

Interaktionen und Wechselwirkungen im Rahmen der Mixed ANOVA im psychotherapiewissenschaftlichen Forschungskontext

– Empfehlungen für Anwendung und Interpretation –

Interactions in the context of mixed anova in psychotherapy science

– Recommendations for application and interpretation –

Sophie Gattermeyer, Iris Enengl & Jan Aden

Kurzzusammenfassung

Im folgenden Artikel der Serie Statistik in der Psychotherapiewissenschaft wird die Anwendung von varianzanalytischen Verfahren zum Vergleich mehrerer k-unabhängiger und k-abhängiger Gruppenfaktoren sowie deren Wechselwirkung bzw. Interaktion in einem Modell (Mixed ANOVA und Mehrfaktorielle ANOVA) im Sinne eines Best-Practice Ansatzes vorgestellt. Darüber hinaus wird ihr Einsatzbereich im Rahmen von RCT (Randomised Controlled Trials = Randomisiert-Kontrollierte-Studien) bei Fragestellungen zur Wirksamkeitsforschung von therapeutischen Interventionen thematisiert. Es werden Empfehlungen für (1) eine optimale Verfahrenswahl, (2) den Einsatz von Effektstärken, (3) die Bestimmung der Ergebnisrelevanz sowie (4) Reportkonventionen für die Erhebungsdarstellung gegeben. Zudem wird ein weiterer Einsatzbereich von Wechselwirkungsfragestellungen innerhalb von mehrfaktoriellen Varianzanalysen mit unverbundenen Gruppenfaktoren kurz dargestellt sowie die wesentlichen Unterschiede zur Mixed ANOVA aufgezeigt. Darüber hinaus wird der Einsatz dieser Verfahrensgruppe im psychotherapiewissenschaftlichen Kontext anhand von Beispielen illustriert.

Schlüsselwörter

Mixed ANOVA, Mehrfaktorielle Varianzanalyse, Reportkonventionen, Effektstärken, Therapieevaluation, Wirksamkeit, RCT, Randomisierte-Kontrollierte-Studien

Abstract

In this article of the series Statistics in Psychotherapy Science, the application of methods for comparing several related k-independent and k-dependent samples and their interactions in one model (mixed ANOVA and multi-factorial) in the sense of a best practice approach is presented. Furthermore, their appliance in RCT (Randomised Controlled Trials) is discussed. 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 multifactorial ANOVA for independent samples is presented as another application area of interactions. Finally, the use of this group of procedures in the context of psychotherapy science is illustrated using practical examples.

Keywords

mixed ANOVA, multi-factorial ANOVA, report conventions, effect-sizes, therapy evaluation, effectiveness, RCT, Randomised Controlled Trials

Mixed ANOVA

Einsatzfeld RCT (Randomisiert-Kontrollierte Studien)

Einsatz-
bereich

Der stetige Nachweis der Wirksamkeit von psychotherapeutischen Interventionen ist für eine gelingende Praxis aber auch die dahinterstehende Forschungslandschaft unerlässlich. Neben dem Anliegen für den/die Klient*in eine subjektive Verbesserung des Leidenszustands zu erzielen, haben Wirksamkeitsstudien darüber hinaus in puncto Qualitätssicherung, Vergleichbarkeit sowie langfristiger Legitimation von Therapieformen einen hohen Stellenwert in der psychotherapeutischen Forschung.

Je nachdem welcher Anspruch an Wirksamkeitsforschung formuliert wird, stehen mehrere Optionen für ein Untersuchungsdesign zur Verfügung. Simplere Zugänge, wie beispielsweise der Nachweis über einen Symptomrückgang zwischen zwei oder mehreren Messzeitpunkten, können methodisch beispielsweise über Varianzanalysen oder t-Tests für verbundene Stichproben abgebildet werden, wie schon in vergangenen Artikeln vorgestellt wurde (vgl. Seistock, Ruso & Aden, 2021; Gattermeyer, Vladarski & Aden, 2020). Soll ein Nachweis einer Symptomveränderung jedoch zur Legitimation einer therapeutischen Intervention dienen, reichen Beobachtungen über einen vorgegebenen Zeitraum keineswegs aus. So definiert sich ein systematischer Therapieeffekt nicht nur durch den Grad der Symptomverbesserung einer*es Patienten*in, sondern auch durch die Tatsache, dass dieser Effekt nicht durch Selektions-, Zufalls- oder Zeiteffekte (engl. „bias“) bedingt, sondern systematisch auf die durchgeführte Intervention zurückzuführen ist.

