SFU Forschungsbulletin SFU Research Bulletin 8/2 (2020), S. 128-146 DOI: 10.15135/2020.8.2.128-146

Der t-, Wilcoxon- und Binomial/Vorzeichen-Test für zwei verbundene Stichproben im psychotherapiewissenschaftlichen Forschungskontext

– Empfehlungen für Anwendung und Interpretation –

t-Test for paired samples, paired sample Wilcoxon sign rank test and the paired sample sign test in psychotherapy science

– Recommendations for application and interpretation –

Sophie Gattermeyer, Clara Vladarski, Jan Aden

Kurzzusammenfassung

In diesem zweiten Beitrag der Serie Statistik in der Psychotherapiewissenschaft wird die Anwendung des t-, Wilcoxon- sowie Binomial/Vorzeichen- Tests bei zwei verbundenen Stichproben im Sinne eines Best-Practice Ansatzes vorgestellt. Neben (1) Empfehlungen für eine optimale Verfahrenswahl, (2) dem Einsatz von Effektstärken, (3) der Bestimmung der Ergebnisrelevanz sowie (4) der Vorstellung von Reportkonventionen für die Ergebnisdarstellung wird vor allem (5) auf das Problemfeld der Zuverlässigkeit statistischer Entscheidungen im psychotherapiewissenschaftlichen Forschungskontext und (6) Möglichkeiten zur aktiven Einflussnahme durch die Forscher*innen eingegangen.

Schlüsselwörter

t-Test für verbundene Stichproben, Wilcoxon-Test, Binomial-/Vorzeichentest, abhängige Stichproben, Messwiederholung, Reportkonventionen, Oversampling, Undersampling, Effektstärken, statistische Entscheidungen

Abstract

In this first contribution to the Statistics series in psychotherapy science, the application use of the t-Test for paired samples, paired sample Wilcoxon sign rank test and the paired sample sign test is presented in the sense of a best practice approach. In addition (1) to recommendations for the optimal choice of procedure, (2) the use of effect sizes, (3) the designation of relevant results and (4) report conventions for the presentation of results, (5) the problem of reliable statistical decision making in the research context of psychotherapy science is addressed and therefore, (6) suggestions for dealing with this potential problem are identified.

keywords

t-Test for paired samples, paired sample sign test, paired sample Wilcoxon sign rank test, reporting conventions, Oversampling, Undersampling, statistical decision making

Einsatzfeld und Background

Die stetige Weiterentwicklung und empirische Evaluation von Therapiemethoden ist ein in den Psychotherapiewissenschaften zentrales Anliegen, welches nicht zuletzt die Voraussetzung für eine langfristige Legitimation von therapeutischen Vorgehensweisen und Handlungspraxen schafft und in der Vergangenheit bereits geschaffen hat. Belege der Wirksamkeit und Effektivität dienen in diesem Zusammenhang neben dem Zweck der allgemeinen Qualitätssicherung (vgl. Grawe, 1992, 2005) ebenso der Stärkung evidenzbasierter Praxis und Förderung der Akzeptanz psychotherapeutischer Behandlungen. Zur Beurteilung der Wirksamkeit oder des Erfolgs einer Therapie können unterschiedliche Indikatoren herangezogen werden. Eine gängige Operationalisierung des Erfolgs und der Effektivität psychotherapeutischer Behandlungen stellt der Rückgang der Ausprägung psychopathologischer Symptome dar (vgl. Spinhoven, Klein, Kennis et al., 2018). Ebenso können die Verminderung subjektiver Leidenszustände oder die Promotion salutogener Faktoren (z.B. Resilienz) als Erfolgskriterien fungieren. In der psychotherapiewissenschaftlichen Forschung kommen in diesem Zusammenhang sogenannte „abhängige Forschungsdesigns“ mit Messwiederholungen (z.B. vor Therapie und nach Therapie) zum Einsatz. Der Artikel in dieser Ausgabe des Bulletins befasst sich mit statistischen Verfahren zur Überprüfung eben solcher abhängiger Studiendesigns. Speziell werden in dieser Ausgabe statistische Testverfahren zur Überprüfung von Unterschieden zwischen zwei abhängigen Stichproben vorgestellt.

Um bei der Anwendung statistischer Testverfahren bei zwei abhängigen Gruppen stichhaltige Ergebnisse zu erhalten, ist ein korrektes methodisches Vorgehen im Sinne der korrekten Wahl und Interpretation des geeigneten statistischen Verfahrens von großer Relevanz.

Einsatzbereiche und Fragestellungen

Wie bereits im ersten Artikel der Serie Statistik des Forschungsbulletins (8/1, 2020) (Seistock, Bunina & Aden, 2020) erwähnt, stellt die quantitative Methodik eine breite Auswahl an statistischen Verfahren zur Ermittlung von relevanten Unterschieden bei psychotherapiewissenschaftlichen Fragestellungen zur Verfügung. Unterschiedstestungen für zwei verbundene Stichproben[1], wie sie in diesem Artikel dargelegt werden, kommen häufig innerhalb der Wirksamkeits- und Effektivitätsforschung von Therapiemethoden zum Einsatz. Unterschiedstestungen zwischen zwei verbundenen Stichproben erscheinen dabei meist in Form einer systematischen Erfassung von Messwertunterschieden zwischen zwei Zeitpunkten. Im Zuge eines solchen Untersuchungsdesgins werden – im Gegensatz zu Unterschiedstestungen bei unverbundenen Stichproben – die intraindividuellen Unterschiede in einem Merkmal untersucht, d.h. wie stark sich ein bestimmter Wert innerhalb einer Person (z.B. von Zeitpunkt 1 zu Zeitpunkt 2) verändert. Daraus folgend wird es beispielsweise möglich, eine wünschenswerte Verbesserung des psychischen Zustands aufdecken und einen signifikanten Unterschied zwischen den Werten vor und nach einer Therapie in einem interessierenden psychologischen Merkmal abbilden zu können.

[1] Es werden die Begriffe „verbunden“ und „abhängig“ synonym verwendet. Ebenso finden die Begriffe „Stichprobe“, „Gruppe“ sowie „Messzeitpunkt“ in diesem Beitrag eine synonyme Verwendung.

Im Rahmen der inferenz-statistischen Verfahren für zwei verbundene Stichproben werden die einzelnen intraindividuellen Veränderungen der gesamten Stichprobe analysiert und dadurch versucht, allgemeine Aussagen über die untersuchte Personengruppe zu treffen.

Exkurs: Fragestellungen bei Verfahren für 2 abhängige Stichproben

Messwiederholungen: Fragestellungen, bei denen dieselbe Person vor und nach einer Intervention/Versuchsbedingung/Therapie getestet wird oder bei denen eine Person zu zwei verschiedenen Interventionen/Therapien/Bedingungen getestet wird und ein Vergleich dieser Werte von Interesse ist.

Natürliche Paare: Untersuchung von Personen, die in einer bestimmten Form eine Verbindung aufweisen (beispielsweise Ehepartner, Geschwister, Patient*in-Therapeut*in etc.)

Test-Zwillinge: Proband*innen werden aufgrund von ähnlichen soziodemographischen Merkmalen hinsichtlich einer interessierenden Variable/Versuchsbedingung/Intervention verglichen.

