Statistik für Psychologen und Sozialwissenschaftler

2., aktualisierte und erweiterte Auflage

Markus Bühner
Matthias Ziegler

Pearson

EXTRAS ONLINE

Befruchtung unterzogen worden zu sein, wenn Zwillinge geboren wurden. Wenn wir dies nun auf die Entscheidung bei Hypothesen übertragen, bedeutet dies: $p(H_0|D) \neq p(D|H_0)$ und $p(H_1|D) \neq p(D|H_1)$. Über den Satz von Bayes (siehe *Abschnitt 3.7.1*) ist es möglich, $p(H_0|D)$ aus $p(D|H_0)$ unter bestimmten Vorannahmen zu berechnen:

$$p(H_1|D) = \frac{p(H_1) \cdot p(D|H_1)}{p(D)} \text{ und } p(H_0|D) = \frac{p(H_0) \cdot p(D|H_0)}{p(D)}$$

Wahrscheinlichkeiten für eine Hypothese, die auf der Basis von bereits vorliegenden Daten ermittelt werden, bezeichnet man als **A-posteriori Wahrscheinlichkeiten**. Diese werden ins Verhältnis gesetzt und ergeben einen **Wettquotienten** bzw. genauer eine **Posterior Odds** Ω (sprich Omega, großer griechischer Buchstabe).

$$\Omega = \frac{p(H_1|D)}{p(H_0|D)} = \frac{f(D|H_1)}{f(D|H_0)} \cdot \frac{p(H_1)}{p(H_0)}$$

Wenn $\Omega = 10$, ist $H_1|D$ zehnmal wahrscheinlicher als $H_0|D$.

Vorannahmen. Um die Odds

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

zu ermitteln, ist es nötig, bestimmte Vorannahmen zu treffen: Diese „stecken" im zweiten Teil der Formel:

$$\frac{p(H_1)}{p(H_0)}.$$

Mit diesem Teil wird vor der Durchführung der Datenerhebung eine Annahme darüber getroffen, wie das Verhältnis von H_0 und H_1 a priori definiert ist. Man spricht von einer **Prior Odds**. Werden im Rahmen eines Experiments mit zwei unabhängigen Gruppen H_0 und H_1 als gleichwahrscheinlich angenommen, ergibt sich der Faktor 1. Wird erwartet, dass die H_1 wahrscheinlicher als die H_0 auftritt, wird ein Faktor größer eins verwendet.

Marginale Likelihoods. Der erste Faktor setzt sogenannte marginale Likelihoods ins Verhältnis:

$$\frac{f(D|H_1)}{f(D|H_0)}.$$

Ein solcher Koeffizient wird als Bayes-Faktor BF_{10} bezeichnet, da im Zähler H_1 und im Nenner H_0 steht. Es gibt auch den Bayes-Faktor BF_{01}, bei dem im Zähler die H_0 und im Nenner die H_1 steht. Man erhält so einen Wert, der angibt, um welchen Faktor die Likelihood im Zähler die des Nenners übersteigt. Im Zähler steht in der Regel die Likelihood der Daten (D) unter H_1: $f(D|H_1)$. Im Nenner steht in der Regel die Likelihood der Daten (D) unter H_0: $f(D|H_0)$. Im Folgenden ist der BF_{10} dargestellt:

$$BF_{10} = \frac{f(D|H_1)}{f(D|H_0)} = \frac{M_1}{M_0}$$

Die Berechnung der Likelihoods ist kompliziert und wird im Exkurs nicht dargestellt. Sie findet sich bei Rouder et al. (2009). Die Berechnung mit R ist hingegen einfach und mit dem R-Paket **Bayes Factor (Morey & Rouder, 2015)** möglich:

```
ttest.tstat(t, n1, n2, rscale = "wide")
```

Dabei sind folgende Eingaben vorzunehmen:

`t` = *t*-Wert des *t*-Tests für unabhängige Stichproben

`n1` = Stichprobengröße Gruppe 1

`n2` = Stichprobengröße Gruppe 2

`medium` = kleine a priori erwartete Effektstärke

`wide` = mittlere bis große a priori erwartete Effektstärke

`ultrawide` = große a priori erwartete Effektstärke

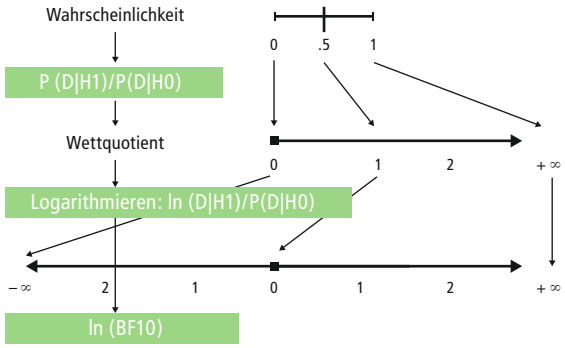
Der aus der Formel resultierende Bayes-Faktor gibt an, um welchen Faktor die beobachteten Daten unter der H_1 wahrscheinlicher sind als unter der H_0. Damit sind **Bayes-Faktoren unabhängig von den Prior Odds**. Unterschiedliche Wissenschaftler können bei gleichem Bayes-Faktor zu unterschiedlichen Posterior Odds kommen, abhängig davon, ob sie a priori die H_0 oder die H_1 für wahrscheinlicher halten und wie stark die jeweilige Präferenz ausfällt.

