SFU Forschungsbulletin SFU Research Bulletin 9/2 (2021), S. 78–99 DOI: 10.15135/2021.9.2.78–99

Die einfaktorielle Varianzanalyse für abhängige Stichproben und der Friedman-Test im psychotherapiewissenschaftlichen Kontext

-Empfehlungen für Anwendung und Interpretation-

The one-way analysis of variance for dependent samples and the Friedman-test in psychotherapy science

-Recommendations for application and interpretation-

David Seistock, Elias Ruso und Jan Aden

Kurzzusammenfassung

In diesem vierten Beitrag der Serie Statistik in der Psychotherapiewissenschaft wird die Anwendung von Verfahren zum Vergleich von k-abhängigen Stichproben (einfaktorielle abhängige ANOVA und der Friedman-Test) im Sinne eines Best-Practice Ansatzes vorgestellt. Es werden Empfehlungen für (1) eine optimale Verfahrenswahl, (2) den Einsatz von Effektstärken, (3) der Bestimmung der Ergebnisrelevanz sowie (4) Reportkonventionen für die Ergebnisdarstellung gegeben. Darüber hinaus wird der Einsatz dieser Verfahrensgruppe im psychotherapiewissenschaftlichen Kontext anhand von Beispielen illustriert.

Schlüsselwörter

Einfaktorielle abhängige Varianzanalyse, Friedman-Test, Reportkonventionen, Effektstärken, post-hoc Test

Abstract

In this fourth contribution in the series Statistics in Psychotherapy Science, the application use of methods for comparing k-dependent samples (one-way dependent ANOVA and the Friedman-test) in the sense of a best practice approach is presented. Recommendations are given for (1) an optimal choice of procedure, (2) the use of effect sizes, (3) the designation of the relevance of the results and (4) report conventions for the presentation of results. In addition, the use of this group of procedures in the context of psychotherapy science is illustrated using examples.

keywords

one-way dependent ANOVA, Friedman test, report conventions, effect-sizes, post-hoc tests

Einsatzfeld und Background

Einsatzbereiche und Fragestellungen

Die Anwendung psychotherapeutischer Interventionsmethoden ohne eine vorangehende Untersuchung der Wirksamkeit ebendieser erscheint heutzutage nahezu undenkbar. Nicht nur die Evaluation der Effektivität bestimmter Therapiemethoden, sondern ebenfalls die Untersuchung damit verbundener psychologischer Merkmale stellen somit zentrale Anliegen der evidenzbasierten Psychotherapiewissenschaft dar. Diese Vorgehensweise dient sowohl dazu, ein Klima des Vertrauens sowie der Akzeptanz hinsichtlich psychotherapeutischer Methoden zu schaffen, als auch, diese im Sinne der Qualitätssicherung langfristig zu legitimieren. Im Rahmen der Untersuchung psychotherapeutischer Interventionsmethoden werden in der Regel Klient*innen zu verschiedenen Zeitpunkten des therapeutischen Prozesses hinsichtlich eines bestimmten Merkmals untersucht (z.B. Ausmaß depressiver Symptomatik zu Beginn der Therapie (Prä), direkt nach Abschluss der Therapie (Post), drei Monate nach Therapieabschluss (Follow-Up/Katamnese). In diesem Zusammenhang wird im Sinne einer statistischen Auswertungslogik von „abhängigen Forschungsdesigns“ mit Messwiederholungen gesprochen. Zur Messung des Therapieerfolges werden üblicherweise entweder der Rückgang psychopathologischer Symptome, die Abnahme des subjektiven Leidensdrucks oder die Stärkung salutogener Faktoren (z.B. adaptive Stressbewältigungsstrategien) herangezogen. In diesem Artikel der Serie Statistik des Forschungsbulletins werden Auswertungsverfahren zur Untersuchung eben solcher abhängiger Forschungsdesigns vorgestellt, anhand derer sich Unterschiede zwischen mehr als zwei verbundenen Stichproben (z.B. Messzeitpunkten) identifizieren lassen.

Wie in den beiden ersten Artikel der Serie Statistik zu den Unterschiedstestungen bei zwei unverbundenen (Seistock, Bunina & Aden, 2020) sowie bei zwei verbundenen Stichproben (Gattermeyer, Vladarski & Aden, 2020) gezeigt wurde, stellt die quantitative Forschungsmethodik eine Vielzahl an verschiedenen Auswertungsverfahren zur Untersuchung von statistisch relevanten Unterschieden im Kontext der psychotherapiewissenschaftlichen Forschung zur Verfügung. In diesem Artikel werden speziell jene Auswertungsverfahren vorgestellt, welche zur Untersuchung von ***mehr als zwei verbundenen Stichproben*** herangezogen werden können. Methoden zur Untersuchung von Unterschieden zwischen verbundenen Stichproben beziehen sich stets auf die Analyse der ***intraindividuellen Veränderungen*** aller Elemente der Gesamtstichprobe (z.B. Veränderung der Ausprägung einer Depression von Person A von Zeitpunkt 1 zu Zeitpunkt 2 usw.; siehe Tabelle 1), auf Basis derer dann allgemeine Aussagen über die untersuchte Gruppe an Personen getroffen werden.

Aufgrund der Komplexität der Veränderung psychologischer Merkmale reicht der Vergleich derselben zwischen bloß zwei Zeitpunkten (Prä vs. Post) oftmals nicht aus. Beispielsweise zeigen sich Veränderungen in bestimmten Merkmalen noch nicht direkt nach Beendigung einer Therapie, sondern werden erst nach einer gewissen Zeit manifest (Abb. 1). Außerdem wird durch die Berücksichtigung von mehr als zwei Zeitpunkten auch die Untersuchung der Nachhaltigkeit therapeutischer Effekte möglich (Abb. 2). Auswertungsverfahren für k-verbundene Stichproben tragen diesem Umstand Rechnung und ermöglichen inferenzstatistische Vergleiche über eine mehr oder weniger beliebige Zahl an Messzeitpunkten.