In diesem Zusammenhang kann es beispielsweise von Interesse sein zu überprüfen, ob eine gewählte Intervention den gewünschten Erfolg erzielt und, z.B. im Falle einer Depression, die Depressionswerte nach der Therapie (T_1) signifikant geringer ausfallen als vor der Therapie (T_0) (z.B. Becks-Depressions-Inventar-V, vgl. Kühner, Bürger, Keller & Hautzinger, 2007) (sh. Abb.1 und Tab. 1).

Als weitere Beispiele lassen sich Fragestellungen zu Trainingserfolgen anführen, welche mittels abhängigen Testungen für zwei Stichproben überprüft werden können. Beispielsweise könnte untersucht werden, ob ein gezieltes Training oder die Aktivierung einer psychologischen Fähigkeit bzw. Ressource (z.B. Emotionsregulation) zu einer signifikanten Verbesserung führt, wobei das Merkmal vor dem Training getestet wird (Vergleichswert) sowie nach Absolvierung des Trainings, um eine potenzielle Verbesserung im Sinne eines Unterschieds beider Werte festzustellen. Im Falle beider Beispiele ist zu beachten, dass zu beiden Zeitpunkten das Merkmal mit demselben Messinstrument erhoben worden sein muss, um die Werte sinnvollerweise vergleichen zu können. Fragestellungen wie diese können als stellvertretende Beispiele von klassischen Prä-Post-Designs herangezogen werden, bei welchen untersucht wird, ob sich Werte eines psychologisch relevanten Merkmals vor (Prä) und nach (Post) einer (therapeutischen) Intervention unterscheiden bzw. sich im Verlauf verändern. Weitere Fragestellungen zu zwei Messzeitpunkten (z.B. Verlaufsevaluation wie beispielsweise von Therapiezufriedenheit über den Therapieverlauf) können selbstverständlich als analog betrachtet werden.

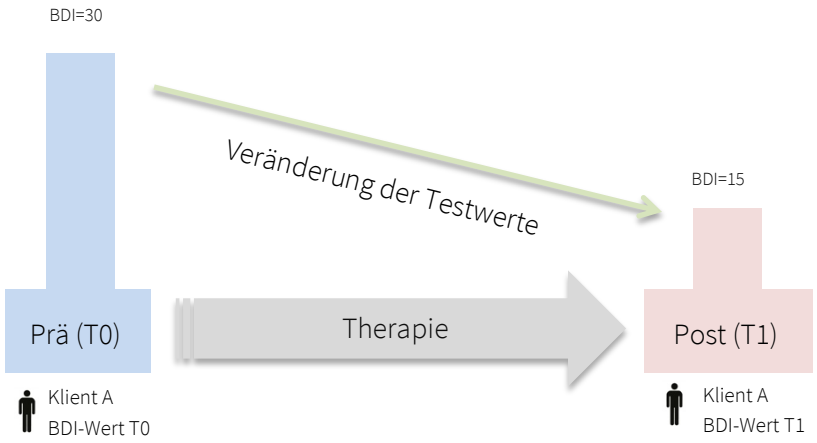


Abb. 1 beispielhafte Darstellung eines Prä-Post-Designs für die erfolgreiche Behandlung einer Major Depression

Tab. 1 Beispiel für eine Unterschiedstestung für zwei verbundene Stichproben

	Messzeitpunkt 1 (Prä= T_0)	***Messzeitpunkt 2 (Post=T_1)***
KlientIn A	*29*	*15*
KlientIn B	*40*	*20*
KlientIn C	*36*	*18*
…	*…*	*…*

Ebenso können Unterschiedstestungen für zwei verbundene Stichproben bei experimentellen Designs zum Einsatz kommen, welche beispielsweise das Ziel haben, Veränderungen in einem psychologischen Merkmal aufgrund einer Manipulation durch eine Versuchungsbedingung zu untersuchen bzw. festzustellen, ob mit der Manipulation einer Versuchsbedingung eine Veränderung der Testwerte einhergeht. Wichtige Handlungsschritte für die korrekte Durchführung eines Experiments finden in einem weiteren Artikel der Serie Statistik gesonderte Betrachtung.

Im Rahmen von Therapieevaluationen ist es gebräuchlich und sinnvoll, nicht nur die Werte von zwei verbundenen Gruppen auf eine Veränderung hin zu untersuchen, sondern ebenso den Vergleich zu einer Kontrollgruppe hinzuzuziehen, um ein bestmögliches Ausmaß an Validität gewährleisten und einen gefundenen Effekt tatsächlich auf die therapeutische Intervention zurückführen zu können (und nicht auf unbekannte Drittvariablen). Der kombinierte Vergleich von abhängigen Messwerten sowie jenen einer Kontrollgruppe lässt sich mittels komplexerer inferenzstatistischer Verfahren (Mixed ANOVA) überprüfen, welche in einem weiteren Artikel der Serie Statistik vorgestellt werden. Folgende Betrachtungen beziehen sich daher nur auf den Einsatz von abhängigen Testungen von zwei verbundenen Gruppen ohne Kontrollgruppe.

Bestimmung der Stichproben-art (verbunden vs. unverbunden)

Eine notwendige Voraussetzung für eine Untersuchung mit zwei verbundenen Stichproben ist im ersten Schritt die Identifikation und **Bestimmung der korrekten Stichprobenart**. Wie bereits im ersten Methodenartikel der Serie Statistik (Seistock, Bunina & Aden, 2020) beschrieben, wird zwischen verbundenen und nicht-verbundenen Gruppen unterschieden, wobei im ersten Artikel Bezug auf die verfügbaren Verfahren hinsichtlich nicht-verbundener Gruppen genommen wird, welche sich dadurch auszeichnen, dass Messwerte nicht paarweise zugeordnet werden können und daher unabhängig voneinander sind (z.B. unterscheiden sich Klient*innen zweier unterschiedlicher Therapieformen (unabhängige Gruppen) hinsichtlich der Zufriedenheit mit der Therapeut*innen-Klient*innenbeziehung?).

In diesem Artikel liegt der Fokus auf **abhängigen Gruppen**, welche daran erkennbar sind, dass deren Werte einander paarweise zugeordnet sind und über einen verbindenden Aspekt verfügen (vgl. Bortz, 2006). Beispielsweise können Geschwisterpaare (verbindender Aspekt = dieselbe Familie) oder Messwiederholungen (verbindender Aspekt= dieselben Personen werden zweimal hintereinander getestet) einander gegenübergestellt werden.

Verfahren für zwei verbundene Stichproben

Um festzustellen, ob sich zwei verbundene Stichproben statistisch signifikant in der Ausprägung eines mindestens rangskalierten (ordinalen) Merkmals unterscheiden, wird in Abhängigkeit bestimmter Voraussetzungen (auf welche folglich genauer eingegangen wird) eines der folgenden Auswertungsverfahren herangezogen:

- **t-Test für abhängige Stichproben**
- **Wilcoxon-Test**
- **Binomial-Test/Vorzeichentest**

Bei den vorliegenden Verfahren handelt es sich ebenfalls um sogenannte inferenzstatistische Verfahren, bei welchen die Formulierung einer Hypothesendarstellung geboten ist.

