ps
psychologie

Markus Bühner
Matthias Ziegler

Statistik für Psychologen und Sozialwissenschaftler

PEARSON
Studium

Markus Bühner
Matthias Ziegler

Statistik für Psychologen und Sozialwissenschaftler

Mit über 480 Abbildungen

PEARSON Studium

Ein Imprint von Pearson Education
München • Boston • San Francisco • Harlow, England
Don Mills, Ontario • Sydney • Mexico City
Madrid • Amsterdam

t-Wert ausfällt, lehnen wir die Nullhypothese ab. Die Überschreitungswahrscheinlichkeit p zu unserem empirischen t-Wert beträgt $p = 0.014$. Auch diese lässt sich mit Programmen oder Tabellen bestimmen (siehe Abschnitt 4.1.6). Damit entscheiden wir uns dafür, anzunehmen, dass der BDI-Mittelwert der psychiatrischen Patienten zu einer anderen als der Grundgesamtheit der Patienten mit einem BDI-Grenzwert von 18 gehört. Wir wissen allerdings noch nicht, ob der beobachtete Unterschied praktisch bedeutsam ist.

Praktische Bedeutsamkeit. Es stellt sich nun die Frage, wie stark die beiden Mittelwerte praktisch[1] voneinander abweichen. Dazu ziehen wir die Formel für Hedges g oder Cohens d_s heran:

$$g = \frac{\bar{x} - \mu}{\hat{\sigma}} = \frac{15 - 18}{4} = -.75$$

g = Hedges g

μ = Mittelwert der Grundgesamtheit

$\bar{x}$ = Mittelwert der Stichprobe

$\hat{\sigma}$ = Standardabweichung geschätzt für Population aus der Stichprobe

Man sieht, dass die Mittelwertsdifferenz stark ausgeprägt ist. Da wir hier nicht die Möglichkeit hatten, eine Stichprobenplanung vorzunehmen, sollten wir noch die post-hoc-Power bzw. Teststärke mithilfe von G*Power berechnen.

Teststärke. Nun stellt sich die Frage, wie wahrscheinlich eine extremere Prüfgröße als die kritische Prüfgröße unter der Alternativhypothesenverteilung beobachtet wird, wenn die empirische Prüfgröße als deren Erwartungswert angenommen wird (Teststärke). Zuerst starten wir G*Power. Dann wählen wir unter der Option TEST FAMILY „T TESTS“ und dann unter STATISTICAL TEST „MEANS: DIFFERENCE FROM CONSTANT (ONE SAMPLE CASE)“. Schließlich wählen wir unter TYPE OF POWER ANALYSIS die Option POST HOC: COMPUTE ACHIEVED POWER – GIVEN α, SAMPLE SIZE, AND EFFECT SIZE. Nun können wir die entsprechenden Werte eintragen: TAIL(S) setzen wir auf TWO, als Effektstärke setzen wir EFFECT SIZE auf 0.75, α auf 0.05 und schließlich die Stichprobengröße TOTAL SAMPLE SIZE auf 15. Klicken wir jetzt auf CALCULATE, dann erhalten wir die Teststärke im Kästchen neben POWER ($1 - \beta$ ERR PROB) angezeigt. Der Wert lautet 0.77. Damit erreicht die Teststärke fast den gewünschten Wert von 80 Prozent.

Fazit. Fasst man die Ergebnisse zusammen, wird ein statistisch bedeutsamer Unterschied zwischen den psychiatrischen Patienten und dem klinisch relevanten Grenzwert von 18 statistisch abgesichert. Der Mittelwertsunterschied fällt groß aus. Die Teststärke ist ausreichend. In einem Artikel würden wir die Ergebnisse wie folgt berichten: $t\,(14) = 2.9$ (zweiseitig), $p < .05$, $g = -.75$, $1 - \beta = .77$ (post-hoc).

1 Die Interpretation von Effektstärken ist nicht unbedingt nötig, wenn die tatsächliche Differenz sinnvoll interpretierbar ist: Beispielsweise liegt eine Mittelwertsdifferenz von 30000 Euro oder ein Temperaturunterschied von 10 Grad vor.

