Mutfried HARTMANN, Nürnberg

Ein Vorschlag zur Verbindung von Signifikanz und Effektstärke zu einer neuen statistischen Kenngröße

Seit Beginn der Beurteilung von Verfahren mittels Hypothesentests besteht ein heftiger Diskurs über deren Sinn bzw. korrekte Anwendung. Bereits Fisher führte mit Neyman und Pearson diesen Diskurs in teils polemischer Form (Gigerenzer et al. 2006). In den letzten Jahrzehnten wurde eine heftige Debatte über die Bedeutung von Signifikanz bzw. Effektstärke geführt (Sedlmeier 1996). Allgemeiner Konsens besteht inzwischen wohl darin, dass weder ein signifikantes Abweichen der Stichprobe, noch die aus der Stichprobe berechnete Effektstärke alleine aussagekräftig für die Bedeutung eines Effekts sind. Es hat sich die Praxis durchgesetzt, beides anzugeben. Im Folgenden soll gezeigt werden, dass auch dieses Vorgehen problematisch ist und wie stattdessen durch die Verbindung beider Größen zu einer einzigen statistischen Kenngröße das Problem gelöst werden könnte.

1. Das übliche Vorgehen

Angenommen mit einem statistischen Test soll ermittelt werden, ob ein in einer Population normalverteiltes Merkmal mit bekanntem Mittelwert μ_{Pop} und Streuung σ durch eine Behandlung auf ein höheres Niveau $\mu_{bePop} > \mu_{Pop}$ gehoben werden kann, so wird üblicherweise nur eine Stichprobe der Größe n behandelt und schließlich geprüft, ob deren Mittelwert $\mu_{Stichprobe}$ mit der Annahme $\mu_{bePop} \leq \mu_{Pop}$ hinreichend unvereinbar erscheint. Dazu wird anhand einer Prüfverteilung – in diesem Fall eine Normalverteilung der Streuung $\sigma_\mu = \frac{\sigma}{\sqrt{n}}$ – geprüft, ob $\mu_{Stichprobe}$ außerhalb eines Konfidenzbereiches liegt. Da σ_μ mit wachsendem n beliebig klein gemacht werden kann, werden bei großem n selbst irrelevant kleine Effekte signifikant. Signifikanz sagt also nichts über die Relevanz einer Behandlung aus. Entscheidend ist vielmehr die sogenannte Effektstärke ε, im Wesentlichen also der Mittelwertsunterschied $\mu_{bePop} - \mu_{Pop}$, der um Skalenunabhängigkeit zu erreichen an der Populationsstreuung relativiert wird: $\varepsilon = \frac{\mu_{bePop} - \mu_{Pop}}{\sigma}$. Diese Effektstärke ist ebenso wenig bekannt wie μ_{bePop}. Deshalb kann nur ein Schätzwert $\frac{\mu_{Stichprobe} - \mu_{Pop}}{\sigma}$ angegeben werden.

2. Das Problem

Der Schätzwert für ε ist ebenso wie $\mu_{Stichprobe}$ ein Artefakt des Zufalls, der durch den Hypothesentest in keiner Weise abgesichert wird. Die Angabe beider Größen erweckt leicht den Eindruck, dass die Signifikanz als Anhaltspunkt dafür dienen könnte, den Schätzwert für die Effektstärke ernst zu nehmen. Das ist aber eine sehr gefährliche Missinterpretation. In Wirklichkeit ist nicht viel erreicht, denn,

- das was nicht interessiert, ein unter Umständen irrelevant kleiner Effekt, wird statistisch abgesichert und
- das was interessiert, die Effektstärke, schätzt man ohne jegliche statistische Absicherung.

