**Univ.-Prof. Dr. Georg Wydra**
**Arbeitsbereich Sportpädagogik**
**Sportwissenschaftliches Institut**
**der Universität des Saarlandes**

**Methoden der Sportwissenschaft 2**

# Arbeitsblatt 2.8: Effektgröße, Stichprobenumfang und Teststärke (aktualisiert am 27. Januar 2007)

## Einleitung

Die Logik der klassischen Signifikanzprüfung geht davon aus, dass der α-Fehler über eine Verringerung des Signifikanzniveaus und/oder eine Vergrößerung der Stichprobenumfänge verringert wird. Das Signifikanzniveau α wird auf 5 %, 1 % bzw. 0.1 % festgelegt.
Insbesondere bei größeren Stichproben wird auch der kleinste Unterschied signifikant. In Tabelle 1 ist dieser Zusammenhang dargestellt.

**Tabelle 1: Zusammenhang zwischen der Stichprobengröße n und der Irrtumswahrscheinlichkeit p für einen t-Test für unabhängige Stichproben bei konstanten Populationswerten ($M_1$=100, $SD_1$ = 10; $M_2$ = 110, $SD_2$ = 10)**

| n = | 5 | 10 | 20 | 50 |
|---|---|---|---|---|
| p = | 0.152 | 0.038 | 0.003 | 0.000 |

„Setzte der Untersuchungsaufwand der Wahl des Stichprobenumfangs keine Grenzen, wäre wohl jede $H_0$ zu verwerfen“ (Bortz & Döring, 1995, S. 563). Das ist im Sinne der klassischen Signifikanzprüfung in Ordnung. Es handelt sich um ein Problem, das vergleichbar ist, mit dem, das man sich einhandelt, wenn man das Signifikanzniveau beliebig erhöht. Bei einer Erhöhung des Signifikanzniveaus steigt die Irrtumswahrscheinlichkeit für die Ablehnung der richtigen $H_0$ (α-Fehler). Bei einer beliebigen Vergrößerung der Stichprobe besteht die Gefahr, dass Alternativhypothesen akzeptiert werden, die zwar signifikant sind (d. h. wahrscheinlichkeitstheoretisch wird ein Irrtum auf dem gewählten Signifikanzniveau ausgeschlossen), aber praktisch ohne jede Bedeutung sind. Es besteht also die Notwendigkeit,

- die Größe eines Unterschiedes zwischen zwei Messwerten unabhängig von den Stichprobengrößen zu beurteilen,
- die Wahrscheinlichkeit, mit der ein Signifikanztest einen tatsächlich existierenden Unterschied (oder Zusammenhang) erkennt, abzuschätzen und
- die optimale Größe der Stichproben zu bestimmen.

## Effektgröße

Um eine Aussage über die Größe eines Unterschiedes unabhängig von der Stichprobengröße machen zu können, benutzt man die sog. **Effektgröße**. Diese gibt an, wie groß ein Unterschied oder ein Zusammenhang sein müssen, damit dieser nicht nur signifikant, sondern auch praktisch bedeutsam ist.
Man orientiert sich dabei an der Streuung der Populationen. Wenn ein Unterschied wesentlich größer als die Streuung der Populationen ist, hat man zweifelsohne einen größeren Effekt, als wenn ein Unterschied sich in Bruchteilen der Streuung der Populationen bewegt.

**Tabelle 2: Zusammenhang zwischen der Standardabweichung SD und der Irrtumswahrscheinlichkeit p für einen t-Test für unabhängige Stichproben bei konstanten Populationswerten ($M_1$=100, $n_1$ = 20; $M_2$ = 110, $n_2$ = 20)**

| SD = | 5 | 10 | 15 | 20 |
|---|---|---|---|---|
| p = | 0.000 | 0.003 | 0.042 | 0.122 |

Es ist deutlich der Einfluss der Streuungen auf die Irrtumswahrscheinlichkeit p ersichtlich: Je größer der Unterschied zwischen zwei Mittelwerten in Relation zu den Streuungen der Populationen ist, umso eher wird ein Unterschied auch signifikant.