Voraussetzungen. Voraussetzungen für t-Tests für eine Stichprobe:

1. Die Messwerte der Personen müssen voneinander unabhängig sein.
2. Die Messwerte müssen intervallskaliert sein.
3. Die untersuchte abhängige Variable muss sich in der Grundgesamtheit (und der Stichprobe) normal verteilen.

Die Unabhängigkeit der Beobachtungen (Messungen) muss angenommen werden, da sonst die nominelle α–Fehlerwahrscheinlichkeit (siehe Kapitel 4) durch den Test nicht eingehalten werden kann. Testen lässt sich diese Annahme nur schwer. Einen Anhaltspunkt für die Unabhängigkeit der Beobachtungen gibt jedoch die sogenannte Intraklassenkorrelation. In Kapitel 4 findet sich hierzu eine weiterführende Erklärung. Da wir bei der Berechnung des t-Tests Differenzen bilden, müssen die Messwerte der AV intervallskaliert sein (siehe Kapitel 2). Man kann auch prüfen, ob Intervallskalenniveau für bestimmte Messwerte vorliegt, z. B. über Item-Response-Modelle wie das Rasch-Modell. Die Normalverteilung muss schließlich angenommen werden, da wir eine t-Verteilung als Prüfgrößenverteilung heranziehen. Ob sich die Daten normal verteilen, lässt sich mit dem Kolmogoroff-Smirnoff-Test oder noch besser per Inspektion des Normalverteilungsdiagramms prüfen (siehe Kapitel 2). Ist die Annahme der Normalverteilung verletzt, approximiert die t-Verteilung nicht mehr den Verlauf der Prüfgröße t. Allerdings ist der t-Test relativ robust gegenüber Verletzungen der Normalverteilungsannahme (Rasch & Guiard, 2004). Problematisch ist jedoch, wenn Ausreißerwerte vorliegen, da Mittelwerte davon stark verzerrt werden.

Ist man sich nicht sicher, ob ein t-Test der richtige Test bei einer Verletzung der Normalverteilungsannahme ist, kann die Einhaltung der α-Fehlerwahrscheinlichkeit durch ein nonparametrisches Verfahren noch einmal überprüft werden (siehe Abschnitt 5.2). Das nonparametrische Verfahren ist der Vorzeichentest, der in diesem Buch nicht dargestellt wird. Interessierte Leser seien hier an Fahrmeir, Künstler, Pigeot und Tutz (2004, S. 436) verwiesen. Im nächsten Abschnitt betrachten wir nun, wie sich der t-Test für eine Stichprobe mithilfe von SPSS durchführen lässt.

t-Test für eine Stichprobe mit SPSS. Auf der Pearson-Website finden Sie einen Datensatz (Buch_t_Test_eine Stichprobe.sav) mit $N = 90$ psychiatrischen Patienten. Die Variable BDI_t1 beinhaltet die BDI-Werte vor einer Therapie. Es soll geprüft werden, ob sich der Mittelwert der psychiatrischen Patienten vom Wert 18 unterscheidet (zweiseitig, $\alpha = 0.05$).

Bitte klicken Sie in SPSS auf ANALYSIEREN → MITTELWERTE VERGLEICHEN und T-TEST BEI EINER STICHPROBE. Es öffnet sich das Fenster in Abbildung 5.2.

Abbildung 5.2: Darstellung der Berechnung eines t-Tests für eine Stichprobe in SPSS

Markieren Sie nun im linken Fenster die Variable BDI_t1 (diese ist in der Darstellung bereits in das rechte Fenster überführt) und klicken Sie dann auf den Pfeil zwischen den beiden Fenstern. Geben Sie dann den festen Wert 18 als TESTWERT ein ❶. Klicken Sie danach auf die Taste OK ❷. Die SPSS-Ausgabe in Abbildung 5.3 ergibt sich.