3. Lösungsvorschlag: Absicherung einer Mindesteffektstärke

Das Problem könnte dadurch gelöst werden, dass anstelle irgendeines, eventuell auch irrelevanten Unterschieds ein relevanter Mindestunterschied statistisch abgesichert wird. Dazu genügt es natürlich nicht zu prüfen, ob die Stichprobe mit der klassischen Nullhypothese, dass die Behandlung keinerlei Verbesserung des Merkmals bewirkt, hinreichend unvereinbar scheint. Vielmehr muss geprüft werden, ob sie sogar mit der Hypothese unvereinbar ist, die Verbesserung sei höchstens irrelevant. Denn dann würde man sinnvoller weise davon ausgehen, dass die Behandlung mindestens eine relevante Verbesserung bewirkt. Um zu prüfen, welche Hypothese gerade noch auf dem 5% -Niveau abgelehnt werden kann, verschiebt man die Prüfverteilung aus ihrer üblichen Position (Abb.1) soweit nach rechts, dass die Grenze des Konfidenzbereichs mit dem Stichrobenmittelwert zur Deckung kommt (Abb. 2).

Da bei einer behandlungsbedingten Erhöhung des Mittelwerts um weniger als $d_{5\%}$ ein solches bzw. höheres Stichprobenergebnis in weniger als 5% der Fälle zu erwarten wäre, geht man sinnvoller weise davon aus, dass für die behandelte Population $\mu_{bePop} \geq \mu_{5\%}$ gilt (vgl. Abb.2). $\varepsilon_{5\%} = d_{5\%} / \sigma$ stellt also einen auf dem 5%-Niveau abgesicherten und damit in gewissem Sinne verlässlichen Mindesteffekt dar, der mit der Behandlung erreicht wird.

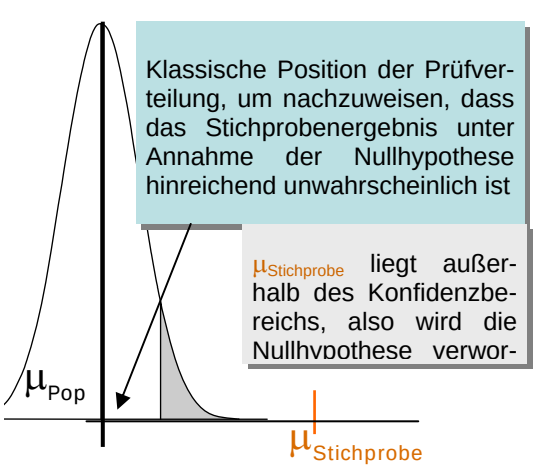


Abb. 1

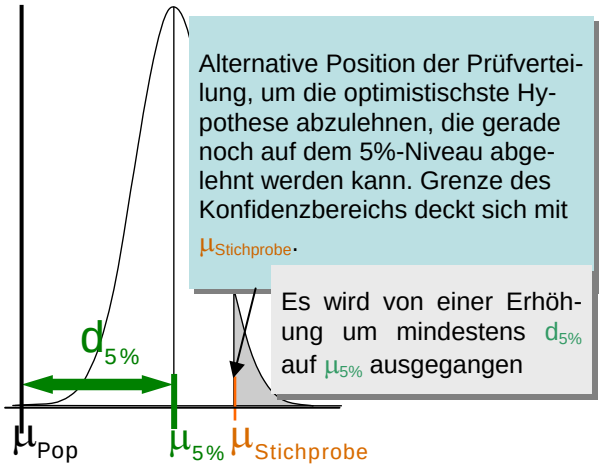
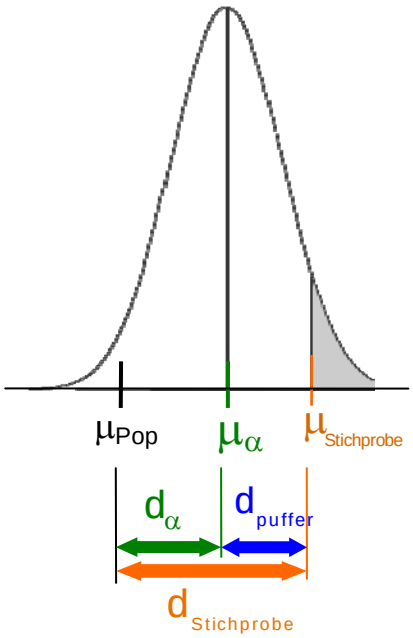