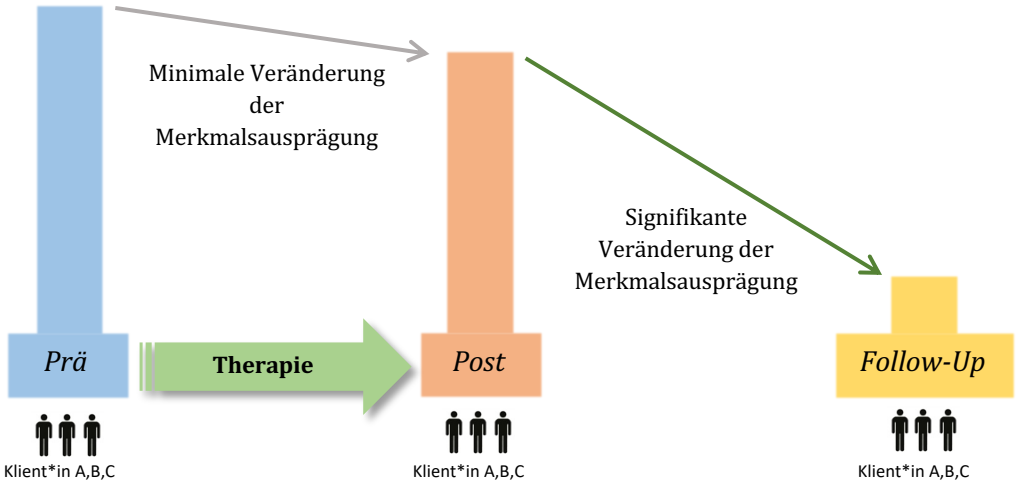


Abb. 1: Beispiel für Verändungen in einem Merkmal nach Abschluss einer Therapie

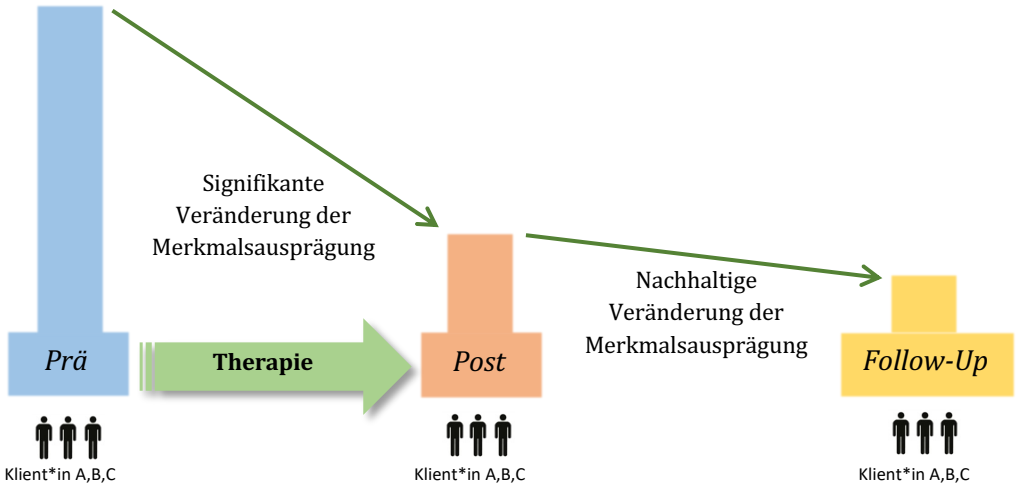


Abb. 2: Beispiel für die Untersuchung der Nachhaltigkeit therapeutischer Effekte

Dafür werden in der Regel die *intraindividuellen* Schwankungen in der Ausprägung eines Störungsbildes (z.B. Depression gemessen anhand des Beck-Depressions-Inventar-II; Kühner, Bürger, Keller & Hautzinger, 2007) über den Therapieverlauf hinweg an mehreren Zeitpunkten gemessen. Ein klassisches, diesbezügliches Design in der Evaluationsforschung umfasst dabei zumeist drei Messzeitpunkte, die als Prä (Messung Depression vor Beginn der Therapie T0), Post (Messung Depression direkt nach Beendigung der Therapie T1) sowie Follow-Up (z.B. Messung Depression drei Monate nach Beendigung der Therapie T2) bezeichnet werden. An dieser Stelle sollte jedoch angemerkt werden, dass es im Zuge der Evaluation der Eignung

psychotherapeutischer Interventionsmethoden gängige Praxis ist, zusätzlich zur Analyse mehrerer Messzeitpunkte simultan einen Vergleich mit einer Kontrollgruppe ohne Intervention/Therapie vorzunehmen. Über diese Vorgehensweise soll eine höhere Validität der Ergebnisse sichergestellt werden, da so genauer untersucht werden kann, ob die Veränderung in der intraindividuellen Merkmalsausprägung tatsächlich auf die Interventionsmethode und nicht auf andere Drittvariablen „zurückzuführen" ist. Ein solches Untersuchungsdesign, welches die Analyse mehrerer Messzeitpunkte (verbundene Stichproben) unter Berücksichtigung des Vergleichs zwischen Interventions- und Kontrollgruppe (unverbundene Stichproben: Personen welche Therapie erhalten vs. Personen welche keine Therapie erhalten) über die Messzeitpunkte hinweg vorsieht, lässt sich anhand der sogenannten *Mixed ANOVA* auswerten, welche in einem späteren Artikel der Serie Statistik gesondert vorgestellt wird.

Tabelle 1: Bsp. für k-verbundene Stichproben

	Messzeitpunkt 1 (*Prä*)	Messzeitpunkt 2 (*Post*)	Messzeitpunkt 3 (*Follow-Up*)
Klient*in A	60	35	47
Klient*in B	55	43	47
Klient*in C	55	48	37
...	...	...	...

Bestimmung der Stichproben-art: verbunden vs. unverbunden

Die Untersuchung von Personen über verschiedene Messzeitpunkte hinweg (Messwiederholungen) stellt nur eine Form verbundener Stichproben dar. Neben sogenannten Messwiederholungen (siehe Tabelle 1) werden ebenfalls natürliche Paare (z.B. Untersuchung von Geschwistertriaden → Erst-, Mittel sowie Letztgeborene) sowie Test-Zwillinge/Matched Pairs (Proband*innen werden auf Basis verschiedener soziodemographischer Merkmale in Paarungen eingeteilt und anschließend verglichen) zu den verbundenen Stichproben gezählt (siehe auch Gattermeyer, Vladarski & Aden, 2020). Die Gemeinsamkeit dieser Formen verbundener Stichproben ist, dass die einzelnen Werte innerhalb der Gruppen/Messzeitpunkte einen verbindenden Aspekt (z.B. Geschwistertriaden → dieselbe Familie) aufweisen und einander zugeordnet werden können (Bortz, 2006). Werden hingegen Stichproben untersucht, welche keinen verbindenden Aspekt aufweisen, wird von sogenannten unverbundenen Stichproben gesprochen. Ein Beispiel für die Untersuchung von unverbundenen Stichproben stellt der Vergleich von Personen verschiedener Diagnosegruppen (Diagnose A, B & C) vor Beginn einer Therapie hinsichtlich ihrer Psychotherapiemotivation (z.B. FPTM (Schulz, Nübling & Rüddel, 1995; Schulz, Lang, Nübling & Koch, 2003; Nübling et al., 2005)) dar (siehe Tabelle 2). Zur näheren Erläuterung von Verfahren zum Vergleich von k-unverbundenen Stichproben siehe Aden, Bunina und Vavrik (2021).

Tabelle 2: Bsp. für k-unverbundene Stichproben

Diagnosegruppe A		Diagnosegruppe B		Diagnosegruppe C	
Klient*in A	65	Klient*in D	58	Klient*in G	22
Klient*in B	48	Klient*in E	19	Klient*in H	77
Klient*in C	59	Klient*in F	46	Klient*in I	31
...	...	...	...	...	...

Um zu überprüfen, ob sich k-verbundene Stichproben statistisch signifikant voneinander hinsichtlich eines ***mindestens*** rangskalierten (ordinal) Merkmals unterscheiden, wird je nach Vorliegen bestimmter Voraussetzungen eines der folgenden Analyseverfahren herangezogen: die einfaktorielle abhängige Varianzanalyse (engl. analysis of variance, ANOVA) mit/ohne anschließender Korrektur der Freiheitsgrade (entweder nach Greenhouse-Geisser,Huynh-Feldt oder Box) oder der Friedman-Test. Bei diesen Auswertungsmethoden handelt es sich um inferenzstatistische Analyseverfahren, anhand derer auf Basis einer Stichprobe identifizierte Effekte für die Gesamtpopulation verallgemeinert werden sollen. Dies setzt die Formulierung sowie Überprüfung eines Hypothesenpaares (Null- und Alternativhypothese) voraus, welche in weiterer Folge auf die Population übertragen werden. Bezugnehmend auf das bereits erwähnte Schema zur Evaluationeiner psychotherapeutischen Intervention zur Reduktion depressiver Symptomatik anhand von drei Messzeitpunkten (Prä, Post, Follow-Up) ergibt sich beispielsweise folgende Formulierung:

Hypothesen-Formulierung

H0: Es besteht kein signifikanter Unterschied in der Ausprägung einer depressiven Symptomatik zwischen den drei Messzeitpunkten (T0, T1, T2).

H1: Es besteht ein signifikanter Unterschied in der Ausprägung einer depressiven Symptomatik zwischen den drei Messzeitpunkten (T0, T1, T2).

Werden Auswertungsverfahren für zwei verbundene/unverbundene Stichproben herangezogen, kann zwischen ein- und zweiseitigen Hypothesen unterschieden werden (Gattermeyer, Vladarski & Aden, 2020; Seistock, Bunina & Aden, 2020). Sollen auch im Rahmen von Verfahren für k verbundene Stichproben spezifische Annahmen bezüglich der Unterschiede zwischen den einzelnen Gruppen überprüft werden (z.B. wenn eine lineare Abnahme der Ausprägung der Depression über die Messzeitpunkte angenommen wird), kann im Rahmen der einfaktoriellen abhängigen

Varianzanalyse beispielsweise auf sogenannte *polynomiale Kontraste* zurückgegriffen werden (diese werden in einem späteren Artikel der Serie Statistik näher erläutert).

Verfahren für k-verbundene Stichproben als „Over-all-Tests“

Analog zu den Verfahren für k-unverbundene Stichproben, welche im dritten Artikel der Serie Statistik (Aden, Bunina & Vavrik, 2021) detailliert beschrieben wurden, handelt es sich bei den Verfahren für k-abhängige Stichproben ebenfalls um „Over-all-Verfahren“/Hauptverfahren“. Demzufolge können bei einem signifikanten Ergebnis zwar globale Unterschiede attestiert werden (z.B. mindestens ein signifikanter Unterschied zwischen den Messzeitpunkten), Aussagen über **die Anzahl der paarweisen Unterschiede** sowie darüber, **welche Gruppen beziehungsweise Messzeitpunkte sich konkret voneinander unterscheiden** (z.B. T0 vs. T1, T0 vs. T2, T0 vs. T3 usw.), können basierend auf dem signifikanten Ergebnis des „Over-All“-Verfahrens jedoch noch nicht getroffen werden. Im obigen Beispiel gilt es die Wirksamkeit einer psychotherapeutischen Interventionsmethode im Hinblick auf die Reduktion depressiver Symptomatik zu untersuchen, wobei drei unterschiedliche Zeitpunkte (Prä, Post, Follow-Up) definiert wurden (siehe z.B. Abb. 1). Um die potentiellen Effekte der Therapie untersuchen zu können, werden zu allen drei Zeitpunkten die Ausprägungsgrade der depressiven Symptomatik aller Klient*innen erhoben (z.B. Depression gemessen anhand des Beck-Depressions-Inventar-II, Kühner, Bürger, Keller & Hautzinger, 2007). Da es sich in jenem Beispiel um Messzeitpunkte, also verbundene Stichproben handelt, müssen Verfahren für k-verbundene Stichproben als statistische Auswertungsverfahren herangezogen werden. In Abhängigkeit bestimmter Voraussetzungen, welche es im Zuge der Analyse zu beachten gilt, können entweder eine einfaktorielle abhängige Varianzanalyse (abhängige ANOVA) mit bzw. ohne Korrekturen oder ein Friedman-Test zur Auswertung herangezogen werden.