Allerdings hängt der Bayes-Faktor davon ab, ob wir unter der H_1 eher große oder eher kleine Effekte erwarten. Der Einfluss dieser Vorannahme fällt jedoch in der Regel klein aus. Die Vornahme ist subjektiv und spiegelt die „Vermutung" des Wissenschaftlers unter der gegebenen Fragestellung wider. In R wird die a priori erwartete Effektstärke indirekt mit dem `rscale`-Argument spezifiziert. Es stehen drei mögliche Voreinstellungen zur Verfügung. Geht man a priori von kleinen Effektstärken aus, empfiehlt sich die Einstellung „`medium`". Werden größere Effekte erwartet, ist die Einstellung „`wide`" angemessen. Je „weiter" `rscale` gewählt wird, desto stärker fällt der Bayes-Faktor zugunsten der H_0 aus. Möchte man also die Vorannahme möglichst so wählen, dass ein gewünschtes Ergebnis am „schwersten" realisiert wird, kann auch die Einstellung „`ultrawide`" sinnvoll sein.

Obwohl der oben beschriebene Bayes-Faktor sehr anschaulich ist, wird er in der Regel logarithmiert und anschließend bewertet. Dies wird im Folgenden kurz dargestellt.

Teilt man zwei Wahrscheinlichkeiten, setzt sie also ins Verhältnis, ergibt sich ein Wert von null bis plus unendlich. Man bezeichnet solche Koeffizienten auch als Wettquotienten oder Odds. Ist ein Wettquotient eins, bedeutet dies in unserem Fall, dass die Wahrscheinlichkeiten der Daten unter der Null- und Alternativhypothese gleich sind, z. B. jeweils $p = 0.50$. Je weiter der Wettquotient unter eins liegt, desto mehr Evidenz liegt für die Nullhypothese vor, je mehr der Wert über eins liegt, desto mehr Evidenz liegt für die Alternativhypothese vor. Werden die Wettquotienten logarithmiert, ergibt sich für

den Logarithmus von eins der Wert null. Werte, die unter null liegen, sprechen für die Alternativhypothese. Werte, die über null liegen, für die Alternativhypothese. Dies ist in ▶*Abbildung 4.24* veranschaulicht.



p(D|H1)	p(D|H0)	p(D|H0) /p(D|H1)	Ln(p(D|H0) /p(D|H1))
0,95	0,05	19,00	2,94
0,9	0,1	9,00	2,20
0,85	0,15	5,67	1,73
0,8	0,2	4,00	1,39
0,75	0,25	3,00	1,10
0,7	0,3	2,33	0,85
0,65	0,35	1,86	0,62
0,6	0,4	1,50	0,41
0,55	0,45	1,22	0,20
0,5	0,5	1,00	0,00
0,45	0,55	0,82	-0,20
0,4	0,6	0,67	-0,41
0,35	0,65	0,54	-0,62
0,3	0,7	0,43	-0,85
0,25	0,75	0,33	-1,10
0,2	0,8	0,25	-1,39
0,15	0,85	0,18	-1,73
0,1	0,9	0,11	-2,20
0,05	0,95	0,05	-2,94

Abbildung 4.24: Veranschaulichung des Prinzips der Bayes-Faktoren

Folgende Richtlinien haben Rouder et al. (2009) für die Interpretation der Bayes-Faktoren vorgeschlagen (Logarithmierte Grenzen des BF_{10}):

Faktor	Interpretation
> 4.6	extreme Evidenz für H_1
3.4 bis 4.6	sehr starke Evidenz für H_1
2.3 bis 3.4	starke Evidenz H_1
1.1 bis 2.3	moderate Evidenz H_1
0 bis 1.1	anekdotische Evidenz (im Sinne eines Trends) für H_1
0	keine Evidenz
−1.1 bis 0	anekdotische Evidenz (im Sinne eines Trends) für H_0
−2.3 bis −1.1	moderate Evidenz für H_0
−3.4 bis −2.3	starke Evidenz für H_0
−4.6 bis −3.4	sehr starke Evidenz für H_0
< −4.6	extreme Evidenz für H_0

Unser Beispiel führt zu folgendem Ergebnis:

```
library(BayesFactor)
ttest.tstat(3.13, 40, 40, rscale = "ultrawide")
$bf
[1] 2.377632
```

Der Wert 2.377632 stellt den logarithmierten BF_{10} dar. Es ergibt sich nach der Tabelle eine starke Evidenz für die H_1. Um den BF_{10} wieder zu erhalten, müssen wir in R folgende Syntax eingeben:

```
exp(2.377632)
```

Der in der Ausgabe resultierende Wert von 10.77935 stellt den BF_{10} dar. Damit ist, bei einer Prior Odds von 1 (H_0 und H_1 sind a priori gleich plausibel), die H_1 knapp 11-mal wahrscheinlicher als die H_0. Wir würden uns also auch mit dieser Methode für die Alternativhypothese entscheiden.