Hypothesen-formulierung

Hierbei werden eine Null- und Alternativhypothese festgelegt, welche für die Population verallgemeinert werden sollen. Um Bezug auf das vorherige Beispiel bezüglich der Wirksamkeit von Psychotherapie bei depressiv erkrankten Personen zu nehmen, könnte demnach ein entsprechendes Hypothesenpaar wie folgt formuliert werden:

Zweiseitige Testungen

H_0: Es besteht kein signifikanter Unterschied zwischen dem Zeitpunkt vor einer Therapie (T_0) und einer Woche nach Beendigung der Therapie (T_1) hinsichtlich der Depressionswerte. ($\mu T_0 = \mu T_1$)

H_1: Es besteht ein signifikanter Unterschied zwischen dem Zeitpunkt vor einer Therapie (T0) und einer Woche nach Beendigung der Therapie (T_1) hinsichtlich der Depressionswerte. ($\mu T_0 \neq \mu T_1$)

Des Weiteren ist es möglich, bestimmte, vorab entwickelte Vorannahmen bezüglich der Richtung des Unterschieds in der Formulierung der Hypothesen zu berücksichtigen. Der Unterschied zwischen solchen zweiseitigen und einseitigen Unterschiedstestungen bezieht sich hierbei darauf, dass bei einseitigen Hypothesen eine bereits vorgegebene Richtung, hinsichtlich welcher der Unterschied angenommen wird, besteht. Dies ist besonders bei Studien und Untersuchungen zur

Therapiewirksamkeit zu beachten, da in diesen Fällen meist eine Vorannahme – nämlich die Annahme der Wirksamkeit der Intervention – besteht.

Um auf das vorliegende Beispiel Bezug zu nehmen, könnte eine **einseitig formulierte Unterschiedshypothese** folglich lauten:

einseitige Testungen

H_0: Es besteht kein signifikanter Unterschied zwischen dem Zeitpunkt vor einer Therapie (T_0) und einer Woche nach Beendigung der Therapie (T_1) hinsichtlich der Depressionswerte, dahingehend, dass die Depressionswerte nach Beendigung der Therapie geringer sind, als zum Zeitpunkt vor der Therapie ($\mu T_0 \leq \mu T_1$).

H_1: Es besteht ein signifikanter Unterschied zwischen dem Zeitpunkt vor einer Therapie (T_0) und einer Woche nach Beendigung der Therapie (T_1) hinsichtlich der Depressionswerte, dahingehend, dass die Depressionswerte nach Beendigung der Therapie geringer sind, als zum Zeitpunkt vor der Therapie ($\mu T_0 > \mu T_1$).

Die korrekte Anwendung einseitiger Testungen

Hinsichtlich der Interpretation von **einseitigen Unterschiedstestungen** gibt es zwei Punkte zu beachten, die bereits im ersten Artikel der Serie Statistik zu statistischen Testverfahren bei zwei unverbundenen Stichproben erläutert wurden (Seistock, Bunina & Aden, 2020). Zum einen gilt es, den Signifikanzwert, der in Form eines p-Wertes angegeben wird, durch zwei zu dividieren, wodurch ein anfangs nicht-signifikanter p-Wert (z.B. p= .060) nach Halbierung auf einen signifikanten Unterschied mit p= .030 hinweisen kann. Zudem müssen die beiden Gruppen im Anschluss eines signifikanten Ergebnisses anhand deskriptiver Kennwerte verglichen werden. Hierfür sollen der Mittelwert oder der Median der jeweiligen Stichproben (hier T_0 und T_1) einander gegenüber gestellt werden, um infolge dessen die Richtung des Unterschieds bewerten zu können (zum Beispiel der Mittelwert des Depressionsmesswertes vor einer Therapie ist größer als der Mittelwert des Depressionsmesswertes nach der Therapie). Hier gilt es zu beachten, welche Werteausprägung im gewählten Verfahren eine Verbesserung und welche eine Verschlechterung des gemessenen Merkmals bedeutet. Erst nach Berücksichtigung der Deskriptivstatistiken und der Feststellung der angenommenen Richtung des Mittelwertunterschiedes kann die H_0 verworfen werden

Signifikanz-bestimmung

Wie auch in im ersten Artikel der Serie Statistik beschrieben, erfolgt die **Signifikanzbestimmung** und Bewertung bei den drei vorgestellten Verfahren (t-Test für verbundene Gruppen, Wilcoxon-Test und Binomial-Vorzeichentest) für zwei verbundene Stichproben auf der Grundlage von Prüfverteilungen. Mittels dieser wird der Unterschied zwischen zwei abhängigen Wertepaaren (z.B. Depressionswert Zeitpunkt 1 und Depressionswert Zeitpunkt 2) berechnet und in einen numerischen Verteilungswert der jeweiligen Prüfverteilung überführt (je nach Verfahren, z.B. t-Verteilung bei t-Test). Um zu bestimmen, ob es sich um einen rein zufälligen Unterschied handelt oder ob dieser überzufällig, d.h. statistisch bedeutsam ist, wird der berechnete Verteilungswert wiederum in einen Wahrscheinlichkeitswert (p-Wert) transformiert, der durch den Vergleich mit der festgelegten Signifikanzgrenze (Alphaniveau) (üblicherweise 5%) letztendlich Aussagen über die statistische Signifikanz des Unterschiedes beider Werte zulässt. Für genauere Ausführungen zur Bewertung der Signifikanz siehe Seistock, Bunina und Aden (2020).

Statistische und inhaltliche Relevanz

Ein statistisch signifikantes Ergebnis gibt jedoch noch keine Information über die inhaltliche Relevanz wieder, da der p-Wert neben der Größe des tatsächlichen Unterschieds auch stark abhängig von Faktoren wie beispielsweise der Stichprobengröße ist (vgl. Jones, Carley & Harrison, 2003; Krzywinski & Altman, 2013). Bei Wirksamkeitsstudien wie im Beispiel der Depression wäre ersteres unter anderem der Fall, wenn der beobachtete Unterschied im klinisch unauffälligen Bereich zu finden wäre (MBDI T_0= 13, MBDI T_1= 7) oder das Intervall einer beispielsweise moderaten Depression nicht verlassen würde (vgl. Beck, Steer & Brown (1996): nicht klinischer Bereich 0-13 Punkte, milde Depression 14-19 Punkte, moderate Depression 20-28 Punkte, schwere Depression 29-63). Bestünde hingegen ein statistisch signifikanter Unterschied dahingehend, dass der Punktewert im BDI vor der Therapie MBDI T_0= 25 und nach der Therapie MBDI T_1= 12 betragen würde, so wäre der Unterschied auch als inhaltlich bzw. klinisch relevant zu bewerten, da der numerische Unterschied auf eine bedeutsame symptomatische Verbesserung schließen lässt.

Ergebnisverzerrungen können ebenso aufgrund (zu) großer Stichprobengrößen (Oversampling) zustande kommen, was sich darin ausdrücken kann, dass zwar der p-Wert einen signifikanten Unterschied zwischen Testzeitpunkt 1 und 2 anzeigt (p≤.05), dieser jedoch in der Population korrekterweise nicht vorzufinden ist (Fehler 1. Art). Die Berechnungen würden somit aufgrund mangelhafter Stichprobenplanung auf die Wirksamkeit einer Intervention schließen lassen, obwohl diese tatsächlich keinen Effekt hat. Handlungsempfehlungen zu dieser Thematik sind in der abschließenden Konklusion zu finden.

Konfidenz-intervalle

Wie im ersten Artikel der Serie Statistik ausführlich beschrieben, so unterliegen Testungen für verbundene Gruppen ebenso Zufallsschwankungen, weshalb es sich auch hier empfiehlt, zur Bewertung des Ergebnisses ein **Konfidenzintervall** hinzuzuziehen, um der Unsicherheit der Schätzung zusätzlich Rechnung zu tragen. Detailliertere Beschreibungen sind hierzu im ersten Methodenartikel zu finden (Seistock, Bunina & Aden, 2020).

