe ASA, 2019; Wasserstein & Lazar, 2016) besteht ein zentrales und grundlegendes Problem daher darin, dass das Verfahren keine Aussage darüber liefern kann, inwiefern mit einem signifikanten Ergebnis auch in Zukunft oder in anderen Untersuchungen gerechnet werden kann.

Zusammengefasst gibt der NHST (die Signifikanz) an, ob das Stichprobenergebnis aus der Untersuchung zu den 5 %, 1 % oder 0.1 % untypischsten bzw. unwahrscheinlichsten gehört – in einer *hypothetischen* Population mit dem Erwartungswert null. Der NHST/die Signifikanz gibt jedoch nicht an, ob das Stichprobenergebnis zu den 5 %, 1 % oder 0.1 % unwahrscheinlichsten bzw. untypischsten in der *tatsächlichen* Population gehört. Genau wie der p-Wert trifft der NHST bzw. die Signifikanz aus einer Einzeluntersuchung also keine Aussage über die Typizität des Stichprobenergebnisses. Nur Replikationen können weitere Indizien über die Typizität des Stichprobenergebnisses liefern und dann haben Signifikanzaussagen gegenüber deskriptiven Analysen, ebenso wie p-Werte keinen Erkenntnismehrwert.

4.3 Hypothesentest nach Neyman und Pearson

Während der NHST *eine* konkrete hypothetische Population heranzieht, nutzt der Hypothesentest von Neyman und Pearson (mindestens) *zwei* konkrete hypothetische Populationen. Im Rahmen des Ansatzes von Neyman und Pearson wird damit eine Wahrscheinlichkeitsaussage über das Stichprobenergebnis für zwei Erwartungswerte getroffen. Ähnlich wie im NHST werden auch im Ansatz von Neyman und Pearson Cut-Off-Werte für den ermittelten p-Wert genutzt. Anders als im NHST werden die Cut-Off-Werte jedoch anhand von Kosten-Nutzen-Überlegungen für die spezifische inhaltliche Fragestellung bestimmt und als Fehler oder Risiken bezeichnet.

Auf die Untersuchung zur Werbewirkung übertragen, könnten wir zusätzlich zu der Nullhypothese, laut der wir als typisches Ergebnis keinen Unterschied in der Kaufbereitschaft erwarten, eine zweite Hypothese formulieren. Diese Alternativhypothese soll eine hypothetische Population beschreiben, in der es im Schnitt einen Skalenpunkt Unterschied zugunsten der Werbemaßnahme gibt. Die Wahrscheinlichkeit für den Unterschied in der Kaufbereitschaft von einem Skalenpunkt (oder mehr) haben wir für die hypothetische Population mit dem Erwartungswert null im Abschnitt weiter oben mit p = .007 bestimmt. Nun bestimmen wir, wie typisch bzw. wahrscheinlich das Stichprobenergebnis wäre in einer hypothetischen Population mit dem Erwartungswert von einem Skalenpunkt mit p = .50. Diese Wahrscheinlichkeiten können wir nun mit zuvor festgelegten Cut-Off-Werten vergleichen. Hätten wir zuvor z. B. einen Cut-Off-Wert von 5 % (Risiko I, α) für die erste Wahrscheinlichkeit und einen Cut-Off-Wert von 20 % (Risiko II, β) für die zweite Wahrscheinlichkeit festgelegt, könnten wir uns anhand dieser Werte für eine der beiden Hypothesen entscheiden. Unser Stichprobenergebnis gehört zu den untypischsten 5 % in der hypothetischen Population mit dem Erwartungswert von null Skalenpunkten Unterschied. Gleichzeitig kommt das Stichprobenergebnis in einer hypothetischen Population mit einem Skalenpunkt Unterschied auf lange Sicht häufiger als in 20 % der Fälle vor und ist daher typischer als per Cut-Off-Wert festgelegt. Wir entscheiden uns daher für die Alternativhypothese.

Grundsätzlich unterscheidet sich der Ansatz von Neyman und Pearson damit nicht von dem NHST, was die Angabe der Typizität für das empirische Ergebnis betrifft. Im Hypothesentest von Neyman und Pearson wird nur angegeben, ob das Stichprobenergebnis zu den untypischsten α % in einer spezifischen *hypothetischen* Population bzw. zu den untypischsten β % in einer anderen spezifischen *hypothetischen* Population gehört. Ob das Ergebnis aus

[5] Sogar wenn 100 von 100 Studien, d. h. 100 von 100 per Zufall erhaltenen Stichprobenkennwerten (z. B. Mittelwertdifferenzen) einen p-Wert unter 1 % ergeben würden, könnte es dennoch sein, dass die nächsten 9 900 Studien Stichprobenkennwerte mit p-Werten deutlich über 1 % ergeben (bekannt als *representativeness misconception*, Kahneman, Slovic & Tversky, 1982; Garfield, 2003).

der Studie zu den α % oder β % der unwahrscheinlichsten bzw. untypischsten in der *tatsächlichen* Population gehört, kann der Ansatz von Neyman und Pearson nicht angeben.

Genau wie die anderen bisher beschriebenen inferenzstatistischen Verfahren trifft der Ansatz also keine Aussage über die Typizität des empirischen Ergebnisses.

Ein sinnvoller Anwendungsbereich des Neyman und Pearson Ansatzes liegt aber in Situationen vor, in denen konkrete Verhaltensentscheidungen wiederholt routinemäßig getroffen werden müssen, in denen die Typizität der Stichprobenergebnisse bereits bekannt ist oder sinnvoll begründet werden kann und in denen die Konsequenzen von falschem Verhalten quantifiziert werden können (für Details siehe Anhang B). In solchen Situationen ist der Nutzen des Neyman-Pearson-Verfahrens deutlich größer als der des NHST. Die allermeisten Fragestellungen in der (Wirtschafts-)Psychologie weisen die genannten Merkmale jedoch nicht auf.

Für unser Werbewirkungsbeispiel haben wir keine Verhaltensentscheidung zu treffen: Wir wollen *Erkenntnis gewinnen* über die Wirkung unserer Werbemaßnahme. Eine zur Anwendung des Neyman-Pearson-Verfahrens passende Fragestellung wäre dagegen etwa: „Sollte die Werbemaßnahme großflächig eingesetzt werden oder nicht?". Hier wäre nicht Erkenntnis von Interesse, sondern ausschließlich *wie wir uns verhalten sollten*.

Während manche wirtschaftspsychologische Fragestellungen Verhaltensentscheidungen enthalten können (z. B. „Soll das eignungsdiagnostische Verfahren A oder B zur Personalauswahl genutzt werden?", „Sollte die Personalmarketingmaßnahme C eingesetzt werden?" etc.), so sind solche Entscheidungen jedoch in der Regel einmalig oder zumindest selten zu treffen. Habe ich die Frage danach, ob eine bestimmte Werbemaßnahme großflächig eingesetzt werden sollte beantwortet, dann werde ich dieselbe Frage in nächster Zukunft vermutlich nicht erneut beantworten müssen/wollen. Im Rahmen der Qualitätskontrolle bei Produktionsprozessen wird dieselbe Verhaltensentscheidung dagegen routinemäßig getroffen. Hier stellt sich z. B. die Frage „Sollte der Produktionsprozess angehalten werden (weil Teile systematisch falsch produziert werden) oder nicht?" wiederholt in derselben Form im Verlauf des gesamten Produktionsprozesses.

Bei Produktionsprozessen ist darüber hinaus die Typizität des Stichprobenergebnisses für beide Hypothesen bekannt: Der Außendurchmesser von korrekt produzierten Metallschrauben muss z. B. 10 mm betragen. Beträgt der Außendurchmesser 9.9 mm oder 10.1 mm, d. h. ist der Außendurchmesser um 0.1 mm kleiner oder größer, dann kann die Schraube nicht mehr verwendet werden. Auch können Konsequenzen von falschem Verhalten in der Produktion recht klar quantifiziert werden. Werden Schrauben mit dem korrekten Durchmesser produziert und die Produzentin hält die Maschinen an, so ergeben sich für sie Opportunitätskosten. Über einen Produktionszyklus hinweg, in dem die Maschinen verschieden häufig angehalten wurden, ergeben sich damit unterschiedlich hohe Opportunitätskosten. Werden Schrauben systematisch fehlerhaft produziert, etwa häufiger als im Vertrag mit der Abnehmerin vereinbart und die Produzentin hält die Produktion nicht an, so können Reklamationskosten entstehen. Auch diese können unterschiedlich hoch ausfallen, je nachdem wie häufig die Maschinen nicht angehalten wurden. Sowohl die Opportunitätskosten als auch die Reklamationskosten können prinzipiell gut quantifiziert werden, z. B. über monetäre Beträge, und als Risiken I und II ausgedrückt werden. In den meisten psychologischen Fragestellungen haben wir jedoch keine Kenntnis darüber, was das typische Stichprobenergebnis ist – hier geht es uns darum, erst in Erfahrung zu bringen, was das typische Stichprobenergebnis ist. In Bezug auf die Werbemaßnahme interessiert uns gerade, wie sehr die Werbung die Kaufbereitschaft steigert. Auch können wir die Kosten bestenfalls nur sehr ungenau schätzen, die damit verbunden wären, wenn wir eine Entscheidung zugunsten oder zuungunsten der Einführung der Werbemaßnahme träfen.

Der Anwendungsbereich des Ansatzes von Neyman und Pearson begrenzt sich daher auf Situationen, in denen der Anwender aufgrund von Kosten-Nutzen-Abwägungen ermitteln kann, wie risikofreudig er Entscheidungen treffen sollte. Zeigt eine Stichprobenziehung einer bestimmten Anzahl von Schrauben einen durchschnittlichen Durchmesser von 10.05 mm

entscheidet sich die Produzentin – je nachdem, wie risikobereit sie ist – dafür, die Maschinen anzuhalten oder weiterlaufen zu lassen. Ob sich die Risikofreudigkeit am Ende der Geschäftsperiode auszahlt, kann die Produzentin überprüfen und ihre Entscheidungskriterien gegebenenfalls anpassen. Eine solche Situation unterscheidet sich jedoch sehr von den üblichen Situationen in der Forschung, wie auch Fisher erläutert:

[…] acceptance procedures [i.e., the Neyman-Pearson approach] are of great importance in the modern world. When a large concern like the Royal Navy receives material from an engineering firm it is […] subjected to sufficiently careful inspection and testing to reduce the frequency of the acceptance of faulty or defective consignments. [The procedure] must also […] keep low both the cost of testing and the frequency of the rejection of satisfactory lots. I am casting no contempt on acceptance procedures, and I am thankful [they exist], whenever I travel by air […]. But the logical differences between such an operation and the work of scientific discovery […] seem to me so wide that the analogy between them is not helpful. […]

[The] great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed at, let us say, speeding production or saving money. (Fisher, 1955, S. 69-70)

Obwohl der Ansatz von Neyman und Pearson also vor allem in Produktionskontexten von enormem Nutzen sein kann, scheint er für Forschungskontexte und die meisten (wirtschafts-)psychologischen Fragestellungen keinen inkrementellen Nutzen gegenüber den bisher dargestellten inferenzstatistischen Verfahren aufzuweisen. Wie die Cut-Off-Wahrscheinlichkeit beim NHST können die Cut-Off-Wahrscheinlichkeiten Risiko I und Risiko II für eine Einzeluntersuchung von nahezu 0 % bis nahezu 100 % variieren, unabhängig davon, ob die p-Werte die Cut-Offs unter- oder überschritten haben. Damit ist der Ansatz nicht im Stande, die Unsicherheit zu quantifizieren, die bei einem Schluss von Stichprobendaten auf die Population verbunden ist, da auch hier die Typizität des Stichprobenergebnisses nicht ermittelt werden kann.