In Abbildung 1 ist der Einfluss der Stichprobengröße und der Streuung auf die Irrtumswahrscheinlichkeit p dargestellt. Gerechnet wurde ein t-Test für unabhängige Stichproben mit den Mittelwerten 100 bzw. 110. Dieser sehr kleine Gruppenunterschied von 10 Punkten wird bei einer Stichprobengröße von jeweils 1000 selbst bei einer Standardabweichung von 40 hochsignifikant.

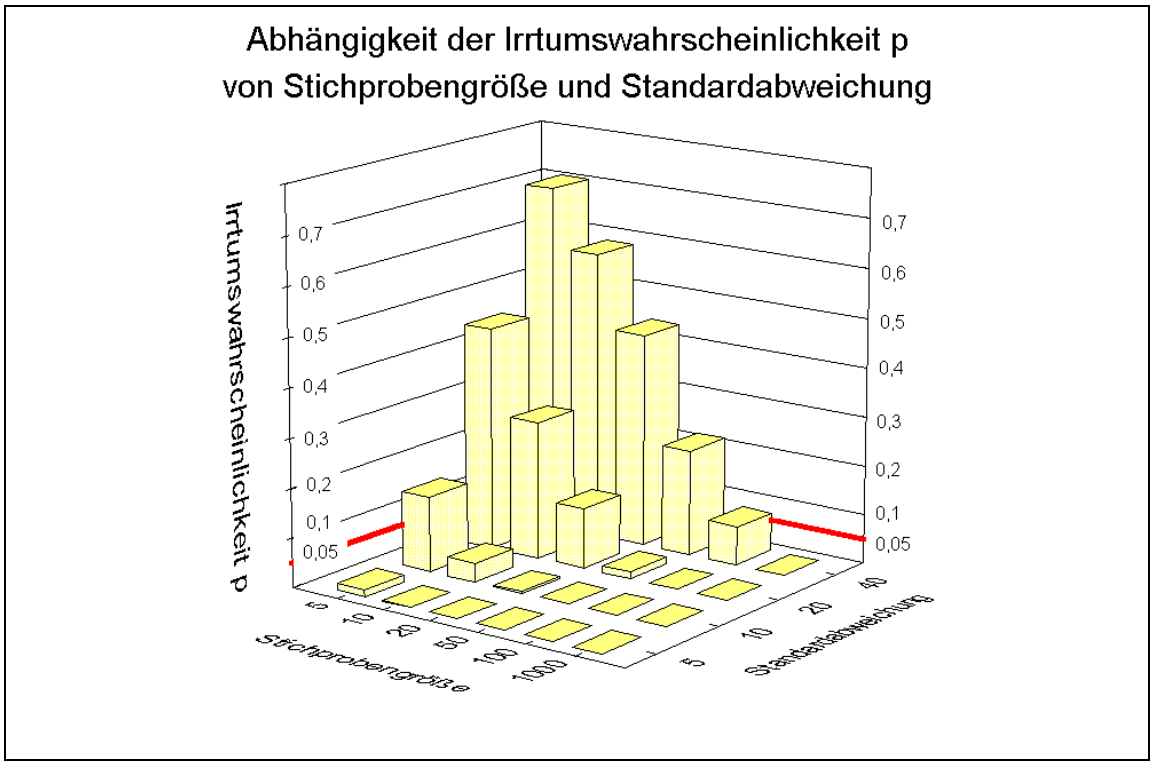


**Abbildung 1: Abhängigkeit der Irrtumswahrscheinlichkeit p von der Stichprobengröße und der Standardabweichung. Ergebnis eines t-Tests mit $M_1$=100 und $M_2$=110. Die Signifikanzgrenze von $\alpha$ = 5 % (p=0.05) ist gesondert eingezeichnet.**

**Formel 1: Effektgröße d für den t-Test für unabhängige Stichproben (Bortz & Döring, 1995, S. 568 - 569)**

$$d = \frac{\overline{X}_{EG} - \overline{X}_{KG}}{s} \qquad s = \sqrt{\frac{s^2_{EG} + s^2_{KG}}{2}}$$

**Formel 2: Effektgröße d für abhängige Stichproben (Bortz & Döring, 1995, S. 569 - 570)**

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s}\sqrt{2} \qquad s = \sqrt{s^2_1 + s^2_2 - 2r \times s_1 \times s_2}$$

In Tabelle 1 sind die Effektgrößen für die der Abbildung 1 zugrunde liegende Beispiel dargestellt.