4.5 Null- oder Alternativhypothese als Wunschhypothese

Im Rahmen der Signifikanztestung müssen zwei Fälle unterschieden werden: zum einen der Fall, bei dem die Alternativhypothese die **Wunschhypothese** darstellt, und zum anderen der Fall, bei dem die Nullhypothese die Wunschhypothese darstellt. Im ersten Fall wünschen wir uns, einen Unterschied zu entdecken, und im zweiten Fall wünschen wir uns, dass kein Unterschied vorliegen möge. In beiden Fällen müssen wir unterschiedliche Fehlerwahrscheinlichkeiten entsprechend mehr beachten: Im ersten Fall muss vor allem der Fehler 1. Art kontrolliert werden und im zweiten Fall der Fehler 2. Art.

Alternativhypothese als Wunschhypothese. Beginnen wir mit dem *ersten Fall* **(Alternativhypothese als Wunschhypothese)**. Wir möchten hier darauf testen, dass es einen Effekt in der Grundgesamtheit gibt. Aus wissenschaftstheoretischer Sicht sollte man es sich besonders schwer machen, sich für die Alternativhypothese, nämlich dass es in der Grundgesamtheit einen Unterschied gibt, zu entscheiden. Ich muss also die Nullhypothese „kein Effekt in der Grundgesamtheit“ so lange wie möglich beibehalten (siehe ▶*Abbildung 4.25*). Wir müssen also den Fehler 1. Art minimieren.

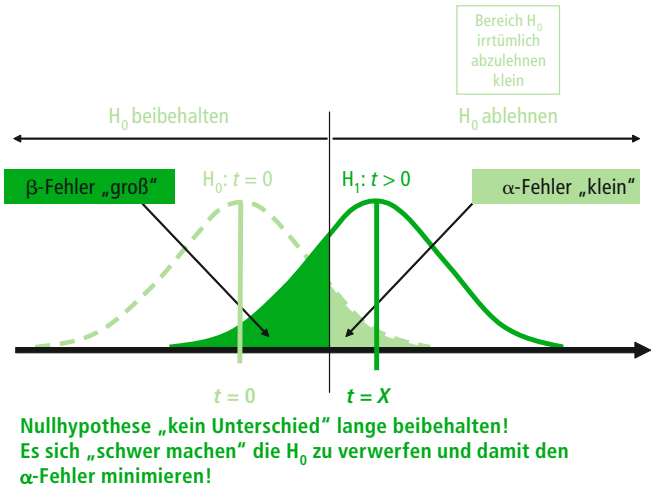


Abbildung 4.25: Alternativhypothese als Wunschhypothese

Nullhypothese als Wunschhypothese. Im *zweiten Fall* **(Nullhypothese als Wunschhypothese)** ist der Wunsch, die Annahme zu prüfen, dass es keinen Effekt in der Grundgesamtheit gibt. Daher sollte in diesem Fall die Alternativhypothese, die ausdrückt, dass ein Effekt in der Grundgesamtheit vorliegt, so lange wie möglich beibehalten werden (siehe ▶*Abbildung 4.26*). Durch dieses Vorgehen wird die Wahrscheinlichkeit für einen Fehler 2. Art auf einem akzeptablen Minimum (z. B. fünf Prozent) gehalten.

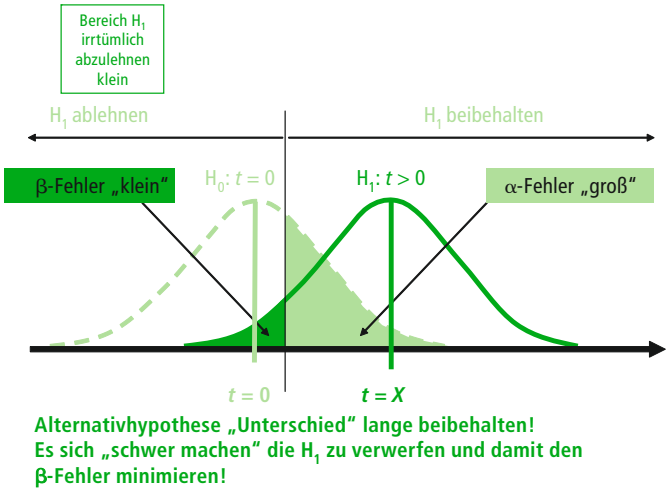


Abbildung 4.26: Nullhypothese als Wunschhypothese

Die Konsequenz ist also, dass im ersten Fall, bei dem die Alternativhypothese unsere Wunschhypothese ist, der Fehler 1. Art kleiner gewählt werden muss als der Fehler 2. Art (z. B. Fehlerwahrscheinlichkeit 1. Art: fünf Prozent, Fehlerwahrscheinlichkeit 2. Art: 20 Prozent). Im zweiten Fall, wenn die Nullhypothese unsere Wunschhypothese ist, sollte der Fehler 2. Art kleiner als der Fehler 1. Art gewählt werden (z. B. Fehlerwahrscheinlichkeit 1. Art: 20 Prozent, Fehlerwahrscheinlichkeit 2. Art: fünf Prozent). Dies ist ein wichtiges Prinzip der Hypothesentestung.

