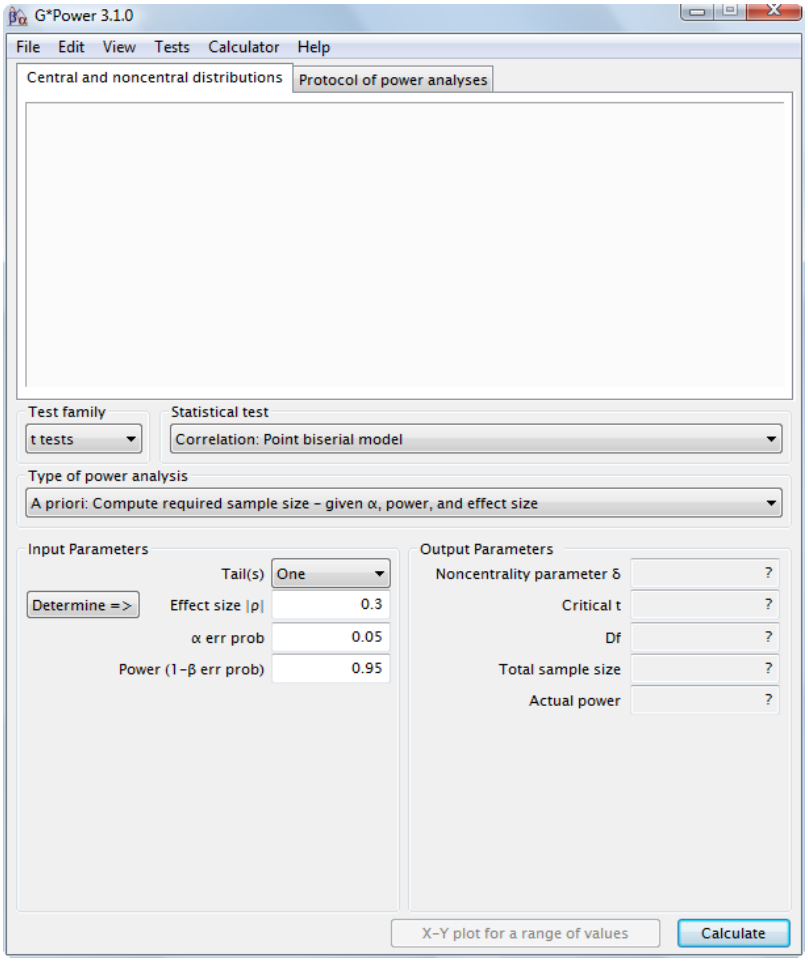
Kapitel 3: Der *t*-Test

t-Test für unabhängige Stichproben

Berechnen der Effektgröße *d*

In Kapitel 3.3.1 haben Sie erfahren, wie sich die Effektgröße *d* aus empirischen Werten berechnen lässt. Dazu haben wir den Vergleich der Erinnerungsleistung der Verarbeitungsgruppen „strukturell“ und „bildhaft“ herangezogen. Wir wollen diese Berechnungen an dieser Stelle mit G*Power nachvollziehen. Starten Sie dazu G*Power, so dass Sie den folgenden Bildschirm vor sich sehen:

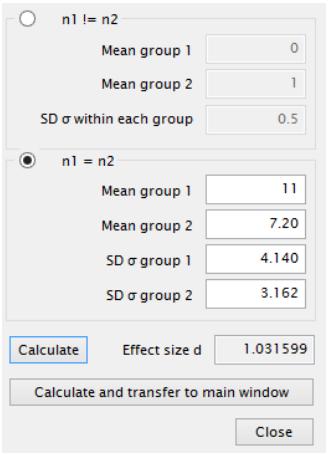


Ein Klick auf das Feld „Test family“ in der Mitte links verrät die verschiedenen Klassen statistischer Verfahren, für die G*Power Berechnungen durchführt. Die Standardeinstellung ist „*t*-tests“. Um den Dialog für den uns interessierenden *t*-Test für unabhängige Stichproben zu erhalten, wählen wir im Drop-Down Menü „Statistical test“ die Option „Means: Difference between two independent means (two groups)“ aus. Wir können von diesem Bildschirm aus mit unserer Analyse starten.

Im Feld „Input Parameters“ klicken Sie auf „Determine“. Dies öffnet ein Fenster an der rechten Seite. In der Mitte am oberen Bildschirmrand öffnen Sie das Feld Calc Effectsize. Dort geben Sie die Mittelwerte und die aus den Daten geschätzte Populationsstreuung an. Aus dem Datensatz bzw. aus Kapitel 3.1 können Sie diese Werte entnehmen: ($\bar{x}_{bildhaft} = 11$; $\bar{x}_{strukturel} = 7{,}2$; $\hat{\sigma}_{bildhaft} = 4{,}140$; $\hat{\sigma}_{strukturel} = 3{,}162$).

Eine Voraussetzung für die Anwendbarkeit des *t*-Tests ist die Homogenität der Varianzen in den Stichproben. Empirisch sind die Varianzen aber praktisch nie vollkommen identisch. Der *t*-Test ist gegen solche Abweichungen robust. Er liefert also weiterhin zuverlässige Ergebnisse, so lange die Abweichungen nicht zu stark werden. Obwohl G*Power die theoretische Annahme der Varianzhomogenität ebenfalls macht, bezieht es auftretende empirische Unterschiede mit ein. Deshalb erwartet das Programm für jede Stichprobe die empirische Streuung, nicht nur den Mittelwert der Streuungen. Dies führt zu einem Wert für *d* von 1,03 (vgl. Kapitel 3.3.1). Dies ist den Konventionen von Cohen (1988) folgend ein großer Effekt.

Achtung: Bitte beachten Sie, dass G*Power für die korrekte Berechnung Punkte an Stelle von Kommata erwartet.