In der Vergangenheit wurden RCTs (Randomised Controlled Trial = randomisierte kontrollierte Studie) immer wieder als „Goldstandard“ der evidenzbasierten Forschung zur Evaluierung der Wirksamkeit von Therapien ausgewiesen, da sie aufgrund ihrer strengen experimentellen Methodik als einziges Design in der Lage seien, die Wirkung einer Therapie auf die Intervention und nicht auf etwaige Drittvariablen zurückzuführen. Das methodische Design des RCTs hat seinen Ursprung in der EBM (Evidenz Basierten Medizin, vgl. Djulbegovic & Guyatt, 2017) und zeichnet sich im Wesentlichen dadurch aus, dass teilnehmende Personen randomisiert einer Interventions- oder Kontrollgruppe zugewiesen werden und über einen vordefinierten Testzeitraum zu mehreren Messzeitpunkten hinsichtlich des interessierenden Merkmals (z.B. wie hier Depression) untersucht werden (sh. Abb. 1) (zur vollständigen Beschreibung vgl. Jadad & Enkin, 2007; West et al., 2008). Dies soll einen möglichen Selektions-Bias bzw. rein zufällige Zeiteffekte ausschließen. In medizinischen Studien erfolgt häufig zusätzlich eine „double-blind“ Bedingung, wodurch weder Testperson noch Untersuchungsleiter*in wissen, welche Person welche Therapie erhält, um vor einer möglichen Beeinflussung zu schützen. Aufgrund des hohen Grades an Kontrolle wird dem RCT ein hohes Maß an interner Validität zugesprochen, durch welche eine beinahe bias-lose Ergebnisqualität und der „kausale“ Schluss von Therapie auf erzielte Wirkung ermöglicht wird. Aus diesem Grund werden auch RCTs beispielsweise reinen Beobachtungsstudien vorgezogen, da nur erstere einen Schluss auf sogenannte Treatment-Effekte zulassen (Pocock & Elbourne, 2000).

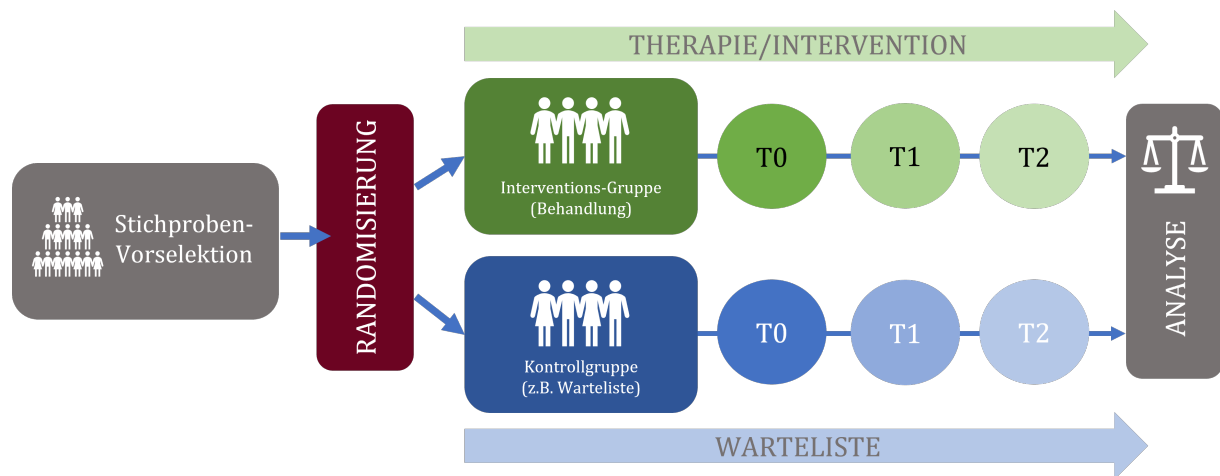


Abb. 1 vereinfachte Darstellung einer Randomisiert-Kontrollierten-Studie (RCT)