Post-hoc Tests

Ein signifikantes Ergebnis des „Over-All“-Verfahrens bedeutete in diesem Beispiel, dass sich die Werte von mindestens zwei der drei Messzeitpunkte paarweise signifikant voneinander unterscheiden. Um in weiterer Folge feststellen zu können, **zwischen wie vielen** und **zwischen welchen** der untersuchten k-verbundenen Stichproben konkret paarweise Unterschiede bestehen, können im Anschluss an das „Over-All“-Verfahren sogenannte **Post-Hoc-Testungen** vorgenommen werden. Im Anschluss an die Berechnung einer einfaktoriellen abhängigen Varianzanalyse können beispielsweise paarweise Post-Hoc-Tests nach Bonferroni berechnet werden (eine detaillierte Übersicht zur Anwendung verschiedener Post-Hoc-Testungen je nach empirischer Datenlage siehe z.B. Bortz, 2006). Zur detaillierten Beschreibung der Unterschiede werden deskriptive Kennwerte (Mittelwert, Standardabweichung) sowie Konfidenzintervalle angegeben (zur Verwendung von Konfidenzintervallen als Hilfsmittel zur Bestimmung der Ergebnisrelevanz siehe z.B. Seistock, Bunina & Aden, 2020).

Im Vergleich zur einfaktoriellen abhängigen Varianzanalyse handelt es sich beim Friedman-Test um ein non-parametrisches Verfahren, welches auf der Verwendung sogenannter Ränge basiert (am Beispiel des U-Tests, siehe Seistock, Bunina & Aden, 2020), weshalb es sich unter anderem empfiehlt, im Anschluss an ein signifikantes Ergebnis je nach Skalenniveau der untersuchten

Variable entweder paarweise Wilcoxon-Tests (bei metrischen, jedoch nicht normalverteilten Merkmalen innerhalb der Stichproben) oder paarweise Binomial-Vorzeichentests (bei rangskalierten Merkmalen) zu berechnen (für weitere Informationen zur Testung von Unterschieden bei zwei verbundenen Stichproben hinsichtlich nicht normalverteilter oder rangskalierter Merkmale siehe Gattermeyer, Vladarski & Aden, 2020). Eine erste Interpretation der Unterschiede kann hier jedoch bereits auch anhand deskriptiver Kennwerte erfolgen.

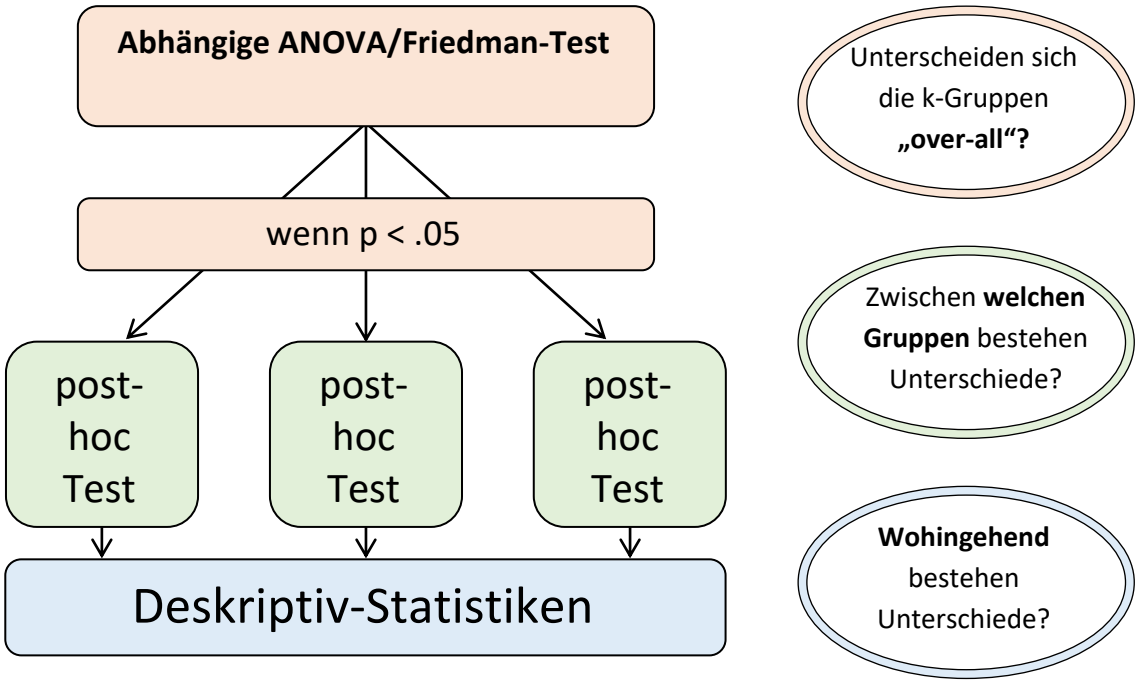


Abbildung 3: Hauptverfahren und post-hoc Tests (modifiziert nach Aden, Bunina & Vavrik, 2021)

Die paarweisen Vergleiche der verbundenen Stichproben (in diesem Beispiel der Messzeitpunkte) anhand der eingesetzten Post-Hoc-Testungen liefern in weiterer Folge konkrete Informationen über die tatsächliche Anzahl signifikanter Unterschiede (z.B.T0 vs. T1, T0 vs. T2 usw.) und erlauben somit eine differenziertere Ergebnisinterpretation.

Multiples Testen

Insgesamt bieten „Over-All"-Verfahren die Möglichkeit, die Anzahl an Testungen zu reduzieren (da nur eine Hypothesenprüfung notwendig ist: Besteht insgesamt ein signifikanter Unterschied zwischen den k-verbundenen Stichproben?). Werden im Anschluss an das Hauptverfahren jedoch paarweise Post-Hoc-Testungen durchgeführt, steigt die Anzahl der durchgeführten Signifikanzprüfungen dennoch, weshalb der damit verbundene Anstieg der Wahrscheinlichkeit einem Alpha-Fehler zu unterliegen (Alpha-Inflation) berücksichtigt werden muss. Eine Möglichkeit mit diesem Umstand umzugehen, besteht beispielsweise in der Verwendung von Post-Hoc-Verfahren, welche automatisch eine Adjustierung der Signifikanzwerte (p-Werte) vornehmen (z.B. Post-Hoc-Bonferroni-Tests). Weitere Möglichkeiten zum Umgang mit dem Problem der Alpha-Fehler-Kumulierung im Kontext multipler Testungen finden sich beispielsweise bei Bender, Lange & Ziegler (2007).

Signifikanz-bestimmung

Die Signifikanzbestimmung der beiden vorgestellten Verfahren (abhängige Varianzanalyse sowie Friedman-Test) erfolgt, wie bereits in den vorangehenden Artikeln der Serie Statistik erläutert, ebenfalls auf der Grundlage bestimmter Prüfverteilungen. Unter Verwendung der Prüfverteilungen (F-Verteilung bei der abhängigen Varianzanalyse, Chi-Quadrat-Verteilung beim Friedman-Test) wird ein empirischer Verteilungwert berechnet, welcher die Größe des Unterschiedes zwischen den k-verbundenen Stichproben (Messzeitpunkten T0, T1, T2) hinsichtlich des untersuchten Merkmales (Ausprägung der depressiven Symptomatik) ausdrückt. Dieser wird im Anschluss mit einem kritischen Verteilungswert, welcher je nach Verteilung basierend auf der Stichprobengröße, der Anzahl an untersuchten Stichproben sowie dem definierten Signifikanzniveau (in den Sozialwissenschaften üblicherweise 5%) festgelegt wird, verglichen und zur Bestimmung der statistischen Relevanz des Unterschiedes (handelt es sich um einen zufälligen bzw. nicht signifikanten oder einen überzufälligen bzw. signifikanten Unterschied) herangezogen. Fällt der empirische Verteilungswert größer aus als der kritische, wird von einem statistisch relevanten (überzufälligen) Unterschied ausgegangen. Üblicherweise wird der berechnete, empirische Verteilungswert ebenfalls in einen Wahrscheinlichkeitswert (p-Wert) transformiert, welcher direkt mit dem festgelegten Signifikanzniveau (i.d.R. 5 %) verglichen werden kann ($p \leq 0.05$ bedeutet ein signifikantes Ergebnis). Genauere Ausführungen zur Signifikanzbewertung eines Ergebnisses finden sich beispielsweise bei Seistock, Bunina & Aden (2020).