Statistik bei einer Stichprobe

	N	❶ Mittelwert	❷ Standard-abweichung	❸ Standardfehler des Mittelwertes
BDI_t1	90	17,3333	10,21565	1,07682

Abbildung 5.3: Darstellung der deskriptiven Statistiken eines t-Tests für eine Stichprobe in SPSS

In Abbildung 5.3 sind der Mittelwert ❶ und die Standardabweichung ❷ angegeben sowie der Standardfehler des Mittelwerts ❸. Es zeigt sich schon deskriptiv, dass der Stichprobenmittelwert ❶ sehr nahe am Testwert (fester Wert) 18 liegt.

In Abbildung 5.4 sieht man, dass die mittlere Differenz zwischen psychiatrischen Patienten und dem Testwert (festen Wert) nur –0.67 ❹ beträgt. Gleichzeitig wird ein Konfidenzintervall ❺ für die mittlere Differenz angegeben. Dieses Konfidenzintervall hat die Grenzen –2.81 und 1.47. Da dieses Konfidenzintervall den Wert „null" enthält, kann man bereits daraus schließen, dass die mittlere Differenz von –0.67 ❹ nicht signifikant vom Wert 18 abweicht.

Der empirische t-Wert ❶ beträgt etwa –0.62 bei df = 89 (90 – 1) Freiheitsgraden ❷. Die Überschreitungswahrscheinlichkeit p ❸ beträgt 0.537. Damit liegt sie über unserer zuvor gewählten α-Fehlerwahrscheinlichkeit von 0.05. Das bedeutet, dass die beobachtete Differenz statistisch nicht signifikant ist.

Test bei einer Stichprobe

	Testwert = 18					
					5 95 % Konfidenzintervall der Differenz	
	T	Df	Sig. (2-seitig)	Mittlere Differenz	Untere	Obere
BDI_t1	1 –,619	2 89	3 ,537	4 –,66667	–2,8063	1,4730

Abbildung 5.4: Darstellung von t-Wert und Überschreitungswahrscheinlichkeit eines t-Tests für eine Stichprobe in SPSS

Nun benötigen wir aber noch die Effektstärke als Anhaltspunkt dafür, ob die gefundene Differenz praktisch bedeutsam ist. Die Effektstärke können wir schnell mit dem Taschenrechner ermitteln: Es ist die mittlere Differenz -0.67 4, dividiert durch die Standardabweichung 10.22 (Abbildung 5.3, 2) und damit ist $|d_s| = 0.07$. Das heißt, dass ein sehr kleiner Effekt vorliegt. Aufgrund dieser Ergebnisse kann man davon ausgehen, dass der mittlere BDI-Wert von psychiatrischen Patienten dem BDI-Wert von 18 entspricht. Die Stichprobe weist im Mittel klinisch auffällige Depressionssymptome auf. Die Teststärke in diesem Beispiel zu berechnen erübrigt sich. Sie wird sehr gering ausfallen. Einen Effekt von 0.07 statistisch abzusichern, erscheint für dieses Beispiel nicht sinnvoll. Ein solcher Effekt ist viel zu klein, um von einer praktisch bedeutsamen Abweichung vom Grenzwert 18 auszugehen, wenn wir zudem berücksichtigen, dass es sich um psychiatrische Patienten handelt. Daher verzichten wir hier auf die Darstellung der a-priori-Stichprobenplanung mit G*Power und verweisen auf Kapitel 4.

5.1.2 t-Test für abhängige Stichproben

Fragestellung. Im Untersuchungsdesign, welches mit einem t-Test für abhängige Stichproben ausgewertet wird, werden mehrere Personen zu zwei Messzeitpunkten mit demselben Instrument untersucht. Die Fragestellung lautet, ob die Differenz zwischen den Mittelwerten zu den beiden Testzeitpunkten null ist. Mithilfe des t-Werts kann eine Entscheidung darüber getroffen werden, ob die Differenz dieser beiden Mittelwerte von null verschieden ist.