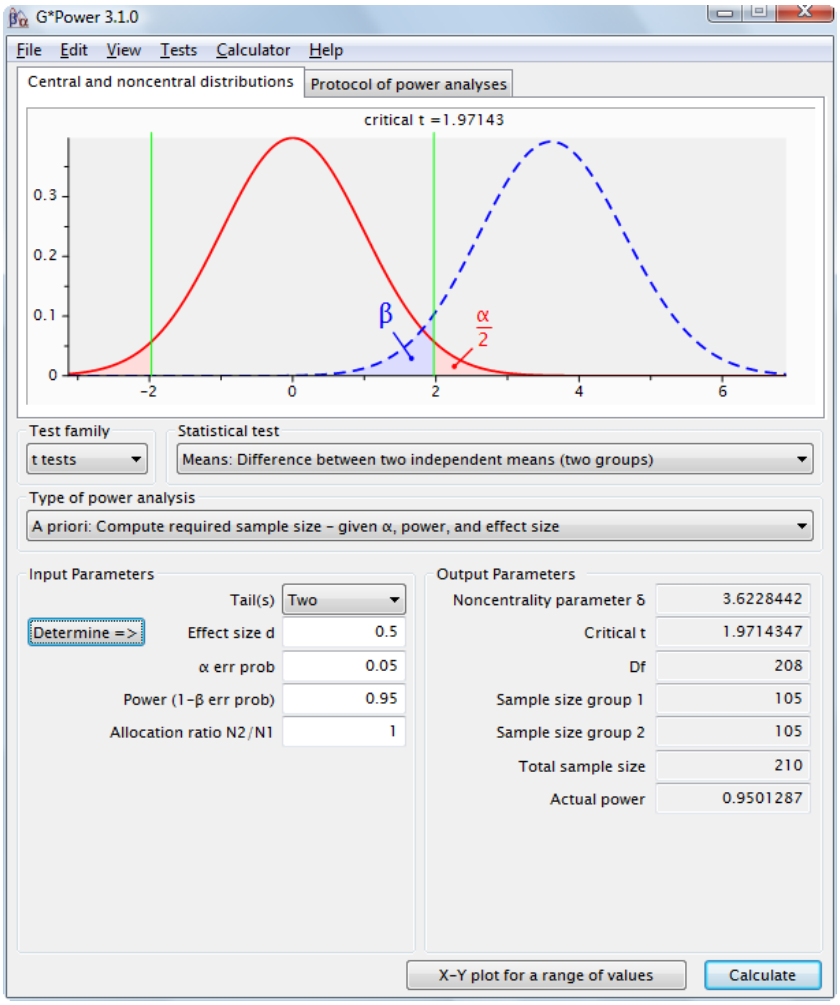


Berechnen der Teststärke a priori bzw. Stichprobenumfangsplanung

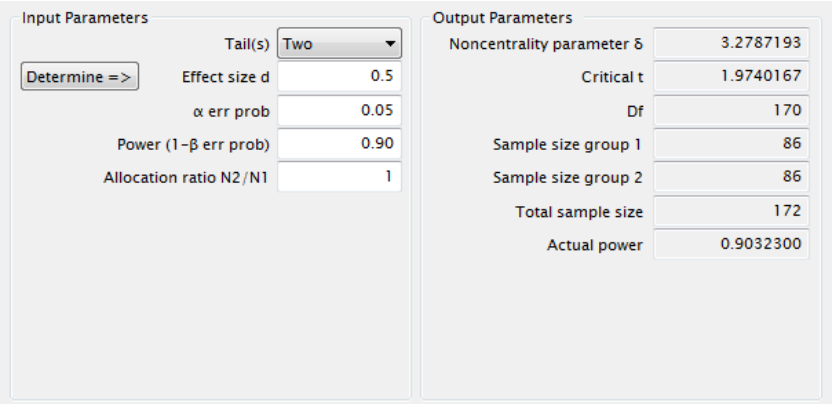
Im Hauptfenster von G*Power sehen Sie die Auswahlmöglichkeit „Type of power analysis“. Die Standardeinstellung ist „A priori“; was gleichzeitig die richtige Option ist, um den Stichprobenumfang vor einer Studie zu berechnen. Etwas weiter unten können Sie einstellen, ob Sie die Stichprobengröße für einen einseitigen oder zweiseitigen Test berechnen möchten.

Die Stichprobenumfangsplanung kommt dann zum Zuge, wenn ein Forscher eine Untersuchung plant und wissen möchte, wie viele Personen er unter den gegebenen Annahmen rekrutieren muss, um auf jeden Fall ein interpretierbares Ergebnis zu erhalten. In unserem Beispiel erwartet der

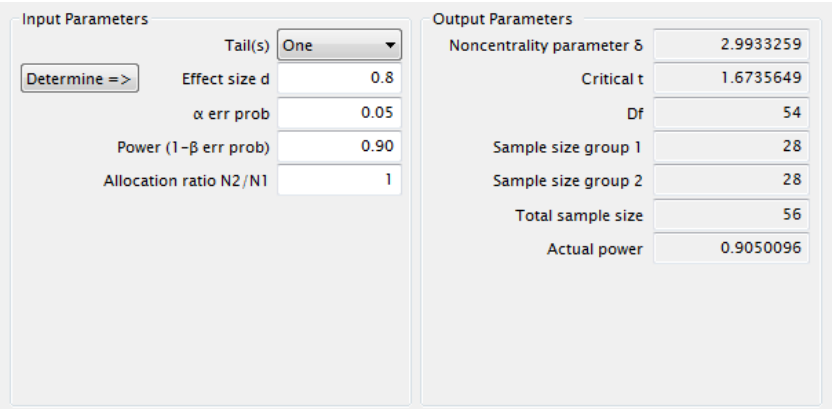
Forscher für seine ungerichtete Fragestellung einen mittleren Effekt von d = 0,5. Ob eine Fragestellung gerichtet oder ungerichtet ist, können Sie im Feld „Tail(s)“ einstellen. „Two tails“ signalisiert, dass es sich um eine ungerichtet Fragestellung handelt. Der Forscher setzt das Signifikanzniveau auf α = 0,05 und möchte außerdem, dass der β-Fehler auch nicht größer ist. Dies würde in einer Teststärke von 1-β = 0,95 resultieren. G*Power errechnet für diese spezifische Konstellation einen Bedarf von 210 Versuchspersonen, also 105 Personen in jeder Gruppe.



Würde sich der Forscher mit einer Teststärke von 90% zufrieden geben, würde sich die benötigte Anzahl Versuchspersonen auf 172 reduzieren.



Ein anderer Forscher nimmt für seine Untersuchung einen großen Effekt (d = 0,8) zwischen den Gruppen an. Er verfolgt eine gerichtete Fragestellung und ist bereit, einen 10%igen β-Fehler zu akzeptieren. G*Power berechnet einen benötigten Stichprobenumfang von 56 Personen.