Statistische vs. inhaltliche Relevanz

Neben der Bestimmung der statistischen Relevanz von beobachteten Unterschieden sollte jedoch ebenfalls die inhaltliche Relevanz des Ergebnisses berücksichtigt und diskutiert werden. Mitunter spielen in der Berechnung der statistischen Relevanz eines Unterschiedes (p-Wert) neben der Größe des tatsächlichen Unterschiedes zwischen den verbundenen Stichproben auch andere Einflussfaktoren wie beispielsweise die Stichprobengröße eine Rolle (Jones, Carley & Harrison, 2003; Krzywinski & Altman, 2013). Aus diesem Grund empfiehlt es sich, die identifizierten Unterschiede in weiterer Folge auf einer inhaltlichen Ebene zu interpretieren. Im bisher verwendeten Beispiel könnte es etwa zu statistisch signifikanten Veränderungen der Werte kommen. Aus klinischer Perspektive könnten diese jedoch zu gering ausfallen, beispielsweise wenn die Verbesserungen in der Ausprägung der Symptomatik das Intervall einer schweren Depression nicht verlassen (z.B. M_{BDI} T0= 55, M_{BDI} T1= 50, M_{BDI} T2= 43). Wäre hingegen eine deutlichere Reduktion der Werte feststellbar (z.B. M_{BDI} T0= 55, M_{BDI} T1= 40, M_{BDI} T2= 22), wäre der statistisch signifikante Unterschied auch aus klinischer Sicht bedeutsam, da die Therapie zu einer erfolgreichen Verbesserung der Symptomatik von einer schweren Depression zu einer moderaten Depression geführt hat (Beck, Steer & Brown (1996): nicht klinischer Bereich 0-13 Punkte, milde Depression 14-19 Punkte, moderate Depression 20-28 Punkte, schwere Depression 29-63). Zusätzlich zur Bestimmung der Ergebnisrelevanz anhand deskriptiver Kennwerte (Mittelwerte und Standardabweichungen) sollten ebenfalls Effektstärken herangezogen werden.

Effektstärken

Wie bereits erläutert, hängt die Berechnung des p-Wertes nicht nur von der eigentlichen Größe des beobachteten Unterschiedes ab. Beispielsweise könnte eine für die Berechnungen nicht optimale, sondern zu große Anzahl an Personen herangezogen worden sein (Oversampling), weshalb ein Unterschied nur aufgrund dessen signifikant ausfallen könnte (Jones, Carley & Harrison, 2003; Krzywinski & Altman, 2013). Ein gegenteiliges Szenario könnte auftreten, wenn zu kleine

Stichprobengrößen (Undersampling) für die Berechnungen herangezogen wurden (nicht signifikantes Ergebnis, obwohl in der Population ein statistisch relevanter Effekt zu beobachten wäre). Um den eigentlichen Effekt eines Unterschieds sinnvoll bestimmen zu können, ist es somit nicht ausreichend, lediglich den p-Wert als Bestimmungsmaß anzugeben. Zusätzlich sollte immer eine Effektstärke berechnet und interpretiert werden, welche Aussagen über die Größe des eigentlichen Effektes (Unterschiedes) sowie einen studienübergreifenden Ergebnisvergleich ermöglichen (Morris & Richler, 2012; Sullivan & Feinn, 2012; Seistock, Bunina & Aden, 2020).

Im Rahmen einfaktorieller varianzanalytischer Verfahren (ANOVAs) ist das *partielle Eta-Quadrat* (η_p^2) die am häufigsten verwendete Effektstärke (Fritz, Morris & Richler, 2012; Döring & Bortz, 2016). Neben dem η_p^2 existieren noch weitere Effektmaße, die im Rahmen der einfaktoriellen ANOVA für verbundene Stichproben berechnet und interpretiert werden können (z.B. η_G^2 siehe Bakeman, 2005). Im vorliegenden Beitrag wird sich die ausführlichere Beschreibung auf das partielle Eta-Quadrat (η_p^2) beschränken.

Partielles Eta-Quadrat

Die Effektstärke η_p^2 gibt an, wie hoch der Anteil an der gesamten „Innersubjektvariation“ (*SSwithin*) ist, der auf den Gruppenfaktor bzw. die Messzeitpunkte zurückzuführen[1] ist (*SStreat*). Anders formuliert: η_p^2 gibt an, welcher Anteil an der gesamten Verschiedenheit über die Zeitpunkte hinweg (Innersubjektvariation *SSwithin*) in einem bestimmten Merkmal allein durch die Zeitpunkte (bzw. die dazwischen stattfindenden Intervention) erklärt werden kann. Bezogen auf das vorangegangene Beispiel zur Evaluation psychotherapeutischer Behandlung von Depressionen gäbe das η_p^2 wieder, wie stark Veränderungen in der depressiven Symptomatik durch die Zeitpunkte bzw. die dazwischen liegende therapeutische Intervention erklärt werden können. Der η_p^2-Wert kann dabei als Prozentsatz interpretiert werden. So bedeutete ein Eta-Quadrat von η_p^2= .23, dass 23% der Innersubjektvariation in der metrischen Zielvariable durch die Zeitpunkte bzw. die darin repräsentierte therapeutische Intervention erklärt werden kann.

Die Formel zur Berechnung des partiellen Eta-Quadrats setzt sich dabei aus dem Quotienten der systematischen Variation zwischen den Messzeitpunkten (*SStreat)* und der gesamten Innersubjektvariation (*SSwithin*). Die SSwithin setzt sich ihrerseits aus der systematischen Variation zwischen den Messzeitpunkten (*SStreat)* und einer zusätzlichen Fehlervariation (*SSerror)* zusammen (Tomczak & Tomczak, 2012).

$$\eta_p^2 = \frac{SS_{treat}}{(SS_{treat} + SS_{error})}$$

Das η_p^2 kann Werte zwischen 0 und +1 annehmen. Eine Möglichkeit zur Beurteilung des Effektes sowie dessen Stärke liefert beispielsweise Cohen (1988), welcher folgende Bewertungsstufen vorschlägt:

[1] Alltagssprachliche Formulierung, keineswegs als Kausalität zu verstehen.

Beurteilung des Effekts

0.01 kleiner Effekt

0.06 mittlerer Effekt

0.14 großer Effekt

Die Bewertungsstufen sollten jedoch eher als Orientierungsrahmen dienen, da selbst Cohen (1988) globale Bewertungskonventionen kritisch bewertete. Als Bezugsrahmen für die Interpretation der Größe des Effektes sollten stets Effektgrößen aus vergleichbaren Studien herangezogen werden – genauere Ausführungen zur Anwendung von Bewertungskonventionen in der psychologischen Forschung sind beispielsweis bei Schäfer und Schwarz (2019) zu finden.

Kendall's W

Im Rahmen des Friedman-Tests ist es zumindest theoretisch möglich eine eigene Effektstärke zu berechnen. Dabei handelt es sich um Kendall's *W* (Tomczak & Tomczak, 2014):

$$W = \frac{\chi^2}{n(k-1)}$$

In der Auswertungspraxis findet dieses Maß jedoch nur in seltenen Fällen Berücksichtigung.