Abb.2

4. Berechnung der abgesicherten Effektstärke ε_α



Die abgesicherte Erhöhung d_α ist vom angestrebten Sicherheitsniveau α abhängig. Je höher dieses ist, umso größer wird der Sicherheitspuffer d_{Puffer} und umso kleiner die abgesicherte Distanz $d_\alpha = d_{Stichprobe} - d_{Puffer}$. Der Sicherheitspuffer beträgt für $\alpha = 5\%$ etwa $1{,}6\cdot\sigma_\mu$, für $\alpha = 1\%$ bereits etwa $2{,}3\cdot\sigma_\mu$. Allgemein liefert die Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung das entsprechende Vielfache des Standardfehlers. Es gilt also $d_{puffer} = -z(\alpha)\cdot\sigma_\mu$.

Damit erhält man für die auf dem α-Niveau abgesicherte Effektstärke:

$$\varepsilon_\alpha = \frac{d_{Stichprobe} - \left(-z(\alpha)\cdot\sigma_\mu\right)}{\sigma}$$

5. Zusammenhang der aus der Stichprobe geschätzten mit der abgesicherten Effektstärke

Setzt man $\sigma_\mu = \frac{\sigma}{\sqrt{n}}$ in obige Gleichung ein, so erhält man:

$$\varepsilon_\alpha = \frac{d_{Stichprobe} - \left(-z(\alpha)\cdot\frac{\sigma}{\sqrt{n}}\right)}{\sigma} = \frac{d_{Stichprobe}}{\sigma} - \frac{-z(\alpha)}{\sqrt{n}}$$

mit $\varepsilon_{Stichprobe} = \frac{d_{Stichprobe}}{\sigma}$ ergibt sich

$$\varepsilon(\alpha) = \varepsilon_{Stichprobe} - \frac{-z(\alpha)}{\sqrt{n}}.$$

Die abgesicherte Effektstärke berechnet sich also als Differenz aus der auf Basis der Stichprobe geschätzten Effektstärke und eines Korrekturgliedes, welches besonders dann bedeutsam wird, wenn n klein bzw. das Sicherheitsbedürfnis groß ist.

6. Ein Beispiel

Angenommen für einen Versuch an 25 Probanden wurde das Signifikanz-Niveau auf 5% festgelegt und der aus den Versuchdaten berechnete p-Wert beträgt 1,3%. Damit einher geht ein aus dem Stichprobenwert geschätzter mittlerer Effekt ($\varepsilon = 0{,}44$). Diese Effektstärke könnte aber leicht ein Zufallsartefakt sein. Auf dem 5%-Niveau ließe sich, wie folgende Rechnung zeigt, nur eine sehr kleine Effektstärke absichern:

$$\varepsilon_{5\%} = \varepsilon_{Stichprobe} - \frac{1,6}{\sqrt{n}} = 0,44 - \frac{1,6}{\sqrt{25}} = 0,44 - 0,32 = 0,12.$$

7. Zusammenfassung

Die abgesicherte Effektstärke stellt eine sowohl leicht zu berechnende als auch leicht zu interpretierende Kenngröße dar. Sie sichert nicht nur irgendeinen eventuell nur irrelevanten insbesondere unbekannten Unterschied, sondern konkrete Effekte mittels des Signifikanztests ab. Insbesondere bei kleinen Effekten oder niedrigen Probandenzahlen vermeidet diese Kenngröße Überinterpretationen der Stichprobenergebnisse.

Literatur

[1] Gerd Gigerenzer, Zeno Swijtink, Theodore Porter u. a.: Das Reich des Zufalls: Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Unschärfen. Spektrum Akademischer Verlag 1999

[2] Peter Sedlmeier: Jenseits des Signikanztest-Rituals: Ergänzungen und Alternativen. In: Methods of Psychological Research Online 1996, Vol.1, No.4