An diesen Beispielen können Sie sehr anschaulich nachvollziehen, wie sich die vier Determinanten eines statistischen Tests gegenseitig bedingen. Sind drei von ihnen festgelegt, ist auch die letzte eindeutig bestimmt. Wir möchten Sie ermutigen, selber einige Beispiele mit G*Power zu rechnen, um zu sehen, wie die Veränderung einer Determinante den benötigten Stichprobenumfang beeinflusst: Ein großer angenommener Effekt verringert den Stichprobenumfang, während ein kleiner ihn erhöht. Eine geringere Teststärke erfordert weniger Versuchsteilnehmer als eine hohe Teststärke. Ein liberaleres α-Niveau verlangt ein kleineres N als ein strenges (siehe Kapitel 3.4.2).

Teststärkebestimmung a posteriori

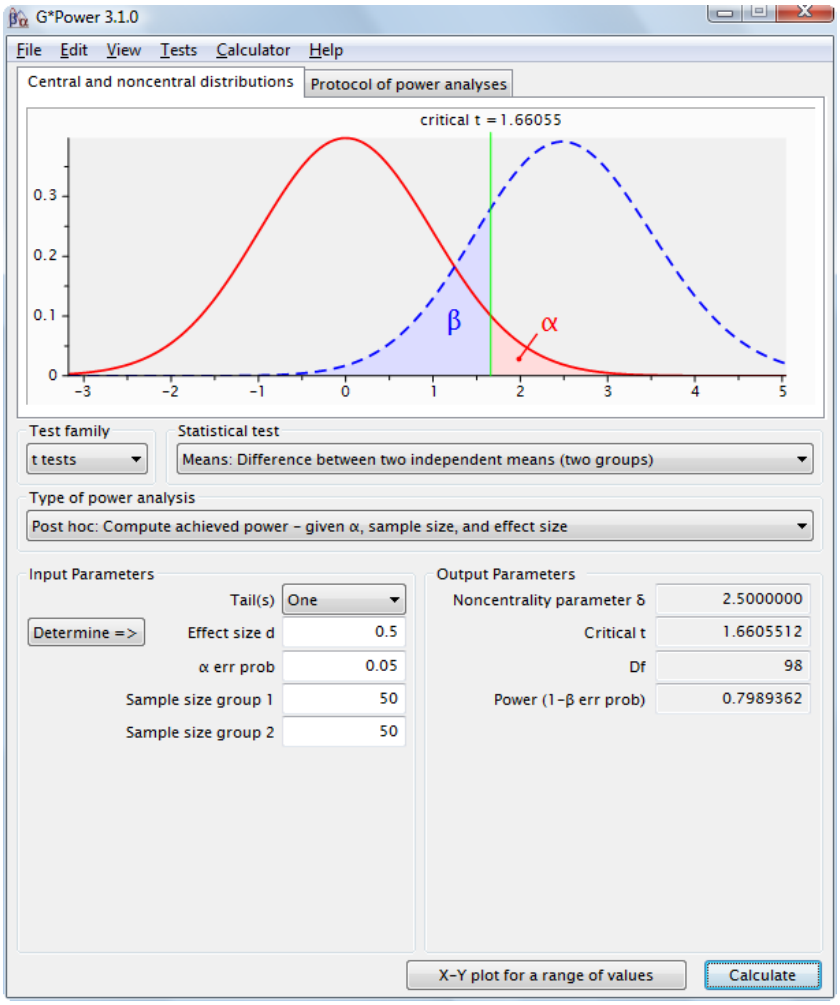
In der Forschungspraxis ist eine Teststärkebestimmung a priori bzw. eine Stichprobenumfangsplanung bedauerlicher Weise noch kein Standard. Häufig wünschen sich Wissenschaftler aber nach einem nicht signifikanten Ergebnis in einer Untersuchung zumindest eine Antwort auf die Frage, wie groß denn die Chance überhaupt war, den vermuteten Effekt zu finden. Die Teststärkebestimmung a posteriori beantwortet diese Frage.

Auch ein weiterer Fall verhilft dieser Analyse zur häufigen Anwendung: In der Realität ist es häufig so, dass Forscher schon vor einer Untersuchung wissen, wie viele Versuchspersonen sie für die Studie erheben können. Gründe dafür sind z.B. ein begrenzter Zugang zu finanziellen Mitteln

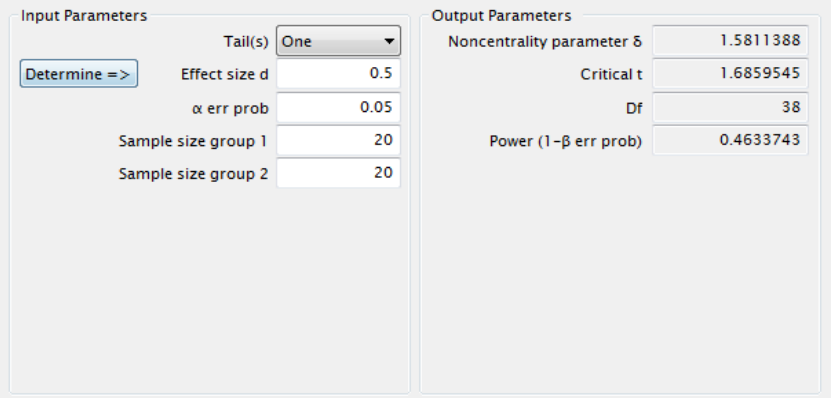
oder Räumen zur Datenerhebung. In diesem Fall bietet die Teststärkebestimmung a posteriori trotz ihres Namens schon vor der Durchführung eine Entscheidungshilfe für die Frage, ob sich die Datenerhebung überhaupt lohnt.

Ein Forscher hat eine Untersuchung mit je 50 Versuchspersonen in zwei Gruppen durchgeführt. Er hatte eine gerichtete Hypothese und vermutete einen mittleren Effekt von d = 0,5 hinter dem untersuchten Phänomen. Das Ergebnis war allerdings auf dem 5%-Niveau nicht signifikant. Mit Hilfe von G*Power ermittelt er eine empirische Teststärke von knapp 80%. Der β-Fehler lag also bei 20%. Sollte er seine Hypothese auf Grund dieser Daten verwerfen und einen Nullunterschied zwischen den Gruppen annehmen, würde er mit 20%iger Wahrscheinlichkeit einen Fehler machen. Eine Power von 80% gilt als gerade noch akzeptabel.

In G*Power können Sie diese Werte eingeben, wenn Sie unter „Type of power analysis“ „Post hoc“ auswählen.



Ein anderer Forscher weiß, dass er nur 20 Versuchspersonen pro Bedingung erheben kann. Er nimmt ebenfalls einen mittleren Effekt an und setzt das α-Niveau auf 5% für seine gerichtete Fragestellung. Wenn er diese Untersuchung durchführen möchte, muss er mit einer Teststärke von weniger als 50% vorlieb nehmen. Er könnte also ebenso gut eine Münze werfen.