Voraussetzungen

Skalenniveau

Auch im Rahmen der Untersuchung von Unterschieden zwischen k-verbundenen Stichproben (einfaktorielle abhängige ANOVA oder Friedman-Test) ist die Wahl des optimalen Auswertungsverfahrens an bestimmte Voraussetzungen gebunden, welche zuerst geprüft werden müssen, bevor eine korrekte Berechnung sowie Ergebnisinterpretation möglich ist. In erster Linie ist hier das Skalenniveau (metrisch oder ordinal) des zu untersuchenden Merkmal (abhängige Variable, im verwendeten Beispiel der Depressionsscore gemessen mit Hilfe des BDI-II) entscheidend. Die einfaktorielle Varianzanalyse mit Messwiederholung (abhängige ANOVA) wird ausschließlich bei metrisch skalierten Variablen angewandt. Weist die zu untersuchende Variable ein ordinales Skalenniveau auf, dient der Friedman-Test als Verfahren der Wahl. Wie bereits bei Aden, Bunina und Vavrik (2021) verdeutlichen, wird das Skalenniveau der untersuchten Variable durch die Art der Datenerhebung bestimmt. Während ordinalskalierte Variablen beispielsweise aus subjektiven Einschätzungen der Untersuchungsteilnehmer*innen resultieren (z.B. Einschätzung der Zufriedenheit mit einer Therapie von 0 bis 100), gelten Merkmale (z.B. Ausprägung einer depressiven Symptomatik, Psychotherapiemotivation), welche auf der Basis von standardisierten Messinstrumenten (z.B. standardisierte Fragebögen wie das Beck-Depressions-Inventar (Kühner, Bürger, Keller & Hautzinger, 2007) oder der FPTM (Nübling et al., 2005)) erlangt werden, als metrisch. Die Bestimmung des korrekten Skalenniveaus der zu untersuchenden Variable (metrisch/mindestens intervallskaliert oder ordinal) stellt somit die Grundvoraussetzung für eine korrekte Verfahrenswahl dar.

Neben dem Skalenniveau des zu untersuchenden Merkmals sollten Überlegungen zur Maximierung der Testmacht eine entscheidende Rolle für die Verfahrenswahl spielen. Handelt es sich um eine metrische Variable, sollte wenn möglich eine einfaktorielle Varianzanalyse für verbundene Stichproben berechnet werden, da diese über eine höhere Testmacht verfügt, somit präziser imstande ist, einen systematischen Unterschied korrekterweise als solchen zu identifizieren. Im Vergleich zum Friedman-Test stellt die einfaktorielle Varianzanalyse für verbundene Stichproben (abhängige ANOVA), analog zur einfaktoriellen Varianzanalyse für unverbundene Stichproben (Aden, Bunina & Vavrik, 2021), das Verfahren mit den restriktiveren Voraussetzungen dar. Da die Berechnung der einfaktoriellen Varianzanalyse für verbundene Stichproben ebenfalls auf Mittelwerten und Varianzen beruht, stellt die Normalverteilung des untersuchten Merkmals in allen Stichproben (in allen Gruppen/zu allen Messzeitpunkten) eine grundlegende Voraussetzung für die Anwendung ebendieser dar. Eine Möglichkeit die Normalverteilung des untersuchten Merkmals (z.B. der BDI-Messwerte) zu allen Zeitpunkten/in allen Gruppen (z.B. Messzeitpunkt T0, T1, T2) liefern die sogenannten *Anpassungstests* (*Goodness-of-Fit-Tests*) nach Kolmogorov-Smirnov beziehungsweise Shapiro-Wilk dar. Detailliertere Ausführungen zur Normalverteilung sowie zu Möglichkeiten der Überprüfung ebendieser finden sich beispielsweise im ersten Artikel der Serie Statistik des Forschungsbulletins (Seistock, Bunina & Aden, 2020). Weist das untersuchte Merkmal keine Normalverteilung zu allen Zeitpunkten/in allen Gruppen auf, kann alternativ der Friedman-Test zur Bestimmung signifikanter Unterschiede herangezogen werden.

Normal-verteilung in jeder Gruppe

Im Gegensatz zur einfaktoriellen Varianzanalyse für unabhängige Stichproben, bei der die Homogenität der Varianzen eine weitere wesentliche Voraussetzung darstellt (Aden, Bunina & Vavrik, 2021), wird diese Voraussetzung bei der einfaktoriellen Varianzanalyse für verbundene Stichproben um einen zusätzlichen Aspekt erweitert. Neben der Varianzhomogenität, also dem Vorhandensein ähnlicher Merkmalsstreuungen zu allen Messzeitpunkten/in allen Gruppen, muss auch eine Korrelationshomogenität zwischen den Merkmalsausprägungen der Gruppen beziehungsweise Messzeitpunkten vorliegen. Die Kombination dieser beiden Aspekte wird im Rahmen der Varianzanalyse für verbundene Stichproben als Sphärizität beziehungsweise Zirkularitätsannahme bezeichnet (Bortz & Schuster, 2010; Rasch, Friese, Hofmann & Naumann, 2014; Janczyk & Pfister, 2020). Wird davon ausgegangen, dass drei Messzeitpunkte im Zuge einer Untersuchung analysiert werden, müssen die Korrelationen zwischen den paarweisen Messzeitpunktskombinationen homogen sein.

Sphärizität

Um die Sphärizitätsbedingung empirisch überprüfen zu können, wird der *Mauchly-Test* als Verfahren herangezogen. Wie auch bei der Überprüfung der Normalverteilung beziehungsweise Varianzhomogenität sollte im Zuge der Sphärizitätsprüfung beim Mauchly-Test ein nicht signifikantes Ergebnis vorliegen, damit die Voraussetzung als gegeben angenommen werden kann ($p>.05$). Sollte die Sphärizitätsbedingung als erfüllt beurteilt werden können, wird als Auswertungsverfahren die einfaktorielle Varianzanalyse mit Messwiederholung (abhängige ANOVA) herangezogen. Bei Verletzung der Sphärizität können spezielle Korrekturen als Kompensation durchgeführt werden. Innerhalb jener Korrekturen werden die Freiheitsgrade (df) im Zuge des F-Tests so adjustiert, dass die Wahrscheinlichkeit, die H0 zu verwerfen, geringer ausfällt. Ohne Korrektur wäre die Testung zu progressiv und das Risiko eines Alpha-Fehlers wäre erhöht. Die

Umgang mit Verletzung der Sphärizitäts-bedingung

Korrekturen im Anschluss an einen signifikanten Mauchly-Tests können damit als konservative Konsequenz betrachtet werden, um die Wahrscheinlichkeit, einem potenziellen α-Fehler zu unterliegen, zu minimieren (Bortz & Schuster, 2010).

Grundsätzlich können Korrekturen nach Greenhouse-Geisser, Box und Huynh-Feldt unterschieden werden (Rasch, Friese, Hofmann & Naumann, 2014; Bortz & Schuster, 2010). Dabei variieren all jene Korrekturen danach, wie konservativ die Adjustierung ausfallen soll. Allen Korrekturen liegt dabei der Faktor Epsilon (ϵ) zugrunde. Bei der konservativsten Adjustierung nach Greenhouse-Geisser wird der kleinstmögliche Wert für den Faktor Epsilon herangezogen. Im Computerprogramm SPSS wird diese Korrektur nicht als „Greenhouse-Geisser", sondern als „Untergrenze" benannt, was bei der Interpretation der Ergebnisse berücksichtigt werden sollte. Die Korrektur nach Box, bei der Epsilon ≤ .75 beträgt, zählt zu der am gängigsten verwendeten Methode, in SPSS wird sie als „Greenhouse-Geisser" bezeichnet. Die Huynh-Feldt Korrektur stellt die am wenigsten konservative Adjustierung im Zuge einer Sphärizitätsverletzung dar. Epsilon beträgt dabei >.75. Hierbei kann es aufgrund der nur geringen Korrektur zu einem fälschlicherweise signifikanten Ergebnis kommen. In der Literatur wird daher die Korrektur nach Box (in SPSS „Greenhouse-Geisser") empfohlen, wenn die Sphärizitätsbedingung nicht als erfüllt angenommen werden kann (Rasch, Friese, Hofmann & Naumann, 2014).