Hypothesen. Die **Nullhypothese** lautet, dass die mittlere Differenz der Messwertpaare in der Grundgesamtheit null ist. Die ungerichtete **Alternativhypothese** lautet, dass die mittlere Differenz der Messwertpaare in der Grundgesamtheit nicht null ist. Bei einer gerichteten Alternativhypothese bedeutet dies, dass die Differenz in der Grundgesamtheit entweder größer oder kleiner als null ist.

Beim t-Test für abhängige Stichproben ist also die Prüfung der folgenden Hypothesen möglich:

Nullhypothesen und ungerichtete Alternativhypothesen:

Nullhypothese:

H_0: $\mu_{Diff} = 0$

Alternativhypothese ungerichtet:

H_1: $\mu_{Diff} \neq 0$

Nullhypothesen und gerichtete Alternativhypothesen:

H_0: $\mu_{Diff} \geq 0$ und H_1: $\mu_{Diff} < 0$

H_0: $\mu_{Diff} \leq 0$ und H_1: $\mu_{Diff} > 0$

μ_{Diff} = mittlere Differenz in der Grundgesamtheit

Bei einer gerichteten Fragestellung ist darauf zu achten, dass der t-Wert auch in der erwarteten Richtung ausfällt. Das heißt, wenn unter der Alternativhypothese eine positive Differenz erwartet wird, wird die Nullhypothese auch dann beibehalten, wenn sich eine signifikante, negative Differenz ergibt.

Beispiel. In einem Unternehmen herrscht Unzufriedenheit über einen Trainingsanbieter. Einzelne Mitarbeiter berichten, ein angebotenes emotionales Intelligenztraining hätte bei einzelnen Teilnehmern sogar negative Auswirkungen gehabt. Im Rahmen eines der nächsten emotionalen Intelligenztrainings wird vor und nach dem Training ein Fragebogen zur Erfassung emotionaler Intelligenz vorgegeben. Von 100 mittleren Führungskräften eines Unternehmens werden zufällig 20 ausgewählt, die dieses Training absolvieren. Die Gruppe erhält den Fragebogen zur emotionalen Intelligenz vor und nach dem Training. Das Ergebnis jeder Person wird mithilfe einer Normtabelle in T-Werte umgewandelt (T-Werte siehe Kapitel 2). Vor dem Training ergibt sich für diese Gruppe von $N = 20$ Führungskräften in dem Fragebogen ein Mittelwert von T = 56 und nach dem Training von T = 58. Die Standardabweichung beträgt zu beiden Messzeitpunkten $\hat{\sigma} = 8$. Die Korrelation zwischen den Messwerten zum Zeitpunkt 1 und 2 beträgt $r = 0.78$ (Korrelation siehe Abschnitt 7.1). Die Frage, die geklärt werden soll, ist, ob das Training zu einer Veränderung der emotionalen Intelligenz führt. In diesem Fall entscheiden wir uns für eine zweiseitige Fragestellung, da aus der Beobachtung heraus auch negative Effekte des Trainings möglich sind.

Berechnung. Zunächst sind die Formeln für die Ermittlung der Freiheitsgrade, für den t-Wert und den Standardfehler der Differenz dargestellt. Das Prinzip entspricht dem gängigen Vorgehen, den Mittelwertsunterschied durch den Standardfehler zu teilen. Die Prüfgröße zeigt somit an, wie viele Standardabweichungen die beobachtete Differenz vom Mittelwert der Nullhypothesenverteilung (Verteilung aus Mittelwertsdifferenzen) entfernt ist. Beim Standardfehler handelt es sich hier um den Standardfehler der Differenz zwischen Testzeitpunkt 1 und 2.

$$df = N - 1$$
$$t_{df} = \frac{\bar{x}_D}{SE_D}$$
$$\bar{x}_D = \bar{x}_1 - \bar{x}_2$$
$$SE_D = \frac{\hat{\sigma}_D}{\sqrt{N}}$$
$$\hat{\sigma}_D = \sqrt{\frac{\sum_{i=1}^{N}(x_{Di} - \bar{x}_D)^2}{N-1}} = \sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2 \cdot r_{12} \cdot \hat{\sigma}_1 \cdot \hat{\sigma}_2}$$

t_{df} = t-Wert der t-Verteilung mit einer bestimmten Anzahl von Freiheitsgraden df