4.4 Konfidenzintervall

Unter anderem aufgrund der eingeschränkten Aussagekraft von Signifikanz- und Hypothesentests mehrten sich in den 80er und 90er Jahren des letzten Jahrhunderts Forderungen, als inferenzstatistische Angabe stattdessen Konfidenzintervalle zu nutzen (z. B. Brandstätter, 1997; Cumming & Finch, 2005; Fidler, 2005; Wilkinson & Task Force on Statistical Inference, 1999).

Ein Konfidenzintervall (KI oder CI für *confidence interval*) gibt an, in welchen hypothetischen Populationen unser Stichprobenergebnis zu den X % typischsten bzw. wahrscheinlichsten gehören würde. Dabei beschreibt X den Konfidenzkoeffizienten (KK oder CC für *confidence coefficient*): Zu wie viel % der typischsten bzw. wahrscheinlichsten Werten soll unser Stichprobenergebnis gehören? Für das Werbewirkungsbeispiel könnte etwa ein KK von 95 % gewählt werden. Das 95 % Konfidenzintervall für das Werbewirkungsbeispiel lautet 95 %-KI [0.21; 1.79] (Details im Anhang C). Dieses Intervall gibt an, dass die Mittelwertdifferenz von einem Skalenpunkt zu den 95 % der typischsten bzw. wahrscheinlichsten gehört in Populationen, die eine wahre Mittelwertdifferenz von 0.21 bis 1.79 Skalenpunkten haben. *Hätte* die wahre Mittelwertdifferenz also tatsächlich einen (beliebigen) konkreten Wert aus dem Zahlenbereich von 0.21 bis 1.79 Skalenpunkten, z. B. 0.98 Skalenpunkte, so würde eine Mittelwertdifferenz von einem Skalenpunkt zu den 95 % der wahrscheinlichsten gehören.

Das Konfidenzintervall wird unter anderem auf folgende, nicht korrekte Weise interpretiert – hier auf das vorliegende Beispiel bezogen: „Die Wahrscheinlichkeit dafür, dass die wahre Mittelwertdifferenz zwischen 0.21 und 1.79 Skalenpunkten liegt, beträgt 95 %" (Belia, Fidler, Williams & Cummings, 2005; Hoekstra et al., 2014; Lyu, Peng & Hu, 2018). Eine solche Interpretation ist jedoch aus verschiedenen Gründen nicht zulässig, unter anderem weil wir nicht wissen, ob die Mittelwertdifferenz von einem Skalenpunkt eine typische ist. Wüsste man, dass die Mittelwertdifferenz von $e = 1$ sich typischerweise in Untersuchungen zeigt, d. h. würden mehrere Untersuchungen ein 95 %-KI von [0.21; 1.79] ergeben, wäre es plausibel (wenn auch mathematisch-logisch nicht folgerichtig), davon auszugehen, dass der Erwartungswert der Mittelwertdifferenz in diesem Zahlenbereich liegt. Gehört die Mittelwertdifferenz von $e = 1$ jedoch zu den untypischen bzw. seltenen, so wären wir

vermutlich nicht bereit, anzunehmen, dass die wahre Mittelwertdifferenz im Bereich zwischen 0.21 und 1.79 Skalenpunkten liegt – je nachdem, welche Mittelwertdifferenzen und dazugehörige Konfidenzintervalle (z. B. [-0.11; 0.35]) sich als typisch erweisen.

Das Konfidenzintervall gibt also lediglich an, in welchen *hypothetischen* Populationen das Stichprobenergebnis zu den KK % typischsten gehört. Eine Aussage darüber, ob das Stichprobenergebnis zu den typischsten KK % in der *tatsächlichen* Population gehört erlaubt das Konfidenzintervall jedoch nicht. Damit kann auch mit Hilfe des Konfidenzintervalls, ebenso wie anhand der bisher dargestellten Inferenzstatistiken, keine Aussage über die Typizität eines Stichprobenergebnisses selbst getroffen werden. Da sich das Konfidenzintervall, ebenso wie die anderen Inferenzstatistiken, direkt aus dem Stichprobenergebnis ergibt, bleibt offen, ob sich das ermittelte Konfidenzintervall auch in anderen Untersuchungen zeigen würde. Um diese Typizitätsfrage zu beantworten, bleibt auch hier nur der Weg der Replikation. Wie bereits angeführt, wird jedoch bei mehreren Untersuchungen zu einer inhaltlichen Hypothese eine Beurteilung der Typizität der inferenzstatistischen Angabe überflüssig. Die Beurteilung der Typizität des Stichprobenkennwerts (z. B. der Mittelwertdifferenz oder der Korrelation) kann durch deskriptive Analyse erfolgen und diese zeigt bereits an, ob es sich bei einem Stichprobenkennwert (eher) um eine Ausnahme oder die Regel handelt.

Ein sehr großer Verwendungsbereich für das Konfidenzintervall besteht in der psychologischen (Individual-)Diagnostik, wobei für dessen Bestimmung zusätzlich zur Streuung noch die Reliabilität des Testinstruments einbezogen wird. Ob eine einzelne Messung ein typisches oder untypisches Messergebnis der Person darstellt, bleibt auch hierbei offen. Diese Tatsache wird in der Individualdiagnostik jedoch – anders als in der üblichen Forschungspraxis – explizit in der Forderung nach einem vollständigen diagnostischen Prozess berücksichtigt (z. B. Schmidt-Atzert & Amelang, 2012). So wird etwa eine enge Bezugsperson gebeten, einzuschätzen, ob ein Kind sein typisches Verhalten in der Testung an den Tag gelegt hat; ein an die Testung anschließendes Interview soll Auskunft geben, ob ein Patient oder Klient typisches Verhalten angegeben hat oder es werden mehrere Tests zu verschiedenen Zeitpunkten vorgelegt, um die Typizität des Messergebnisses besser beurteilen zu können. Insbesondere bei diagnostischen Prozessen, die eine schwerwiegende Entscheidung zur Folge haben (z. B. Empfehlung für einen Wechsel auf eine Förderschule, langfristiger Führerscheinentzug, Entzug des Sorgerechts, Bestätigung der Schuldfähigkeit usw.) müssen verschiedene Informationen aus unterschiedlichen Quellen herangezogen werden, um Fehlentscheidungen zu vermeiden. Für die Vollständigkeit des diagnostischen Prozesses ist es dabei weitgehend unerheblich, ob das Konfidenzintervall einen bestimmten Cut-Off-Wert (z. B. Intelligenztestgrenzwert von 85 Punkten) nicht enthält – es werden auch in diesem Fall weitere diagnostische Informationen herangezogen. Enthält das Konfidenzintervall den entscheidungskritischen Cut-Off-Wert, so werden ebenfalls zusätzliche Informationen eingeholt. Damit ist die Interpretation des Testwerts unabhängig von dem Konfidenzintervall für diesen Testwert, aber abhängig von anderen (deskriptiven) diagnostischen Informationen.

4.5 Bayesianische Ansätze

Die bisher dargestellten Inferenzstatistiken wurden innerhalb der frequentistischen Statistikschule entwickelt. Sowohl Fisher, der als Begründer des p-Werts und (wenn auch unfreiwilliger) Vater des NHST angesehen werden kann, als auch Neyman und Pearson, Urheber der Entscheidungstheorie und Weiterentwickler des Konfidenzintervalls, sahen keinen Sinn darin, Populationsparametern (Hypothesen bzw. Erwartungswerten von Stichprobenkennwerten) Wahrscheinlichkeiten zuzuschreiben (z. B. Fisher[6], 1924, 1956; Lehmann, 2011; Neyman, 1977, siehe Anhang D für Erläuterung).

Die Bayesianische Interpretation von Wahrscheinlichkeit, die zu den epistemischen Interpretationen gehört, lässt jedoch explizit Wahrscheinlichkeitsaussagen über Hypothesen zu. Wahrscheinlichkeit wird innerhalb dieser Denkschule als Zuversicht oder Glaube verstanden und entspricht unserem Alltagsverständnis von Wahrscheinlichkeit daher sehr viel mehr. Aussagen der Art „ich bin mir zu 99 % sicher, dass ich zu Besuch komme", „die Chancen dafür, dass ich zu Besuch komme, stehen fifty-fifty" oder „ich bin mir zu 80 % sicher, dass die Werbemaßnahme einen kleinen Effekt hat" geben epistemische

[6] Fisher war zwar nicht immer konsistent in seiner Meinung, stellte sich jedoch durchweg gegen die Angabe von posterior-Wahrscheinlichkeiten für Populationsparameter.

Wahrscheinlichkeiten an. Die verwendeten Zahlen werden nicht im Sinne einer Auftretenshäufigkeit, sondern im Sinne eines *subjektiven Glaubens* auf einer Skala reeller Zahlen von 0 bis 1 bzw. 0 % bis 100 % angegeben. Vor dem Hintergrund dieses Glaubens und der Sammlung zusätzlicher Informationen, die als Evidenz oder Indizien fungieren, wird nun neu bestimmt, inwiefern sich der subjektive Glaube ändert. Diese Änderung des subjektiven Glaubens nach der Berücksichtigung neuer Informationen wird als Wahrscheinlichkeitsrevision oder Aktualisierung des Glaubens (*updated belief*) bezeichnet (Lambert, 2018; Sedlmeier & Renkewitz, 2018).

4.5.1 Wahrscheinlichkeitsrevision

Nehmen wir für das Werbewirkungsbeispiel an, dass wir uns maximal unsicher sind über das Zutreffen der Nullhypothese „Die Werbemaßnahme hat keinen Effekt auf die Kaufbereitschaft". Damit weisen wir dieser Hypothese die Wahrscheinlichkeit von 50 % zu. Nun sammeln wir Informationen als Evidenz zu dieser Hypothese, indem wir unsere Untersuchung durchführen. Unsere Untersuchung ergibt mit einem Skalenpunkt Unterschied zugunsten der Werbemaßnahme einen p-Wert von 0.7 % – ein solcher Unterschied (oder ein noch größerer) ist sehr unwahrscheinlich, wenn es tatsächlich keinen Effekt der Werbemaßnahme gibt. Es stellt sich nun die Frage, wie sicher wir uns angesichts dieser neuen Information sein sollten, dass die Nullhypothese zutrifft. Die Wahrscheinlichkeit für, d. h. der Glaube an die Nullhypothese soll also angesichts der neuen Informationen aktualisiert werden. Die Wahrscheinlichkeitsrevision erfolgt anhand des Bayes-Theorems (für Erläuterung siehe Anhang D 1) und ergibt die Wahrscheinlichkeit 0.7 % (siehe Anhang D 1 für Details). Vor der Untersuchung waren wir uns maximal unsicher über das Zutreffen der Nullhypothese. Angesichts der Untersuchungsergebnisse reduziert sich unsere Unsicherheit deutlich: Nun sind wir sehr sicher, dass die Nullhypothese nicht zutrifft – wir schätzen die Wahrscheinlichkeit für die Nullhypothese angesichts des Stichprobenergebnisses als sehr gering ein.