Da sich der medizinische Anwendungskontext von RCTs von jenem in der Psychotherapie in einigen zentralen Aspekten grundlegend unterscheidet, erfährt der Standpunkt des RCTs als Goldstandard für Interventionsforschung häufig auch kritische Stimmen, auf welche hingewiesen, jedoch im Rahmen dieses Artikels nicht näher eingegangen wird (vgl. Grossman & Mackenzie, 2005; Sanson-Fisher et al., 2007; Ware & Hamel, 2011; Philips & Falkenström, 2011). Unabhängig ihrer Kritik sind RCTs ein fester Bestandteil im Rahmen von quantitativer Therapieforschung, die aufgrund ihrer Komplexität einer korrekten Anwendung bedürfen, um möglichen methodischen Bias gering zu halten. Im Folgenden wird ein statistisches Verfahren zur Auswertung eines RCTs dargestellt und auf relevante Voraussetzungen und Kennwerte eingegangen.

Mixed ANOVA im Rahmen von RCTs

Um die Wirksamkeit einer therapeutischen Intervention nun im Rahmen einer randomisiert kontrollierten Studie nachweisen zu können, ergeben sich 3 wesentliche Fragestellungen, die beantwortet (und im Falle eines tatsächlichen Therapieeffekts bestätigt) werden sollten:

- 1. Die untersuchte Symptomatik bzw. das Merkmal sollten sich über den Untersuchungszeitraum der Testzeitpunkte reduzieren**
- 2. Die Interventionsgruppe sollte sich von der Kontrollgruppe unterscheiden**
- 3. Der Unterschied zwischen Interventions- und Kontrollgruppe sollte im zeitlichen Verlauf der Untersuchungszeitpunkte sichtbar werden (d.h. die Interventionsgruppe sollte einen systematisch stärkeren Symptommrückgang über die Zeitpunkte aufweisen im Gegensatz zur Kontrollgruppe)**

Sofern eine Therapie bzw. gesetzte Intervention tatsächlich „wirksam“ ist, sollten jene drei Aspekte in Kombination simultan zutreffen, was mit der Notwendigkeit einhergeht, sie auch methodisch in einem Modell integriert zu testen (und nicht beispielsweise in separaten Varianzanalysen).

Dieser Komplexität wird das Verfahren der Mixed ANOVA gerecht, da sie zur simultanen Überprüfung von Unterschieden zwischen mehreren Gruppenfaktoren, k-unabhängige (Interventions- vs. Kontrollgruppe) und k-abhängige (Messzeitpunkte), in einem Modell herangezogen werden kann. Dies bringt den zusätzlichen Vorteil, dass durch die kombinierte Testung der Gefahr einer Alphafehlerkummulierung entgegengewirkt wird.

Wie aus dem Namen *Mixed ANOVA* bereits hervorgeht, handelt es sich hierbei um ein Konglomerat der einfaktoriellen Varianzanalyse für unabhängige Stichproben und der einfaktoriellen Varianzanalyse für abhängige Stichproben, die bereits in der Serie Statistik zu den Unterschiedstestungen vorgestellt wurden (Aden, Bunina & Vavrik, 2021; Seistock, Ruso & Aden, 2021). Der wesentliche Unterschied zwischen der Mixed ANOVA und der einfaktoriellen Varianzanalyse besteht darin, dass auf Basis inhaltlich-theoretischer Erwägungen k-unverbundene und k-verbundene Stichproben und deren Interaktion bzw. Wechselwirkung im selben Modell statistisch untersucht werden. Im Rahmen der Berechnung einer Mixed ANOVA können daher Effekte über vorhandene Wechselwirkungen deutlich werden, die sich aus der Kombination mehrerer k-abhängigen und k-unabhängigen Gruppen ergeben. Folgernd führen die Ergebnisse über vorhandene Interaktionen der Variablen zu weiteren Erkenntnissen, welche bei Berechnungen einfaktorieller Varianzanalysen latent bleiben (Bortz, 2006). Nachfolgend wird das Verfahren der Mixed ANOVA im Detail dargestellt sowie auf wichtige Aspekte bei der Durchführung hingewiesen, die einen praxisnahen Einsatz im Rahmen von beispielsweise RCTs in den Therapiewissenschaften ermöglichen soll.

