Kritik am Standardvorgehen des Nullhypothesen-Testens

Kritik am Nullhypothesentesten

Einführende Literatur

📖 Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Beltz. [Kap. 8.1 & 8.2]

📖 Kline, R. B. (2005). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington: American Psychological Association. [Kap. 3]

Weiterführende Literatur

📖 Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587-606.

📖 Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-301.

NHST

- Das dargestellte Vorgehen des Nullhypothesentestens (NHST= Null Hypothesis Significance Testing) ist in der Psychologie eine Standardvorgehensweise bei der Auswertung von Studien. Cumming et al. (2007) kommen bei einer Analyse der Artikel von 10 führenden internationalen Zeitschriften zu dem Schluss, dass in ca. 97% das NHST-Vorgehen zum Einsatz kam.
- Das NHST Standard-Vorgehen ist aber nicht ohne Kritik geblieben (von einigen als „Null-Ritual“ bezeichnet). Einige Probleme sollen im Folgenden dargestellt werden.

„*Sir Ronald had befuddled us, mesmerized us, and led us down the primrose path. I believe the almost universal reliance on merely refuting the null hypothesis … is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.*“ (*Paul Meehl, 1978, p. 817*)

gemeint ist Ronald Fisher, einer der „Väter“ des NHST

„*The earth is round (p < .05)*“ (*Titel eines Artikels von Jacob Cohen, 1994*)

NHST

- Einige Kritikpunkte am Vorgehen des NHST sind:
  - Die meisten Studierenden und Lehrenden interpretieren ein statistisch signifikantes Ergebnis fehlerhaft.
  - Viel wichtiger als die Signifikanz eines Ergebnisses ist, wie stark der Effekt ist. Oder: Bei sehr großem n können auch triviale, minimale Effekte statistisch signifikant werden.
  - Die Voraussetzungen vieler statistischer Verfahren (Normalverteilung und Varianz-homogenität) sind häufig verletzt, was zu fehlerhaften Entscheidungen führen kann. (Außerdem werden die Voraussetzungen zu selten geprüft.)
  - Die in der Regel getesteten Nullhypothesen sind unrealistisch, kleine Effekte sind eigentlich immer zu erwarten.
  - Viele Studien haben so geringe Stichprobengrößen, dass sie bei der zu erwartenden Stärke der Effekte gar keine Chance haben, die Nullhypothese zurückzuweisen (d.h. die Power ist zu gering)

NHST

- Die folgende Befragung geht auf Oakes (1986) zurück und wurde seither mehrfach mit Studierenden und Forschern wiederholt (z.B. Haller & Krauss, 2002).

„Sie stellen mittels eines t-Tests fest, dass der zwischen zwei Stichproben gefundene Unterschied auf dem 1% Niveau statistisch signifikant ist. Welche der folgenden Aussagen lassen sich nun aus dieser Tatsache folgern?“ (Es können dabei alle oder auch keine Aussage zutreffen.)	Prozent Bejahung		
	O86	HKS	HKW
1. Es ist eindeutig bewiesen, dass die Nullhypothese falsch ist.	1	34	15
2. Es ist eindeutig bewiesen, dass die Alternativhypothese wahr ist.	3	20	13
3. Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden.	46	32	26
4. Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativhypothese richtig ist.	43	59	33
5. Entscheidet man sich nun die Nullhypothese zu verwerfen, dann kennt man jetzt die Wahrscheinlichkeit dafür, dass diese Entscheidung falsch ist.	69	68	67
6. Wenn man das Experiment sehr oft wiederholen würde, würde man in 99% der Fälle ein signifikantes Ergebnis bekommen.	34	41	49

O86: 70 Wissenschaftler in England (Oakes, 1986)
HKS: 44 Psychologie-Studierende verschiedener Universitäten in Deutschland (Haller & Krauss, 2002)
HKW: 39 Wissenschaftler in der Psychologie, nicht aus der Methodenlehre (Haller & Krauss, 2002)

- Alle Aussagen sind falsch!
  - Behauptungen 1 und 2 sind am offensichtlichsten falsch: Mit Sicherheit können in der Inferenzstatistik überhaupt keine Aussage getroffen werden; es werden immer nur Wahrscheinlichkeitsaussagen getroffen.
  - Die anderen Behauptungen sind falsch, weil die Inferenzstatistik Aussagen über die Wahrscheinlichkeit bestimmter Daten (bzw. daran bestimmter Prüfgrößen) macht, nicht von Hypothesen. Sie erlaubt auch keine direkten Aussagen darüber, wie wahrscheinlich ein Ergebnis erneut eintritt (sich repliziert). Der p-Wert ist nicht die Wahrscheinlichkeit, dass die Nullhypothese wahr ist! Vielmehr ist er die bedingte Wahrscheinlichkeit des Auftretens der Daten D (oder noch extremerer Daten), wenn die H_0 richtig ist, also $P(D \mid \mathrm{H}_0)$. Und nicht $P(\mathrm{H}_0 \mid D)$!