Die einfaktorielle abhängige Varianzanalyse

einfakt. abhängige ANOVA

Analog zur einfaktoriellen unabhängigen Varianzanalyse basiert die abhängige ANOVA auf der grundlegenden Idee der *Quadratsummenzerlegung* (Partition of Sum of Squares) zur Erklärung der Variation/Varianz eines untersuchten Merkmals. Dabei wird versucht, die Gesamtvariation des untersuchten Merkmals in einzelne Varianzbestandteile, nämlich systematische sowie unsystematische, zu gliedern, welche in weitere Folge zur Untersuchung, ob ein signifikanter Unterschied zwischen den verbundenen Stichproben besteht, in Relation gesetzt werden. Im Vergleich zur einfaktoriellen unabhängigen Varianzanalyse, im Rahmen derer die Gesamtvarianz in Varianzanteile zwischen den Gruppen (Zwischensubjektvariation, Quadratsumme zwischen, SS_{factor}) sowie innerhalb der Gruppen (Innersubjektvariation, Quadratsumme innerhalb, SS_{error}) aufgeteilt wird (Aden, Bunina & Vavrik, 2021), fokussiert die abhängige ANOVA ausschließlich die Analyse der Variation innerhalb der Stichproben (Gruppen/Messzeitpunkten). Dieser als *Quadratsumme Innerhalb* ($SS_{within} = \sum\sum(\bar{\boldsymbol{x}}_{\boldsymbol{im}} - \bar{\boldsymbol{P}}_{\boldsymbol{m}})^2$) bezeichnete Varianzanteil setzt sich wiederum aus zwei unterschiedlichen Varianzquellen, nämlich der systematischen Variation, welche auf Treatmenteffekte (z.B. Effekt durch psychotherapeutische Intervention) zurückgeführt wird (*Quadratsumme Treatment*, $SS_{treat} = \mathrm{n} * \sum(\bar{A}_i - \bar{G})^2$) sowie des unsystematischen Varianzanteils, welcher auf Grund von Interaktionseffekten beziehungsweise Messfehlereffekteentsteht (*Quadratsumme error*, $SS_{error} = SS_{within} - SS_{treat}$) zusammen. Alltagssprachlich formuliert, wird im Kontext der einfaktoriellen abhängigen Varianzanalyse somit jener Anteil an Unterschiedlichkeit (Varianz) des untersuchten Merkmals analysiert, welcher auf die Verschiedenheit INNERHALB der Personen über die verschiedenen Messzeitpunkte hinweg zurückzuführen ist. Umgelegt auf das bereits erläuterte Beispiel der Untersuchung der Effektivität einer Therapiemethode zur Behandlung von Depressionen (drei Messzeitpunkte BDI T0, T1, T2), bedeutet dies, dass nicht untersucht wird, inwiefern sich die einzelnen Personen voneinander unterscheiden, sondern ob

sich ein Unterschied in der Werteausprägung innerhalb der einzelnen Personen über die drei Messzeitpunkte hinweg identifizieren lässt. Um zu überprüfen, ob sich die Messzeitpunkte signifikant voneinander unterscheiden, wird der systematische Anteil der Verschiedenheit (SS_{treat}) mit dem unsystematischen Anteil an Verschiedenheit (SS_{error}) in Relation gesetzt. Dafür werden die beiden Quadratsummen (SS_{treat}, SS_{error}) durch die Relativierung an den Freiheitsgraden in Varianzen umgerechnet:

$$\hat{\sigma}^2{}_{treat} = \frac{\boldsymbol{SS_{treat}}}{df_{treat}}$$

$$\hat{\sigma}^2{}_{error} = \frac{\boldsymbol{SS_{error}}}{df_{error}}$$

Anschließend werden die beiden Varianzanteile unter Verwendung des sogenannten F-Tests in Relation gesetzt:

$$F = \frac{\hat{\sigma}^2{}_{treat}}{\hat{\sigma}^2{}_{error}}$$

Der als Resultat des F-Tests erhaltene empirische Verteilungswert F wird in weiterer Folge als Basis zur Bestimmung der Signifikanz des Ergebnisses herangezogen (siehe Signifikanzbestimmung weiter oben).

Angenommen man erhält beim Vergleich von drei Messzeitpunkten (Prä, Post, Follow-Up) hinsichtlich einer metrischen Variable mittels einfaktorieller Varianzanalyse für verbundene Stichproben einen F-Wert von 197,87 bei df_{treat} = 2 und df_{error} = 58, mit einem p-Wert von <.001 sowie einer Effektstärke von η_p^2 = .87.

Für die statistische Reportlegung des Ergebnisses des Over-All-Verfahrens (ANOVA abhängig) empfiehlt sich die Angabe folgender Kennwerte:

Reportkonventionen einfakt. abhängige ANOVA

$$\textit{(F(2,58)= 197,87, p< .001, } \eta_p^2 \textit{= 0.87)}$$

Im Anschluss an das signifikante Ergebnis des Hauptverfahrens werden dann zur differenzierteren Betrachtung der paarweisen Unterschiede post-Hoc-Tests durchgeführt. Die Ergebnisse der paarweisen Vergleiche können dann beispielsweise wie folgt berichtet werden:

Im Anschluss an das signifikante Ergebnis der abhängigen Varianzanalyse, konnte im Rahmen der paarweisen Vergleiche festgestellt werden, dass sich Messzeitpunkt Prä $M_{Prä}$= 22.10 (SD= 5.71), signifikant von Messzeitpunkt Post M_{Post}= 20.17 (SD= 5.66) sowie von Messzeitpunkt Follow-Up $M_{FollowUp}$= 13.70 (SD= 5.84) unterscheidet (jeweils p< .001). Darüber hinaus zeigt sich eine signifikante Differenz zwischen Messzeitpunkt Post und Messzeitpunkt Follow-Up (p< .001)[2].

Im Falle der Verletzung der Sphärizitätsbedingung können, wie bereits erläutert, die Freiheitsgrade korrigiert werden (z.B. nach Greenhouse-Geisser oder Huynh-Feldt). Die Reportlegung folgt in einem solchen Fall im Wesentlichen jener der abhängigen ANOVA ohne Korrektur, es werden lediglich die korrigierten Freiheitsgrade statt den unkorrigierten angegeben.

Der Friedman-Test

Friedman-Test

Beim Friedman-Test handelt es sich ebenfalls um ein inferenzstatistisches Auswertungsverfahren zur Berechnung von signifikanten Unterschieden zwischen mehr als zwei verbundenen Stichproben (Schuchmann & Sanns, 2018). Im Vergleich zur Varianzanalyse für verbundene Stichproben wird im Rahmen des Friedman-Test die Chi^2-Verteilung für die Bestimmung der Signifikanz eines Unterschiedes herangezogen. Der Friedmantest findet seine Anwendung, wenn k-verbundene Stichproben a) hinsichtlich eines metrischen, jedoch nicht normalverteilten Merkmals oder b) hinsichtlich eines rangskalierten Merkmals verglichen werden. Ähnlich zum Kruskal-Wallis-H-Test, einem Testverfahren zur Überprüfung von Unterschieden zwischen k-unverbundenen Stichproben (Aden, Bunina & Vavrik, 2021), werden im Zuge der Berechnung des Friedman-Tests die einzelnen Werte in sogenannte Ränge transfomiert (detaillierte Informationen zur Rangtransformation siehe z.B. Seistock, Bunina & Aden, 2020). Die Verwendung von Rängen erlaubt in weiterer Folge eine Berechnung des Verfahrens ohne spezifische Anforderungen an die Verteilungslage der erhobenen Daten, weshalb der Friedmantest sowohl für nicht normalverteile Daten als auch für rangskalierte Merkmale berechnet werden kann.

$$\chi^2 = \frac{12}{n*k*(k+1)} * \sum_i^k T^2{}_i - 3 * n * (k+1)$$

Analog zur Ergebnisinterpretation der einfaktoriellen Varianzanalyse für verbundene Stichproben erfolgt die Beurteilung der Signifikanz der Gruppenunterschiede/Unterschiede der Messzeitpunkte anhand des berechneten Verteilungswertes (χ^2). Je nach Skalenniveau der untersuchten Variable können im Anschluss an das Hauptverfahren entweder a) bei einem metrischen, nicht normalverteilten Merkmal paarweise Post-Hoc Wilcoxon-Tests oder b) bei einem ordinalen Merkmal paarweise Post-Hoc- Binomial-Vorzeichentests berechnet werden.

Angenommen man erhält beim Vergleich von drei Messzeitpunkten (Prä, Post, Follow-Up) hinsichtlich einer ordinalen Variable mittels Friedman-Test einen χ^2-Wert von 52.47 bei df = 2, mit einem p-Wert von <.001.

[2] Bei spezifischem Interesse empfiehlt es sich die Effektstärken *d* zu den jeweiligen Vergleichen zu ergänzen.

Für die statistische Reportlegung des Ergebnisses des Over-All-Verfahrens (Friedman-Test) empfiehlt sich die Angabe folgender Kennwerte:

Reportkonventionen Friedman-Test

$$(\chi^2(df= 2, n= 30)= 52.47, p< .001)$$

Im Anschluss an das signifikante Ergebnis des Hauptverfahrens können dann zur differenzierteren Betrachtung der paarweisen Unterschiede post-Hoc-Tests (Wilcoxon/Binomial-Vorzeichentests) durchgeführt werden. Die Ergebnisse der paarweisen Vergleiche können dann beispielsweise wie folgt berichtet werden:

Im Anschluss an das signifikante Ergebnis des Friedman-Tests konnte im Rahmen der paarweisen Vergleiche festgestellt werden, dass sich Messzeitpunkt Prä $M_{Prä}$= 22.10 (SD= 5.71) signifikant von Messzeitpunkt Post M_{Post}= 20.17 (SD= 5.66) (Z= -5.32, p<.001) sowie von Messzeitpunkt Follow-Up $M_{FollowUp}$= 13.70 (SD= 5.84) unterscheidet (Z= -4.77, p<.001). Darüber hinaus zeigt sich eine signifikante Differenz zwischen Messzeitpunkt Post und Messzeitpunkt Follow-Up (z= -4.72, p<.001)[3].