Bestimmung der Stichproben/Gruppen und Fragestellungen

Zu Beginn ist vor allem die korrekte Identifikation der jeweiligen Stichprobenart und der damit einhergehenden Fragestellung von großer Bedeutung, da innerhalb der Mixed ANOVA, wie bereits erwähnt, mehrere unterschiedliche Stichproben miteinander verglichen werden. Grundsätzlich wird bei Stichprobenarten zwischen abhängigen (verbundenen) und unabhängigen (unverbundenen) Stichproben unterschieden, deren Definition bereits im ersten Methodenartikel der Serie Statistik (Seistock, Bunina & Aden, 2020) näher erläutert wurde. Abhängige Stichproben zeichnen sich dadurch aus, dass sich die Elemente einander paarweise zuordnen lassen, wie beispielsweise bei Messwiederholungen, Parallelisierungen, Paarungen oder Matched Pairs. Im Rahmen einer Evaluationsforschung, zur Überprüfung der Wirksamkeit einer Therapie, handelt es sich bei der abhängigen Stichprobe meist um eine Messwiederholung, bei welcher eine bestimmte Merkmalsausprägung (z.B. Depression) mittels eines störungsspezifischen Messinstruments (z.B. gemessen anhand des Beck-Depressions-Inventar-II; Kühner, Bürger, Keller & Hautzinger, 2007) zu mehreren Testzeitpunkten erfasst wird. Die Erhebung einer Merkmalsausprägung findet in diesem Zusammenhang üblicherweise zu drei Messzeitpunkten statt, die sich mit den Begriffen Prä (Depression vor der Therapie), Post (Messung der Depression nach der Therapie) und Follow-Up (Messung der Depression drei Monate nach Beendigung der Therapie) abkürzen lassen (sh. Abb.2 bzw. Tab.1).

Bestimmung
der Stich-
probenart

Tab. 1 beispielhafte Werteübersicht von Test- und Gesamtmittelwerten im Rahmen eines Prä-Post-FollowUp Designs für die erfolgreiche Behandlung einer Major Depression im Vergleich mit einer Kontrollgruppe

		Messzeitpunkt 1 (Prä= T ₀)	Messzeitpunkt 2 (Post=T ₁)	Messzeitpunkt 3 (Follow Up =T ₂)
INTERVENTIONS-GRUPPE	KlientIn A	29	24	15
	KlientIn B	30	27	14
	KlientIn C	34	24	18
	...	...	...	...
	Mw	30	21	15
KONTROLL-GRUPPE	KlientIn D	31	34	30
	KlientIn E	25	30	28
	KlientIn F	27	33	31
	...	...	...	...
	Mw	28	35	31
GESAMT	Mw	29	28	23

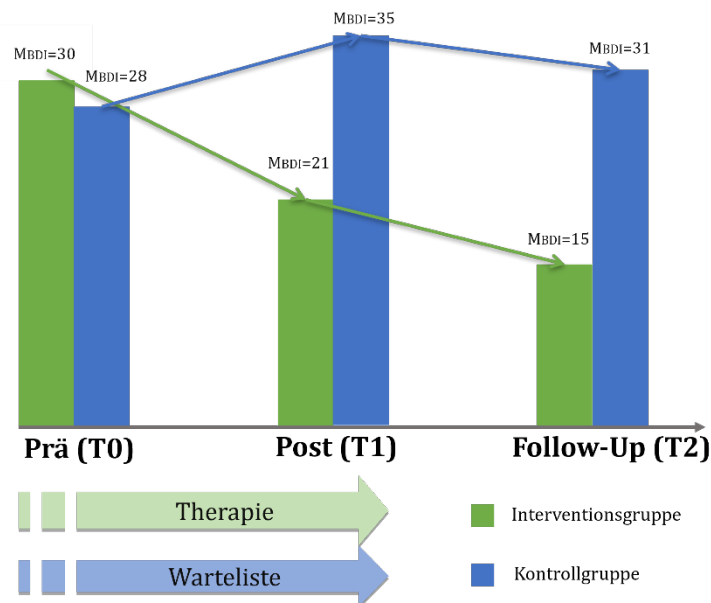


Abb. 2 beispielhafte Darstellung von Mittelwertsveränderungen im Rahmen eines Prä-Post-FollowUp Designs für die erfolgreiche Behandlung einer Major Depression im Vergleich mit einer Kontrollgruppe