**Tabelle 1: Veränderung der Effektgröße d bei unterschiedlich hohen Standardabweichungen (SD) und konstant gehaltenen Mittelwerten von 100 bzw. 110. Die Werte wurden mit Formel 1 berechnet.**

| **SD** | 5 | 7.5 | 10 | 12.5 | 15 | 17.5 | 20 | 25 | 30 | 40 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **d** | 2 | 1.33 | 1 | 0.8 | 0.66 | 0.57 | 0.5 | 0.4 | 0.33 | 0.25 | 0.2 |

Die Effektgröße allein ist aber kein hinreichendes Kriterium für den Erfolg einer Maßnahme, sondern muss in einem größeren Interpretationszusammenhang gesehen werden. Hierbei spielen Kosten-Nutzen-Analysen (Frage der Effizienz), Nebenwirkungen (bei diagnostischen und therapeutischen Maßnahmen) und ethische Fragen eine Rolle (vgl. Jacobs, 1998).
Auch minimale Effektgrößen sind als Erfolge zu werten, wenn durch derartige Treatments (z. B.: ein bestimmtes Medikament) einige Leben gerettet werden könnten. Aus der Diagnostik ist bekannt, das selbst geringe Effekte für bestimmte praktische Entscheidungsprobleme einen hohen Nutzen haben können: „A small effect can make a large difference“. Man sollte deshalb vor jeder entsprechenden Untersuchung versuchen, die Minimaleffekte festzulegen.

**Tabelle 2: Die vier Möglichkeiten des statistischen Entscheidungsproblems im Überblick (Jacobs, 1998).**

| | |
|---|---|
| **α-Fehler** | Nicht existierender Unterschied (oder Zusammenhang) wird als Effekt ausgegeben |
| **β-Fehler** | Vorhandener Unterschied (oder Zusammenhang) wird nicht entdeckt. |
| **1 - α** | Nicht existierender Unterschied (oder Zusammenhang) wird auch erkannt. |
| **1 - β** | Vorhandener Unterschied (oder Zusammenhang)wird entdeckt (=Teststärke oder -power). |

Die Gefahr eines α-Fehlers wird über das Signifikanzniveau kontrolliert. Je niedriger α angesetzt wird, um so größer ist die Gefahr, dass ein β-Fehler entsteht, d. h. ein vorhandener Unterschied nicht entdeckt wird. Darunter leidet die **Teststärke (Testpower)** (1 - β), d. h. die Wahrscheinlichkeit, mit der ein Signifikanztest einen tatsächlich existierenden Unterschied (oder Zusammenhang) erkennt, d. h. einen Fehler 2. Art ausschließt.
Über die Zahl der Versuchspersonen kann man Einfluss nehmen auf die Auftretenswahrscheinlichkeit eines Fehlers 2. Art. Bei einer Erhöhung der Probandenzahl werden Unterschiede bzw. Zusammenhänge zwischen zwei Messungen viel leichter als signifikant eingestuft als bei einer geringeren Probandenzahl (siehe Abbildung 1).
Somit kann man über eine Erhöhung der Stichprobengröße dafür sorgen, dass vorhandene Unterschiede auch tatsächlich entdeckt werden, d. h. die Testpower erhöht wird. Die Gefahr, dass ein nicht existierender Unterschied als Effekt ausgegeben wird (α-Fehler) wird, wie oben gezeigt wurde, über das Signifikanzniveau kontrolliert (siehe Abbildung 2).