Auch an diesen Beispielen sehen Sie, wie sich die vier Determinanten des *t*-Tests gegenseitig bedingen. Probieren Sie ein wenig aus, welche Auswirkungen es auf die Teststärke hat, wenn Sie die Effektgröße, das α-Niveau und/oder die Stichprobengröße verändern!

t-Test für abhängige Stichproben

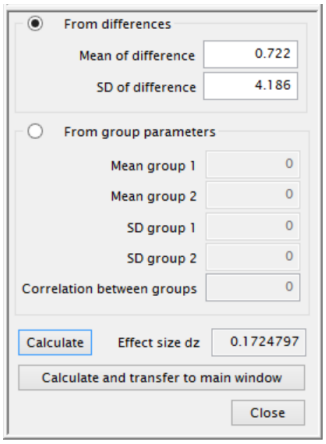
Berechnen der Effektgröße d_z

Auch für abhängige Stichproben lässt sich eine Effektgröße aus empirischen Werten ermitteln. Um deutlich zu machen, dass sie die Effektgröße für abhängige Stichproben ist, heißt sie d_z. Ebenso wie die Effektstärke d bei unabhängigen Stichproben ist d_z wie eine Streuungseinheit zu interpretieren. Im Unterschied zu d geht in die Berechnung von d_z allerdings noch die Stärke der Abhängigkeit der Messwerte mit ein (vgl. nachfolgenden Abschnitt). Daher lassen sich d_z und d nicht direkt miteinander vergleichen, und es liegen auch keine Konventionen für d_z vor.

Die Effektgröße d_z lässt sich auf zwei Arten berechnen, zum einen über die Differenzen zwischen den beiden Messwertereihen, zum anderen über die Kennwerte der beiden Gruppen. Für die Berechnung über die Differenzen der Messwertereihen gilt:

$d_z = \frac{\bar{x}_d}{\sigma_d}$, wobei $\bar{x}_d$ der Mittelwert der Differenzen ist und σ_d die Streuung dieser Differenzen (Cohen, 1988). Sowohl der Mittelwert der Differenzen als auch die Streuung der Differenzen sind dem SPSS-Output eines t-Tests für abhängige Stichproben zu entnehmen. Mathematisch lässt sich zeigen, dass d_z eng verwandt ist mit der in Kapitel 3.5 diskutierten Effektgröße $f^2_{s(abhängig)}$ ist. Es gilt $d_z = \sqrt{f^2_{s(abhängig)}} = f_{s(abhängig)}$. Außerdem gilt $d_z = \sqrt{\frac{2}{1-r}} \cdot f_{unabhängig} = \sqrt{\frac{2}{1-r}} \cdot \frac{d_{unabhängig}}{2}$.

In G*Power können Sie auf diese Weise d_z berechnen, indem Sie „Mean: Difference between two dependent means (matched pairs)“ sowie „Post hoc“ einstellen. Durch einen Klick auf „Determine“ öffnet sich rechts ein Seitenfenster zur Berechnung von d_z. Wenn Sie die obere Option aktivieren, können Sie dort den Mittelwert und die Standardabweichung der Differenzen eingeben. Mit den Daten aus dem Beispiel in Kapitel 3.5 ergibt sich die Effektstärke d_z = .17.

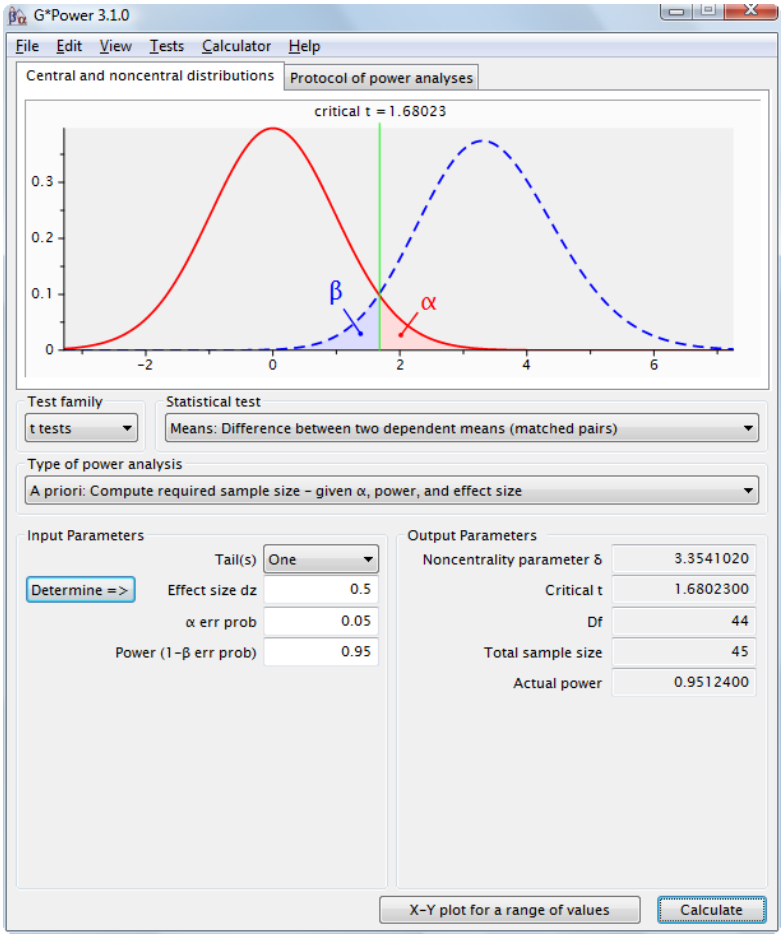


So wie die Berechnung des *t*-Werts für abhängige Stichproben zur Prüfung auf Signifikanz ist auch die Berechnung des Effektstärkenmaßes d_z beeinflusst durch die Abhängigkeit der beiden Messzeitpunkte. Diese muss also bei der Berechnung des Maßes berücksichtigt werden. Im Fall der Berechnung von d_z über die Differenzen der Messwertereihen ist die relevante Information über die Abhängigkeit der Daten in der Streuung der Differenzen verwoben.

Wenn Sie d_z lieber über die Parameter der Stichproben berechnen möchten, aktivieren Sie im Seitenfenster recht dazu einfach die entsprechende Option und tragen die Werte ein. Diese finden Sie in jedem SPSS Output für einen *t*-Test für abhängige Stichproben (siehe SPSS-Ergänzungen).