Im Rahmen der Untersuchung von verbundenen/abhängigen Stichproben lassen sich dementsprechend Fragestellungen zu Unterschieden bzw. Veränderungen im betreffenden Merkmal (z.B. Depression) über den Verlauf der Testzeitpunkte untersuchen. Dieser Bereich der Mixed ANOVA

wird als *Innersubjekteffekt* bezeichnet, da (Veränderungs- bzw. Zeit-) Effekte „innerhalb“ der untersuchten Individuen erfasst werden können. Methodisch bildet dies eine einfaktorielle Varianzanalyse für verbundene Stichproben ab. Um die Wirksamkeit einer Therapie belegen und die erste Fragestellung zufriedenstellend beantworten zu können, bleibt hier zu erwähnen, dass im Rahmen des Innersubjekteffekts die Mittelwertsunterschiede zwischen den Testzeitpunkten selbstverständlich in der vermutenden Richtung bestehen und laut Testverfahren einen Symptomrückgang verzeichnen müssten (z.B. Werte nehmen ab).

Im Gegensatz dazu wird im Rahmen der Mixed ANOVA der Vergleich zwischen der Merkmalsausprägung (z.B. Depression) einer Interventionsgruppe (Klient*innen, welche die zu testende Therapieform erhalten) mit jenen einer Kontrollgruppe (z.B. Klient*innen auf der Warteliste für einen Therapieplatz) durch die unverbundene Stichprobe dargestellt, da die erhobenen merkmalspezifischen Ausprägungen der beiden Gruppen keinen verbindenden Aspekt aufweisen und nicht einander zugeordnet werden können (Aden, Bunina und Vavrik, 2021; Bortz, 2006). Da anzunehmen ist, dass Personen in der Interventionsgruppe im Gegensatz zur Kontrollgruppe eine Form der Therapie erhalten, durch welche sich die Symptome des Merkmals (in diesem Fall Depression) verändern sollten, müsste daher ein „Overall“ - Unterschied zwischen der Interventionsgruppe und der Kontrollgruppe im untersuchten Merkmal zu erwarten sein. Diese Art der Fragestellung trägt im Rahmen der Mixed ANOVA den Namen *Zwischensubjekteffekt* und spiegelt eine einfaktorielle Varianzanalyse für unverbundene Stichproben wider.

Es bleibt jedoch festzuhalten, dass sowohl Innersubjekt-, als auch Zwischensubjekteffekt sich nicht wechselseitig berücksichtigen, dh. jeweils nur Aussagen zu Unterschieden bzw. Veränderungen zwischen den Testzeitpunkten für *alle* Individuen (unabhängig von Interventions- & Kontrollgruppenzugehörigkeit) (*Innersubjekteffekt*), als auch zu allgemeinen Mittelwertsdifferenzen zwischen Interventions- und Kontrollgruppe (unabhängig von den Testzeitpunkten) (*Zwischensubjekteffekt*) getroffen werden können.

Um jedoch die Kombination beider Stichproben über die Testzeitpunkte berücksichtigen zu können, bietet die Mixed ANOVA die Möglichkeit eine Interaktion bzw. Wechselwirkung zu berechnen. Aufgrund der wiederholten Vorgabe des Messinstruments, mittels dessen die Ausprägung eines spezifischen Merkmals sowohl bei der Interventionsgruppe als auch bei der Kontrollgruppe erhoben wird, lassen sich über den Verlauf der Therapie hinweg intraindividuelle Schwankungen wie auch durch den hinzugezogenen Vergleich mehrerer unabhängiger Gruppen interindividuelle Differenzen in der Ausprägung eines Störungsbildes bei Klient*innen untersuchen (Seistock, Ruso und Aden, 2021). Besteht ein signifikanter Unterschied zwischen den intraindividuellen Veränderungen einer Merkmalsausprägung der Interventionsgruppe zu denen der Kontrollgruppe zu mindestens einem Messzeitpunkt, handelt es sich um eine statistisch relevante Wechselwirkung, deren Kennwerte in weiterer Folge für eine korrekte Interpretation noch genauer betrachtet werden müssen. Im Rahmen der Mixed ANOVA wird die *Wechselwirkung/Interaktion* neben den *Inner-* und *Zwischensubjekteffekt* als dritte Modellkomponente analysiert.