$\bar{x}_D$ = Differenz der Mittelwerte zu Testzeitpunkt 1 und 2

SE_D = Standardfehler der Differenz

N = Stichprobengröße

$\hat{\sigma}_D$ = Standardabweichung der Differenz für die Grundgesamtheit aus der Stichprobe geschätzt

x_{Di} = Differenz der Werte zwischen Messzeitpunkt 1 und 2 für eine Person i

r_{12} = Korrelation zwischen den Messwerten zu Testzeitpunkt 1 und 2

$\hat{\sigma}_1$ = Standardabweichung zu Testzeitpunkt 1

$\hat{\sigma}_2$ = Standardabweichung zu Testzeitpunkt 2

Für unser Beispiel bedeutet dies:

$$df = N - 1 = 20 - 1 = 19$$
$$\bar{x}_D = 56 - 58 = -2$$
$$\hat{\sigma}_D = \sqrt{\frac{\sum_{i=1}^{N}(x_{D_i} - \bar{x}_D)^2}{N-1}} \; oder \; \hat{\sigma}_D = \sqrt{\hat{\sigma}_{t_1}^2 + \hat{\sigma}_{t_2}^2 - 2 \cdot r_{12} \cdot \hat{\sigma}_{t_1} \cdot \hat{\sigma}_{t_2}}$$
$$= \sqrt{8^2 + 8^2 - 2 \cdot .78 \cdot 8 \cdot 8} = \sqrt{64 + 64 - 99.84} = \sqrt{28.16} = 5.31$$
$$SE_D = \frac{\hat{\sigma}_D}{\sqrt{N}} = \frac{5.31}{\sqrt{20}} = 1.19$$
$$t_{19} = \frac{\bar{x}_D}{SE_D} = \frac{-2}{1.19} = -1.68$$

N = Stichprobengröße

$\bar{x}_D$ = Differenz der Mittelwerte zu Testzeitpunkt 1 und 2

$\hat{\sigma}_D$ = Standardabweichung der Differenz für die Grundgesamtheit aus der Stichprobe geschätzt

x_{Di} = Differenz der Werte zwischen Messzeitpunkt 1 und 2 für eine Person i

SE_D = Standardfehler der Differenz

t_{19} = t-Wert der t-Verteilung bei 19 Freiheitsgraden

Entscheidung. Der kritische t-Wert bei df = 19 Freiheitsgraden beträgt –2.093 (zweiseitig, $\alpha = 0.05$). Der empirische t-Wert beträgt –1.68. Die Überschreitungswahrscheinlichkeit p für den empirischen t-Wert beträgt $p = 0.11$ (zur Bestimmung des kritischen t-Werts und der Überschreitungswahrscheinlichkeit siehe Kapitel 4). Damit besteht kein statistisch signifikanter Unterschied in der emotionalen Intelligenz vor und nach dem Training. Die Nullhypothese wird also beibehalten. Betrachten wir nun die praktische Bedeutsamkeit und die Teststärke.

Praktische Bedeutsamkeit. Auch hier soll die Effektstärke berechnet werden, um die praktische Bedeutsamkeit des Mittelwertsunterschieds abschätzen zu können.

$$d_Z = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_D} = \frac{-2}{5.31} = -.38$$

d_Z = Effektstärke für abhängige Stichproben von Cohen

$\bar{x}_1$ = Mittelwert zu Testzeitpunkt 1

$\bar{x}_2$ = Mittelwert zu Testzeitpunkt 2

$\hat{\sigma}_D$ = Standardabweichung der Differenz für die Grundgesamtheit aus der Stichprobe geschätzt