Effektstärke

Um die Aussagekraft von Ergebnissen zu erhöhen, ist es empfehlenswert, neben Konfidenzintervallen ebenso **Effektstärken** heranzuziehen, wobei letztere im Falle eines signifikanten Unterschiedes erst wirklich imstande sind, die Wirksamkeit einer Intervention statistisch zuverlässig zu bewerten.

Kleine Stichproben und Drop-Out Raten

Vor allem bei **kleinen Stichproben** ist die zusätzliche Bewertung des Ergebnisses mittels Effektstärken (z.B. Cohens d beim t-Test) von besonderer Bedeutung, da aufgrund eines sogenannten Undersamplings (zu wenige Personen in der Stichprobe) die Wahrscheinlichkeit eines Fehlers 2. Art (ein Effekt wird statistisch nicht erkannt, obwohl dieser aber tatsächlich vorliegt) größer ist. Dieser Punkt ist im Rahmen von Evaluationsstudien insofern von zentraler Bedeutung, da Verlaufstestungen mit mehreren Messzeitpunkten (bzw. Längsschnittstudien) häufig von einer großen Drop-out-Rate betroffen sind, was wiederum die Stichprobengröße erheblich reduzieren kann. Die Interpretation der jeweiligen Effektstärken ist folgend im jeweiligen Unterkapitel der einzelnen Tests zu finden.

Es bleibt zu erwähnen, dass wie bei den Verfahren für zwei unverbundene Gruppen auch die Wahl der hier vorgestellten Tests für verbundene Stichproben vom Skalenniveau des zu untersuchenden Merkmals (in welchem Merkmal wird die Veränderung untersucht, z.B. Depression) abhängig ist. Wie auch im ersten Artikel der Serie Statistik dargelegt, kommen die hier vorgestellten Tests bei

mindestens rangskalierten (ordinalen) Daten zum Einsatz. Nachfolgend ist eine Übersicht der Tests für zwei verbundene Gruppen und der Skalenniveaus der Testvariable zu finden.

Tab. 1 Übersicht der Verfahren je nach Skalenniveau

Übersicht Skalenniveau und Tests

Test	Skalenniveau	Beispiel
t-Test für 2 verbundene Stichproben	Metrisch (& normalverteilt)	z.B. standardisierteTestinstrumente und Fragebögen (z.B. Becks Depressions Inventar; Beck, Steer & Brown (1996))
Wilcoxon-Test	Metrisch (& nicht normalverteilt)	z.B. Daten aus standardisierten Fragebögen, die aufgrund der Stichprobe nicht normalverteilt sind, bei metrischen Daten mit Ausreißern >2 Standardabweichungen
Binomial Vorzeichen-Test	Ordinal	z.B. subjektive Einschätzungen wie etwa Zufriedenheitseinschätzungen, jegliche Fragebögen, die nicht standardisiert sind

Aufgrund der größeren Testmacht ist bei metrischen und normalverteilten Daten der t-Test für verbundene Stichproben immer vorzuziehen, da damit die Wahrscheinlichkeit steigt, einen Effekt (z.B. die Verbesserung der depressiven Symptomatik) auch tatsächlich als solchen statistisch nachweisen zu können.

Unterschiedstestungen für zwei verbundene Stichproben

Um nun auf die einzelnen Verfahren im Detail einzugehen, werden im folgenden Abschnitt notwendige empirische Voraussetzungen für die korrekte Wahl und Durchführung der drei Unterschiedstestungen (t-Test für verbundene Stichproben, Wilcoxon Test, Binomial Vorzeichentest) vorgestellt sowie dafür wichtige mathematische Grundlagen erklärt. Als erste Voraussetzung gilt die Identifikation des Skalenniveaus des zu untersuchenden Merkmals, welche im vorherigen Abschnitt ausführlich beschrieben wurde. Um eine korrekte Anwendung der Testverfahren gewährleisten zu können, gilt es, die fachgerechte Handhabung und Einhaltung der empirischen Voraussetzungen zu beachten.

t-Test bei verbundenen Stichproben

Der t-Test für verbundene Stichproben ist, wie bereits erwähnt, von den drei Verfahren jenes mit der größten Testmacht. Allerdings ist dessen Anwendung auch an das Vorliegen der meisten empirischen Voraussetzungen gebunden. Zunächst muss die Variable, hinsichtlich derer die beiden Gruppen/Zeitpunkte miteinander verglichen werden sollen, metrisch, also mindestens intervallskaliert, sein. Erläuterungen zur Bestimmung der Skalenniveaus sind in Tabelle 2 und im ersten Artikel der Serie Statistik im Forschungsbulletin (Seistock, Bunina & Aden, 2020) zu finden.

Voraussetzungen

Normal-verteilung des Differenz-scores

Im Anschluss an die Bestimmung des Skalenniveaus des untersuchen Merkmals, sollte die Normalverteilung des Differenzscores ($di=T_1-T_0$, Depressionswert nach der Therapie – Depressionswert vor der Therapie) überprüft werden. Um die Normalverteilung des Differenzscores zu prüfen, kann als graphische Orientierung ein Histogramm gewählt werden, in welchem die Messwerte (Differenzscores) in Form von Balken dargestellt sowie eine Normalverteilungskurve als Referenz zur Bewertung herangezogen werden kann.

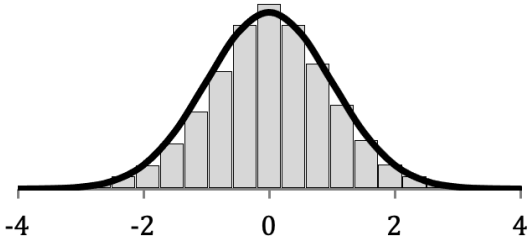


Abb. 3 Histogramm (Normalverteilung)

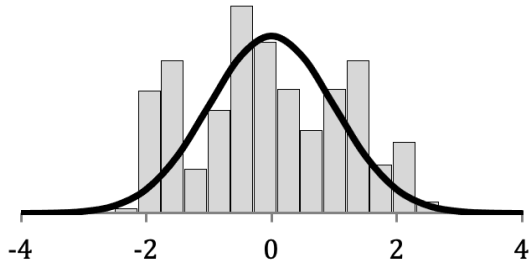


Abb. 2 Histogramm (keine Normalverteilung)

Prüfung der Normal-verteilung

Um die Normalverteilung als gegeben anzunehmen, sollten die vorliegenden Balken hierfür einen möglichst symmetrischen Verlauf abbilden (siehe Abb. 2). Kommt es zu größeren Abweichungen und keiner symmetrischen Verteilung der Balken (siehe Abb. 3), kann geschlossen werden, dass die Normalverteilung nicht gegeben ist. Ebenso kann die Normalverteilung entlang deskriptiver Kennwerte ermittelt werden, wobei die Kennwerte der Schiefe und der Kurtosis heranzuziehen sind (vgl. Seistock, Bunina & Aden, 2020). Zudem besteht die Möglichkeit, diese Voraussetzung mithilfe statistischer Anpassungstests (Goodness-of-Fit-Tests) zu überprüfen. Hierfür wird der Test nach Kolmogorov-Smirnov und/oder nach Shapiro-Wilk angewendet, wobei das Vorliegen mindestens eines nicht-signifikanten Tests ($p > .05$) eine Normalverteilung annehmen lässt.

