# Planung des Stichprobenumfangs

**Die Abhängigkeit des Ergebnisses eines Signifikanztests vom Stichprobenumfang führt zu einem Dilemma: Ist der Stichprobenumfang hoch, werden Effekte – zum Beispiel Unterschiede zwischen zwei Mittel- oder Anteilswerten – als signifikant ausgewiesen, obwohl sie aus praktischer Sicht irrelevant sind. Ist er gering, bleiben praktisch relevante Effekte durch einen statistischen Test unerkannt. Gibt man vor, ab welcher Stärke Effekte relevant sind, lässt sich ein optimaler Stichprobenumfang bestimmen, sodass diese Effekte mit einer hohen Wahrscheinlichkeit identifiziert werden.**

## Die Autoren

**Johannes Lüken**, Diplom-Psychologe, ist Leiter des Bereichs Data Sciences bei IfaD.

*jlueken@ifad.de*

**Prof. Dr. Heiko Schimmelpfennig**, Diplom-Kaufmann, ist Projektleiter für Data Sciences bei IfaD.

*hschimmelpfennig@ifad.de*

**Der optimale Stichprobenumfang.** Einen zu hohen Stichprobenumfang kann es aus statistischer Sicht im Grunde nicht geben. Mit steigender Stichprobengröße wird der zufällige Fehler kleiner und damit die Schätzung von beispielsweise Mittel- oder Anteilswert genauer. Alle signifikanten Effekte können im Nachhinein anhand der Effektstärke in die relevanten und irrelevanten aufgeteilt werden. Um aber unnötige Kosten zu vermeiden, stellt sich die Frage nach dem kleinsten und damit optimalen Stichprobenumfang, der ausreicht, um die praktisch relevanten Effekte zu identifizieren.

Die Power eines Tests bezeichnet die Wahrscheinlichkeit, einen in der Grundgesamtheit vorhandenen Effekt durch den statistischen Test zu erkennen. Diese ist abhängig von Effektstärke, Signifikanzniveau und Stichprobenumfang. Im Allgemeinen gibt man das Signifikanzniveau vor und definiert eine untere Grenze, ab der ein Effekt relevant ist. Zu jedem Stichprobenumfang ergibt sich dann eine bestimmte Power. Der optimale Stichprobenumfang ist der kleinste, der zur Erreichung einer gewünschten Power notwendig ist.

**Bestimmung des Stichprobenumfangs.** Die Power eines Tests lässt sich einfach mit dem Tool G*Power von Faul et al. (2007) berechnen, indem Signifikanzniveau (zumeist gleich 5 Prozent), Effektstärke und Stichprobenumfang angegeben werden. Für den Vergleich der Mittelwerte $\bar{x}_1$ und $\bar{x}_2$ von zwei unabhängigen Stichproben ist die Effektstärke definiert durch $d = \frac{\bar{x}_1 - \bar{x}_2}{s}$ . Ein kleiner Effekt liegt nach Cohen (1988) in diesem Fall ab d = 0,2 vor. Dieser entspricht beispielsweise bei einer 7-stufigen Rating-Skala einer Differenz der Mittelwerte von 0,3, wenn die beiden Gruppen gleich groß sind und die Standardabweichung der Antworten s = 1,5 beträgt.

Abbildung 1 zeigt die Abhängigkeit der Power vom Stichprobenumfang für einen 2-seitigen t-Test auf Mittelwertunterschiede bei gleich großen Gruppen. Häufig wird als Power 80 Prozent gefordert. Um mit dieser Wahrscheinlichkeit auch kleine Effekte aufzudecken, ist 788 der optimale Stichprobenumfang.

Abbildung 2 veranschaulicht den Einfluss der Effektstärke auf den optimalen Stichprobenumfang für eine Power von 80 Prozent. Dieser sinkt mit zunehmender Effektstärke. Ein mittlerer Effekt liegt ab d = 0,5 vor, ein großer ab d = 0,8. Unter gleichen Bedingungen wie zuvor entsprechen diese einer Differenz der Mittelwerte von 0,75 beziehungsweise 1,2. Um einen mittleren Effekt mit einer Wahrscheinlichkeit von 80 Prozent zu entdecken, ist ein Stichprobenumfang von 128 nötig. Für einen großen Effekt genügen 52 Befragte.

Für den Vergleich von zwei Anteilswerten $p_1$ und $p_2$ können dieselben Grenzen wie beim Vergleich der Mittelwerte zur Kategorisierung der Effektstärke herangezogen werden, wenn diese durch $h = 2 \cdot \arcsin(\sqrt{p_2}) - 2 \cdot \arcsin(\sqrt{p_1})$ mit $p_2 \geq p_1$ gemessen wird. Einem kleinen Effekt von h = 0,2 können zum Beispiel $p_1 = 0,5$ und $p_2 = 0,6$ oder $p_1 = 0,1$ und $p_2 = 0,17$ zugrunde liegen. Um einen kleinen, mittleren oder großen Effekt mittels 2-seitigem z-Test mit einer Wahrscheinlichkeit von 80 Prozent zu entdecken, bedarf es derselben Stichprobenumfänge wie beim Vergleich der Mittelwerte.

*In Ausgabe 1/2017: Varianzanalyse*

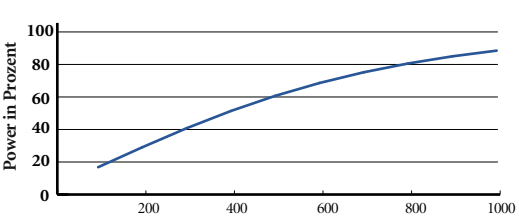
## Literatur

Cohen, J.: ***Statistical Power Analysis for the Behavioral Sciences***, 2. Auflage, Hillsdale, 1988.

Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A.: ***G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences.*** In: Behavior Research Methods, Jg. 39/2007, 2, S. 175-191.

Platz, F.; Kopiez, R.; Lehmann, M.: ***Statistische Poweranalyse als Weg zu einer ‚kraftvolleren' Musikpsychologie im 21. Jahrhundert.*** In: Auhagen, W.; Bullerjahn, C.; Höge, H. (Hrsg.): Populäre Musik, Göttingen, 2012, S. 165-179.

Entwicklung der Power für den 2-seitigen t-Test mit d = 0,2 — 1




Optimale Stichprobenumfänge für den 2-seitigen t-Test — 2