Prinzipiell kann man beide bedingten Wahrscheinlichkeiten mittels des Bayes-Theorems ineinander umrechnen:

$$P(H_0 \mid D) = \frac{P(D \mid \mathrm{H}_0) \cdot P(\mathrm{H}_0)}{P(D \mid \mathrm{H}_0) \cdot P(\mathrm{H}_0) + P(D \mid \neg \mathrm{H}_0) \cdot P(\neg \mathrm{H}_0)}$$

Da man aber in der Regel die apriori Wahrscheinlichkeit $P(\mathrm{H}_0)$ nicht kennt, kann man die interessantere Wahrscheinlichkeit $P(\mathrm{H}_0 \mid D)$ so nicht bestimmen. (Anhänger der Bayes-Statistik greifen dann auf bestimmte Schätzungen zurück.)

NHST

- Kritikpunkt: Viel wichtiger als die Signifikanz eines Ergebnisses ist, wie stark der Effekt ist. Oder: Bei sehr großem n können auch triviale, minimale Effekte signifikant werden.

- Beispiel: Zur Prüfung der Wirksamkeit eines Raucherentwöhnungsprogramms wurden 800 Raucher per Zufall auf die Trainings- (TG) und Kontrollgruppe (KG, ohne Behandlung) zugewiesen und vier Wochen nach dem Training zeitgleich in beiden gleich großen Gruppen die Zahl der durchschnittlich pro Tag gerauchten Zigaretten erfasst. In der TG ergab sich ein Mittelwert von 61.5 ($s = 14.6$) und in der KG von 63.5 ($s = 14.2$). Im t-Test für unabhängige Gruppen resultiert bei zweiseitiger Fragestellung ein statistisch signifikanter Unterschied: $t = 1.96$, $df = 798$, $p < .05$. Wir schließen, dass durch die Teilnahme am Training statistisch signifikant weniger geraucht wird (obwohl die mittlere Abnahme nur 2 Zigaretten beträgt).

- Variante: Die gleiche Studie wie oben, allerdings wurden nur 10 Raucher pro Gruppe untersucht. Alle Statistiken sind gleich, bis auf den Mittelwert der TG, der 50.0 beträgt. Im t-Test resultiert hier kein statistisch signifikanter Unterschied: $t = 2.10$, $df = 18$, ns. Wir schließen also, dass nach dem Training nicht statistisch signifikant weniger geraucht wird als vorher (obwohl die mittlere Abnahme hier immerhin 13.5 Zigaretten beträgt).

- Erwiderung: Die Kritik ist berechtigt. Der Signifikanztest prüft nur, ob sich die Mittelwerte unterscheiden (bzw. der Unterschied von 0 verschieden ist).

- Konsequenz: Deshalb ist es sinnvoll, zusätzlich zur statistischen Signifikanz Effektstärkemaße (Effektgrößen, manchmal auch als „Maße der praktischen Signifikanz“ bezeichnet) zu bestimmen und berichten. Wie diese Maße berechnet werden, hängt vom Kontext ab.
- Anwendung: Ein Maß für die Stärke eines Mittelwertunterschieds (zwischen zwei unabhängigen Gruppen) ist Hedges g. Bei diesem Maß wird der Mittelwertunterschied beider Gruppen an der gepoolten Standardabweichung relativiert (um den Unterschied unabhängig vom Maßstab von X zu machen). Im Falle von $n_1 = n_2$ gilt:

$$g = \frac{\bar{x}_{TG} - \bar{x}_{KG}}{s_g} \text{ mit } s_g = \sqrt{\frac{s_{TG}^2 + s_{KG}^2}{2}}$$

Hier: $$s_g = \sqrt{\frac{s_{TG}^2 + s_{KG}^2}{2}} = \sqrt{\frac{14.6^2 + 14.2^2}{2}} = 14.40$$

Für die Studie mit großer Stichprobe:

$$g = \frac{63.5 - 61.5}{14.40} = 0.14$$

Für die Studie mit kleiner Stichprobe:

$$g = \frac{63.5 - 50.0}{14.40} = 0.94$$

Eine grobe (!) Orientierung gibt Cohen (1988). Er klassifiziert Werte von g in diesem Kontext von

- ≈ 0.20 als „kleinen“ Effekt
- ≈ 0.50 als „mittleren“ Effekt
- ≈ 0.80 als „starken“ Effekt

- Effektstärken können über verschiedene Studien hinweg aggregiert werden (Metaanalysen).

NHST

- Kritikpunkt: Die Voraussetzungen vieler statistischer Verfahren (Normalverteilung und Varianzhomogenität) sind häufig verletzt, was zu fehlerhaften Entscheidungen führen kann. (Außerdem werden die Voraussetzungen zu selten geprüft.)
- Beobachtung: Micceri (1989) untersuchte 440 große Datensätze aus der psychologischen und pädagogischen Forschung und fand heraus, dass sich nur 7% der Variablen hinreichend einer Normalverteilung annäherten. Keselman et al. (1998) untersuchten in pädagogischen und entwicklungspsychologischen Zeitschriften (u.a.), wie stark sich Varianzen zwischen Gruppen unterscheiden und fanden ein durchschnittliches Verhältnis von 2:1 (zwischen den extremsten Gruppen, wobei 1:1 gefordert ist).
- Erwiderung: Dass Daten häufig Voraussetzungen nicht erfüllen ist unbestritten. Aber …
  - unter bestimmten Bedingungen erweisen sich auch parametrische Tests als robust gegenüber solchen Verletzungen.
  - daneben stehen parametrische Verfahren zur Verfügung, die voraussetzungsärmer sind (z.B. der Welch-Test relativ zum t-Test für unabhängige Gruppen) und darüber hinaus nonparametrische Verfahren.
  - Darüber hinaus stehen noch weitere Strategien zur Verfügung, die im Falle von Voraussetzungsverletzungen eingesetzt werden können.