Grundsätzlich ist festzuhalten, dass von einer Testung der Normalverteilung bei größeren Stichproben ($n \geq 30$) gemäß dem zentralen Grenzwertsatz abgesehen werden kann. Hingegen ist bei kleinen Stichproben ($n < 30$) sehr wohl auf die Voraussetzung der Normalverteilung Rücksicht zu nehmen.

Sollte die Voraussetzung der Normalverteilung des Differenzscores nicht gegeben sein, kann auf das Ausweichverfahren des Wilcoxon-Tests, welcher in weiterer Folge noch genauer beschrieben wird, zurückgegriffen werden.

Vorsicht bei Ausreißern

An dieser Stelle ist ebenfalls die Berücksichtigung von sogenannten Ausreißern zu betonen, welche vereinzelte Extremwerte (extrem geringe oder extrem hohe Werte) darstellen und zu Verzerrungen des Mittelwerts und infolgedessen zu potenziellen Fehlinterpretationen führen können. Diese können beispielsweise im Rahmen der Normalverteilungstestung bzw. durch Häufigkeitsanalysen (z.B. mittels Boxplot) eruiert werden.

Sollte eine der beiden beschriebenen Voraussetzungen (NV und Ausreißer) verletzt werden, reagiert der t-Test (für verbundene Stichproben) relativ robust (vgl. Wilcox, 2012). Dennoch sollte bei nicht gegebener Normalverteilung der Wilcoxon-Test herangezogen werden, da dieser in jenem Fall die Teststärke erhöht.

Sind jene Voraussetzungen jedoch gegeben, kann der t-Test für verbundene Stichproben angewandt werden.

Berechnung des empirischen t-Wertes

Dem Ziel des abhängigen t-Tests, Unterschiede zwischen beispielsweise zwei abhängigen Messzeitpunkten im Sinne einer Mittelwertdifferenz sichtbar zu machen, wird auch in der mathematischen Berechnung Ausdruck verliehen.

Die Formel zur Berechnung des empirischen t-Werts lautet hierbei wie folgt:

$$t = \frac{\bar{d}}{\frac{\hat{\sigma}_d}{\sqrt{n}}}$$

$$\hat{\sigma}_{\bar{x}_d} = \sqrt{\frac{1}{n-1}} * \sum (d_i - \bar{d})^2$$

Berechnung des empirischen t-Werts

Im Rahmen der Berechnung wird folglich für jede Person ein Differenzwert (di) von Zeitpunkt 1 und Zeitpunkt 2 gebildet und daraus der Gesamtmittelwert der Differenzscores der Stichprobe berechnet ($\bar{d}$). Daraus folgend kann der t-Wert als standardisierter Mittelwertscore der Differenzen zwischen Testzeitpunkt 1 und 2 bezeichnet werden, wobei die Standardisierung an der gepolten Standardabweichung der Differenzscores ($\hat{\sigma}_{\bar{x}_d}$) erfolgt, um einen Rückschluss auf die Population zuzulassen und damit verallgemeinerbare Aussagen treffen zu können (bei repräsentativen Stichproben).

Ist der empirische t-Wert errechnet, stellt sich die Frage nach der Beurteilung dieses Ergebnisses. Zu diesem Zweck müssen zusätzlich die Freiheitsgrade oder auch Degrees of Freedom (df) berücksichtigt werden, mittels derer ein kritischer t-Wert aus der t-Verteilung als Referenzwert zur Bewertung des Ergebnisses bestimmt werden kann. Die Freiheitsgrade werden mit der vorliegenden Formel berechnet, wobei n hier die Anzahl der Messwertpaare (Anzahl der Personen, die zu beiden Zeitpunkten getestet wurden) abbildet.

$$df = n-1$$

Liegen diese vor, kann unter Berücksichtigung des veranschlagten Alphaniveaus (in den Sozialwissenschaften meist bei α= 5%) ein zugehöriger (kritischer) Verteilungswert der t-Verteilung eruiert werden (z.B. mittels t-Tabellen aus einschlägigen Lehrbüchern (z.B. Bortz & Schuster, 2010)), der eine Signifikanzbeurteilung des errechneten Ergebnisses zulässt. Ist der errechnete t-Wert größer als jener (kritische) Verteilungswert, so ist die Mittelwertsdifferenz als statistisch signifikant zu beschreiben. Liegt der empirische (berechnete) Wert jedoch unter dem kritischen Verteilungswert, so ist das Ergebnis statistisch nicht signifikant. Bei der üblicheren Berechnung des t-Tests für verbundene Gruppen mittels einer Statistik-Software (z.B. IBM SPSS) ist der p-Wert zur Bewertung der Signifikanz heranzuziehen und zu reportieren (sh. Reportkonventionen weiter unten).

Berechnung der Effektstärke

Zusätzlich sollte für die Interpretation die **Effektstärke Cohens d** berücksichtigt werden, um neben der Signifikanz auch die Größe des Mittelwertunterschieds bewerten zu können. Zur Vereinfachung der Interpretation dieser definiert Cohen (1988) drei Intervalle, welche unterschiedliche Effektkategorien widerspiegeln. Ein Wert von d= +/-.20 entspricht einem kleinen, d= +/-.50. einem mittleren, und d= +/- .80 einem großen Effekt.

$$d = \frac{M_1 - M_2}{S_{Differenz}}$$

Um eine nachvollziehbare Interpretation der Ergebnisse zu ermöglichen, schreibt die American Psychological Association (2020) hierbei vor, dass der t-Test für verbundene Gruppen unter Angabe der folgenden Kennwerte dargestellt werden soll: der empirische t-Wert, die Degrees of Freedom, der p-Wert, die Effektstärke d, die Mittelwerte und Standardabweichungen der beiden Gruppen/Zeitpunkte sowie optional die Konfidenzintervalle der Mittelwertdifferenz.

Report-konventionen (APA) t-Test für verbundene Stichproben

Als Beispiel könnte hierfür angenommen werden, dass ein t-Wert von 6.53, bei Degrees of Freedom von 24, mit einem p-Wert von p<.001, einem Mittelwert von Gruppe eins ($MBDIT_0$) von 26.00 und einer Standardabweichung von 5.76, sowie einem Mittelwert von Gruppe zwei ($MBDIT_1$) von 21.84 und einer Standardabweichung von 4.62, wobei die Effektstärke .75 beträgt. Demnach wären die vorliegenden Kennwerte in Form eines statistischen Reports anzuführen:

($t(24)= 6.53$, $p< .001$, $M_{BDIT0}= 26.00$ ($SD= 5.76$), $M_{BDIT1}= 21.84$ ($SD= 4.62$), $d= .75$, 95% CI [2.85, 5.47])

Im Folgenden wird auf den Wilcoxon-Test Bezug genommen, welcher ein Ausweichverfahren im Falle der Verletzung einer der Voraussetzungen darstellen kann.

Wilcoxon – Test

Wilcoxon-Test beim Vorliegen keiner NV

Der Wilcoxon-Test (vgl. Wilcoxon, 1945, 1947) gilt in diesem Kontext als Verfahren, welches im Falle der nicht vorliegenden Normalverteilung des Differenzscores bei zwei abhängigen Stichproben hinsichtlich eines metrischen Merkmals verwendet wird. Dieses Testverfahren zählt zu den non-parametrischen Verfahren und ist somit nicht an eine spezifische Verteilungsanforderung (wie die der Normalverteilung) gebunden. An dieser Stelle ist jedoch zu erwähnen, dass der Wilcoxon-Test gegenüber dem t-Test eine geringere Testmacht aufweist, weshalb unter Erfüllung der Voraussetzungen der t-Test vorzuziehen ist.