Je nachdem, wie sicher wir uns sind über das Zutreffen der (Null-)Hypothese vor der Datensammlung, kann die aktualisierte Wahrscheinlichkeit deutlich von dem p-Wert abweichen. Halten wir die Nullhypothese im vornhinein für recht unplausibel[7] (10 % Wahrscheinlichkeit für die Nullhypothese), so beträgt die aktualisierte Wahrscheinlichkeit nur noch 0.1 %. Haben wir es mit einer hochspekulativen Alternativhypothese zu tun und halten wir die Nullhypothese für wesentlich wahrscheinlicher (z. B. mit 95 %), so ist die aktualisierte Wahrscheinlichkeit mit 12 % vergleichsweise hoch.

Die aktualisierte Wahrscheinlichkeit gibt also an, wie sehr wir daran glauben sollten, dass eine Hypothese stimmt angesichts der *bisher vorliegenden Daten* und des vergangenen Glaubens an die Hypothese. Die aktualisierte Wahrscheinlichkeit gibt jedoch nicht an, wie sicher wir uns sein können, dass das Stichprobenergebnis aus unserer Untersuchung ein typisches ist. Wie sicher wir uns sein können, dass Folgeuntersuchungen ein ähnliches Stichprobenergebnis zeigen, vermag die aktualisierte Wahrscheinlichkeit also nicht anzugeben. Wie bei den bisher dargestellten inferenzstatistischen Ansätzen liefert die aktualisierte Wahrscheinlichkeit daher keinen Erkenntnismehrwert gegenüber Replikationen und deskriptiven Analysen für die (wirtschafts-)psychologische Forschung.

4.5.2 Bayes-Faktor

Eine gängige Statistik der Bayesianischen Ansätze ist der Bayes-Faktor, der angibt, wie viel wahrscheinlicher die Daten auf der Alternativhypothese als auf der Nullhypothese sind bzw. wie viel Mal besser die Daten zur Alternativhypothese als zur Nullhypothese passen. Auf das Werbewirkungsbeispiel bezogen ergibt sich z. B. ein Bayes-Faktor von 6.57 (siehe Anhang D 2 für mehr Details). Dieser Bayes-Faktor zeigt an, dass die Daten rund sieben Mal besser zu der Alternativhypothese passen als zur Nullhypothese. Dieser Bayes-Faktor ist jedoch nur dann aussagekräftig bzw. zuverlässig, wenn wir damit rechnen können, dass er nicht nur für die vorliegenden Untersuchungsdaten gilt, sondern auch darüber hinaus, z. B. für eine nächste Untersuchung. Das ist wiederum eine Frage nach der Typizität des Bayes-Faktors aus dieser Untersuchung und damit nach der Typizität des Stichprobenergebnisses aus der Untersuchung. Hätte die Untersuchung eine Mittelwertdifferenz von einem halben

[7] Dies ist sehr viel realistischer: wir würden vermutlich zumindest von einem kleinen Werbewirkungseffekt mit hoher Zuversicht ausgehen. Insgesamt werden im Forschungsalltag vermutlich überwiegend plausible Alternativhypothesen getestet, um diese zu bestätigen, insbesondere in angewandten psychologischen Disziplinen.

Skalenpunkt zugunsten der Werbemaßnahme ergeben, so betrüge der Bayes-Faktor 0.36 und würde für die Glaubhaftigkeit der Nullhypothese sprechen. Auch in diesem Fall könnten also nur Replikationen Auskunft über die Typizität des Stichprobenergebnisses liefern.

Der zentrale Anwendungsbereich und enorme Nutzen des Bayes-Theorems sind vor allem in diagnostischen Fragestellungen zu finden. Besonders hervorzuheben ist hier der Nutzen bei Bildgebungsverfahren. Die Diagnose einer Krankheit, einer Störung, der Eignung etc. ist zwangsläufig mit mehr oder weniger Unsicherheit verbunden, je nachdem wie groß die Basisrate und die Spezifität und Sensitivität des Messinstruments sind. Für Diagnostiker und Diagnostikerinnen ist es daher enorm wichtig, die Wahrscheinlichkeit dafür, dass die Person eine bestimmte Diagnose fälschlicherweise erhalten hat, nicht aus dem Blick zu verlieren. Eine Person, die z. B. ein positives Ergebnis eines HIV-Tests erhält, der 100 % von HIV-infizierten Personen als solche identifiziert, wird mindestens einmal erneut getestet, da die Wahrscheinlichkeit dafür, dass die Person tatsächlich den HI-Virus in sich trägt, nach einem ersten positiven Testergebnis bei nur 5 % liegen kann (Gigerenzer & Weiler, 2019). Eine Person, die im Personalauswahlverfahren, das eine Trefferquote von 90 % hat, als geeignet identifiziert wurde, könnte nur mit einer Wahrscheinlichkeit von 49 % tatsächlich geeignet sein (für Details siehe Anhang D 1). Inwiefern der Bayes-Ansatz jedoch einen gegenüber Replikationen und deskriptiven Analysen inkrementellen Nutzen für die übliche sozialwissenschaftliche Forschung hat, ist fraglich.

5 Fazit

Inferenzstatistische Verfahren sollen die Quantifizierung der Unsicherheit, die bei einem Schluss von den Stichprobendaten auf die Population verbunden ist, ermöglichen. Solche Quantifizierungsmöglichkeiten sind p-Werte, dichotome Signifikanzaussagen, Konfidenzintervalle oder Bayes-Faktoren. Diesen Inferenzstatistiken ist jedoch eine gravierende Einschränkung der Aussagekraft gemeinsam: Sie liefern nur dann eine pragmatisch-hilfreiche Quantifizierung, wenn sie typisch bzw. reliabel sind. Ein p-Wert, ein signifikantes Ergebnis, ein Konfidenzintervall oder ein Bayes-Faktor sind dann für den Schluss auf die Population hilfreich, wenn wir uns darauf verlassen können, dass sie keine Besonderheit dieser einen empirischen Erhebung sind. Ist die inferenzstatistische Angabe eine untypische, so kann die quantitative Angabe in Form des p-Werts, der Signifikanz, des Konfidenzintervalls oder Bayes-Faktors extrem von den tatsächlichen Gegebenheiten in der Population abweichen. Da die vorgestellten Inferenzstatistiken direkt aus dem Stichprobenergebnis ermittelt werden, ist die Forderung nach der Typizität oder Verlässlichkeit der Inferenzstatistik äquivalent mit der Forderung nach der Typizität oder Verlässlichkeit des Stichprobenergebnisses. Damit scheint die zentrale Frage bei dem Schluss von Stichprobendaten auf die Population diejenige nach der Typizität der Stichprobendaten bzw. des Stichprobenergebnisses zu sein.

Die Typizität eines Stichprobenergebnisses allerdings kann nach heutigem Erkenntnisstand grundsätzlich nicht beurteilt werden. Unter dem Schlagwort des Induktionsproblems findet die Tatsache Berücksichtigung, dass wir keine Aussage über Beobachtungen treffen können, die wir (noch) nicht gemacht haben. Sieht man über diese grundlegende epistemische Einschränkung hinweg, so lässt sich die Typizität nur durch Wiederholungen von Stichprobenziehungen auf Basis pragmatischer Überlegungen besser beurteilen. Die Beurteilung der Typizität eines Stichprobenergebnisses kann damit praktisch betrachtet nur über Replikationen von Untersuchungen erfolgen. Liegen aber Replikationen als Behelf zur Beurteilung der Typizität des Stichprobenergebnisses vor, so bringen inferenzstatistische Aussagen, da sie sich direkt aus den Stichprobendaten ergeben, hierfür keinen inkrementellen Nutzen. Damit scheinen inferenzstatistische Quantifizierungen für eine Einzeluntersuchung im besten Fall unzuverlässig, im schlimmsten Fall irreführend (Stichwort „Replikationskrise") und für mehrere Untersuchungen überflüssig zu sein.

Als ein plakatives Beispiel soll der zweifache Lottogewinn von David Long dienen (Pidd, 2015). Der Brite gewann zwei Mal 1 Million Pfund, was laut der Camelot Group, dem Betreiber der britischen National Lottery, einem p-Wert von $p = .000000004$ entspricht (1 zu 283 Milliarden, Pidd, 2015). Sicherlich wäre man bei einem solch niedrigen p-Wert berechtigt, die Nullhypothese „David Long hat die Lottozahlen geraten" zugunsten einer Alternativhypothese „David Long kann hellsehen" oder „David Long hat betrogen" abzulehnen. Unsere Intuition rät uns in diesem Fall jedoch höchstwahrscheinlich korrekterweise, diesen p-Wert nicht überzuinterpretieren bzw. ganz außer Acht zu lassen: David Long hat einfach Glück gehabt.

Auch wenn zu den Motiven und Beweggründen von Forscherinnen und Forschern für die Nutzung von Inferenzstatistik einige Vermutungen geäußert wurden (z. B. Nickerson, 2000), so fehlen dazu empirische Belege weitestgehend. Es liegt wohl nahe zu vermuten, dass ein menschliches Bedürfnis nach Sicherheit oder zumindest deren Quantifizierung eine große Rolle spielt als Grund für die ungebrochen intensive Nutzung der Inferenzstatistik in „soften" Wissenschaften, wie der Psychologie. Wir müssen jedoch einsehen, dass es weder eine Sicherheit bezüglich des Eintretens eines Stichprobenergebnisses noch bezüglich des Zutreffens einer Hypothese geben kann, wie man sich anhand von zwei einfachen Beispielen vor Augen führen kann. In mehreren Stichproben von Studierenden in ähnlichen Studiengängen wurden Korrelationen zwischen logischen und mathematischen Fertigkeiten ermittelt. Diese betrugen $r = .32$ $(n = 104)$, $r = .44$ $(n = 29)$, $r = -.34$ $(n = 13)$ und $r = .29$ $(n = 16)$ (Sawatzky, 2020). Solche teilweise widersprüchlichen Stichprobenergebnisse sind der Alltag der empirischen Forschung, den wir akzeptieren müssen. Es gibt keine Kennzahl, die uns eine Aussage darüber erlaubt, welche Korrelation wir in einer nächsten Untersuchung erhalten werden. Wie (un-)sicher wir uns darüber sein sollten, eine beliebige Korrelation zu finden, kann ebenfalls nicht quantifiziert werden, weil wir keine Kenntnis darüber haben, wie hoch der Zusammenhang tatsächlich ist. Es kann weder Sicherheit noch eine Quantifizierung

der Unsicherheit für das Eintreten eines bestimmten empirischen Ergebnisses in der üblichen psychologischen Forschung geben. Das zweite Beispiel ist, aus Ermangelung von etablierten Hypothesen in der Psychologie, der Physik entnommen. Über mehr als ein Jahrtausend galt das ptolemäische Weltbild als gültige Hypothese in der Astronomie (z. B. Kuhn, 2016) und erwies sich doch als falsch, zumindest soweit uns heute bekannt ist. Auch wenn wir uns also sehr sicher sind über das Zutreffen einer Hypothese und eine überwältigende Menge an (scheinbaren) Belegen für die Hypothese vorliegt, ist dies keine Garantie dafür, dass die Hypothese tatsächlich zutrifft. In der Physik gibt es natürlich noch viele weitere Beispiele für solche überraschenden Entwicklungen (z. B. Newtons Gravitationstheorie, das Konzept des Äthers usw.). Auch kann eine (Un-)Sicherheit über das Zutreffen der Hypothese, die objektiv betrachtet angebracht sein sollte, nicht quantifiziert werden, da wir einfach nicht wissen, was wir nicht wissen. Ungewissheit liegt damit im Kern aller empirischen Forschungen und diese Tatsache sollte akzeptiert werden.