Nun haben wir uns sehr ausführlich mit den Grundlagen des Hypothesentestens auseinandergesetzt. Im Folgenden soll nun darauf eingegangen werden, wie man für verschiedene Fragestellungen eine Stichprobenplanung mithilfe eines Anwendungsprogramms durchführt.

4.6 Versuchsplanung mit G*Power und R

In diesem Unterkapitel soll an verschiedenen Beispielen gezeigt werden, wie man eine **Stichprobenplanung** durchführen und dabei die Art der Wunschhypothese berücksichtigen kann.

Praxistipp Das Programm G*Power kann unter der Internetadresse *http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/* kostenlos heruntergeladen werden.

Stichprobenplanung und Berechnung der Post-hoc-Teststärke bei zwei unabhängigen Stichproben

Fall 1: Stichprobenplanung, wenn die Alternativhypothese die Wunschhypothese darstellt (G*Power). Wir möchten zunächst eine Untersuchung zur Effektivität eines Interviewtrainings für Führungskräfte planen. Eine Voruntersuchung hat folgende Ergebnisse erzielt: Der Mittelwert der $n = 5$ Führungskräfte, die das Training absolviert haben, beträgt in einem Wissenstest $\hat{\mu} = 18.5$ mit $\hat{\sigma} = 6$. Der Wert einer zweiten Gruppe von Führungskräften $n = 8$, die in der Zeit des Interviewtrainings einen Betriebsausflug absolvierten, war $\hat{\mu} = 11.5$ mit $\hat{\sigma} = 9$.

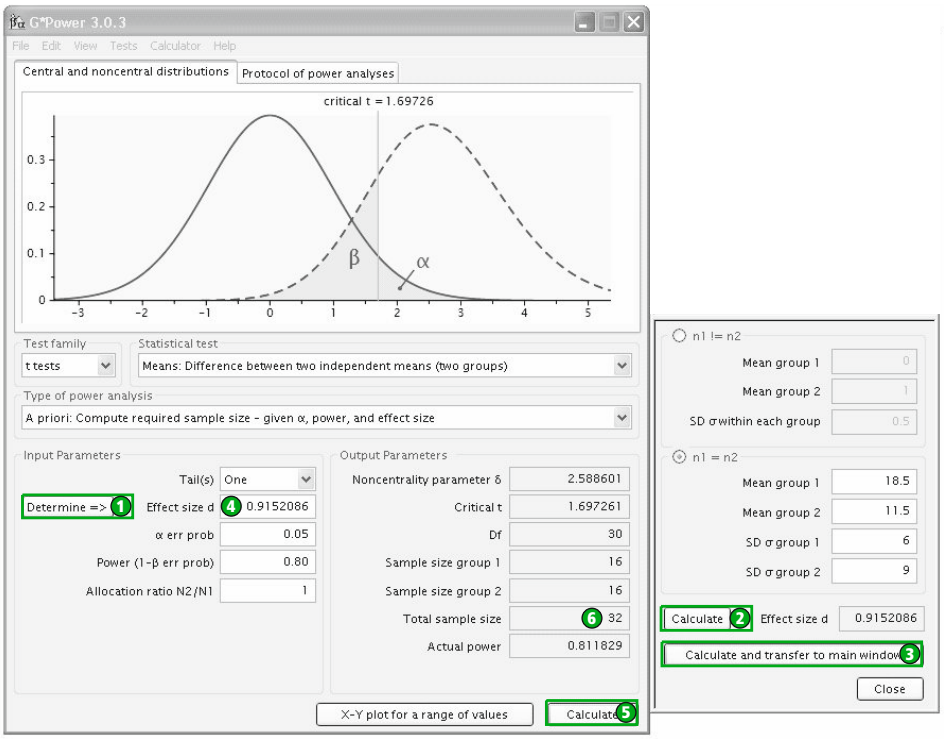


Abbildung 4.27: Berechnung der Effektstärke und der notwendigen Stichprobengröße beim *t*-Test für unabhängige Stichproben

In ▶*Abbildung 4.27* ist die Oberfläche von G*Power dargestellt. Zunächst ist in dem Pull-Down-Menü TEST FAMILY die Option T TESTS zu wählen. In dem Menü STATISTICAL TEST wird nun MEANS: DIFFERENCE BETWEEN TWO INDEPENDENT MEANS (TWO GROUPS) ausgewählt. Danach wählen wir unter TYPE OF POWER ANALYSIS die Option A PRIORI: COMPUTE REQUIRED SAMPLE SIZE – GIVEN α, POWER, AND EFFECT SIZE. Im Anschluss daran müssen folgende Einstellungen vorgenommen werden: TAIL(S), EFFECT SIZE D (gemeint ist d_s bzw. g), α ERR PROB, POWER ($1 - \beta$ ERR PROB) und ALLOCATION RATIO N2/N1. Wenn man, wie oben empfohlen, Glass Delta verwenden möchte, muss die Effektstärke per Hand berechnet werden und dann unter EFFECT SIZE D eingesetzt werden. Im Folgenden stellen wir Überlegungen an, welche Einstellungen für unsere Fragestellung geeignet sind. Wir gehen davon aus, dass unser Training die Wissensleistung erhöht, also von einer einseitigen Testung, und setzen