Wilcoxon-Test und Rangtrans-formation

Bei diesem Verfahren wird im Vergleich zum t-Test nicht mit den tatsächlichen Messwerten gerechnet, sondern es werden diese durch eine Rangtransformation in Rangplätze überführt. Im Rahmen der Berechnung werden im ersten Schritt die Differenzen der Messwerte von Zeitpunkt 1 und 2 jeder Person ermittelt, welche sich, um auf das bereits erwähnte Beispiel der Überprüfung der Wirksamkeit von Psychotherapie, aus den Depressionswerten vor Beginn der Therapie und nach Beendigung der Therapie ergeben. Nachdem diese Differenzen für jede Person ermittelt wurden, werden diese folglich einem Rang je nach Ausprägung des Differenzwertes zugeordnet, wobei der kleinste Wert den Rang 1 und der größte Wert den Rang n erhält (vgl. Wilcoxon, 1945). Aus diesen Rangplatzzuweisungen wird analog zum t-Test eine Testgröße (T^2) berechnet. Bei ausreichend großen Stichproben (n > 20) lässt sich die Testgröße des Wilcoxon-Tests unter Verwendung einer Normalverteilungsapproximation in eine z-Verteilung mit z-Werten überführen, mittels welcher die Signifikanzbeurteilung der Rangplatzunterschiede erfolgen kann.

Effektstärke r

Die Effektgröße r für den Wilcoxon-Test wird dabei auf Basis der Testkenngröße z gemäß der Formel

$$r = \frac{z}{\sqrt{N}}$$

ermittelt (vgl. Fritz, Morris & Richler, 2012) und nach Cohen (1988) hinsichtlich der Stärke (kleiner Effekt: r>.1, mittlerer Effekt: r>.3, großer Effekt: r>.5) interpretiert, wobei N hierbei die Anzahl der Messwerte angibt und nicht die Anzahl der Personen (da von jeder Person Messwerte zu zwei Zeitpunkten bestehen, gibt es immer zweimal so viele Messwerte wie es Personen gibt). Darüber hinaus werden die Mittelwerte und Standardabweichungen oder optional die Mediane der beiden Messzeitpunkte als deskriptive Kennwerte angeführt.

[2] Bei diesem T-Wert handelt es sich nicht um die Testgröße des t-Tests, sondern um die Summe der Rangzuweisungen mit demselben Vorzeichen (vgl. Bortz & Schuster, 2010, S. 133)

Report konvention (APA) Wilcoxon-Test

Im Hinblick auf die Ergebnisdarstellung soll der statistische Report ebenso wie beim t-Test für verbundene Stichproben bestimmte Kennwerte enthalten (APA, 2020). Konkret handelt es sich um den empirischen z-Wert (bei ausreichend großer Stichprobe von n > 20), den p-Wert und ergänzend dazu die Effektstärke r, um eine verlässliche Interpretation der Ergebnisse zu ermöglichen.

Ergibt sich beispielsweise ein z-Wert von -4.05, mit einem p-Wert von p< .001 und einer Effektstärke von r= .52, könnte der Report wie folgt angegeben werden:

$$(Z= -4.05, p< .001, r= .52, M_{BDIT0}= 19.60 (SD= 5.62), M_{BDIT1}= 15.80 (SD= 5.19))^{3}$$

Binomial-Test /Vorzeichentest

Vorzeichen-test beim Vorliegen ordinaler Merkmale

Sollten zwei abhängige Gruppen hinsichtlich eines ordinal skalierten Merkmals miteinander verglichen werden, wird hierfür der Binomial-Test/Vorzeichentest angewandt. Dieser zählt ebenfalls zu den nonparametrischen Verfahren und ist demnach nicht auf ein kardinales Messniveau der Daten angewiesen (vgl. Messer & Schneider, 2019). In diesem Kontext könnte beispielsweise die subjektive Belastungseinschätzung auf einer Skala von 0-20 vor sowie nach einer Therapie/Intervention bemessen werden, um zu überprüfen, inwiefern sich die subjektive Einschätzung verbessert bzw. verändert. Subjektive Einschätzungsskalen sind in diesem Zusammenhang grundsätzlich als ordinale Daten einzustufen, da sie über keinerlei testtheoretische Basis verfügen und daher nicht als Messung im statistischen Sinne gelten.

Entscheidung für den Binomial-Test

Entlang des Binomial-/Vorzeichentests gilt es die Differenz der Messwerte pro Gruppe/Zeitpunkt zu ermitteln, wobei überprüft werden muss, ob der Differenzwert hierbei ein negatives oder positives Vorzeichen aufweist. Je nachdem, welches Vorzeichen seltener über alle Proband*innen hinweg vorkommt, wird daraufhin die Prüfgröße x definiert. In diesem Kontext wird ersichtlich, dass für die Bestimmung der Prüfgröße x die exakten Messwerte der Belastungseinschätzung vor und nach der Therapie nicht unbedingt vorgegeben sein müssen, da die Bestimmung des Vorzeichens anhand der Entscheidung von entweder x_Ai>x_Bi (positives Vorzeichen) oder x_Ai<x_Bi (negatives Vorzeichen) getroffen werden kann und sich die Einsatzmöglichkeiten des Verfahrens dadurch erhöhen (vgl. Bortz & Döring, 2008). Aus der Prüfgröße x wird – analog zum t-Test für verbundene Stichproben und zum Wilcoxon-Test – eine Teststatistik (basierend auf dem Differenzwert) bestimmt, welche wiederum zur Signifikanzbeurteilung des Ergebnisses mittels p-Wert dient. Bei Stichprobengrößen von N>25 wird eine X^2-Teststatistik für Alternativdaten bestimmt, welche in einen z-Wert der Standardnormalverteilung überführt wird und Rückschlüsse über die Signifikanz eines Ergebnisses zulässt.

[3] Bei ordinalen Daten sind bei einem signifikanten Ergebnis anstatt von Mittelwerten die Mediane anzugeben.

Report-konvention (APA) Binomial/ Vorzeichen-Test

In Anlehnung an die Reportempfehlung der American Psychological Association (APA, 2020) kann bei der Verwendung des Binomial-Vorzeichentests der p-Wert als entscheidender Kennwert für die Festlegung eines Unterschieds angegeben werden, wobei ein signifikantes Ergebnis ($p \leq .05$) zur Verwerfung der H_0 führt. Ebenso kann der z-Wert im Reporting integriert werden. Als Beispiel könnte der statistische Report unter der Annahme eines nicht signifikanten Ergebnisses folglich lauten:

Es bestehen keine signifikanten Unterschiede in der subjektiven Einschätzung der Belastung vor und nach der Therapie ($z=-1.16$; $p= .246$). Die H_0 wird beibehalten.

Im Falle eines signifikanten Ergebnisses können zur Orientierung ebenso zusätzlich zum p-Wert die Mediane der Zeitpunkte/Gruppen angegeben werden.