Ein weiterer Grund für die verbreitete Verwendung inferenzstatistischer Angaben liegt möglicherweise darin, dass hierbei verschiedene Stichprobenkennwerte, wie Mittelwertdifferenzen, Korrelationen, Regressionsgewichte, Anteile aufgeklärter Varianz usw. in eine einheitliche Kennzahl, wie den p-Wert umgerechnet werden (Greenwald, Gonzales, Harris & Guthrie, 1996). Damit entfällt für die Verwenderin und den Verwender die Notwendigkeit, verschiedene Statistiken zu verstehen bzw. zu interpretieren. Stattdessen kann lediglich eine Angabe, z. B. der p-Wert hinsichtlich eines Cut-Off-Werts beurteilt werden. Als anekdotische Evidenz hierfür soll die Reaktion eines Kollegen/einer Kollegin auf die Forderung der American Statistical Association, das Konzept der statistischen Signifikanz nicht mehr zu verwenden (Wasserstein, Schirm & Lazar, 2019), dienen: „Oh nein, $p < .05$ ist das Einzige, was ich verstehe!"

Insgesamt betrachtet scheinen Inferenzstatistiken als Maß der Evidenz und als Grundlage für das Beurteilen des Zutreffens einer Hypothese in den allermeisten Anwendungssituationen der (wirtschafts-)psychologischen Forschung ungeeignet zu sein. Was ist die Alternative? Eine Antwort gibt John Tukey in seiner Ansprache vor der American Psychological Association im Jahr 1968:

Data analysis needs to be both exploratory and confirmatory. In exploratory data analysis there can be no substitute for flexibility, for adapting what is calculated – and, we hope, plotted – both to the needs of the situation and the clues that the data have already provided. In this mode, data analysis is detective work – almost an ideal example of seeking what might be relevant. […] Confirmatory data analysis must be the means by which we adjust optimism and pessimism, not only ours but those of our readers. […] [But the Roman Catholic Church] has long held that sanctification was only for the dead – indeed only for those already dead for an appropriate period. I believe and I urge you to feel, that sanctification of data [d. h. konfirmatorische Datenanalyse zur Beurteilung der Gültigkeit einer Hypothese] is equally only for dead data – data that are only of historical importance, like Newton's apple. If we could all live by this precept, we might have to think more – painful though that might be – but we would, by the same token, accomplish more. (Tukey, 1969, S. 90)

Eine detaillierte deskriptive und grafisch gestützte Analyse der jeweils vorliegenden Untersuchungsdaten in Verbindung mit einer subjektiven Einordnung der Ergebnisse in einen größeren Forschungskontext – aber ohne den Anspruch auf eine abschließende Beurteilung einer Hypothese – erscheint für die meisten (wirtschafts-)psychologischen Forschungsinhalte sinnvoller zu sein als die Verwendung von Inferenzstatistik in der vergeblichen Hoffnung auf eine Quantifizierung der (Un-)Sicherheit über die tatsächlichen Gegebenheiten in der Population.

6 Literatur

Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A. & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian academic psychologists. Frontiers in Psychology, 7, Article 1247. https://doi.org/10.3389/fpsyg.2016.01247

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423–437. https://doi.org/10.1037/h0020412

Belia, S., Fidler, F., Williams, J. & Cumming, G. (2005). Researchers misunderstand Confidence Intervals and standard error bars. *Psychological Methods, 10*(4), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin, 16*(10), 335–338. https://doi.org/10.1037/h0074554

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods of Psychological Research, 4*(2), 33–46. https://www.dgps.de/fachgruppen/methoden/mpr-online/issue7/art2/brandstaetter.pdf

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. & Finch, S. (2005). Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist, 60*(2), 170–180. https://doi/10.1037/0003-066X.60.2.170

Eid, M., Gollwitzer, M. & Schmitt, M. (2017). *Statistik und Forschungsmethoden* (5., korrigierte Aufl.). Weinheim: Beltz.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I. & Tutz, G. (2016). *Statistik. Der Weg zur Datenanalyse* (8., überarbeitete und ergänzte Aufl.). Berlin: Springer Spektrum.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology*. Department of History and Philosophy of Science, University of Melbourne. Retreived from http://www.botany.unimelb.edu.au/envisci/docs/fidler/fidlerphd_aug06.pdf

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J. & Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments & Computers, 36*(2), 312–324. https://doi.org/10.3758/BF03195577

Fisher, R. A. (1924). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 3, 329-332. Retrieved from https://hekyll.services.adelaide.edu.au/dspace/bitstream/2440/15169/1/14.pdf

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Oliver and Boyd: Edinburgh.

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Hafner Publishing Co.

Fisher, R. A. (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society. Series B (Methodological), 17*(1), 69-78. Retrieved from http://www.jstor.org/stable/2983785

Fritz, A., Scherndl, T. & Kühberger, A. (2012). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology, 23*(1), 98-122. https://doi.org/10.1177/0959354312436870

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22–38. Retrieved from http://jse.amstat.org/v10n3/garfield.html

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Gigerenzer, G., Swijtink, Z. G., Porter, T. M., Daston, L., Beatty, J. & Krüger, L. (1989). *Ideas in context. The empire of chance: How probability changed science and everyday life.* Cambridge University Press.

Gigerenzer, G. & Weiler, S. (2019, 8. Januar). *„Sie sind wahrscheinlich HIV-Positiv*". Unstatistik des Monats. https://www.rwi-essen.de/unstatistik/86/

Greenwald, A. G., Gonzalez, R., Harris, R. J. & Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology*, *33*(2), 175–183. https://doi.org/10.1111/j.1469-8986.1996.tb02121.x

Henderson, L. (2020). The Problem of Induction. In E. N. Zalta (ed.) *The stanford encyclopedia of philosophy* (Spring 2020 Edition). Retrieved from https://plato.stanford.edu/archives/spr2020/entries/induction-problem

Hoekstra, R., Morey, R. D., Rouder, J. N. & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hubbard, R. & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—And its future prospects. *Educational and Psychological Measurement, 60*(5), 661–681. https://doi.org/10.1177/00131640021970808

Hume, D. (1748/2004). *Enquiry concerning human understanding*. In the version by Jonathan Bennett presented at www.earlymoderntexts.com

Hume, D. (1739-40/2017). *Enquiry concerning human understanding, book 1.* In the version by Jonathan Bennett presented at www.earlymoderntexts.com

Inferenzstatistik. (2020). In M. A. Wirtz (Hrsg.), Dorsch – Lexikon der Psychologie. Abgerufen am 09.02.2020, von https://portal.hogrefe.com/dorsch/inferenzstatistik/

JASP Team (2020). JASP (Version 0.13.1) [Computer software].

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573–603. https://doi.org/10.1037/a0029146

Kuhn, W. (2016). *Ideengeschichte der Physik. Eine Analyse der Entwicklung der Physik im historischen Kontext* (2. Aufl.). Berlin: Springer.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology, 22*(1), 67–90. https://doi.org/10.1177/0959354311429854

Lambert, B. (2018). *A student's guide to Bayesian statistics*. Sage.

Lehmann, E. L. (2011). Fisher, Neyman, and the creation of classical statistics. New York: Springer. https://doi.org/10.1007/978-1-4419-9500-1

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science* 57(1), 69–91. https://doi.org/10.1093/bjps/axi152

Lyu, Z., Peng, K. & Hu, C.-P. (2018). P-value, confidence intervals, and statistical inference: A new dataset of misinterpretation. *Frontiers in Psychology, 9,* Article 868. https://doi.org/10.3389/fpsyg.2018.00868

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese 36*(1), 97-131. https://doi.org/10.1007/BF00485695

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301. https://doi.org/10.1037/1082-989X.5.2.241

Oakes, M. (1986) Statistical inference: A commentary for the social and behavioural sciences. Chichester: Wiley.

Pidd, H. (2015, 1. April). British couple celebrate winning second £1m EuroMillions prize. The Guardian. Retrieved from https://www.theguardian.com/uk-news/2015/apr/01/british-couple-celebrate-winning-second-1m-euromillions-prize

Popper, K. (1935). *Logik der Forschung.* Wien: Springer.

Popper, K. (2002). *The logic of scientific discovery* (2nd Ed.). Routledge.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Romeijn, J.-W. (2017). Philosophy of statistics. In E. N. Zalta (ed.) *The Stanford encyclopedia of philosophy* (Spring 2017 Edition). Retrieved from https://plato.stanford.edu/archives/spr2017/entries/statistics/

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA. http://www.rstudio.com/

Russell, B. (1912). *The problems of philosophy*. New York: Henry Holt and Company. Retrieved from https://ia802804.us.archive.org/5/items/problemsofphilo00russuoft/problemsofphilo00russuoft.pdf

Sawatzky, A. (2020). *Messung der statistischen Kompetenz in der Hochschulausbildung am Beispiel des Statistical Reasoning Assessment* (Unveröffentlichte Dissertation). Universität zu Köln.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5., vollständig überarbeitete und erweiterte Aufl.). Berlin: Springer. https://doi.org/10.1007/978-3-642-17001-0

Sedlmeier, P., & Renkewitz, F. (2018). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (3., aktualisierte und erweiterte Auflage). Hallbergmoos: Pearson.

Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician*, *73*(Supplement 1), 246-261. https://doi.org/10.1080/00031305.2018.1518264

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*(2), 83–91. https://doi.org/10.1037/h0027108

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M. & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217–239. https://doi.org/10.1037/met0000100

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018a). Bayesian inference for psychology. Part 1: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., . . . Morey, R. D. (2018b). Bayesian inference for psychology. Part 2: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133 https://doi.org/10.1080/00031305.2016.1154108.

Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. (Eds.). (2019). Statistical Inference in the 21st Century: A World Beyond p < 0.05 [special issue sup1]. *The American Statistician*, *73*(Supplement 1).

Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05." *The American Statistician*, *73*(Supplement 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wilkinson, L., & 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. https://doi.org/10.1037/0003-066X.54.8.594

Anhang

A: p-Wert

Die grundlegende Idee bei diesem Verfahren besteht darin, dass wir uns den deskriptiven Stichprobenkennwert (hier die Mittelwertdifferenz von einem Skalenpunkt) als Zufallsergebnis vorstellen. Aus einer Population, deren Elemente alle möglichen Mittelwertdifferenzen sind, die auf Basis von je 50 Personen pro Gruppe zustande kommen könnten, erhalten wir per Zufall eine. Je nachdem, wie „oft" eine Mittelwertdifferenz wie unsere (oder eine noch größere oder kleinere) in dieser Population vorkommt, ist es mehr oder weniger wahrscheinlich, dass wir eine solche Mittelwertdifferenz (oder eine noch größere bzw. noch kleinere) als Ergebnis einer Zufallsziehung erhalten. Die „Häufigkeit", mit der bestimmte Mittelwertdifferenzen in einer Population von Mittelwertdifferenzen vorkommen, wird durch sogenannte Parameter der Population festgelegt, wobei die Population hier als Wahrscheinlichkeitsverteilung mathematisch definiert ist. Für das dargestellte Beispiel wird die Population der Mittelwertdifferenzen durch die t-Wahrscheinlichkeitsverteilung mit den beiden Parametern *Freiheitsgrade*[8] und *Nonzentralitätsparameter*[9] wiedergegeben. Während der Parameter der Freiheitsgrade durch die Untersuchung festgelegt ist (hier $df = 98$), spielen für die Festlegung des Nonzentralitätsparameters inhaltliche Überlegungen eine Rolle: Wie stellen wir uns die Population aller möglichen Mittelwertdifferenzen vor? Gehen wir im Sinne der informationsärmsten Auslegung des Falsifikationsprinzips davon aus, dass die tatsächliche Mittelwertdifferenz (d. h. der Erwartungswert der Mittelwertdifferenz μ_e) 0 beträgt? Wenn ja, würden wir mit dem p-Wert die Wahrscheinlichkeit für eine Mittelwertdifferenz von einem Skalenpunkt (und solchen Mittelwertdifferenzen, die noch größer oder noch kleiner sind) ermitteln – wenn die tatsächliche Mittelwertdifferenz 0 Skalenpunkte beträgt. Nehmen wir an, dass Unterschiede von bis zu 0.5 Skalenpunkten aus praktischen Gesichtspunkten eine zu vernachlässigende Werbewirkung wiedergeben? Wenn ja, erhalten wir mit dem p-Wert die Wahrscheinlichkeit für eine Mittelwertdifferenz von einem Skalenpunkt (und solchen, die noch größer oder noch kleiner sind), wenn die tatsächliche Mittelwertdifferenz 0.5 Skalenpunkte beträgt. Oder gehen wir davon aus, dass die Werbemaßnahme tatsächlich eine recht hohe Wirkung, von z. B. 2 Skalenpunkten hat? Dann interessiert uns die Wahrscheinlichkeit für eine Mittelwertdifferenz von einem Skalenpunkt (und solchen, die noch größer oder noch kleiner sind), wenn die tatsächliche Mittelwertdifferenz 2 Skalenpunkte beträgt.

Der Nonzentralitätsparameter hat nur bei Erwartungswerten ungleich 0 eine Relevanz ($\lambda = \frac{\mu_e}{SE_e} = \frac{0}{SE_e} = 0$) und wird darüber hinaus in Datenauswertungsprogrammen oft nicht berücksichtigt. Daher soll die Berechnung des p-Werts für das oben beschriebene Beispiel mit einer auf lange Sicht erwarteten Mittelwertdifferenz von 0 dargestellt werden. Um die zu erwartende Häufigkeit (frequentistische Wahrscheinlichkeit) des Vorkommens der Mittelwertdifferenz von einem (oder mehr bzw. oder weniger) Skalenpunkt in einer Population, die eine durchschnittliche Mittelwertdifferenz von 0 hat zu bestimmen, wird die Mittelwertdifferenz zunächst in einen t-Wert umgerechnet:

$$t_{e=1} = \frac{e - \mu_e}{SE_e} = \frac{e - \mu_e}{\sqrt{\frac{{SD_{EG}}^2}{n_{EG}} + \frac{{SD_{KG}}^2}{n_{KG}}}} = \frac{1 - 0}{\sqrt{\frac{2^2}{50} + \frac{2^2}{50}}} = \frac{1}{0.4} = 2.5 \qquad (1)$$

mit SE_e: Standardfehler (*standard error*) der Mittelwertdifferenz (die durchschnittlich zu erwartende Unterschiedlichkeit aller möglichen Mittelwertdifferenzen, die auf Basis von je 50 Personen pro Gruppe zustande kommen könnten oder die Standardabweichung der möglichen Mittelwertdifferenzen[10]). Der Standardfehler von 0.4 Skalenpunkten besagt, dass

[8] Für dieses Beispiel gibt der Parameter Freiheitsgrade an, auf Basis von wie vielen Einzelwerten, die sich durch Zufall unterscheiden könnten, sich die Mittelwertdifferenz ergibt, hier: $df = n_{EG} - 1 + n_{KG} - 1 = 49 + 49 = 98$.

[9] Für dieses Beispiel gibt der Nonzentralitätsparameter vereinfacht gesagt, die Mitte der Wahrscheinlichkeitsverteilung bzw. genaugenommen, das Verhältnis des Erwartungswerts der Mittelwertdifferenz zu dem Standardfehler $\lambda = \frac{\mu_e}{SE_e}$ an.

[10] Hierbei ist zu beachten, dass es sich um die Standardabweichung eines statistischen Kennwerts, hier der Mittelwertdifferenz, handelt (SE) und nicht um die Standardabweichung einzelner Personenwerte (SD).

wenn die Mittelwertdifferenz tatsächlich im Schnitt 0 beträgt ($\mu_e = 0$), wir auf lange Sicht im Schnitt mit Mittelwertdifferenzen zwischen -0.4 und 0.4 Skalenpunkten ($\mu_e \pm 1 \cdot SE_e = 0 \pm 1 \cdot .04 = 0 \pm 0.4$ Skalenpunkte) rechnen könnten, wenn wir eine Mittelwertdifferenz aus dieser Population (d. h., anhand von je 50 per Zufall ausgewählten Personen) ermitteln würden. Die Mittelwertdifferenz aus unserer Untersuchung liegt nun klar außerhalb dieses Durchschnittsbereichs. Sie ist um 2.5 Standardfehler größer als die im Schnitt erwartete Mittelwertdifferenz. Oder anders ausgedrückt, statt eines erwarteten t-Werts von ungefähr 0 (± 0.4) erhalten wir einen t-Wert von 2.5. Da die theoretische Auftretenshäufigkeit von t-Werten durch das Vorliegen der entsprechenden t-Wahrscheinlichkeitsverteilungsfunktion bekannt ist, kann der prozentuale Anteil von t-Werten von 2.5 oder größer ($t \geq 2.5$) auf einer t-Verteilung mit 98 Freiheitsgraden und dem Erwartungswert 0 (bzw. einem Nonzentralitätsparameter von 0) durch Integralrechnung bestimmt werden:

$$p(t_{e=1} \geq 2.5 | t_{\mu_e=0}, df = 98) = .007$$

mit p: relativer Anteil bzw. frequentistische Wahrscheinlichkeit.

Der p-Wert von .007 drückt Folgendes aus:

- wenn wir in einer Untersuchung (mit je 50 zufällig ausgewählten Personen in zwei Versuchsgruppen) auf lange Sicht einen t-Wert von 0 erwarten (d. h., von einer Mittelwertdifferenz von 0 Skalenpunkten ausgehen),

- dann erhalten wir einen t-Wert von 2.5 oder größer (d. h., einer Mittelwertdifferenz von einem Skalenpunkt bei einer Standardabweichung von zwei Skalenpunkten in beiden Gruppen) – auf lange Sicht – in 0.7 % solcher Untersuchungen erhalten würden.

Der Wahrscheinlichkeitsverteilungswert, der den p-Wert festlegt, kombiniert Informationen über den unstandardisierten Effekt, die Streuung innerhalb der Stichprobe und die Stichprobengröße auf multiplikative Weise (siehe Formel 1). Die Höhe des p-Werts kann damit auf eine oder mehrere der genannten Dateneigenschaften zurückzuführen sein. Zusätzlich enthält der p-Wert keine weiteren Informationen außer dem standardisierten Effekt, z. B. dem Mittelwertunterschied, der Korrelation, dem Anteil aufgeklärter Varianz usw., der Streuung in der Stichprobe und der Stichprobengröße. Wie die Bestimmung eines Volumens nicht mehr als die Bestandteile Höhe, Breite und Tiefe enthält, so enthält auch der p-Wert grundsätzlich nicht mehr als die Informationen aus der Stichprobe. Diese multiplikative Verknüpfung der drei Informationen gilt für alle gängigen deskriptiven Stichprobenkennwerte wie Korrelation, Regressionsgewicht, Anteil der aufgeklärten Varianz usw., auch wenn sich die Formeln auf den ersten Blick deutlich unterscheiden.

B: Hypothesentest nach Neyman und Pearson

Der NHST entwickelte sich historisch vermutlich als Reaktion auf die Jahrzehnte überdauernde, oft hoch polemische Auseinandersetzung zwischen Fisher, der als Vater der modernen Inferenzstatistik angesehen werden kann, und Neyman, einem der Begründer einer Reihe von inferenzstatistischen Verfahren (z. B. statistische Entscheidungstheorien, Konfidenzintervalle) (Gigerenzer et.al., 1989; Gigerenzer, 2004; Lehmann, 2013; Lenhard, 2006). Hierbei fanden Elemente aus Fishers Signifikanztest und dem Hypothesentest von Neyman und Pearson Eingang, darunter der Begriff der Irrtumswahrscheinlichkeit, auch als Fehler oder Risiko bezeichnet. Die Irrtumswahrscheinlichkeit im NHST hat jedoch keine sinnvolle Bedeutung, d. h. die tatsächliche Irrtumswahrscheinlichkeit kann zwischen nahezu 0 % und nahezu 100 % liegen, unabhängig davon, welches Signifikanzniveau in einer Einzeluntersuchung festgestellt wird. Im Hypothesentest von Neyman und Pearson kann die Irrtumswahrscheinlichkeit in geeigneten Situationen dagegen sinnvoll interpretiert werden.

Die grundlegende Idee des Hypothesentests von Neyman und Pearson besteht darin, dass vor Beginn einer Untersuchungsreihe eine Entscheidungsschablone konstruiert wird. Daher heißt dieser Ansatz auch *Entscheidungstheorie*. In die Konstruktion der Schablone gehen vier Arten von Informationen ein: 1) die Entscheidungsoptionen hinsichtlich eines konkreten Verhaltens, 2) die maximal tolerierbaren Risiken für falsche Entscheidungen, 3) die zu erwartende Streuung der Einzelwerte und 4) die Stichprobengröße. Diese Informationen

werden so aufeinander abgestimmt, dass eine für das jeweilige Entscheidungsproblem optimale, d. h. gleichzeitig maximal verlässliche und im Sinne des Stichprobenumfangs sparsamste Untersuchungsreihe angelegt werden kann. Nutzt man die Entscheidungsschablone *langfristig* für das Treffen von Entscheidungen, so ist garantiert, dass die maximal tolerierbaren Risiken für falsche Entscheidungen auf lange Sicht nicht überschritten werden.