Konklusion

Auswertungsverfahren zum Vergleich mehrerer verbundener Stichproben (z.B. Messzeitpunkte) geben Forscher*innen einer quantitativ ausgerichteten Psychotherapiewissenschaft die Möglichkeit, die Effektivität psychotherapeutischer Interventionen differenzierter und umfangreicher zu untersuchen. Beispielsweise können durch einen Vergleich von drei Messzeitpunkten (Prä, Post, Follow-Up) im Vergleich zu klassischen Prä-Post-Designs mit nur zwei Messungen (vor Beginn und nach Abschluss der Therapie) Aussagen über nachhaltige Effekte von Therapien getroffen werden. Des Weiteren ermöglicht der Vergleich mehrerer Messzeitpunkte, der Komplexität der Veränderung psychologischer Merkmale Rechnung zu tragen, da Veränderungen in manchen Fällen erst einige Zeit nach Abschluss eines therapeutischen Programms manifest werden können (z.B. erlernte Strategien zum Umgang mit Belastungen müssen längerfristig geübt werden, bevor diese zu einer merklichen Reduktion des Belastungsgrades führen). Darüber hinaus ermöglichen inferenzstatistische Verfahren zum Vergleich mehrerer verbundener Stichproben prinzipiell die Auswertung von Längsschnittstudien, wodurch sich ein Einsatz für eine breite inhaltliche Vielfalt psychotherapeutischer Fragestellungen ergibt. Beispielsweise können diese Auswertungsverfahren auch für die Evaluation von Zufriedenheitswerten der Klient*innen über den Verlauf der Therapie oder zur Untersuchung der Veränderung der Klient*innen-Therapeut*innen-Beziehung eingesetzt werden.

[3] Bei spezifischem Interesse empfiehlt es sich die Effektstärken *r* zu den jeweiligen Vergleichen zu ergänzen.

Insbesondere im Bereich der Evaluation psychotherapeutischer Interventionen zum Zweck der Qualitätssicherung sowie der Förderung der Akzeptanz von Therapieprogrammen ist eine korrekte Verfahrenswahl sowie Interpretation der Ergebnisse essenziell. Die wichtigsten Empfehlungen im Umgang mit Auswertungsverfahren zur Untersuchung von Unterschieden zwischen k-verbundenen Stichproben werden abschließend nochmals zusammenfassend dargestellt:

→ *Alpha-Kumulierung & Forschungsökonomie*

Der paarweise Vergleich von mehr als zwei verbundenen Stichproben führt zu einer Kumulierung der Alpha-Fehler-Wahrscheinlichkeit und schränkt dadurch die statistische Entscheidungssicherheit ein. Der Einsatz von „Over-All"-Verfahren wie der abhängigen ANOVA sowie des Friedman-Tests kann diese Problematik einschränken, indem nur eine „übergeordnete" Hypothese überprüft wird (Besteht zumindest ein paarweiser Unterschied zwischen den k-verbundenen Stichproben?). Neben dem Vorteil der Vermeidung der Alpha-Fehler-Inflation stellen Over-All-Verfahren prinzipiell ökonomischere Auswertungsverfahren dar, da bei einem nicht signifikanten Ergebnis keine weiteren (potenziell zahlreichen) paarweise Testungen berechnet werden müssen. Bei einem signifikanten Ergebnis ermöglicht die anschließende Berechnung sogenannter paarweiser Post-Hoc-Testungen Aussagen über die konkrete Anzahl an paarweisen Unterschieden:

→ *Post-Hoc-Testungen*

Ein signifikantes Ergebnis im Rahmen einer abhängigen ANOVA oder eines Friedman-Tests bedeutet lediglich, dass zumindest ein paarweiser Unterschied zwischen den untersuchten k-verbundenen Stichproben (z.B. Messzeitpunkten) besteht. Dieses Ergebnis erlaubt jedoch noch keinen Rückschluss darauf, welche Messzeitpunkte sich in welcher Form voneinander unterscheiden. Um diese Frage zu beantworten, werden sogenannte paarweise Post-Hoc-Testungen durchgeführt. Durch die Berechnung paarweiser Post-Hoc-Testungen steigt die Anzahl der durchgeführten Signifikanzprüfungen dennoch, weshalb der damit verbundene Anstieg der Wahrscheinlichkeit, einem Alpha-Fehler zu unterliegen (Alpha-Inflation), berücksichtigt werden muss. Eine Möglichkeit mit diesem Umstand umzugehen, besteht beispielsweise in der Verwendung von Post-Hoc-Bonferroni-Tests (abhängige ANOVA), welche automatisch eine Adjustierung der Signifikanzwerte (p-Werte) vornehmen (Bender, Lange & Ziegler, 2007).

Darüber hinaus können im Kontext der abhängigen ANOVA auch sogenannte Kontraste berechnet werden, welche die Überprüfung spezifischer Annahmen bezüglich der Unterschiede zwischen den einzelnen Messzeitpunkten ermöglichen (z.B. wenn eine lineare Abnahme der Ausprägung der Depression über die Messzeitpunkte angenommen wird). Die Anwendung polynomialer sowie orthogonaler Kontraste wird in einem späteren Artikel der Serie Statistik behandelt.

→ *Effektstärken*

Effektstärken wie das partielle Eta-Quadrat (η_p^2) ermöglichen eine Interpretation der Größe des übergeordneten Unterschiedes (alltagssprachlich formuliert: Wie groß ist der insgesamte

Unterschied zwischen den Messzeitpunkten?) und sollten stets zur Interpretation herangezogen werden. Dieser übgeordnete Effekt kann jedoch nicht auf die paarweisen Unterschieden übertragen werden, weshalb die Möglichkeit besteht, gesonderte Effektstärken (z.B. d oder r) für die paarweisen Vergleiche anzugeben.

Die Berechnung der Signifikanz eines Ergebnisses kann durch Stichprobeneffekte, beispielsweise bedingt durch eine nicht optimale Samplingstrategie (Over- oder Undersampling), beeinflusst werden. Insofern ist die Interpretation eines Ergebnisses nur auf der Basis des Signifikanzwertes (p-Wertes) unzureichend, weshalb stets Effektstärken angegeben werden sollten.

→ *Kalkulation des optimalen Stichprobenumfangs*

Um die beschriebenen Stichprobeneffekte (Over- oder Undersampling) zu vermeiden, empfiehlt es sich, eine Kalkulation des optimalen Stichprobenumfangs vorzunehmen (z.B. Chow, Shao, Wang & Lokhnygina, 2018). Darüber hinaus kann so ein adäquates Verhältnis von Testmacht und Fehlerwahrscheinlichkeit sichergestellt werden. Kalkulationen des optimalen Stichprobenumfangs könne beispielsweise mit den Programmen „G*Power“ oder „R“ vorgenommen werden.

→ *Statistische und inhaltliche Bedeutsamkeit*

Speziell im Bereich der Evaluation von psychotherapeutischen Interventionen ist es essenziell, neben der statistischen Bedeutsamkeit der Ergebnisse (anhand des p-Wertes) auch die inhaltliche Relevanz ebendieser zu diskutieren. Neben der Angabe und Interpretation von Effektstärken sollten hier vor allem die Skalierung des Merkmals sowie etwaige Grenzwerte beachtet werden. Beispielsweise kann unter Betrachtung deskriptiver Kennwerte (Mittelwert, Standardabweichung, Median) festgestellt werden, ob eine therapeutische Intervention tatsächlich einen bedeutsamen Effekt auf die Veränderung in der Ausprägung eines Merkmals hat. Sinkt z.B.der Depressionswert (BDI-II) nach der Therapie in einem Ausmaß, welches als substanzielle Symptomverbesserung interpretiert werden kann (M_{BDI} T0= 55, M_{BDI} T1= 40, M_{BDI} T2= 22).