Konklusion

Im Forschungsalltag über ein solides methodisches Wissen zu verfügen, stellt nicht nur die Grundbedingung für zuverlässige statistische Befunde dar, sondern schützt ebenso vor der Anfechtbarkeit von Ergebnissen. Zu diesem Zweck wurde im vorgestellten Artikel ein Überblick gegeben, bei welchen Fragestellungen die vorgestellten Testverfahren zur Anwendung kommen können, auf welcher mathematischen Grundlage diese basieren, welche zu prüfende Voraussetzungen für eine korrekte Anwendung beachtet werden müssen, welche Ausweichverfahren bei unvollständigen Voraussetzungen zu wählen sind und welche Darstellung der Ergebnisse gebräuchlich ist.

Insbesondere bei abhängigen Testungen wie z.B. Prä-Post-Designs, wie sie in der psychotherapeutischen Interventionsforschung häufig zur Anwendung kommen, ist das Problem der Verlässlichkeit der Ergebnisinterpretation von großer Bedeutung. Dies ist nicht zuletzt der Fall, da auf Basis solcher Ergebnisse Implikationen für die Praxis abgeleitet werden. Es ist an dieser Stelle festzuhalten, dass Wirksamkeitsmessungen, im Sinne eines Kausalschlusses von einer Intervention auf eine Veränderung in einem psychischen Merkmal, mit den hier vorgestellten Verfahren nicht zulässig sind. Ist jedoch ein „einfacherer“ Vorher-Nachher Vergleich von Interesse, sowie andere Fragestellungen mit abhängigen Stichproben (z.B. Unterschiede zwischen Geschwistern, Ehepaaren oder zwischen Selbst- und Fremdeinschätzung einer Person), so ist unter Berücksichtigung der hier beschriebenen methodischen Konventionen mit verlässlichen statistischen Ergebnissen zu rechnen. Letzteren ist jedoch nur dann Unanfechtbarkeit beizumessen, wenn neben der statistischen Genauigkeit und Relevanz auch die Frage nach der inhaltlichen Bedeutsamkeit geklärt ist. Zu diesem Zweck sind nachfolgend die wesentlichsten Empfehlungen zur statistischen sowie inhaltlichen Reflexion und Bewertung von Ergebnissen bei abhängigen Testungen abschließend zusammengefasst.

Wahl des korrekten Verfahrens

→ *Wahl des korrekten Verfahrens*

Wie schon zuvor festgehalten, ist die Wahl des korrekten Verfahrens maßgeblich verantwortlich für den Erhalt von verlässlichen Ergebnissen. Damit einhergehend ist ebenso die Testmacht (Power) von der Wahl des Verfahrens beeinflusst. Dies ist insbesondere zu bedenken, da mit höherer Testmacht die Wahrscheinlichkeit steigt, einen vorhandenen Effekt auch statistisch nachweisen zu können. Daher empfiehlt es sich bei gegebenen Voraussetzungen immer dem Verfahren mit der höchsten Testmacht (in diesem Artikel t-Test für verbundene Stichproben) den Vorzug zu geben.

Stichproben größe & Planung

→ *Stichprobengröße & -Planung*

Die psychotherapeutische und klinische Forschung ist häufig mit einer geringen Stichprobengröße konfrontiert, welche zusätzliche Risikoquellen mit sich bringen kann. Neben der oft eingeschränkten Repräsentativität von kleinen Stichproben (Under-Sampling) besteht ebenso die Gefahr, einen Fehler 2. Art (Beta-Fehler) zu begehen und dementsprechend einen vorhandenen Effekt nicht statistisch als signifikant nachweisen zu können. Zusätzlich ist bei Verlaufstestungen oft mit teils hohen Drop-Out-Raten zu rechnen, die wiederum die Stichprobengröße deutlich verringern können. Es empfiehlt sich, dies schon bei der Studienplanung mit zu berücksichtigen (z.B. mit Incentives oder Erinnerungen entgegenzuwirken).

Ähnliche Schwierigkeiten bestehen bei einer zu großen Stichprobe (Over-Sampling), welche die Gefahr birgt, einen Fehler 1. Art (Alpha-Fehler) zu begehen, wobei fälschlicherweise eine Systematik angenommen wird, obwohl in der Population keine besteht. Um jene Risiken zu vermeiden, ist es ratsam, im Rahmen der Studienplanung eine für das Forschungsvorhaben passende Stichprobengröße zu kalkulieren (beispielsweise mittels „G*Power“ oder „R“), um Effekte auch als richtigerweise signifikant nachweisen zu können.

Effektstärke als Indikator

→ *Effektstärke als relevanter Indikator*

Neben der korrekten Wahl des Verfahrens und der Bestimmung der statistischen Signifikanz von Ergebnissen ist ebenso die Bewertung der Größe eines vorhandenen Mittelwertunterschiedes für die Interpretation notwendig. Für diesen Zweck ist im Falle eines signifikanten Ergebnisses immer auch die jeweilige Effektstärke hinzuziehen, da diese erst Aufschluss über den Umfang des Effekts zulässt. Vor allem beim Vorliegen von kleinen Stichproben, in denen Effekte unter Umständen nicht als signifikant nachweisbar sind, ist ein Blick auf die Effektstärke zur Identifizierung bzw. Bewertung einer möglichen Tendenz sinnvoll.

inhaltlich bedeutsam ≠ statistisch bedeutsam

→ *inhaltlich bedeutsam ≠ statistisch bedeutsam*

Ist ein Effekt korrekterweise als signifikant nachgewiesen, so lässt dies nicht gleichermaßen einen Rückschluss auf die inhaltliche Ergebnisrelevanz zu. Zur Bestimmung der inhaltlichen Bedeutsamkeit empfiehlt es sich, neben der Interpretation der Effektstärke (sh. oben) ebenso die Eigenschaften des Testinstruments und der Skalierung des Merkmals bzw. etwaiger Grenzwerte zu beachten. Beispielsweise kann eine statistische Veränderung in einem Testscore zwar signifikant sein, jedoch in

einem klinisch unbedeutsamen Veränderungsbereich variieren (z.B. ist der Depressionswert nach der Therapie zwar geringer als vorher, jedoch noch immer in einem überdurchschnittlichen Bereich und stellt somit keine substantielle Symptomverbesserung dar).

Dahingehend ist eine gute Kenntnis über das zu untersuchende Merkmal sowie die verwendeten Testinstrumente unabdingbar für die Bestimmung von inhaltlich relevanten Ergebnissen.

Ergänzende Handlungsempfehlungen

Werden die eben beschriebenen Ratschläge im Rahmen einer quantitativen Forschung befolgt, so kann mit statistisch zuverlässigen Ergebnissen gerechnet werden. Nichtsdestotrotz ist es ratsam, die folgenden Aspekte während des Forschungsprozesses zu reflektieren.