Entscheidungsoptionen als Effekt

Zunächst werden die Entscheidungsoptionen hinsichtlich konkreten Verhaltens inhaltlich und statistisch formuliert. Auf das Werbewirkungsbeispiel übertragen gäbe es etwa die Verhaltensoption, die Werbemaßnahme einzuführen und die Verhaltensoption, die Werbemaßnahme nicht einzuführen. Statistisch knüpft man die Option, die Werbemaßnahme einzuführen an eine hohe Werbewirkung (z. B. $\mu_{e_1} = 3$ Skalenpunkte) und die Option, die Werbemaßnahme nicht einzuführen an eine nur geringe Werbewirkung (z. B. $\mu_{e_2} = 1$ Skalenpunkt). Es wird hierbei deutlich, dass die statistische Formulierung der Entscheidungsoptionen auf eine individuelle Schwerpunktsetzung ausgelegt ist. Handelt es sich um eine kostengünstige Werbemaßnahme, sind wir vermutlich gewillt, diese auch bei einer geringeren Wirkungskraft einzusetzen (z. B. ab $\mu_{e_1} = 1$ Skalenpunkt) und verzichten auf den Einsatz der Maßnahme nur im Fall vollständig fehlender Wirkung ($\mu_{e_2} = 0$ Skalenpunkte). Muss bei der Einführung der Werbemaßnahme sehr viel Geld in die Hand genommen werden, sind wir möglicherweise nicht gewillt, diese einzusetzen, wenn die Wirkung nur gering bis mittelmäßig ist (z. B. wenn $\mu_{e_2} \leq 1$ Skalenpunkte gegenüber $\mu_{e_1} = 3$ Skalenpunkte). Das Ausmaß des Unterschieds zwischen den beiden statistisch formulierten Entscheidungsoptionen wird als Effekt bezeichnet (z. B. hoher oder mittlerer Effekt bei $\mu_{e_{Option\,1}} = 3$ und $\mu_{e_{Option\,2}} \leq 1$ und geringer Effekt bei $\mu_{e_{Option\,1}} = 1$ und $\mu_{e_{Option\,2}} = 0$).

Maximal tolerierbaren Risiken für falsche Entscheidungen

Als nächstes werden die maximal tolerierbaren Risiken für eine jeweils falsche Entscheidung festgelegt. Da eine Entscheidung zugunsten der Einführung der Werbemaßnahme, obwohl diese tatsächlich keine (oder eine zu geringe) Wirkung zeigt (Risiko I), finanzielle Einbußen (sunk costs oder versunkene Kosten) mit sich bringt, soll eine solche Fehlentscheidung natürlich möglichst vermieden werden. Ganz ausgeschlossen werden kann dieses Risiko aufgrund der der Empirie inhärenten Unsicherheit zwar nicht, aber es ist möglich, das Risiko sehr klein zu wählen, z. B. 0.01 %. Zu beachten ist allerdings auch das Risiko für die umgekehrt falsche Entscheidung (Risiko II). Die Werbemaßnahme wird nicht eingeführt, obwohl sie eine (hohe) Wirkung hat und in der Folge entstehen Opportunitätskosten. Je weniger wir bereit sind, uns für die Einführung der Werbemaßnahme zu entscheiden (geringes Risiko I bzw. geringe versunkene Kosten), desto größer werden die Opportunitätskosten (höheres Risiko II). Je weniger wir bereit sind, auf die Einführung der Werbemaßnahme zu verzichten (geringes Risiko II bzw. geringe Opportunitätskosten), desto größer werden die versunkenen Kosten (höheres Risiko I). Um eine möglichst ideale Balance zwischen den versunkenen und den Opportunitätskosten herzustellen, können die Risiken durch Kosten-Nutzen-Szenarien festgelegt werden: Wie hoch könnten die versunkenen und die Opportunitätskosten unter Berücksichtigung des maximal und minimal zu erwartenden Effekts werden? Diese monetären Beträge können dann im Idealfall als Risiken ausgedrückt werden.

Auch die Festlegung der Risiken erfolgt also anhand subjektiver (Kosten-Nutzen-) Überlegungen. So sind mit den jeweils falschen Entscheidungen zugunsten und zuungunsten der Einführung einer Werbemaßnahme andere Kosten verbunden als etwa bei der Produktion und dem Vertrieb medizinischer Produkte auf der einen und Möbelschrauben auf der anderen Seite. Während des Produktionsprozesses[11] von Arteriengefäßstützen (Stents) etwa ergeben sich für die Qualitätssicherung die beiden Entscheidungsoptionen „Maschinen anhalten und warten" (2) und „Maschinen nicht anhalten und die Produktion weiterlaufen lassen" (1). Diese Entscheidungsoptionen können sich statistisch abbilden in den Hypothesen (1) Stents haben den geforderten Durchmesser von 2mm und (2) Stents haben einen zu großen Durchmesser von mindestens 2.5mm. Wird in einer Zufallsstichprobe von produzierten Stents im Rahmen der Qualitätskontrolle ein durchschnittlicher Durchmesser von 2.3mm

[11] Da sich für die weiteren Komponenten des Neyman-Pearson-Ansatzes ein psychologisches Beispiel nicht sinnvoll formulieren lässt, wird ein Beispiel aus der Produktion angeführt.

gemessen, stellt sich nun die Frage, ob die betreffende Maschine angehalten und gewartet werden muss oder ob es sich um eine noch tolerierbare bzw. zufällige Abweichung von dem produktionsübergreifend geforderten Durchmesser von 2mm handelt. Ein entsprechendes Szenario ergibt sich auch für die Schraubenproduktion. Da die (monetären und rechtlichen) Reklamationskosten für fehlerhaft produzierte Stents wesentlich höher sein dürften als für Möbelschrauben und die Opportunitätskosten bei der Schraubenproduktion die Reklamationskosten vermutlich weit übersteigen, werden die jeweiligen Produzentinnen die Risken I und II[12] jeweils sehr unterschiedlich wählen. In der Stentproduktion könnte das Risiko II (Reklamationskosten) etwa mit 0.000001 % sehr niedrig gewählt werden und entsprechend müsste hier ein wesentlich höheres Risiko I (Opportunitätskosten) hingenommen werden müssen. In der Schraubenproduktion würde man dagegen möglicherweise zunächst ein recht niedriges Risiko I (Opportunitätskosten) festsetzen (z. B. 0.01%), dann über die nächsten Monate beobachten, wie hoch die Reklamationskosten ausfallen und auf dieser Basis beide Risiken so aufeinander abstimmen, dass sich Reklamations- und Opportunitätskosten die Waage halten.

Stichprobengröße

Die letzte Information, die für die Entscheidungsschablone nach subjektiven Gesichtspunkten festgelegt werden kann, wenn auch in eingeschränktem Ausmaß, ist die Stichprobengröße. Je nachdem, wie kostenintensiv eine Untersuchung der Stichprobe ist, kann man eine möglichst kleine Stichprobengröße anstreben. Während etwa für die Werbewirkungsforschung die Stichprobengröße keinen besonderen Kostenpunkt darstellt, ist dies bei der Crash-Testung von Fahrzeugen (oder anderen Untersuchungen, bei denen die Elemente der Stichprobe zerstört werden) anders. Auch bei sehr aufwendigen Untersuchungen der Stichprobenelemente, z. B. der Überprüfung der Funktionstüchtigkeit von Magnetresonanztomographen kann es von hohem Interesse sein, die Stichprobengröße möglichst klein zu halten.

Zu erwartende Streuung

Als letzte Information für die Entscheidungsschablone wird schließlich die zu erwartende Streuung der einzelnen Messwerte berücksichtigt. Wenn diese Populationsstreuung σ nicht bekannt ist, was bei fast allen nicht-produktionsbezogenen Fragestellungen der Fall sein dürfte, kann alternativ der Effekt (der Unterschied zwischen den beiden statistisch formulierten Entscheidungsoptionen) in standardisierter Form formuliert werden. Für die Mittelwertdifferenz ist der standardisierte Effekt zum Beispiel:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{2}$$

In der standardisierten erwarteten Effektgröße δ ist die Streuung in der Population damit enthalten.

Anwendung der Entscheidungsschablone

Da die vier Informationen Effekt, Risiken, Streuung und Stichprobengröße multiplikativ miteinander verknüpft sind (siehe Formel 1) und damit verschiedene, aber nicht beliebig wählbare Ausprägungen der vier Komponenten zu demselben Endwert (Wahrscheinlichkeitsverteilungswert) führen können, müssen in der subjektiven Festsetzung einzelner Schabloneneigenschaften teilweise Kompromisse formuliert werden. Will man etwa zwischen zwei ähnlichen Entscheidungsoptionen wählen (d. h., ist der Effekt klein), gleichzeitig aber beide Risiken geringhalten und möglichst wenige Elemente untersuchen, so kann man entweder darauf hoffen, dass die Streuung der Einzelwerte in der Population sehr gering ist oder aber man muss doch eine größere Stichprobe in Kauf nehmen (auch *a priori Poweranalyse* oder Bestimmung des *optimalen Stichprobenumfangs* genannt). Oder aber man ist doch gezwungen, die Risiken höher zu setzen. Hat man eine akzeptable Entscheidungsschablone erstellt, d. h. die *Testplanung* abgeschlossen, kann der Entscheidungsprozess starten. Auf die oben genannten Produktionsbeispiele bezogen

[12] Je nach der inhaltlichen Fragestellung können die Risiken jeweils unterschiedliche Arten von Kosten repräsentieren, z. B. kann das Risiko I Reklamations- oder auch Opportunitätskosten oder andere Arten von Kosten ausdrücken.

können nun in regelmäßigen Abständen Stichproben derselben Größe gezogen und untersucht und entsprechend p-Werte bestimmt werden. Landet der p-Wert im Risiko- (bzw. Ablehnungs-) Bereich einer Entscheidungsoption, verhält man sich entsprechend der anderen oder *alternativen* Entscheidungsoption (siehe Abb. 1, berechnet und grafisch dargestellt mit den kostenlosen Programmen R, R Core Team, 2020 und RStudio, RStudio Team, 2020). Erhält man etwa bei der Kontrolle von 20 zufällig ausgewählten Stents einen durchschnittlichen Durchmesser von mehr als 2.18mm (dem *kritischen Wert*), so hält man die Produktion der entsprechenden Maschine an und wartet das Gerät. Ergibt sich bei der nächsten Zufallsstichprobe von 20 Stents ein Durchmesser von weniger als 2.18mm, wird die Produktion nicht angehalten.

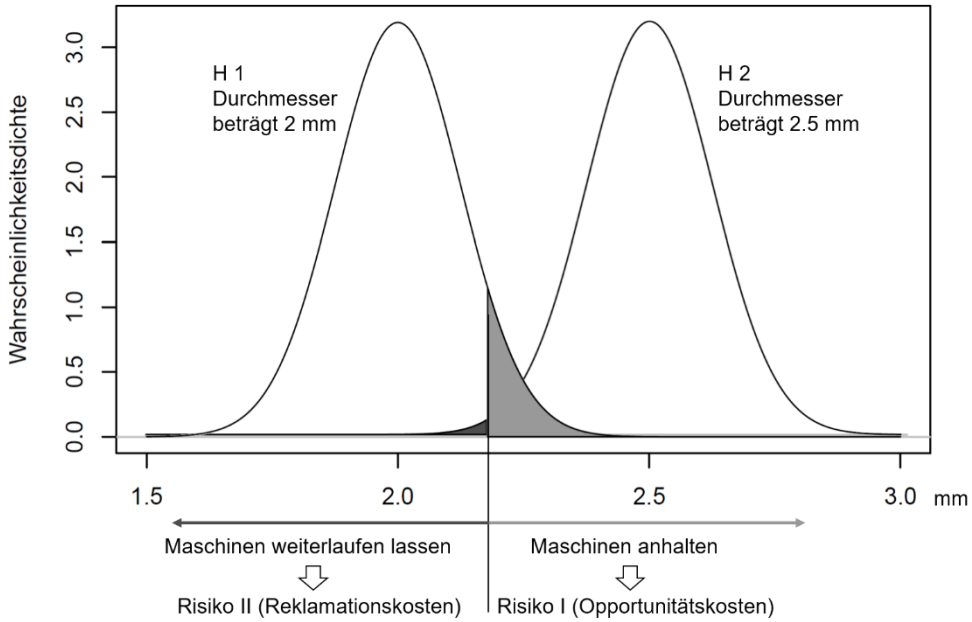

Abbildung 1. Beispiel für eine Entscheidungsschablone laut der Neyman-Pearson-Entscheidungstheorie für die Produktion von Stents mit Risiko I von 7.7 % und Risiko II von 0.5 % und einem Entscheidungsgrenzwert von 2.18mm.