NHST

- Zu solchen Strategien gehören auch (bisher noch wenig eingesetzt und z.T. von den Gegnern des klassischen NHST propagiert):
  - Transformationen der Daten können die Verstöße verringern (z.B. kann durch eine Logarithmierung von X deren Schiefe verringert werden, also eine linkssteile Verteilung „normalisiert“ werden)
  - Bei Ausreißern können z.B. robustere Statistiken berechnet werden und mit diesen dann die bekannten Tests durchgeführt werden. Z.B. können statt des Mittelwertes getrimmte Mittelwerte (d.h. auf beiden Seiten der Verteilung werden z.B. die 5% extremsten Werte abgeschnitten) oder winsorisierte Mittelwerte (d.h. auf beiden Seiten der Verteilung werden z.B. die 20% extremsten Werte abgeschnitten und durch die extremsten gerade nicht abgeschnittenen Werte ersetzt) verwendet werden.
  - Statt die Stichprobenverteilung unter bestimmten Voraussetzungen theoretisch abzuleiten, kann diese auch auf der Basis der bestehenden Daten per Computersimulation erzeugt werden (Bootstrapping).

NHST

- Kritikpunkt: In den Nullhypothesen, die in der Regel getestet werden, wird behauptet, dass Unterschiede, Zusammenhänge etc. exakt Null sind, d.h. es wird das vollständige Fehlen irgendeines Effektes postuliert (von den Kritikern auch als „nil hypothesis“ bezeichnet). Dies ist unrealistisch, kleine Effekte sind eigentlich immer zu erwarten. (Damit benötigt man nur eine ausreichend große Stichprobe, um ein signifikantes Ergebnis zu erzielen.)

- Erwiderung: Es lassen sich prinzipiell auch Nullhypothesen aufstellen und testen, die einen von Null abweichenden Effekt postulieren, z.B. bei der Produkt-Moment Korrelation H_0: $\rho = .30$ statt H_0: $\rho = 0$. Außerdem geht es ja auch um die Frage, wie stark der Effekt ist, was man mittels Effektstärke-Maßen quantifizieren kann.

- Kritikpunkt: Viele Studien haben so geringe Stichprobengrößen, dass sie bei der zu erwartenden Stärke der Effekte gar keine Chance haben, die Nullhypothese zurückzuweisen (d.h. es werden viele β-Fehler gemacht und die Power ist zu gering).

- Erwiderung: Es ist möglich die erforderliche Stichprobengröße bei einer angenommenen Effektstärke und gewünschten Power für eine Studie im Vorhinein abzuschätzen, so dass dieses Problem prinzipiell gelöst werden kann.

- Weitere Kritikpunkte:
  - Die Inferenzstatistik geht von der Annahme von Zufallsstichproben aus, die aber in der Regel in der Forschung nicht gegeben sind.
  - Bei der Publikation von Studien sind überproportional Studien repräsentiert, die statistisch signifikante Effekte gefunden haben (verursacht durch Selbstselektion der Forscher oder/und den Reviewprozess). Dies stellt eine Verzerrung dar; die Effekte werden überschätzt und die Rate von Fehler erster Art ist höher als durch die kontrollierte Irrtumswahrscheinlichkeit α suggeriert.
  - NHST ist eine Kombination (hybride Logik) von zwei verschiedenen partiell unvereinbaren statistischen Vorgehensweisen, die auf Ronald Fisher (1925, 1935, 1955, 1956) sowie Jerzy Neyman und Egon Pearson (1928, 1933, 1936) zurückgehen und von ihnen (aus jeweils unterschiedlichen Gründen) abgelehnt würden (vgl. einführend Eid, Gollwitzer & Schmitt, 2010 , Kap. 8.1 & 8.2 und weiterführend z.B. Gigerenzer, 2004)
  - Inferenzstatistik suggeriert, dass es immer um Entscheidungen für oder gegen eine Theorie geht. Tatsächlich sollte Forschung ein kumulativer Prozess sukzessiver Approximation an die Wahrheit sein, in dem die Replikation von Befunden eine wichtigere Rolle spielen sollte.

NHST: Zitierte Quellen

Zitierte Quellen

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*, 997-1003.
- Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A. et al. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, 18, 230–232.
- Haller, H. & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers? *Methods of Psychological Research*, *7*, 1–20. http://www.mpr-online.de.
- Meehl, P. (1978). Theoretical risks and tabular asteriks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 469-505.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester: Wiley.