TAIL(S) damit auf ONE. In diesem Fall ist die Alternativhypothese die Wunschhypothese, denn wir wollen zeigen, dass das Training wirkt. Daher müssen wir es uns schwer machen, die Nullhypothese zu verwerfen. Der schwerwiegendere Fehler wäre, zu sagen, das Training wirkt, obwohl es das nicht tut. Daher setzen wir die Fehlerwahrscheinlichkeit 1. Art (α ERR PROB) auf fünf Prozent (0.05) und nach Konvention die POWER ($1-\beta$ ERR PROB) auf 80 Prozent (0.80). Wir gehen davon aus, dass auf Kontroll- und Experimentalgruppe dieselbe Anzahl an Führungskräften aufgeteilt wird. Dies ist notwendig, da für die Eingabe entweder die Gruppengrößen oder die Standardabweichungen beider Gruppen übereinstimmen müssen. Da wir die Effektstärke noch nicht berechnet haben, können wir dies auch in G*Power tun. Um die Effektstärke zu berechnen, klicken wir auf DETERMINE => (1). Es öffnet sich ein weiteres Menü. In diesem Menü wählen wir N1 = N2 und tragen unsere Mittelwerte und Standardabweichungen aus der Voruntersuchung in die entsprechenden Platzhalter ein. Anschließend klicken wir auf CALCULATE (2). Dadurch wird die Effektstärke berechnet. Durch Klick auf die Schaltfläche CALCULATE AND TRANSFER TO MAIN WINDOW (3) wird dort die berechnete Effektstärke EFFECT SIZE D (4) automatisch an die entsprechende Stelle im Hauptfenster übertragen. Im Hauptfenster klicken wir nun auf die Schaltfläche CALCULATE (5).

Im Fensterbereich OUTPUT PARAMETERS erscheinen nun die Stichprobengrößen (TOTAL SAMPLE SIZE) (6), die erforderlich sind, um eine Teststärke von mindestens 80 Prozent und eine Fehlerwahrscheinlichkeit 1. Art von fünf Prozent zu erzielen: $n = 32$ insgesamt, dabei für $n_1 = 16$ und für $n_2 = 16$ Personen. Unter der Voraussetzung, dass sich die Effektgröße in einer weiteren Untersuchung nicht ändert, reicht eine Stichprobengröße von 32 Führungskräften (davon jeweils 16 in einer Gruppe) für die statistisch optimale Absicherung aus. Es ist jedoch ratsam, geringfügig mehr Personen zu untersuchen (z. B. zusätzlich zwei Personen je Gruppe), da mit dem Ausfall von Teilnehmern immer gerechnet werden muss.

Fall 1: Stichprobenplanung, wenn die Alternativhypothese die Wunschhypothese darstellt (R). In R lässt sich die A-priori-Stichprobenplanung, wie bereits erwähnt (*Abschnitt 4.2*, *Formel 4.12*), mit dem Package `stats` durchführen. Zuvor muss jedoch die Effektstärke, die sich in der Vorstudie ergeben hat, geschätzt werden. Dazu nutzen wir das Paket **compute.es** (Del Re, 2014):

```
library(compute.es)
mes(m.1=18.5, m.2=11.5, sd.1=6, sd.2=9, n.1=5, n.2=8, level=95)
```

Die Funktion, die wir verwenden, ist **mes** (effect size from mean). Nacheinander werden nun die Mittelwerte (`m.1` und `m.2`), Standardabweichungen (`sd.1` und `sd.2`) und Stichprobengrößen (`n.1` und `n.2`) angegeben. Schließlich kann man durch den Zusatz `level=95` ein Konfidenzintervall um die Effektstärke legen lassen. Neben Cohens *d* gibt dieser Befehl noch eine Reihe weiterer Effektstärkemaße aus, auf die wir nicht weiter eingehen. Ein Unterschied im Vergleich zu G*Power ist zu beachten. Zur Berechnung der Effektstärke verwenden beide Programme eine gepoolte Standardabweichung ($\hat{\sigma}_{\text{pooled}}$) als Schätzung der Populationsvarianz. Diese berechnet sich allgemein so:

$$\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(n_1-1)\cdot\hat{\sigma}_1^2 + (n_2-1)\cdot\hat{\sigma}_2^2}{df}}$$

$\hat{\sigma}_{pooled}$ = gepoolte Standardabweichung

n_1 = Stichprobengröße Stichprobe 1

n_2 = Stichprobengröße Stichprobe 2

$\hat{\sigma}_1$ = Standardabweichung Stichprobe 1 geschätzt für die Population

$\hat{\sigma}_2$ = Standardabweichung Stichprobe 2 geschätzt für die Population

df = Freiheitsgrade

Das Ergebnis ist, dass die geschätzte Effektstärke in G*Power größer ausfällt als in R (0.91 versus 0.87). Dies wiederum bewirkt im weiteren Verlauf, dass die geschätzte Stichprobengröße in R größer sein wird. Die entsprechende Syntax lautet:

```
library(stats) (Team, 2014)
power.t.test(n=NULL, delta=.87, sig.level=.05, power=.8, type="two.sample",
   alternative="one.sided")
```