Dieses Vorgehen garantiert nun *auf lange Sicht*, dass von allen Entscheidungen, die ich anhand dieser Entscheidungsschablone treffe, höchstens Risiko I % bzw. Risiko II % falsch getroffen wurden. Für die Entscheidungsschablone aus Abb. 1 wäre etwa garantiert, dass, von allen Fällen, in denen ich die Maschinen angehalten habe (allgemeiner: mich gemäß der Hypothese 2 verhalten habe), es sich in höchstens 7.7 % um eine falsche Entscheidung gehandelt hat (in Wirklichkeit produziert die Maschine im Schnitt Stents mit einem Durchmesser von 2mm). Entsprechend wäre garantiert, dass, von allen Fällen, in denen ich die Maschinen nicht angehalten habe (allgemeiner: mich gemäß der Hypothese 1 verhalten habe), es sich in höchstens 0.5 % um eine falsche Entscheidung gehandelt hat (in Wirklichkeit produziert die Maschine im Schnitt Stents mit einem Durchmesser von 2.5mm oder mehr). Diese garantierte Einhaltung der beiden Risiken gilt jedoch nicht für eine Einzeluntersuchung oder beliebig viele Untersuchungen, sondern nur für den unendlichen Fall. Nur aus Plausibilitätsüberlegungen, nicht in mathematisch-logisch begründbarer Weise, kann man erwarten, dass die Risiken auch im endlichen Fall, z. B. bei 1 000 oder 10 000 Stichprobenziehungen eingehalten werden.

Entscheidend ist nun die empirische Kontrolle und ggf. Anpassung der Entscheidungsschablone, die zunächst keine Besonderheiten einzelner Maschinen (Alter, Empfindlichkeit gegenüber Temperaturschwankungen usw.) oder sonstiger Faktoren (z. B. Reklamationsbereitschaft von Abnehmern) beachten kann. Nach einem Produktionszyklus, z. B. am Ende eines Quartals oder eines Geschäftsjahres, kann nun überprüft werden, wie groß die Reklamations- und Opportunitätskosten jeweils tatsächlich gewesen sind und ob die Entscheidungsschablone entsprechend angepasst werden muss. So könnten etwa höhere Opportunitätskosten als angenommen beobachtet worden sein und in der Folge sollte das entsprechende Risiko niedriger angesetzt werden. Wurden mehr (oder weniger) Reklamationskosten beobachtet als in der Entscheidungsschablone angesetzt, kann entsprechend das dazu passende Risiko verringert (oder erhöht) werden. Auch könnten

Anpassungen hinsichtlich der Stichprobengröße oder des Effekts vorgenommen werden. Sämtliche Anpassungen müssen dann jedoch für eine bestimmte Periode konstant gehalten werden, denn jede Veränderung einer einzelnen Information führt dazu, dass sich die Entscheidungsschablone „verzieht": Beschließt ein Produktionsschichtleiter etwa, statt 20 Stents 21 oder 19 per Zufall auszuwählen, so ändert dies den Standardfehler und damit den kritischen Wert und in der Folge auch die beiden tatsächlichen Risiken. Jede Veränderung dieses aufeinander abgestimmten Systems von Effekt, Risiken, Stichprobengröße und Streuung resultiert dann in einer neuen, etwas veränderten Entscheidungsschablone.

Aussagekraft: Wissen über die Typizität des Stichprobenergebnisses

Die Frage nach der Typizität des Stichprobenergebnisses wird im Neyman-Pearson-Ansatz zum einen durch wiederholte Testungen von Zufallsstichproben umgesetzt. Zum anderen kann die Typizität des empirischen Ergebnisses durch die Referenz auf gehaltvolle Erwartungswerte eingeschätzt werden. Sowohl der Durchmesser von 2mm als auch der Durchmesser von 2.5mm können sehr konkret und relativ eindeutig (physiologisch und technisch) begründet werden. Beide Aspekte, die wiederholten Testungen von Zufallsstichproben aus derselben Population und recht eindeutig begründbare rivalisierende Erwartungswerte, sind jedoch in den meisten (wirtschafts-) psychologischen Forschungskontexten sehr selten, wenn überhaupt gegeben oder umsetzbar. Als Produktionsleiterin weiß ich, dass Stents mit einem Durchmesser von 2mm typisch sind oder zumindest sein sollten – weil das der Hersteller der Produktionsmaschinen garantiert. In der Werbewirkungsforschung kann ich keine vergleichbare Annahme über die Typizität der Werbewirkung formulieren – meine Frage richtet sich gerade darauf, *was* das typische Ergebnis ist. Die Bedeutung der Nullhypothese bzw. des Effekts im Produktionskontext („Stents werden mit Durchmesser von 2mm produziert" – das weiß ich) unterscheidet sich damit deutlich von der im Forschungskontext („Es gibt keine Wirkung" – das weiß ich nicht). Ein untypisches Ergebnis kann daher im Produktionskontext wesentlich besser identifiziert werden: Ein einzelner Stent mit einem Durchmesser von 2.5mm oder eine Stichprobe von Stents mit einem durchschnittlichen Durchmesser von 2.15mm sind definitiv untypische Ergebnisse. Hier stellt sich lediglich die Frage, wie risikofreudig ich bin, um mit dieser Information weiter umzugehen. Hat die Werbemaßnahme aber die Kaufbereitschaft bei einer Person um einen Skalenpunkt oder bei einer Gruppe von Personen um 0.1 Skalenpunkte erhöht, so weiß ich nicht, ob es sich hierbei um ein typisches oder untypisches Ergebnis handelt.

Es gibt Vorschläge, die Risiken in einer Einzeluntersuchung als Angabe über mögliche Welten und damit die Einzeluntersuchung als eine unendlich lange Untersuchungsreihe zu verstehen (z. B. Lambert, 2018): Wenn man annimmt, dass wir in einem Multiversum nach der Viele-Welten-Interpretation der Quantentheorie leben, so würden wir in 0.5 % (= Risiko) aller möglichen, unendlich vielen Welten, in denen diese Untersuchung durchgeführt worden ist, eine falsche Entscheidung treffen. Eine solche Interpretation scheint jedoch keinen inkrementellen praktischen Nutzen für die Beurteilung der Typizität des beobachteten Ergebnisses oder die Beurteilung der inhaltlichen Hypothese in unserem Universum zu bringen.

C: Konfidenzintervall

Ein Konfidenzintervall (KI oder CI für *confidence interval*) gibt den Bereich von solchen Erwartungswerten an, auf deren Wahrscheinlichkeitsverteilungen der Stichprobenkennwert aus einer konkreten Untersuchung zu den X % häufigsten (d. h. wahrscheinlichsten) gehört. Um das Konfidenzintervall mit dem Konfidenzkoeffizienten von 95 % für die Mittelwertdifferenz $e = 1$ aus der fiktiven Untersuchung zu bestimmen, wird jeweils der kleinstmögliche und der größtmögliche Erwartungswert berechnet, in dessen Population diese Mittelwertdifferenz noch gerade zu den 95 % häufigsten gehört (bei einer Zufallsziehung von jeweils 50 Personen aus einer Population mit der angenommenen Populationsstreuung). Hierfür finden der Standardfehler und die Wahrscheinlichkeitsverteilungswerte der passenden Wahrscheinlichkeitsverteilung Verwendung:

$$\mu_{e_{untere\,Grenze}} = e - t_{2.5\%;\,df=98} \cdot SE_e = 1 - 1.98 \cdot 0.4 = 0.21$$

$$\mu_{e_{obere\,Grenze}} = e + t_{97.5\%;\,df=98} \cdot SE_e = 1 + 1.98 \cdot 0.4 = 1.79$$

Die in diesem Fall passenden Wahrscheinlichkeitsverteilungswerte sind diejenigen t-Werte einer t-Verteilung mit 98 Freiheitsgraden und einem Nonzentralitätsparameter von 0, die auf dieser Verteilung jeweils nach unten und nach oben 2.5 % der Fläche abschneiden und damit 95 % der Fläche in der Mitte der Verteilung abgrenzen. Das Konfidenzintervall für das Werbewirkungsbeispiel lautet damit: 95 %-KI [0.21; 1.79]. Dieses Intervall gibt an, dass in allen (mathematischen) Populationen von Mittelwertdifferenzen, deren Erwartungswerte jeweils 0.21 bis 1.79 Skalenpunkte betragen, die Mittelwertdifferenz von einem Skalenpunkt zu den 95 % wahrscheinlichsten (auf lange Sicht häufigsten) gehört. Hätte die wahre Mittelwertdifferenz also tatsächlich einen (beliebigen) konkreten Wert aus dem Zahlenbereich von 0.21 bis 1.79 Skalenpunkten ($0.21 \leq \mu_e \leq 1.79$, z. B. $\mu_e = 0.98$ Skalenpunkte), so würde eine Mittelwertdifferenz von einem Skalenpunkt zu den 95 % der am häufigsten in einer unendlichen Reihe von Untersuchungen auftretenden Mittelwertdifferenzen gehören (wenn man per Zufall je 50 Personen aus Populationen mit solchen Erwartungswerten ziehen würde):

$$p(e = 1\ ^{13} | 0.21 \leq \mu_e \leq 1.79) = .95$$

D: Bayesianische Ansätze

Innerhalb der frequentistischen Statistikschule werden Populationsparametern (Hypothesen bzw. Erwartungswerten von Stichprobenkennwerten) keine Wahrscheinlichkeiten zugeschrieben. Der Grund dafür liegt in dem Wahrscheinlichkeitsverständnis innerhalb der frequentistischen Schule: Hier wird Wahrscheinlichkeit als Auftretenshäufigkeit auf lange Sicht verstanden, d. h. für den unendlichen Fall. Dies hat zur Folge, dass zwar die Aussage über die „Auftretenshäufigkeit einer Mittelwertdifferenz von einem oder mehr Skalenpunkten in unendlich vielen Untersuchungen, wenn die Nullhypothese zutrifft" sinnvoll gemacht werden kann, jedoch nicht eine Aussage der Art „Auftretenshäufigkeit der Nullhypothese in unendlich vielen (Null-) Hypothesen". Da die Nullhypothese diesem Verständnis nach entweder zutrifft oder nicht, gibt es hier auch keine Vorstellung darüber, was die „Population der möglichen (Null-) Hypothesen" sein sollte. Eine solche Population wäre jedoch erforderlich, um Hypothesen eine frequentistische Wahrscheinlichkeit zuzuweisen. Auf das Werbewirkungsbeispiel bezogen, würde die Aussage „es ist sehr unwahrscheinlich (zu 0.7 %), eine um einen Skalenpunkt höhere Kaufbereitschaft in der Werbung-Gruppe zu beobachten, wenn die Werbung keinen Effekt hat" sinnvoll formuliert werden können. D. h., eine Mittelwertdifferenz von einem Skalenpunkt oder mehr würde – auf lange Sicht – in 0.7 % der Zufallsziehungen von Mittelwertdifferenzen aus einer Population aller möglichen Mittelwertdifferenzen mit dem Erwartungswert 0 auftreten. Eine Aussage der Art „es ist sehr unwahrscheinlich (zu 0.7 %), dass die Werbung keinen Effekt hat" wäre jedoch nicht sinnvoll zu treffen: Auf die Auftretenshäufigkeit von was beziehen sich die 0.7 %? Da die Auftretenshäufigkeit hier also nicht sinnvoll gedacht werden kann, können Wahrscheinlichkeitsaussagen über Hypothesen (oder allgemeiner gesagt, Populationsparameter) nicht sinnvoll getroffen werden.