Die Effektstärke beträgt –0.38. Es handelt sich also um einen geringen bis moderaten Effekt. Die Effektstärke bei abhängigen Stichproben ist insofern schwierig zu interpretieren, da die Standardabweichung der Differenz nicht normiert ist. Im Falle von Normwerten beträgt z. B. die Standardabweichung für IQ-Werte 15 und bei T-Werten 10 in der Normgruppe (nicht unbedingt in der Stichprobe). Bildet man mithilfe standardisierter Werte die Differenz, unterscheidet sich die Standardabweichung der Differenz von diesen normierten Standardabweichungen. Betrachten wir dazu kurz die Formel für die Standardabweichung der Differenz:

$$\hat{\sigma}_D = \sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2 \cdot r_{12} \cdot \hat{\sigma}_1 \cdot \hat{\sigma}_2}$$

$\hat{\sigma}_D$ = Standardabweichung der Differenz für die Grundgesamtheit aus der Stichprobe geschätzt

r_{12} = Korrelation zwischen den Messwerten zu Testzeitpunkt 1 und 2

$\hat{\sigma}_1$ = Standardabweichung zu Testzeitpunkt 1

$\hat{\sigma}_2$ = Standardabweichung zu Testzeitpunkt 2

Nehmen wir vereinfachend an, die Varianzen zu beiden Messzeitpunkten seien gleich: $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}^2$. Dann vereinfacht sich die Formel:

$$\hat{\sigma}_D = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2 - 2 \cdot r_{12} \cdot \hat{\sigma} \cdot \hat{\sigma}} = \sqrt{2 \cdot \hat{\sigma}^2 - 2 \cdot r_{12} \cdot \hat{\sigma}^2} = \sqrt{2 \cdot \hat{\sigma}^2 (1 - r_{12})} = \sqrt{2} \cdot \hat{\sigma} \cdot \sqrt{(1 - r_{12})}$$

$\hat{\sigma}_D$ = Standardabweichung der Differenz für die Grundgesamtheit aus der Stichprobe geschätzt

$\hat{\sigma}_1$ = Standardabweichung zu Testzeitpunkt 1

$\hat{\sigma}_2$ = Standardabweichung zu Testzeitpunkt 2

Gehen wir diese Formel einmal für drei verschiedene Korrelationen durch, ($r_{12} = 0$, $r_{12} = 0.19$ und $r_{12} = 0.84$).

$$r_{12} = 0 \qquad \hat{\sigma}_D = \sqrt{2} \cdot \hat{\sigma} \cdot \sqrt{(1-0)} = \sqrt{2} \cdot \hat{\sigma}$$

$$r_{12} = .19 \qquad \hat{\sigma}_D = \sqrt{2} \cdot \hat{\sigma} \cdot \sqrt{.81} = \sqrt{2} \cdot \hat{\sigma} \cdot .90$$

$$r_{12} = .84 \qquad \hat{\sigma}_D = \sqrt{2} \cdot \hat{\sigma} \cdot \sqrt{.16} = \sqrt{2} \cdot \hat{\sigma} \cdot .40$$

Korrelieren die beiden Messwertreihen also gar nicht, ergibt sich die Standardabweichung der Differenz einfach durch Multiplikation der Standardabweichung zu beiden Messzeitpunkten mit der Wurzel aus 2. Nimmt die Korrelation nun zu, wird dieser Wert mit einer zweiten Konstanten multipliziert. Je höher die Korrelation ist, desto geringer wird diese zweite Konstante.

Das heißt, je stärker zwei Messwertreihen korrelieren[2], desto geringer wird die Standardabweichung der Differenz. Damit sinkt auch der Standardfehler und der empirische t-Wert wird größer. So wird es leichter, ein signifikantes Ergebnis zu erzielen, da der kritische t-Wert einfacher überschritten werden kann. Die Korrelation hat also einen Einfluss auf die Teststärke. Eine andere, weniger erfreuliche Konsequenz ist, dass die Effektstärken zwischen verschiedenen Studien nicht mehr vergleichbar sind, wenn sich die Korrelationen der Messwertreihen unterscheiden.

Um die *Effektstärken* zwischen Studien interpretierbar zu halten, gibt es nun drei Möglichkeiten:

(1) Die erste Möglichkeit besteht darin, den ersten Messzeitpunkt als Baseline-Messung anzunehmen. Das ist vor allem in Therapiestudien der Fall. Nehmen wir nun die Standardabweichung dieses Messzeitpunkts zum Standardisieren der Mittelwertsdifferenz, drückt die Effektstärke aus, um wie viele Standardabweichungen (*vor* der Therapie) sich der Wert verschiebt.