→ *Sphärizität und Korrektur der Freiheitsgrade*

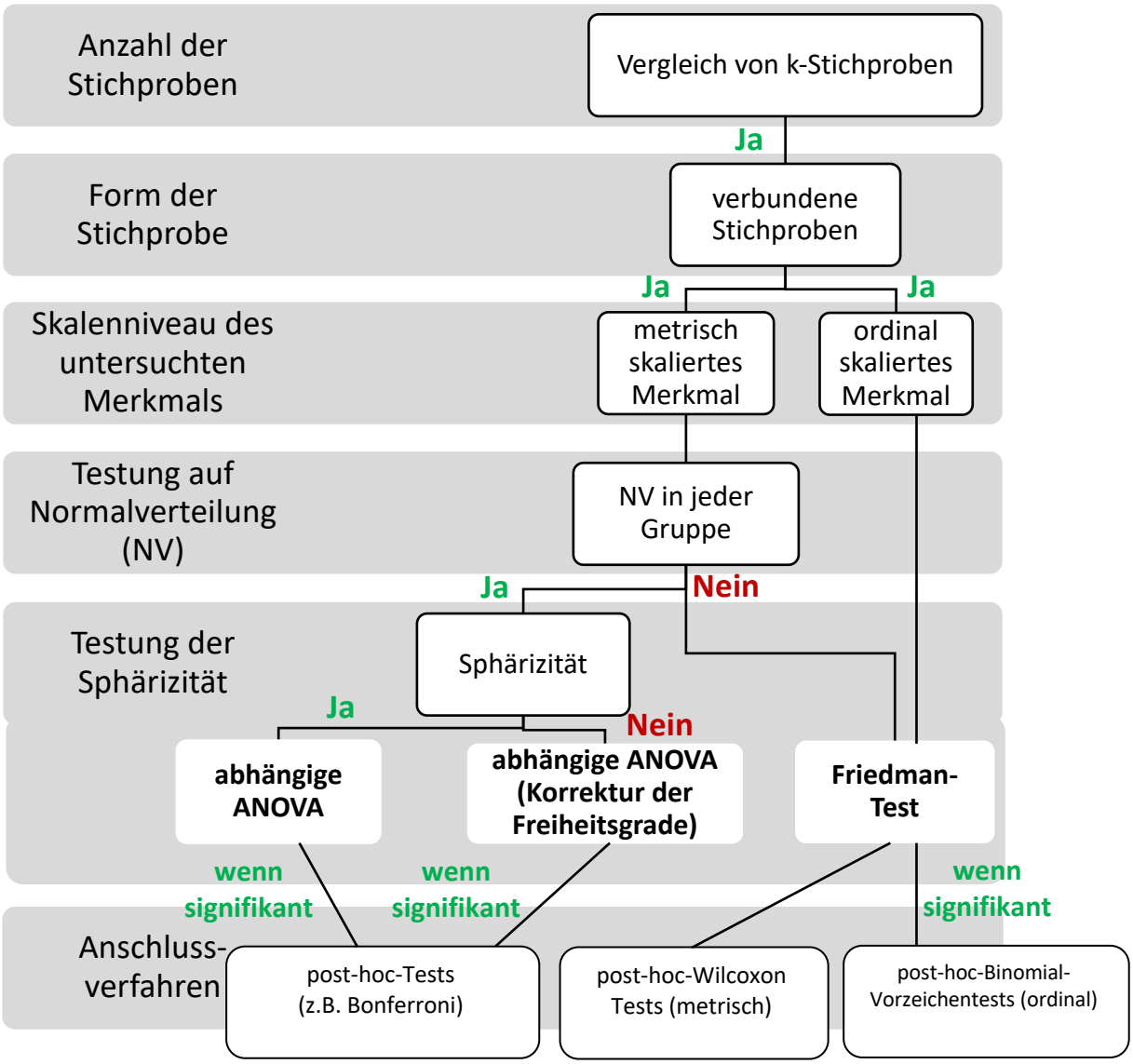
Die Bedingung der Sphärizität spielt im Kontext der Varianzanalyse für verbundene Stichproben eine entscheidende Rolle, da eine Verletzung ebendieser in der Regel dazu führt, dass die Wahrscheinlichkeit fälschlicherweise einen Unterschied zu konstatieren ansteigt (Bortz & Schuster, 2010). Die korrekte Überprüfung der Bedingung der Sphärizität anhand des Mauchly-Tests ist somit essenziell. Ist die Bedingung nicht gegeben, können die Freiheitsgrade korrigiert werden (Greenhouse-Geisser-, Huynh-Feldt-, Box-Korrektur), wodurch die Wahrscheinlichkeit minimiert wird, einem potenziellen α-Fehler zu unterliegen.

→ *Vergleich mit einer Kontrollgruppe*

Im Kontext der Wirksamkeitsforschung psychotherapeutischer Interventionen und Therapieprogrammen gelten heutzutage sogenannte randomisiert kontrollierte Studien (engl.: randomized controlled trial, RCT) als Goldstandard. Im Zuge solcher Studien werden neben der Analyse von Unterschieden zwischen mehreren Messzeitpunkten (z.B. Prä, Post, Follow-Up) ebenfalls Unterschiede zwischen einer Interventions- und Kontrollgruppe (unverbundene Stichproben) über die Messzeitpunkte hinweg evaluiert. Diese Vorgehensweise dient dazu, eine elaboriertere Interpretation der Effekte der Therapie zu gewährleisten, d.h. in einem höheren Maße sicherzustellen, dass die Veränderung in der Merkmalsausprägung auch wirklich auf die Intervention „zurückzuführen“[4] ist. Um die Validität der Evaluationsergebnisse zu erhöhen, empfiehlt sich sich somit, auf die Verwendung von Kontroll- und Interventionsgruppen zurückzugreifen. Untersuchungsdesigns dieser Art lassen sich anhand der Mixed ANOVA auswerten, welche in einem späteren Artikel der Serie Statistik vorgestellt wird.

[4] Eine derartige Untersuchung erlaubt dennoch keine Implikation einer Kausalbeziehung im klassischen Sinne, da varianzanalytische Verfahren lediglich Aussagen über Kovariation ermöglichen.

Entscheidungsbaum



Literatur

Aden, J., Bunina, A. & Vavrik, C. (2021). Die einfaktorielle Varianzanalyse für unabhängige Stichproben und der Kruskal-Wallis-Test im psychotherapiewissenschaftlichen Kontext. Empfehlungen für Anwendung und Interpretation. *SFU Forschungsbulletin*, *9*(1), 68–86.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods*, *37*(3), 379-384.

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

Bender, R., Lange, St., & Ziegler, A. (2007). Multiples Testen. *Dtsch Med Wochenschr, 132,* 26–29.

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

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollständig überarbeitete und erweiterte Auflage). Berlin, Heidelberg: Springer.

Chow, S.C., Shao, J., Wang, H., & Lokhnygina, Y. (2018). *Sample Size Calculations in Clinical Research*. New York: Chapman and Hall/CRC.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York, NY: Academic Press.

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

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General*, *141*(1), 2–18.

Gattermeyer, S., Vladarski, C., & Aden, J. (2020). Der t-, Wilcoxon- und Binomial/Vorzeichen-Test für zwei verbundene Stichproben im psychotherapiewissenschaftlichen Forschungskontext. Empfehlung für Anwendung und Interpretation. *SFU Forschungsbulletin*, *8*(2), 128–146.

Janczyk, M., & Pfister, R. (2020). *Inferenzstatistik verstehen*. Heidelberg: Springer.

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.

Nübling, R, Schulz, H., Schmidt, J., Koch, U. & Wittmann, W. W. (2005). Fragebogen zur Psychotherapiemotivation (FPTM) – Testkonstruktion und Gütekriterien. In R. Nübling, F. A. Muthny & J. Bengel (Hrsg.), *Reha-Motivation und Behandlungserwartung* (S. 252-270). Bern: Huber.

Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2014). Varianzanalyse mit Messwiederholung. *Quantitative Methoden Band 2: Einführung in die Statistik* (S. 99–141). Heidelberg: Springer.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10(813), 1–13.

Schuchmann, M., & Sanns, W. (2018). 10. Die Friedman Rang-Varianzanalyse. In *Nichtparametrische Statistik mit Mathematica* (S. 60–64). Oldenbourg: Wissenschaftsverlag.

Schulz, H., Nübling, R. & Rüddel, H (1995). Entwicklung einer Kurzform eines Fragebogens zur Psychotherapiemotivation. *Verhaltenstherapie*, *5*(2), 89–95.

Schulz, H., Lang, K., Nübling, R., & Koch, U. (2003). Psychometrische Überprüfung einer Kurzform des Fragebogens zur Psychotherapiemotivation – FPTM-23. *Diagnostica*, *49*(29), 83–93.

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.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of graduate medical education*, *4*(3), 279–282.

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *TRENDS in Sport Sciences, 1*(21), 19–25.

Angaben zu den Autoren

David Seistock, Elias Ruso, Jan Aden
Institut für Statistik, Fakultät für Psychologie
Sigmund Freud PrivatUniversität Wien
Adresse: Freudplatz 1, 1020 Wien, Raum 6011
E-Mail: jan.aden@sfu.ac.at, david.seistock@sfu.ac.at