D1: Grundgedanke und Bayes-Theorem

Die Wahrscheinlichkeitsrevision erfolgt anhand des Bayes-Theorems. Da das Bayes-Theorem mit beliebigen Wahrscheinlichkeiten (d. h. auch frequentistischen) anwendbar ist, soll zunächst ein (fiktives) klassisches Beispiel für die Anwendung des Theorems gegeben werden. Nehmen wir an, dass eine Person in einem Personalauswahlverfahren als geeignet für eine Position klassifiziert wurde. Es stellt sich nun die Frage, wie groß die Wahrscheinlichkeit dafür ist, dass diese als geeignet klassifizierte Person tatsächlich für die Stelle geeignet ist (sogenannter Posterior). Die Anwendung des Bayes-Theorems in Verbindung mit der Information darüber, wie viele geeignete Personen mit dem Personalauswahlverfahren als geeignet klassifiziert werden (hier 90 %), wie viele nicht geeignete Personen als geeignet klassifiziert werden (hier 5 %) und wie wahrscheinlich sich

[13] Diese Darstellungsform ($e = 1$) ist formal betrachtet aus verschiedenen Gründen nicht korrekt. Streng genommen wird hier ebenfalls ein Bereich von Werten angegeben und lediglich spezifiziert, dass dieser Bereich die Zahl 1 enthält. Der Einfachheit halber und für die maximale Vergleichbarkeit mit der Darstellung des p-Werts wird jedoch diese formal fehlerhafte Darstellung genutzt.

eine beliebige Person überhaupt eignet (hier 50 %, sogenannter Prior) ergibt diese Wahrscheinlichkeit:

$$p(G|K+) = \frac{p(K+|G) \cdot p(G)}{p(K+)} = \frac{p(K+|G) \cdot p(G)}{p(K+|G) \cdot p(G) + p(K+|NG) \cdot p(NG)} = \tag{3}$$

$$\frac{.90 \cdot .50}{.90 \cdot .50 + .05 \cdot .50} = .95$$

mit G: geeignet, NG: nicht geeignet und $K+$: als geeignet klassifiziert. Die Wahrscheinlichkeit, dass eine als geeignet klassifizierte Person tatsächlich geeignet ist, ist also recht hoch. In diesem Beispiel ist die Basisrate, d. h. der Anteil generell geeigneter Personen (unter den Bewerberinnen und Bewerbern) bei 50 %. Ist die Basisrate jedoch besonders hoch oder besonders niedrig, können die Wahrscheinlichkeiten $p(K+|G)$ (Anteil der als geeignet Klassifizierten unter den Geeigneten) und $p(G|K+)$ (Anteil der Geeigneten unter den als geeignet Klassifizierten) erheblich voneinander abweichen. Nehmen wir etwa eine Basisrate von 5 % an, dann beträgt die Wahrscheinlichkeit nur ca. 49 %, dass die als geeignet klassifizierte Person tatsächlich geeignet ist – obwohl in beiden Fällen die Wahrscheinlichkeit, geeignete Personen als solche zu klassifizieren 90 % beträgt.

Auf das Werbewirkungsbeispiel – extrem vereinfacht – übertragen würde man aus der

- Wahrscheinlichkeit, eine Mittelwertdifferenz von einem Skalenpunkt (oder einer noch größeren) zu erhalten, wenn die wahre Mittelwertdifferenz 0 Skalenpunkte beträgt

- die Wahrscheinlichkeit (d. h. die Zuversicht oder den Glauben) dafür, dass die wahre Mittelwertdifferenz 0 Skalenpunkte beträgt angesichts des in der Untersuchung beobachteten Mittelwertdifferenz von einem Skalenpunkt ermitteln:

$$p(\mu_e = 0|e \geq 1) = \frac{p(e \geq 1|\mu_e = 0) \cdot p(\mu_e = 0)}{p(1 \geq 2)} = \tag{4}$$

$$\frac{p(e \geq 1|\mu_e = 0) \cdot p(\mu_e = 0)}{p(e \geq 1|\mu_e = 0) \cdot p(\mu_e = 0) + p(e \geq 1|\mu_e \neq 0) \cdot p(\mu_e \neq 0)} =$$

$$\frac{p(D|H_0) \cdot p(H_0)}{p(D|H_0) \cdot p(H_0) + p(D|H_1) \cdot p(H_1)}$$

$$\frac{.007 \cdot .50}{.007 \cdot .50 + .99 \cdot .50} = .007$$

mit D: Daten, H_0: Nullhypothese (ungerichtet) und H_1: Alternativhypothese (ungerichtet), die Wahrscheinlichkeit für die Daten, wenn es eine Werbewirkung gibt, halten wir für sehr hoch (99 %).

In Wirklichkeit gestaltet sich der numerische Umgang mit den Wahrscheinlichkeiten für die Hypothesen (der Faktor im Zähler des Bruchs in Gleichung 4) und die Wahrscheinlichkeiten für die Daten (der Nenner des Bruchs in Gleichung 4) wesentlich komplexer, da hier unter anderem nicht einzelne Wahrscheinlichkeiten, sondern Likelihoods und deren Verteilungen geschätzt werden müssen. Als Likelihoods werden Wahrscheinlichkeiten bezeichnet, die sich nicht zu 1 aufsummieren. Ein einfaches Beispiel hierfür lässt sich für den Münzwurf geben. Stellen wir uns vor, wir haben eine Münze geworfen und als Ergebnis „Kopf" erhalten. Wenn die Münze fair, d. h. nicht gezinkt ist, dann beträgt die Wahrscheinlichkeit für „Kopf" 50 %. Ist die Münze unfair und fällt häufiger auf die „Kopf"-Seite, dann beträgt die Wahrscheinlichkeit für „Kopf" mehr als 50 %, z. B. 70 %. Ist die Münze unfair und fällt häufiger auf die „Zahl"-Seite, dann beträgt die Wahrscheinlichkeit für „Kopf" weniger als 50 %, z. B. 30 %. Diese Wahrscheinlichkeiten – 50 %, 70 % und 30 % – beschreiben alle die Wahrscheinlichkeit für „Kopf" für Szenarien, in denen die Münze unterschiedliche „Gezinktheitsgrade" aufweist. Da sich 50 %, 70 % und 30 % nicht auf 100 % summieren, werden solche Wahrscheinlichkeiten Likelihoods genannt. Die genaue Bestimmungsweise der Likelihood-Funktionen ist mathematisch recht komplex, so dass auf eine Darstellung der technischen Details verzichtet wird. Die interessierte Leserschaft sei für ausführlichere

Darstellungen auf Wagenmakers et al. (2018a, 2018b), Sedlmeier & Renkewitz (2018), Kruschke (2013) oder insbesondere Lambert (2018) verwiesen. Das grundlegende Argument dieses Beitrags bleibt von diesen technischen Details jedoch unberührt.

D2: Bayes-Faktor

Der Bayes-Faktor (BF) vergleicht die Likelihood (p') für die Daten auf der Nullhypothese mit der Likelihood für die Daten auf der Alternativhypothese:

$$BF_{10} = \frac{p'(D|H_1)}{p'(D|H_0)}$$

mit BF_{10}: „1" Alternativhypothese im Zähler und „0" Nullhypothese im Nenner. Auf das Werbewirkungsbeispiel bezogen ergibt sich ein Bayes-Faktor von 3.32 bei ungerichteten Hypothesen und ein Bayes-Faktor von 6.57 bei gerichteten Hypothesen (Abb. 2, berechnet und grafisch dargestellt mit dem kostenlosen Programm JASP, Version 0.13.1, JASP, 2020). Beide Bayes-Faktoren zeigen an, dass die Alternativhypothese (gerundet) drei bzw. sieben Mal besser zu den Daten passt im Vergleich zur Nullhypothese. Multipliziert man den Bayes-Faktor mit dem Verhältnis der Priors (für die Null- und die Alternativhypothese), erhält man das Verhältnis der posterior-Wahrscheinlichkeiten (für die Null- und die Alternativhypothese) (Wagenmakers et al., 2018a).

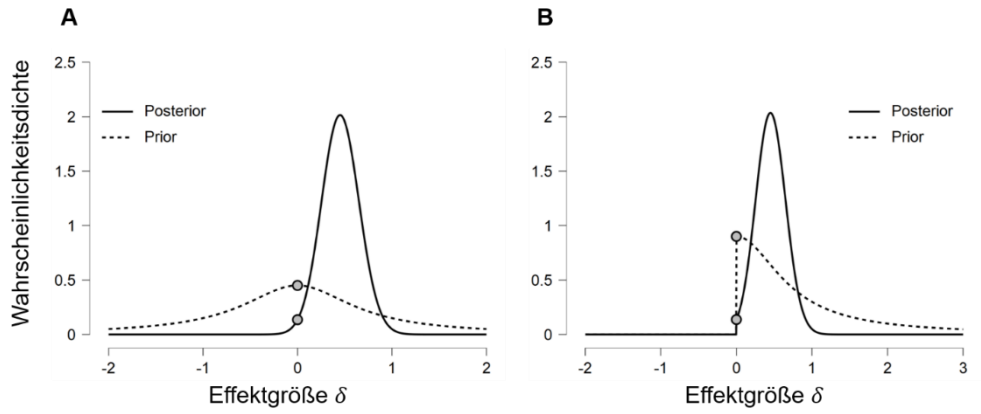

Abbildung 2. Beispiel für einen Bayes-basierten Test mit einer ungerichteten (A) und gerichteten (B) Alternativhypothese, mit Prior: Wahrscheinlichkeit für die Hypothesen vor der Datensammlung, Posterior: Wahrscheinlichkeit für die Hypothesen nach der Datensammlung, A: Bayes-Faktor $\mathbf{BF_{10} = \frac{p'(e\leq|1||H_1)}{p'(e\leq|1||H_0)} = 3.32}$**, B: Bayes-Faktor** $\mathbf{BF_{10} = \frac{p'(e=1|H_1)}{p'(e\geq1|H_0)} = 6.57}$ **wobei der Bayes-Faktor das Verhältnis der Likelihood der Daten auf der Alternativhypothese (grauer Kreis auf der Posterior-Verteilung, „1") zu der Likelihood der Daten auf der Nullhypothese (grauer Kreis auf der Prior-Verteilung, „0") angibt.**

Korrespondenzadresse:
Dr. Alla Sawatzky
Hochschule Fresenius
Im Mediapark 4c
50670 Köln
DEUTSCHLAND
alla.sawatzky@hs-fresenius.de