- **Welches Merkmal wird gemessen?**

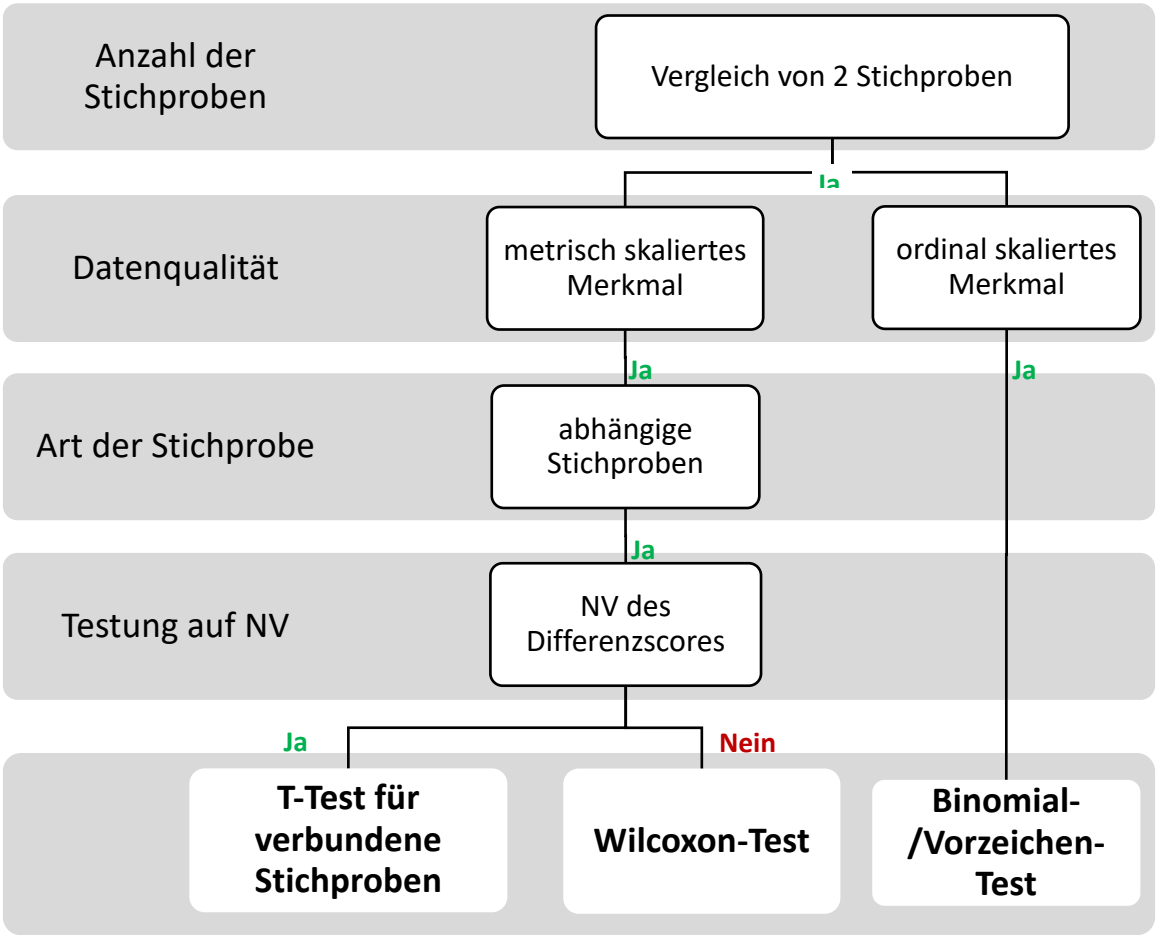
Psychologische Merkmale gelten als unterschiedlich veränderungssensitiv. Dies ist nicht nur eine grundsätzliche inhaltliche Frage, die es zu berücksichtigen gilt, sondern beeinflusst ebenso die Studienplanung maßgeblich, da sich danach die Planung der Post-Testung – wann nach einer Therapie/Intervention sinnvollerweise mit einer Veränderung im zu messenden Merkmal zu rechnen ist – richten kann.

- **Lerneffekte/Übungseffekte/Störvariablen**

Zusätzlich ist vor allem bei Messwiederholungen die Verzerrung durch potenzielle Übungs- bzw. Lerneffekte zu reflektieren, welche eine verlässliche Interpretation der Ergebnisse erschweren kann. Dieser Umstand kann beispielsweise bei der Wahl des Testinstruments mitbedacht werden. Zudem muss bei Untersuchungen dieser Art auch mit Störvariablen gerechnet werden, die aus Lebensereignissen im Zeitraum zwischen den Testzeitpunkten entspringen und die Ergebnisse beeinflussen können. Hier ist es ratsam, schon vorab vermutete Störvariablen mit zu erheben, um im Zweifelsfall diese in die Berechnungen integrieren zu können. Ist ein aufwändigeres Studiendesign von Interesse, so ist zu einer komplexeren Testung mittels Kontrollgruppe zu raten.

Im Rahmen der abhängigen Testungen für zwei verbundene Stichproben, wie sie hier vorgestellt wurden, können die vorgeschlagenen Handlungsempfehlungen dazu beitragen, die Anwendung dieser Verfahren auf ein hohes statistisches Niveau zu heben sowie die korrekte Interpretation sowie Stichhaltigkeit der Ergebnisse zu verbessern.

Entscheidungsbaum



Literatur

American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7. Aufl.). Washington, DC: APA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, N.J: L. Erlbaum Associates.

Beck, A.T., Steer, R.A., & Brown, G.K. (1996). *Beck Depression Inventory* (2. Aufl.). San Antonio: The Psychological Corporation.

Bortz, J. (2006). *Statistik: Für Human-und Sozialwissenschaftler*. Heidelberg: Springer Medizin.

Bortz, J., & Döring, N. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin: Springer.

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.

Fritz, O. F., Morris, P. E., & Richler, J. J. (2012). Effect Size Estimates: Current Use, Calculations, and Interpretation. *Journal of Experimental Psychology, 141*(1), 2–18.

Grawe, K. (2005). (Wie) kann Psychotherapie durch empirische Validierung wirksamer werden? *Psychotherapeutenjournal, 4*(1), 4–11.

Grawe, K. (1992). Psychotherapieforschung zu Beginn der neunziger Jahre [Psychotherapy research at the beginning of the nineties]. *Psychologische Rundschau, 43*(3), 132–162.

Jones, S. R., Carley, S., & Harrison, M. (2003). An introduction to power and sample size estimation. *Emergency Medicine Journal, 20*, 453–458.

Krzywinski, M., & Altman, N. (2013). Power and sample size. *Nature Methods, 10*, 1139–1140.

Kühner, C., Bürger, C., Keller, F., & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II). Befunde aus deutschsprachigen Stichproben. *Der Nervenarzt, 78*, 651–656.

Messer, M., & Schneider, G. (2019). *Statistik. Theorie und Praxis im Dialog*. Berlin: Springer Spektrum.

Seistock, D., Bunina, A., & Aden, J. (2020). Der t-, Welch-und U-Test im psychotherapiewissenschaftlichen Forschungskontext. Empfehlungen für Anwendung und Interpretation. *SFU Forschungsbulletin, 8*(1), 87–105.

Spinhoven, P., Klein, N., Kennis, M., Cramer, A. O., Siegle, G., Cuijpers, P., ... & Bockting, C. L. (2018). The effects of cognitive-behavior therapy for depression on repetitive negative thinking: a meta-analysis. *Behaviour research and therapy, 106*, 71–85.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin, 1*, 80–83.

Wilcoxon, F. (1947). Probability Tables for Individual Comparisons by Ranking Methods. *Biometrics, 3*(3), 119–122.

Wilcox, R. (2012). Chapter 5 - Comparing Two Groups. In R. Wilcox (Hrsg.), *Statistical Modelling and Decision Science. Introduction to Robust Estimation and Hypothesis Testing* (S.137–213) (Third Edition). San Diego: Academic Press.

*Angaben zu den Autor*innen*

Institut für Statistik
Sophie Gattermeyer, Clara Vladarski, Jan Aden
Adresse: Freudplatz 1, 1020 Wien, Raum 6011
E-Mail: jan.aden@sfu.ac.at , sophie.gattermeyer@sfu.ac.at