| **Maßnahme** | | **β-Fehler** | | **Testpower (1 - β)** |
|---|---|---|---|---|
| Verringerung der Probandenzahl | ⇨ | ↑ | ⇨ | ↓ |
| Erhöhung der Probandenzahl | ⇨ | ↓ | ⇨ | ↑ |

**Abbildung 2: Auswirkungen der Veränderung der Probandenzahl auf das Auftreten von β-Fehlern und die Testpower. ↓ = Reduktion; ↑ = Erhöhung.**

Hierbei geht es nicht um eine Maximierung der Testpower, sondern um einen optimalen Wert. Eine Maximierung der Testpower ginge wiederum zu Lasten der α-Fehler-Wahrscheinlichkeit. Man geht im Allgemeinen davon aus, dass die Folgen eines Fehler 1. Art viermal so gravierend sind wie die Folgen eines Fehlers 2. Art (Bortz & Döring, 1995, S. 567). Für den α-Fehler geht man von einer Irrtumswahrscheinlichkeit von $p \leq 0.05$ aus. Für den β-Fehler erachtet man einen Wert von 0,2 als akzeptabel. Für die Testpower (1 - β) geht man deshalb von einem optimalen Wert von 0.8 aus. Die Zusammenhänge zwischen Testpower, Stichprobengröße und Signifikanzniveau ergehen aus Abbildung 3. Für die Berechnung der Testpower steht im Internet von der UCLA (University of California, Los Angeles) ein sog. Power-Calculator zur Verfügung.

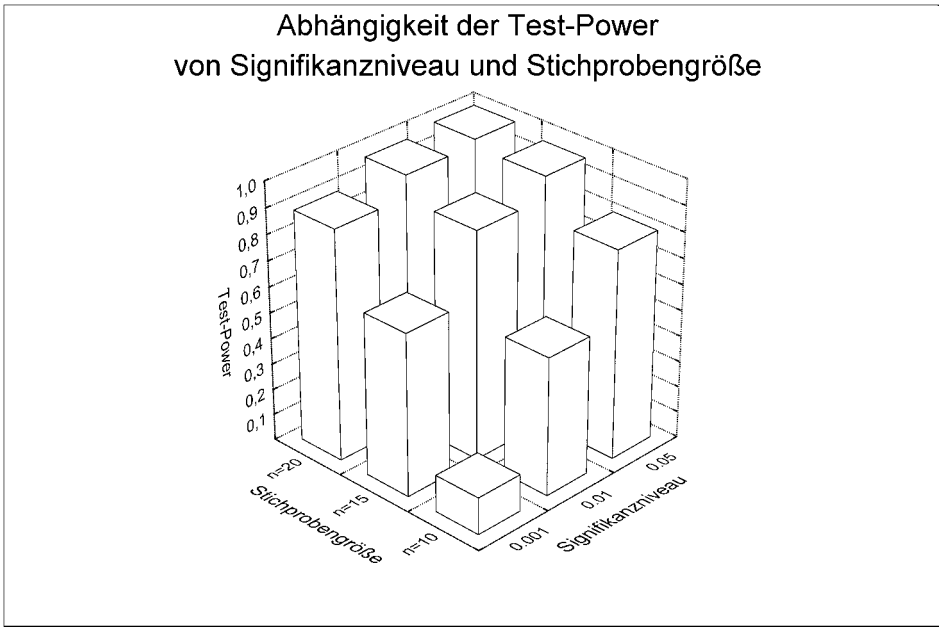


**Abbildung 3: Simulierte Veränderung der Testpower in Abhängigkeit vom Signifikanzniveau α und Stichprobengröße n bei ansonsten konstant gehaltenen Werten. Optimal ist eine Testpower von 0,8.**

Es wird deutlich, wie sich die Testpower bei einer Vergrößerung des Stichprobenumfangs und einer Erhöhung des Signifikanzniveaus erhöht. Für die optimale Testpower von 0,8 gilt es deshalb, in Abhängigkeit von der gewählten Irrtumswahrscheinlichkeit, eine optimale Stichprobengröße zu wählen.