(2) Die zweite Möglichkeit besteht darin, zum Standardisieren der Mittelwertsdifferenz nicht die Standardabweichung der Differenz zu nutzen, sondern einfach

2 Korrelieren heißt, dass mit höheren Merkmalsausprägungen zu Testzeitpunkt 1 im Durchschnitt auch höhere Ausprägungen zu Messzeitpunkt 2 einhergehen. Das heißt für den t-Test für abhängige Stichproben: wenn Mittelwertsunterschiede auftreten, gleichen sich die Differenzwerte der Personen zwischen Messzeitpunkt 1 und 2 mehr und mehr an, je höher die Korrelation ausfällt.

die beiden Messzeitpunkte wie unabhängige Stichproben zu behandeln und die Streuung zu poolen (siehe t-Test für unabhängige Stichproben). Dies empfiehlt sich vor allem, wenn die Veränderung der Messzeitbedingungen nicht auch zu einer Veränderung der Varianzen führt. Das heißt, dass die Messzeitpunktvarianzen ungefähr gleich ausfallen. Die Mittelwertsdifferenz wird dann einfach anhand der gemittelten Standardabweichung der Messzeitpunkte normiert.

(3) Die dritte Möglichkeit kann nur genutzt werden, wenn die Messwerte in Normwerte überführt werden können. Dann kann die Standardabweichung des Normwerts im Nenner der Effektstärke verwendet werden. In unserem Fall würde dies bedeuten, dass wir –2 durch die Standardabweichung für den T-Normwert (σ = 10) teilen. Damit erhalten wir eine kleine Effektstärke von $d_z = -0.20$. Dieses Vorgehen ist jedoch die Ausnahme, da nicht immer geeignete Normstichproben zur Verfügung stehen. Keine der erörterten Methoden ist frei von Problemen und wir empfehlen daher, zum Standardisieren dennoch die Standardabweichung der Differenz zu nutzen. Dies entspricht immer noch der Konvention und wird in den meisten Statistikprogrammen und Lehrbüchern so gehandhabt. Betrachten wir nun weiter, wie hoch die Teststärke für unseren beobachteten Effekt ausfällt.

Teststärke. Auch hier kann die Teststärke post-hoc berechnet werden. Nachdem wir G*Power gestartet haben, wählen wir unter der Option TEST FAMILY „T TESTS" aus. Anschließend entscheiden wir uns unter STATISTICAL TEST für MEANS: DIFFERENCE BETWEEN TWO DEPENDENT MEANS (MATCHED PAIRS). Schließlich wählen wir unter TYPE OF POWER ANALYSIS die Option POST HOC: COMPUTE ACHIEVED POWER – GIVEN α, SAMPLE SIZE, AND EFFECT SIZE. Nun können wir die entsprechenden Werte eintragen: TAIL(S) setzen wir auf TWO, als Effektstärke setzen wir EFFECT SIZE auf 0.38, α auf 0.05 und schließlich die Stichprobengröße TOTAL SAMPLE SIZE auf 20. Klicken Sie jetzt auf die Schaltfläche CALCULATE und Sie erhalten die Teststärke im Kästchen neben POWER (1 – β ERR PROB) angezeigt. Der Wert lautet 0.36. Wenn wir dennoch an den Effekt glaubten, würden wir hier eine sehr viel größere Stichprobe benötigen, um den Effekt von 0.38 statistisch absichern zu können.