Zur Beurteilung des Inner- und Zwischensubjekteffekts sowie der Interaktion im Rahmen einer Mixed ANOVA soll festgehalten werden: Im Rahmen von RCTs ist etwa ein signifikantes Ergebnis im

Un-
verbundene
Stichprobe

Interaktion

Zwischensubjekteffekt bzw. in der Interaktion von höherer Priorität, um eine Wirkung einer therapeutischen Intervention ableiten zu können, als ein signifikanter Innersubjekteffekt, da letzterer allgemeine Veränderungen über die Zeit abbildet, unabhängig von der Gruppenzugehörigkeit und daher keine Information über die dezidierte Wirksamkeit einer Intervention vermittelt.

Hypothesenbildung im Rahmen der Mixed ANOVA

Um im Rahmen einer Mixed ANOVA zu überprüfen, ob zwischen k-unverbundenen Stichproben und k-verbundenen Stichproben statistische Abweichungen sowie eine Interaktion zwischen den unabhängigen und abhängigen Stichproben in Hinblick auf ein intervallskaliertes (metrisches) Merkmal bestehen, müssen in einem nächsten Schritt des Verfahrens die Hypothesenpaare aufgestellt werden. Bezugnehmend auf das bereits angeführte Beispiel werden (A) die Unterschiede zwischen verbundenen Stichproben (Depression erhoben mit dem BDI-II zu drei Messzeitpunkten: Prä, Post und Follow-up) und (B) zwischen unverbundenen (Kontroll- vs. Interventionsgruppe) sowie deren Interaktion (A x B) in dem Verfahren Mixed ANOVA untersucht. Um die drei zu überprüfenden Effekte zu testen, müssen demnach auch drei Hypothesenpaare aufgestellt werden. Die Formulierung des ersten Hypothesenpaars (A) erfolgt analog zu jener einer einfaktoriellen Varianzanalyse für abhängige Stichproben (Aden, Bunina & Vavrik, 2021). Im Zuge des zweiten Hypothesenpaars (B), dem Zwischensubjektfaktor, werden die Mittelwertsunterschiede zwischen den unverbundenen Gruppen abgebildet und die Formulierung der Hypothesen erfolgt nach den Regeln einer einfaktoriellen Varianzanalyse für unabhängige Stichproben (Seistock, Ruso & Aden, 2021). Das dritte Hypothesenpaar (A x B) untersucht die Interaktion zwischen dem Innersubjektfaktor und dem Zwischensubjektfaktor (Witte, 2019). Aufgrund inhaltlich-theoretischer Erwägungen kann es auch sinnvoll sein, die Unterschiede und Interaktionen von mehr als nur zwei Gruppenfaktoren in einem Modell zu überprüfen (z.B. Komorbiditäten vorhanden ja/nein). Hierbei gilt es pro Faktor, der untersucht wird, eine weitere Hypothese aufzustellen (C). Bei drei Gruppenfaktoren in einem Modell, müssten daher sieben Hypothesen formuliert und überprüft werden (A, B, C, A x B, A x C, B x C, A x B x C; vgl. Janczky, 2020). Mit der Anzahl der Hypothesen steigt jedoch auch die Gefahr des Alpha-Fehlers, auf welchen in späterer Folge nochmals genauer eingegangen wird. Die Hypothesen für das bereits angeführte Beispiel, in dem die Wirksamkeit einer Intervention durch den Vergleich der Ausprägung der depressiven Symptomatik zweier unverbundener Gruppen (Interventionsgruppe und Kontrollgruppe) erhoben zu drei Messzeitpunkten (Prä, Post, Follow-Up) evaluiert werden soll, lassen sich die Hypothesenpaare exemplarisch wie folgt formulieren:

Innersubjektfaktor (A)

H0a: Es besteht kein signifikanter Unterschied in der Ausprägung einer depressiven Symptomatik zwischen den drei Messzeitpunkten (T0, T1, T2).

H1a: Es besteht ein signifikanter Unterschied in der Ausprägung einer depressiven Symptomatik zwischen den drei Messzeitpunkten (T0, T1, T2).