Berechnen der Teststärke a priori bzw. Stichprobenumfangsplanung

Nehmen wir an, ein Forscher erwartet einen Effekt von d_z = 0,5 und möchte geringe Fehlerwahrscheinlichkeiten von jeweils 5% für sowohl α als auch β sicherstellen. Er verfolgt eine gerichtete Fragestellung. G*Power errechnet für diesen Fall eine benötigte Stichprobengröße von N = 45.

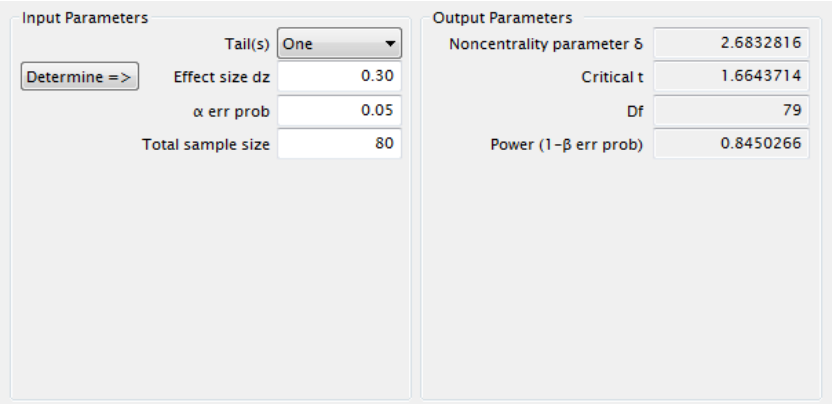


Um die erwarteten Effektstärke d_z zu bestimmen, ist es nicht nur notwendig, Annahmen über den Mittelwertsunterschied zwischen den beiden Gruppen/Messwertreihen und ihre Streuungen zu machen, sondern auch über die zu erwartende Korrelation zwischen den Messwertereihen. Zu beachten ist dabei, dass die empirisch auftretende Korrelation zwischen zwei Messwertereihen vor einer Untersuchung nie bekannt ist. Sie kann lediglich auf Basis vorheriger eigener Studien oder in der Literatur berichteter Studien geschätzt werden. Dies beeinträchtigt die Aussagekraft der Stichprobenumfangsplanung für den *t*-Test für abhängige Stichproben, denn auch wenn Schätzungen der Korrelation zwischen den Messwertereihen auf Basis vorheriger Studien möglich sind, kann die empirisch auftretende Korrelation im Einzelfall immer von den Erwartungen abweichen.

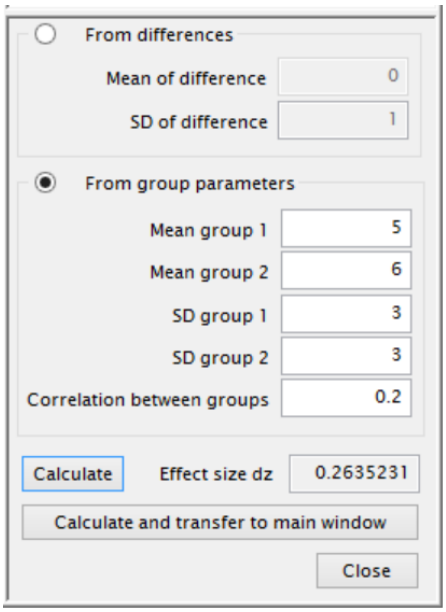
Teststärkebestimmung a posteriori

Um die Teststärke für einen *t*-Test für abhängige Stichproben a posteriori zu bestimmen, ist „Mean: Difference between two dependent means (matched pairs)“ sowie „Post hoc“ die richtige Option.

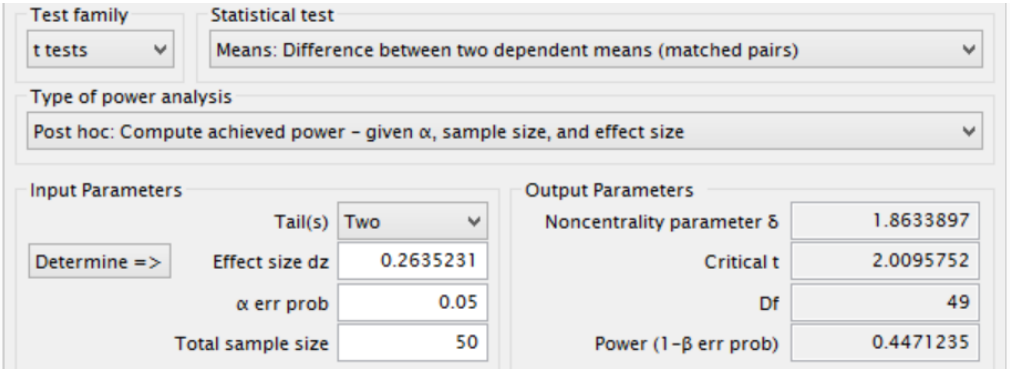
Eine Untersuchung mit einer gerichteten Fragestellung und 80 Versuchspersonen liefert bei einem α-Niveau von 5% ein nicht signifikantes Ergebnis. Wie groß war die Teststärke, einen Effekt von d_z = 0,3 in dieser Studie zu entdecken? G*Power zeigt an, dass die Teststärke bei nahezu 85% lag. Das ist kein überragender, aber ein guter Wert.



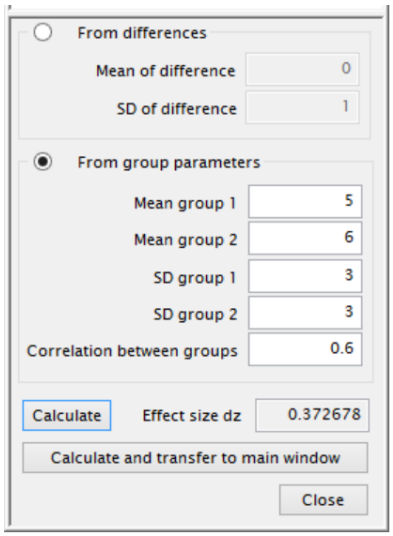
In diesem Zusammenhang lässt sich sehr gut der Einfluss der Höhe der Korrelation zwischen den Messwertereihen auf die Teststärke verdeutlichen. Nehmen wir an, in einer Studie wäre der Mittelwert der abhängigen Variablen zu Zeitpunkt t_1 $\bar{x}_1 = 5$ mit einer Streuung von $\sigma_1 = 3$. Die Werte für Zeitpunkt t_2 lauten $\bar{x}_2 = 6$ und $\sigma_2 = 3$. Die Korrelation der Messwertereihen sei im ersten Fall r = 0,2. Es ergibt sich eine Effektstärke von d_z = 0,26.