Es kann hier auch `power` auf null (`power=Null`) gesetzt werden und n (`n=18`) auf einen bestimmten Wert festgelegt werden, dann gibt R die Power an ($1 - \beta = 0.798814$). Aus der Schätzung mit R geht hervor, dass eine Stichprobengröße von 17 Personen in jeder Gruppe, also $n = 34$ Personen insgesamt, nötig ist. In der Praxis ist ein Vorgehen mit einer kleinen Voruntersuchung nicht ohne Weiteres möglich, da nicht beliebig viele Personen zur Verfügung stehen. Manchmal kann man aber auch Effekte aufgrund von Vorüberlegungen festlegen. Solche Beispiele sind im folgenden Kasten zur Vertiefung aufgeführt.

Beispiele für Planung nach Kosten-Nutzen-Überlegungen

Abschätzbares Kosten-Nutzen-Verhältnis auf Basis der Populationsparameter. In einem Unternehmen soll ein teures Verkaufstraining durchgeführt werden. Eine Kontrollgruppe erhält das bisherige Standardverkaufstraining. Eine Experimentalgruppe das neue, teure Verkaufstraining. Man weiß, dass ein durchschnittlicher Verkäufer einen Jahresumsatz von 120000 Euro vorweisen kann. Die Standardabweichung der Jahresumsätze beträgt 20000 Euro. Die Verweildauer eines Verkäufers im Unternehmen beträgt durchschnittlich zwei Jahre. Ein Training kostet pro Mitarbeiter 5000 Euro plus 5000 Euro für Zeitausfall und zusätzliche Kosten (insgesamt 10000 Euro). Das Training würde sich also lohnen, wenn der Jahresumsatz mit dem Training durchschnittlich um mehr als 10000 Euro pro Person zunähme. Wir haben hier also eine einseitige Fragestellung. Da die Standardabweichung 20000 Euro ist, entspricht dies genau einer halben Standardabweichung. Damit muss eine Effektstärke von mindestens $\delta = 0.50$ (10000 Euro/20000 Euro) bei einseitiger Testung statistisch abgesichert werden. Legt man die genannten Einflussgrößen (neben der Effektgröße auch eine Fehlerwahrscheinlichkeit 1. Art von fünf Prozent und eine Fehlerwahrscheinlichkeit 2. Art von 20 Prozent) zugrunde, zeigen G*Power und R an, dass man in jeder Gruppe 51 Personen bei einseitiger Testung benötigen wird, um diesen Effekt statistisch abzusichern.

Abschätzbares Kosten-Nutzen-Verhältnis auf Basis kleiner Stichproben. Was würden wir aber tun, wenn die Firma nur 20 Mitarbeiter für Trainings zur Verfügung stellen könnte. Eine Möglichkeit bestünde darin, zu fordern, dass das Training deutlich effektiver sein muss, als nur die Unkosten zu decken. Und man müsste willkürlich einen Effekt annehmen, z. B. $\delta = 0.80$, ab dem man bereit ist, das Training durchzuführen. Dieser Effekt müsste jedoch sehr hoch sein, um überhaupt die Chance auf ein signifikantes Ergebnis zu erhalten, wenn die Stichprobe klein ist. Eine andere Möglichkeit bestünde darin, die nominellen Fehlerwahrscheinlichkeiten zu verändern, was durch eine Erhöhung des Fehlers 1. Art erzielt werden könnte. Man könnte also die Fehlerwahrscheinlichkeit 1. Art von fünf Prozent auf 25 Prozent erhöhen sowie die Teststärke auf 75 Prozent senken. Das bedeutet, wir nehmen eine höhere Irrtumswahrscheinlichkeit bzw. Fehlerwahrscheinlichkeit 1. Art und eine geringere Teststärke in Kauf. Die Auswirkungen beider Änderungen können mit G*Power oder R überprüft werden. Bei dem vorgeschlagenen Vorgehen würde man sechs Personen pro Gruppe benötigen, um den Effekt von $\delta = 0.80$ statistisch abzusichern.

Stellen wir uns an dieser Stelle weiter vor, eine Untersuchung mit $n = 20$ ($n = 10$ je Gruppe) ohne Stichprobenplanung und ohne Berechnung der Effektstärke hätte Folgendes ergeben: Durchschnittlicher Umsatz nach einem Jahr in der Experimentalgruppe 136000 Euro und in der Kontrollgruppe 120000 Euro, die Standardabweichung betrage 20000 Euro in beiden Gruppen. Das Ergebnis eines t-Tests würde hier wie folgt ausfallen: $t = 0.89$; $p = 0.38$. Ein Psychologe in der Human-Ressource-Abteilung kommt zu dem Ergebnis, das Training hat keinen signifikanten Effekt. Die eigentliche Effektstärke wird zwar nicht angegeben, beträgt aber

$$\frac{16000}{20000} = 0.80$$

und fällt somit groß aus. Die bloße Betrachtung der Signifikanz und die fehlende Stichprobenplanung führen dazu, dass ein sinnvolles Training möglicherweise verworfen wird. Die negative Konsequenz ist offensichtlich.