Fazit. Fasst man die Ergebnisse zusammen, unterscheiden sich die Mittelwerte vor und nach dem Training statistisch nicht signifikant. Der Unterschied ist als gering bis moderat zu bewerten. Die Teststärke ist ebenfalls gering. Das heißt, um einen so geringen Effekt statistisch abzusichern, wäre eine weitaus größere Stichprobe nötig gewesen. Glaubte man an den Effekt, sollte nun mittels Kosten-Nutzen-Überlegungen geprüft werden, ob es sich lohnt, einen solchen Effekt statistisch abzusichern. Wie die a-priori-Stichprobenplanung mit G*Power funktioniert, haben wir in Kapitel 4 erklärt. In einem Artikel würden wir die Ergebnisse wie folgt berichten: $t\,(19) = 1.68$ (zweiseitig), *n.s.*, $g = -.38$, $1 - \beta = .36$.

Voraussetzungen. Voraussetzungen für t-Tests für abhängige Stichproben:

1. Die Messwerte der Personen zu jedem einzelnen Messzeitpunkt müssen unabhängig sein.

2. Die Messwerte müssen intervallskaliert sein.

3. Die Differenzen zwischen Messzeitpunkt 1 und 2 müssen sich in der Grundgesamtheit normal verteilen.

Die ersten beiden Voraussetzungen kennen wir in ähnlicher Form bereits vom t-Test für eine Stichprobe. Die Unabhängigkeit der Messungen bezieht sich hier jeweils auf einen Messzeitpunkt. Das bedeutet, dass die Personen sich nicht gegenseitig in der abhängigen Variablen beeinflussen dürfen bzw. beeinflusst werden. Wiederum lässt sich dies eigentlich nicht testen. Das Intervallskalenniveau begründet sich hier wieder mit der Differenzbildung. Auch besteht hier die Möglichkeit für Tests der Annahme (z. B. Rasch-Analysen), die jedoch selten genutzt werden. Die Normalverteilungsannahme bezieht sich hier nicht mehr auf die Verteilung der Messwerte, sondern vielmehr auf die Verteilung der Messwertdifferenzen. Die Prüfgrößenverteilung ist wieder eine t-Verteilung, die normalverteilt ist. Daher ist diese Voraussetzung notwendig. Um sie zu testen, müssen zunächst für jede Person die Messwertdifferenzen berechnet werden (siehe Abbildung 5.12). Anschließend lässt sich mithilfe eines Histogramms per Inspektion überprüfen, ob eine Normalverteilung vorliegt. Betrachten wir nun das Vorgehen in SPSS.

t-Test für abhängige Stichproben mit SPSS. Auf der Pearson-Website finden Sie den Datensatz „Buch_t_Test_abhängig.sav“ mit $N = 90$ Personen, die vor und nach einer Behandlung untersucht wurden. Die Variable BDI_t1 beinhaltet die BDI-Werte zu Testzeitpunkt 1 und die Variable BDI_t2 die BDI-Werte zu Testzeitpunkt 2. Wir gehen davon aus, dass sich der Mittelwert im BDI von Testzeitpunkt 1 zu Testzeitpunkt 2 reduziert (einseitige Hypothese, $\alpha = 0.05$).

Bitte klicken Sie in SPSS auf ANALYSIEREN → MITTELWERTE VERGLEICHEN und T-TEST BEI VERBUNDENEN STICHPROBEN. Es öffnet sich das Fenster in Abbildung 5.5.

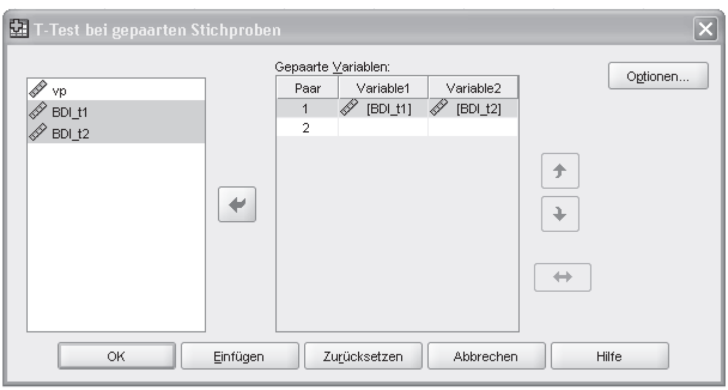

Abbildung 5.5: Darstellung der Berechnung eines t-Tests für abhängige Stichproben in SPSS