Zwischensubjektfaktor (B)

Hypothesen-
formulierung

*H0b: Klient*innen unterscheiden sich nicht signifikant in Abhängigkeit der Therapiegruppe (Interventionsgruppe vs. Kontrollgruppe) hinsichtlich einer depressiven Symptomatik.*

*H1b: Klient*innen unterscheiden sich signifikant in Abhängigkeit der Therapiegruppe (Interventionsgruppe vs. Kontrollgruppe) hinsichtlich einer depressiven Symptomatik.*

Interaktion (AxB)

*H0c: Es besteht keine Interaktion zwischen der Ausprägung einer depressiven Symptomatik zu den drei Messzeitpunkten (T0, T1, T2) und den Therapiegruppen (Interventionsgruppe vs. Kontrollgruppe) der Klient*innen.*

*H1c: Es besteht eine Interaktion zwischen der Ausprägung einer depressiven Symptomatik zu den drei Messzeitpunkten (T0, T1, T2) und den Therapiegruppen (Interventionsgruppe vs. Kontrollgruppe) der Klient*innen.*

Allgemeine Informationen zu Verfahren für k- Stichproben („Over-All“ Verfahren)

Over-All Verfahren

Kongruent zu den einfaktoriellen Varianzanalysen, die Gegenstand des dritten und vierten Artikels der Serie Statistik (Aden, Bunina & Vavrik, 2021; Seistock, Ruso & Aden, 2021) waren, handelt es sich bei dem Verfahren Mixed ANOVA ebenso um ein sogenanntes „Over-all-Verfahren“. Pro aufgestellter Hypothese, die je einen Effekt, der im Zuge des Hauptverfahrens berechnet wird, abbildet, resultiert demnach ein Signifikanzniveau, das darüber Auskunft gibt, ob sich „over-all“ mindestens zwei Gruppen im Zwischensubjektfaktor oder zwei Testzeitpunkte im Innersubjektfaktor voneinander unterscheiden bzw. in einer Wechselwirkung miteinander stehen. Resultiert bei der Berechnung von einer der drei Hypothesen (Innersubjektfaktor, Zwischensubjektfaktor, Interaktion) ein signifikantes Ergebnis, lässt sich auf Basis dessen noch keine Schlussfolgerung darüber treffen, zwischen welchen und wie vielen konkreten Gruppen/Testzeitpunkten ein signifikanter Unterschied oder eine Interaktion begründet liegt (und in weiterer Folge, ob eine Therapie wirksam ist). Im Anschluss an ein signifikantes Ergebnis des Zwischensubjektfaktors oder Innersubjektfaktors müssen daher sogenannte Post-Hoc-Testungen vorgenommen werden, um zu identifizieren, welche Gruppen oder Testzeitpunkte sich konkret voneinander unterscheiden. Zusätzlich dient die Betrachtung deskriptiver Kennwerte zur Feststellung, bei welchem Testzeitpunkt welche Gruppe eine höhere oder niedrigere Ausprägung des spezifischen Merkmals aufweist. Die Betrachtung der deskriptiven Statistik wie auch eines Wechselwirkungsdiagramms eignet sich im Anschluss an die Post-Hoc-Testungen zur inhaltlichen Interpretation der drei berechneten Effekte (Innersubjekteffekt, Zwischensubjekteffekt und Interaktion; Bortz, 2006; Witte, 2019). (sh. Abb.3)

Post-hoc Testung

Signifikanz- bestimmung

Die Signifikanzbestimmung und Bewertung erfolgt pro Faktor (Innersubjekteffekt, Zwischensubjekteffekt und Interaktion) mittels einer Prüfverteilung, analog zu der von klassischen einfaktoriellen Varianzanalysen.

Statistische Signifikanz/Relevanz

Statistische
vs.
inhaltliche
Relevanz

Neben der Bewertung der Signifikanz spielt bei der Evaluierung von therapeutischen Interventionen vor allem aus klinischer Perspektive die deskriptive Betrachtung der Mittelwertsunterschiede eine erhebliche Rolle, um die Relevanz der vorgefundenen Ergebnisse bewerten zu können. Wäre eine Intervention beispielsweise hoch wirksam, müsste sich dies nicht nur in der Signifikanz des Innersubjekteffekts, des Zwischensubjekteffekts und der Interaktion zeigen, sondern vor allem auch in der deskriptiven Veränderung der Mittelwertsunterschiede. Beispielsweise hätte die Intervention therapeutischen Erfolg gehabt, wenn sich die Mittelwerte der Interventionsgruppe derart verändern würden, dass sie von einer schweren Depression ($T_{\text{Intervention}} = 43$) auf eine milde Depression ($T_{\text{Intervention}} = 18$) sinken, während die Werte der Kontrollgruppe auf dem Niveau einer schweren Depression ($T_{\text{Kontrolle}} = 35$) blieben (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). Da auch geringere Mittelwertsunterschiede, welche wie im Beispiel hier z.B. den Bereich einer schweren Depression nicht verlassen, signifikant werden können, ist es daher insbesondere für die Bewertung der klinischen Relevanz ratsam, immer die deskriptiven Statistiken in die Interpretation zu integrieren, bevor ein abschließendes Resümee über die Wirksamkeit einer untersuchten Intervention getroffen wird.