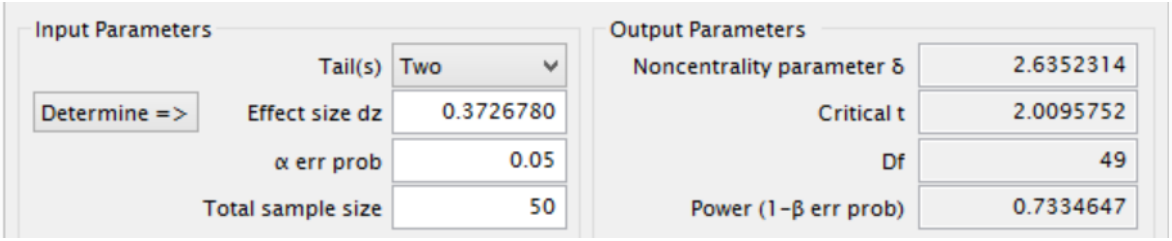


Diese Konfiguration führt zu einer Teststärke von etwa 45% bei einem angenommen α-Fehler von 5%, zweiseitiger Testung und einer Stichprobengröße von $N = 50$.



Wäre die Korrelation deutlich höher ausgefallen, beispielsweise $r = 0{,}60$, wäre auch d_z und damit die Teststärke deutlich gestiegen, nämlich auf $d_z = .37$ bzw. $1\text{-}\beta = .83$.

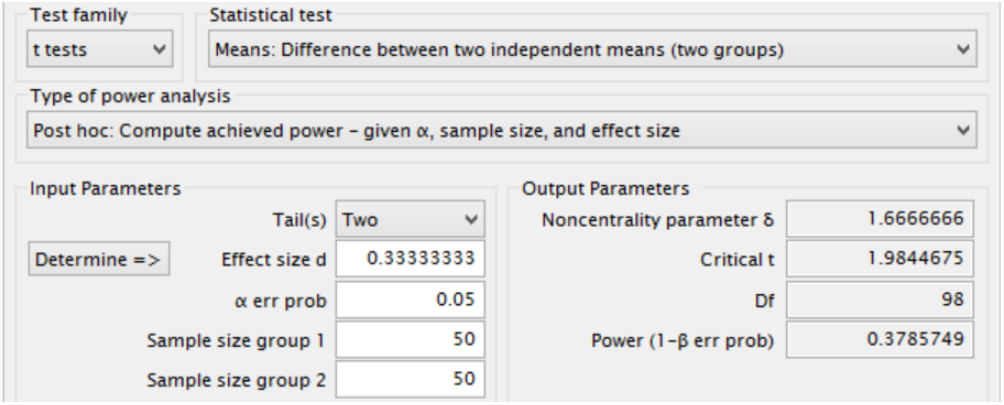




Läge keine Korrelation ($r = 0$) zwischen den Stichproben vor, so ergäbe sich mit $d_z = 0{,}2357$ eine Teststärke von nur 37% (bitte nachrechnen). Weiter oben haben wir gesehen, dass $d_z = \sqrt{\frac{2}{1-r}} \cdot \frac{d_{unabhängig}}{2}$ und damit $d_{unabhängig} = \frac{2d_z}{\sqrt{\frac{2}{1-r}}}$.

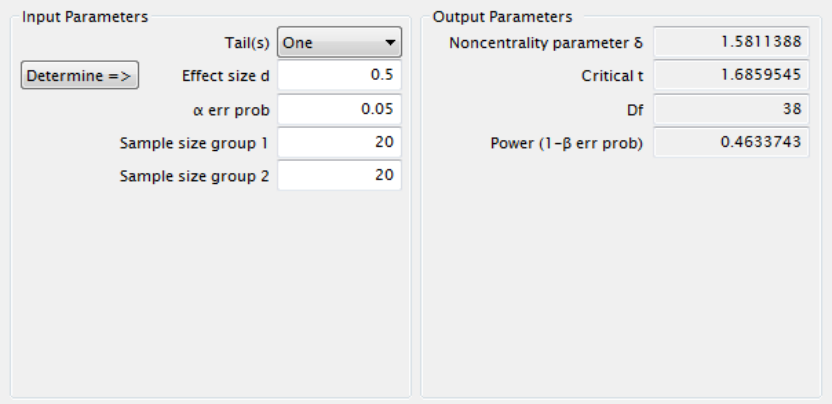
Bei r = 0: $d_z = \frac{d_{unabhängig}}{\sqrt{2}}$ bzw. $d_{unabhängig} = \sqrt{2} \times d_z$

Daraus ergibt sich, dass dieser Wert für die Teststärke bei $r = 0$ der Teststärke eines *t*-Tests für *un*abhängige Stichproben für einen Effekt von $d_{unabhängig} = 0{,}33$ bei einer identischen Anzahl von Messwerten entspricht. Im Fall abhängiger Stichproben geben die 50 Versuchspersonen jeweils zwei Messwerte ab. Die Teststärke entspricht also der eines *t*-Tests für unabhängige Stichproben mit 100 Personen, also 50 Personen pro Gruppe.



Vergleich von *t*-Test für unabhängige und abhängige Stichproben sowie Vertiefung des Konzeptes „Abhängigkeit“

Im Abschnitt über den *t*-Test für unabhängige Stichproben haben wir ein Beispiel betrachtet, in dem ein Forscher wusste, dass er nur 40 Versuchspersonen zur Verfügung hatte. Für seine Studie mit einem angenommenen Effekt mittlerer Größe und einem Signifikanzniveau von 5% ergab sich bei einseitiger Fragestellung eine Teststärke von 46% (siehe Graphik). Würde er die Studie in dieser Form durchführen, wäre das mit einem großen Risiko verbunden, am Ende ohne interpretierbares Ergebnis dazustehen. Gibt es eine Alternative für den Forscher?