Nicht abschätzbares Kosten-Nutzen-Verhältnis ohne Restriktion der Stichprobengröße. Es soll überprüft werden, ob ein zusätzliches Entspannungstraining neben einer Angsttherapie zu einer Reduktion von Angstsymptomen führt. Die Kurztherapie dauert zwei Stunden und kostet 80 Euro. Zur Evaluation benutzt man eine Fragebogenuntersuchung, um die akute Angstsymptomatik zu erfassen. Nach der Standardtherapie liegt der Mittelwert von Patienten bei zwölf Punkten mit einer Standardabweichung von drei Punkten. Das Entspannungstraining kostet wenig Geld und ist frei von Nebenwirkungen, daher würde man es schon anwenden, wenn nur ein geringer positiver Effekt aufträte. Ein geringer Effekt läge nach Cohen (1988) bei $\delta = 0.20$ Standardabweichungen der Mittelwertsdifferenz vor. Das heißt, dass die Angstsymptomatik in einer Experimentalgruppe, die das zusätzliche Entspannungstraining bekommt, um 0.20 Standardabweichungen im Vergleich zur Kontrollgruppe, die nur die Standardtherapie bekommt, absinken müsste. Damit ist eine Reduktion von zwölf auf 11.4 Punkte nötig ($0.20 \cdot 3 = 0.60$), damit man das Entspannungstraining zusätzlich empfehlen kann:

$$\hat{\delta} = \frac{\hat{\mu}_1 - \hat{\mu}_1}{\hat{\sigma}} = \frac{12 - 11.4}{3} = 0.60\ /3 = 0.20.$$

Wenn wir von einer einseitigen Fragestellung ausgehen und die Wahrscheinlichkeit für den Fehler 1. Art auf fünf Prozent und für den Fehler 2. Art auf 20 Prozent festlegen, benötigen wir eine Stichprobe von $n = 310$ pro Gruppe, um einen solchen Effekt statistisch abzusichern.

Fall 2: Stichprobenplanung, wenn die Nullhypothese die Wunschhypothese darstellt (G*Power). Wir normieren einen psychologischen Konzentrationstest. Es sollen Frauen und Männer im Alter von 18 bis 40 Jahren untersucht werden. Wir legen a priori fest, dass standardisierte Mittelwertsunterschiede zwischen Frauen und Männern erst ab $\delta = 0.15$ praktisch bedeutsam sein sollen. Aus Kostengründen wünschen wir uns natürlich, dass es tatsächlich keinen Unterschied gibt, da andernfalls getrennte Normen für jedes Geschlecht erstellt werden müssten, was insgesamt aufwendiger ist. Wir können verschiedenen Studien aus der Literatur entnehmen, dass bis jetzt keine signifikanten Mittelwertsunterschiede aufgrund des Geschlechts gefunden wurden. Im Gegensatz zum vorherigen Beispiel wählen wir jetzt eine zweiseitige Testung, da wir für keines der Geschlechter einen Vorteil annehmen. Wir wählen dann eine Fehlerwahrscheinlichkeit 1. Art von 20 Prozent. Schließlich ist in unserem Fall der Fehler 2. Art entscheidend. Wir wollen möglichst vermeiden, die Alternativhypothese abzulehnen, obwohl sie in der Grundgesamtheit gilt. Aus diesem Grund wählen wir eine Teststärke von 95 Prozent. Dies entspricht einer Fehlerwahrscheinlichkeit 2. Art von fünf Prozent. Das heißt, dass wir die Alternativhypothese so lange wie möglich beibehalten, bevor wir die Nullhypothese akzeptieren. Belassen wir das Verhältnis der beiden Stichproben bei eins, dann zeigt G*Power eine Stichprobengröße von 762 Personen pro Gruppe an. In diesem Fall beträgt die Teststärke 95 Prozent und die Fehlerwahrscheinlichkeit 1. Art 20 Prozent.

Das Vorgehen entspricht weitestgehend dem Vorgehen, wie es in *Abbildung 4.27* für das vorangegangene Beispiel bereits dargestellt ist, jedoch mit jetzt unterschiedlichen Werten für den Fehler 1. bzw. 2. Art. Zunächst sind die vorzunehmenden Programmeinstellungen gleich und es muss unter dem Pull-Down-Menü TEST FAMILY die Option T TESTS gewählt werden. Unter dem Menüpunkt STATISTICAL TEST wird dann MEANS: DIFFERENCE BETWEEN TWO INDEPENDENT MEANS (TWO GROUPS) ausgewählt. Danach wählen wir unter TYPE OF POWER ANALYSIS die Option A PRIORI: COMPUTE REQUIRED SAMPLE SIZE – GIVEN α, POWER, AND EFFECT SIZE. Abweichend vom obigen Beispiel wählen wir hier eine zweiseitige Fragestellung, da wir keine Aussagen über eine Richtung von Unterschieden treffen können (Hypothese ist, dass es keine praktisch bedeutsamen Unterschiede gibt): TAIL(S) wird auf TWO gesetzt. Die Fehlerwahrscheinlichkeit 1. Art unter α ERR PROB wird auf 20 Prozent (0.20) festgelegt und die Fehlerwahrscheinlichkeit 2. Art auf fünf Prozent durch die Einstellung POWER ($1 - \beta$ ERR PROB) auf 95 Prozent (0.95). Die Stichprobengröße wird nun durch CALCULATE berechnet.