## Optimale Stichprobengröße

Bei Kenntnis von Signifikanzniveau und Effektgröße lässt sich die **optimale Stichprobengröße** berechnen (siehe hierzu Bortz & Döring, 1995). Diese gewährleistet, dass ein Signifikanztest bei Gültigkeit der $H_1$ mit einer Wahrscheinlichkeit von 80 % zu einem signifikanten Ergebnis führt. Das Risiko einer Fehlentscheidung ($\alpha$-Fehler) entspricht hierbei dem gewählten Signifikanzniveau. Für die wichtigsten Signifikanztests und Effektgrößen geben Bortz & Döring (1995, S. 568) die optimalen Stichprobenumfänge an. Unterschieden werden kleine, mittlere und große Effektgrößen.
Während bei einer Vergrößerung der Stichprobe die optimale Testpower überschritten wird, stößt man bei einer Verringerung der Probandenzahl an Grenzen der Statistik. Zum einen erhöht sich bei einer Verringerung der Strichprobenzahl der $\alpha$-Fehler. Des weiteren führt eine zu kleine Probandenzahl dazu, dass man keine Normalverteilung mehr nachweisen kann und dann auf parametrische Testverfahren (z. B. t-Test oder Varianzanalyse) verzichten muss.

**Tabelle 3: Optimale Stichprobengröße für große, mittlere und kleine Effektgrößen für den t-Test (Bortz & Döring, 1995).**

| **Effektgröße** | groß | mittel | klein |
|---|---|---|---|
| **d** | ≥ 0.80 | ≥ 0.50 | ≥ 0.20 |
| **Stichprobengröße** | n=20 | n=50 | n=310 |

## Aufgaben

Berechnen Sie mit dem Powerkalkulator (http://calculators.stat.ucla.edu/powercalc/) für die der Abbildung 1 zugrunde liegenden Angaben ($M_1$=100, $SD_1$=10 und $M_2$=110, $SD_2$=10) unter der Voraussetzung 2 Stichproben mit Normalverteilung und gleichen Varianzen und zweiseitige Fragestellung:

- die Testpower und
- die optimale Stichprobengröße für eine Testpower von 0,8 für die Signifikanzniveaus von 5 %, 1 % und 0,1 %.

## Zusammenfassung

Die klassische Signifikanzprüfung ist nicht ohne Probleme. Insbesondere bei größeren Stichproben wird auch der kleinste Unterschied signifikant. Man sollte deshalb neben der Frage der Signifikanz auch die Größe der Effekte, die Testpower und die optimale Stichprobengröße bei der Planung und Interpretation einer Untersuchung berücksichtigen.

## Hilfen im Internet:

Im Internet stehen für die statistischen Prozeduren eine Reihe von Hilfsmitteln zur Verfügung (vgl. Jacobs, 1998). Empfehlen möchte ich folgende Seiten:

**http://www.phil.uni-sb.de/~jakobs/seminar/vpl/bedeutung/bedeut.htm#Effektstärke** - zur Berechnung der Effektgröße.

**http://www.health.ucalgary.ca/~rollin/stats/ssize/n2.html** - zur Berechnung der optimalen Stichprobengröße bzw. zur Berechnung der Testpower bei gegebener Stichprobengröße (Centre of Advancement of Health der University of Calgary)

**http://calculators.stat.ucla.edu/powercalc/** - Powerkalkulator von Jason Bond, zur Berechnung der Testpower bei verschiedenen Signifikanztests (Department of Statistics der Univesity of California, Los Angeles)

## Literatur:


Bös, K., Hänsel, F. & Schott, N. (2000). *Empirische Untersuchungen in der Sportwissenschaft. Planung - Auswertung - Statistik.* Hamburg: Czwalina.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation für Sozialwissenschaftler* (2. erweiterte Aufl.). Berlin: Springer.

Jacobs, B. (1998). *Einführung. in die Versuchsplanung* (Version 1.0). Internetauszug vom 28. November 2001, Homepage von B. Jacobs, Medienzentrum der Philosophischen Fakultät der Universität des Saarlandes: http://www.phil.uni-sb.de/~jakobs/seminar/vpl/index.htm