Bisher haben wir abhängige Daten als solche bezeichnet, die auf derselben abhängigen Variablen an zwei unterschiedlichen Messzeitpunkten von derselben Person produziert wurden. Das Konzept der Abhängigkeit von Daten greift aber noch weiter. Die Messwiederholung ist nur einer von vielen möglichen Fällen abhängiger Daten. Denken Sie an unser Gedächtnisexperiment. Dort ging es darum, positive, negative und neutrale Wörter zu erinnern. Eine Person, die besonders viele positive Wörter erinnert, wird in aller Regel auch viele negative Wörter erinnern. Die dahinter stehende Eigenschaft „Gutes Gedächtnis von Person X“ wirkt sich auf alle drei Wortarten aus. Die Werte für positive, negative und neutrale Wörter kommen also nicht unabhängig voneinander zu Stande, sondern werden alle von der Fähigkeit derselben Person beeinflusst. Sie sind abhängig voneinander. (Beachten Sie, dass diese Ausführungen nichts mit der Einteilung in bildhafte, emotionale und strukturelle Verarbeitung zu tun hat, die wir bislang thematisiert haben.)

Wenn das Ziel darin besteht, herauszufinden, ob es Unterschiede in der Erinnerungsfähigkeit positiver und negativer Adjektive gibt, hat ein Forscher mehrere Möglichkeiten, dieses Ziel zu verfolgen. Zum einen kann er zwei Gruppen bilden, die entweder positive oder negative Adjektive präsentiert bekommen und später abrufen sollen. Die adäquate Auswertungsstrategie für diesen Versuchsaufbau wäre ein *t*-Test für unabhängige Stichproben. Eine von mehreren anderen Möglichkeiten wäre aber, allen Personen beide Arten von Adjektiven zu präsentieren und später die Daten mit einem *t*-Test für abhängige Stichproben auszuwerten.

Betrachten Sie ein anderes Beispiel abhängiger Daten: Ein Sozialpsychologe möchte die Einstellung gegenüber der SPD und den Grünen erfassen. Dafür hat er mehrere Möglichkeiten. Zum einen könnte er eine Personengruppe zu ihrer Einstellung gegenüber der SPD befragen und eine andere Gruppe zu ihrer Einstellung gegenüber den Grünen. Von diesen Gruppen könnte der Forscher mit einem *t*-Test für unabhängige Stichproben die Mittelwerte vergleichen und somit

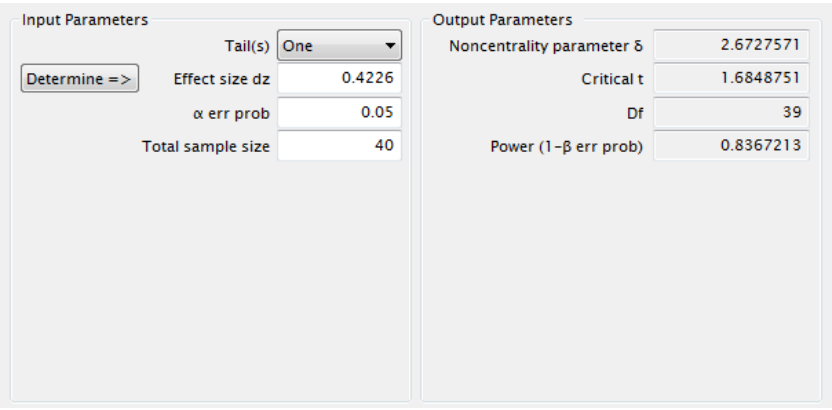
überprüfen, ob er einen Unterschied in den Einstellungen gegenüber beiden Parteien feststellen kann. Eine andere Möglichkeit bestünde darin, *alle* Versuchspersonen zu *beiden* Parteien zu befragen. Allerdings ist es plausibel anzunehmen, dass die Einstellungen einer Person zu den beiden Parteien nicht unabhängig voneinander sind, denn sie sind politisch zwar verschieden, aber verwandt. Eine Person, welche die eine Partei positiv bewertet, hat vermutlich auch eine ähnliche Einstellung gegenüber der anderen Partei. Die dahinter liegende Eigenschaft „politische Einstellung" würde sich also auf beide Einstellungsangaben dieser Person positiv auswirken. Eine andere Person hingegen ist möglicherweise deutlich konservativer eingestellt und gibt deshalb bei beiden Gruppen wenig positive Einstellungen an. Auch hier kämen also die Daten zu beiden Einstellungsmaßen nicht unabhängig voneinander zu Stande. Mit anderen Worten: Sie sind korreliert (siehe Kapitel 4, Band 1).

Das Konzept der Abhängigkeit von Daten kann sogar noch weiter gefasst werden. Stellen Sie sich eine Untersuchung mit Zwillingen vor, die im selben Elternhaus aufgewachsen sind. Auch wenn diese sich natürlich voneinander unterscheiden, ist es doch plausibel anzunehmen, dass Zwillingspaare häufig ähnliche Werthaltungen und Ansichten teilen. In diesem Fall lassen sich sogar die Daten von zwei *unterschiedlichen* Personen als abhängig betrachten. Noch einmal im Kontrast dazu der Fall von unabhängigen Stichproben: Hier geht man davon aus, dass sich in den zwei Gruppen unterschiedliche Personen befinden, die in keinem besonderen Verhältnis zueinander stehen. Ihre Daten sind unkorreliert, denn keine zwei Datenpunkte sind von dem selben dahinter stehenden Konstrukt beeinflusst, wie z.B. der Intelligenz einer Person, den motorischen Fähigkeiten einer Person oder auch nur dem gemeinsamen Elternhaus mit ähnlicher Erziehung etc.

Es gibt also wissenschaftliche Fragestellungen, die sowohl mit unabhängigen als auch mit abhängigen Stichproben untersucht werden können. Welche Auswirkungen hat die Entscheidung für die eine oder andere Vorgehensweise auf die Teststärke? Betrachten wir das obige Beispiel noch einmal, in dem ein Forscher eine schwache Teststärke von 46% mit den ihm zur Verfügung stehenden Mitteln erzielen konnte. Welche Teststärke würde erzielt, wenn die Untersuchung an abhängigen Stichproben durchgeführt würde, die zu $r = 0{,}30$ miteinander korrelieren? (*t*-Test für abhängige Stichproben: $N = 40$, $\alpha = 5\%$, angenommenes $d_{unabhängig} = 0{,}5$, einseitige Fragestellung.)

Zunächst müssen wir d_z ermitteln, um dann die Werte in G*Power einzutragen.