Fall 2: Stichprobenplanung, wenn die Nullhypothese die Wunschhypothese darstellt (R). Die entsprechende Befehlszeile in R lautet für dieses Beispiel:

```
power.t.test(n=NULL, delta=.15, sig.level=.20, power=.95, type="two.sample",
   alternative="two.sided")
```

Auch hier ergeben sich 762 Personen pro Gruppe.

Fall 3: Ermittlung der Teststärke post hoc (G*Power). G*Power lässt sich nicht nur dazu nutzen, um a priori eine Stichprobenplanung durchzuführen. Es lässt sich zudem auch berechnen, wie groß die Teststärke in einer bereits durchgeführten Untersuchung ausgefallen ist (Hybridmodell). Betrachten wir dazu das folgende Beispiel: Wir evaluieren ein Personalentwicklungstraining. Im Rahmen dieses Führungskräftetrainings haben 15 Führungskräfte ein soziales Kompetenztraining in der Experimentalgruppe absolviert und 15 Führungskräfte in der Kontrollgruppe an einem Vortrag zur Mitarbeiterführung teilgenommen. Zur Evaluation wurde ein Assessment Center (AC) mit drei Übungen durchgeführt. Der Mittelwert aus den drei AC-Übungen wurde für jede Gruppe ermittelt. Das heißt, dass in diesem Beispiel die Daten schon erhoben sind. Zur Leistungsbeurteilung der Übungen wurde jeweils eine verhaltensverankerte Ratingskala von 1 „starker Entwicklungsbedarf“ bis 4 „Kompetenz stark ausgeprägt“ verwendet und angenommen, dass für diese Skala Intervallskalenniveau vorliegt. Folgende deskriptive Statistiken haben sich ergeben: Mittelwert mit Training $\hat{\mu} = 3.58$ und $\hat{\sigma} = 0.71$ sowie Mittelwert ohne Training aber mit Vortrag $\hat{\mu} = 3.39$ und $\hat{\sigma} = 0.69$.

In ▶*Abbildung 4.28* ist wieder die Oberfläche von G*Power dargestellt. Zunächst ist unter dem Pull-Down-Menü TEST FAMILY wie im Fall 1 die Option T TEST zu wählen. Unter STATISTICAL TEST wird nun wiederum MEANS: DIFFERENCE BETWEEN TWO INDEPENDENT MEANS (TWO GROUPS) ausgewählt. Danach wird unter TYPE OF ANALYSIS die Option POST HOC: COMPUTE ACHIEVED POWER – GIVEN α, SAMPLE SIZE, AND EFFECT SIZE gewählt. Wir gehen davon aus, dass unser Kompetenztraining die Leistung gegenüber dem Vortrag erhöht, führen also eine einseitige Testung durch und setzen daher TAIL(S) auf ONE. Die Fehlerwahrscheinlichkeit 1. Art α ERR PROB wird auf fünf Prozent (0.05) festgelegt und für beide Stichproben wird die Stichprobengröße von 15 in die Felder SAMPLE SIZE GROUP 1 und SAMPLE SIZE GROUP 2 eingetragen. Danach wählen wir DETERMINE => (1) und tragen die Mittelwerte und Standardabweichungen in die entsprechenden Felder MEAN GROUP 1 und MEAN GROUP 2 bzw. SD σ GROUP 1 und SD σ GROUP 2 ein. Anschließend klicken wir auf die Schaltfläche CALCULATE AND TRANSFER TO MAIN WINDOW (2) und danach im Hauptfenster auf CALCULATE (3).

Die geschätzte Effektstärke (EFFECT SIZE D) beträgt 0.27 (4). Der empirische t-Wert lautet für dieses Beispiel $t = 1.44$ (Berechnung ist hier nicht dargestellt), was bei 28 Freiheitsgraden im Vergleich mit dem kritischen t-Wert (CRITICAL T) (5) nicht signifikant ist. Die Überschreitungswahrscheinlichkeit beträgt $p = 0.23$ (nicht dargestellt). Die Teststärke POWER beträgt in diesem Fall nur knapp 18 Prozent (6). Das heißt, dass der kritische oder ein extremerer t-Wert unter der Alternativhypothesenverteilung mit dem empirischen t-Wert als Erwartungswert nicht sehr wahrscheinlich ist. Damit ist die Wahrscheinlichkeit eines Fehlers 2. Art sehr hoch ($\beta \approx 0.82$). Dies ist insgesamt kein befriedigendes Ergebnis. Die Stichprobe war also viel zu klein, um den gefundenen Effekt statistisch optimal abzusichern. Nun könnten wir aufgrund der vorangegangenen Überlegungen die Wahrscheinlichkeit für den

Copyright