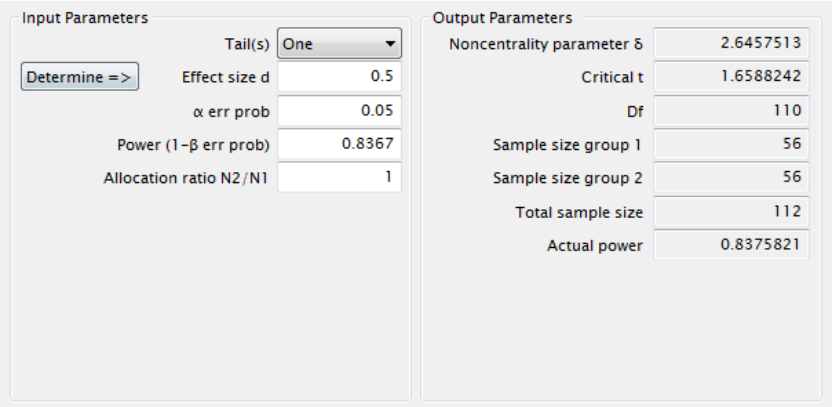
$$d_z = \sqrt{\frac{2}{1 - r}} \times \frac{d_{unabhängig}}{2} = \sqrt{\frac{2}{1 - 0{,}30}} \times \frac{0{,}5}{2} = 0{,}4226$$

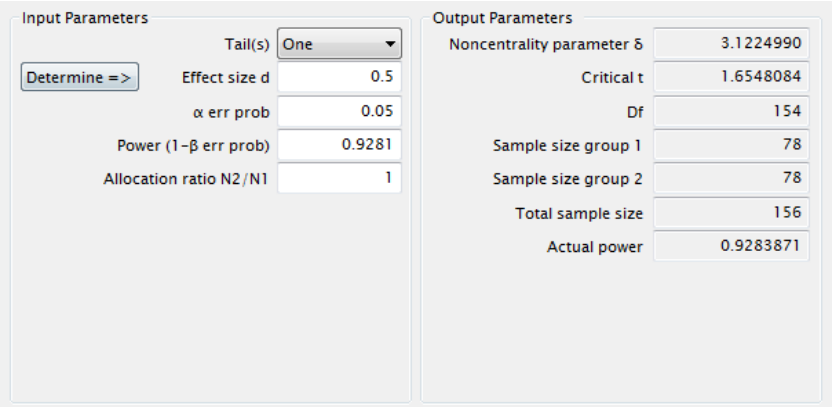


Während der Forscher bei einem Untersuchungsdesign mit unabhängigen Stichproben eine Teststärke von 46% erzielt hat, liegt die Teststärke bei abhängigen Stichproben für r = 0,30 bei wesentlich höheren 84%!

Zwei Gründe führen zu dieser hohen Teststärke: Zum einen gehen in den Test für abhängige Stichproben bei gleicher Gesamtanzahl Probanden doppelt so viele Messwerte ein, da jede Person zwei Werte abgibt, während bei dem Vergleich von unabhängigen Gruppen jede Person nur einen Messwert liefert. In der Formel für d_z zeigt sich dieser Einfluss an der Zahl 2 im Zähler unter der Wurzel, die d_z im Vergleich zu d um den Faktor $\sqrt{2}$ erhöht, und so die Teststärke vergrößert. Zum anderen bewirkt die positive Korrelation einen Anstieg der Teststärke. Dies lässt sich ebenso an der Formel veranschaulichen: je höher die Korrelation, desto kleiner die Zahl im Nenner unter der Wurzel, desto größer wird d_z und damit die Teststärke.

Der Unterschied in dem Beispiel zwischen einer Teststärke von 46% und 84% bei gleicher Versuchspersonenzahl ist beachtlich, wenn man bedenkt, wie viele Kosten unterschiedlicher Natur mit der Rekrutierung und Datenerhebung von Versuchspersonen in der Regel verbunden sind. Läge die Korrelation zwischen den Messwertreihen bei r = 0,50 (ein durchaus realistischer Wert für viele Fragestellungen), würde sich die Power sogar auf nahezu 93% erhöhen! (Rechnen Sie dieses Beispiel nach!) Würde der Forscher auf eine ähnlich hohe Teststärke im Fall unabhängiger Stichproben abzielen, bräuchte er 112 Versuchspersonen für knapp 84% Power und sogar 156 für nahezu 93% Power! (siehe folgende Graphiken)





Der *t*-Test für abhängige Stichproben ist also bei gleicher Anzahl von Personen teststärker als der *t*-Test für unabhängige Stichproben, da jede Versuchsperson zwei Werte abgibt, und weil eine

(positive) empirische Korrelation zwischen den Messwertreihen der abhängigen Stichproben die Teststärke zusätzlich erhöht[1]. Die genaue Höhe der Teststärke hängt letztlich von der Größe der Korrelation ab. Je größer diese ist, desto größer ist der Vorteil der abhängigen Stichproben gegenüber den unabhängigen. Abgesehen von den weiter oben erwähnten eher mathematischen Gründen, warum sich die Korrelation zwischen abhängigen Daten positiv auf die Teststärke auswirkt, gibt es auch noch andere, eher inhaltlich fassbare. Diese Überlegungen sind stark mit dem Konzept der Varianz verbunden. Mehr darüber erfahren Sie in Kapitel 7, Band 2.

Wenn die inhaltliche Fragestellung es zulässt, ein Untersuchungsdesign zu wählen, das ohne Einbußen in der Aussagekraft mit abhängigen Stichproben arbeitet, so hat dies also Vorteile für die Teststärke und damit für die Effizienz der Forschung! In diesen Fällen bestimmt die Planung einer Untersuchung, ob am Ende abhängige oder unabhängige Stichproben vorliegen. Einschränkend sei allerdings gesagt, dass in vielen Fällen unveränderbare Umstände die Frage bestimmen, ob man seine Daten an zwei unabhängigen oder abhängigen Gruppen untersucht. Wenn z.B. Geschlechterunterschiede im Fokus einer Untersuchung stehen, gibt es keine Möglichkeit, die Daten beider Ausprägungen des Merkmals Geschlecht von ein und derselben Person zu erhalten. Etliche andere versuchplanerische Erwägungen schließen die Erhebung mehrer Messwerte pro Person bei bestimmten Fragestellungen ebenfalls aus (Reihenfolgeeffekte, Übungseffekte, Ermüdungseffekte etc.). In diesen Fällen sind unabhängige Stichproben erforderlich.

Anmerkungen: Zum Vergleich von einem *t*-Test für unabhängige und abhängige Stichproben in SPSS, siehe Datei „Kapitel_3_SPSS_Ergaenzungen.pdf“.

[1] Anmerkung: Interessanter Weise würde eine negative Korrelation zwischen abhängigen Stichproben zu einer *Verringerung* der Teststärke führen, wie Sie leicht an der Formel für d_z nachvollziehen können. Allerdings tritt dieser Fall in der Praxis sehr selten auf.

Literatur

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NY: Erlbaum.