Die notwendigen empirischen Voraussetzungen für die korrekte Wahl und Durchführung zur Überprüfung der aufgestellten Hypothesen, die korrekte Interpretation der Kennwerte sowie wichtige mathematische Grundlagen im Rahmen der Berechnung einer Mixed ANOVA, werden in den nächsten Abschnitten noch detaillierter für die einzelnen Effekte erklärt und anhand eines Beispiels exemplarisch dargestellt.

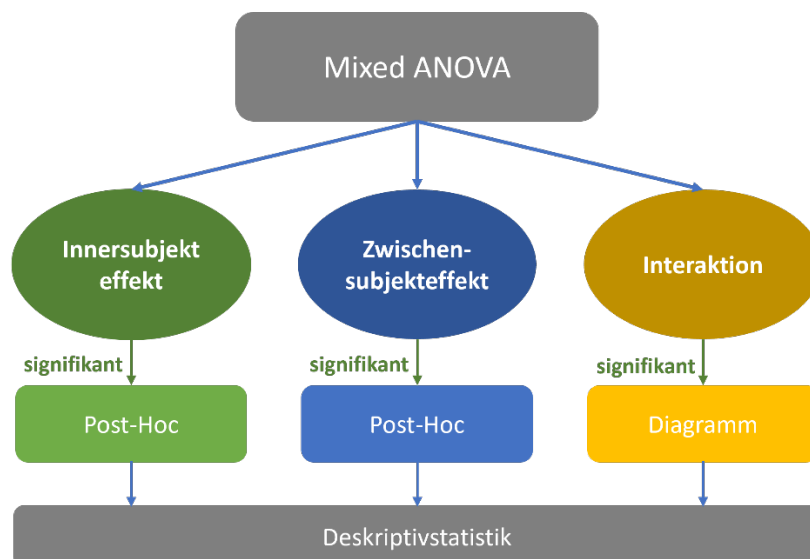


Abb.3 Übersicht Mixed ANOVA: Hauptverfahren und post-hoc Tests (modifiziert nach Aden, Bunina & Vavrik, 2021)

Innersubjekteffekt

Die Berechnung des Innersubjekteffekts erfolgt analog zur Berechnung einer einfaktoriellen Varianzanalyse für abhängige Stichproben, die im vierten Beitrag der Serie Statistik im Forschungsbulletin (sh. Seistock, Ruso & Aden, 2021) beschrieben wurde und findet daher in diesem Artikel keine weitere Explikation. Erinnerung sei jedoch daran, dass die Anwendung varianzanalytischer Verfahren an das Vorliegen bestimmter Voraussetzungen gebunden ist. Zunächst muss die abhängige Variable, hinsichtlich derer die k-verbundenen Stichproben miteinander verglichen werden, metrisch, dh. mindestens intervallskaliert, sein (z.B. Depression gemessen mit BDI-II). Neben dem Skalenniveau stellen die Normalverteilung des untersuchten Merkmals in allen Gruppen/zu allen Messzeitpunkten und die Sphärizitätsbedingung (Zirkularitätsannahme) wesentliche Voraussetzungen dar. Zur Prüfung der Normalverteilung können die Anpassungstests (Goodness-of-Fit-Tests) nach Kolmogorov-Smirnov und Shapiro Wilk als Verfahren herangezogen werden. Die Bedingung der Sphärizität lässt sich mit dem Mauchly-Test überprüfen und gilt als erfüllt, wenn ein nicht signifikantes Ergebnis ($p > .05$) vorliegt und demnach die Varianzen der Merkmalsstreuung zu allen Messzeitpunkten/in allen Gruppen und die Korrelationen zwischen den Merkmalsausprägungen der Gruppen/Messzeitpunkten homogen sind. Resultiert ein signifikantes Ergebnis, so muss eine Korrektur nach Greenhouse-Geisser (wenn $\epsilon \leq .75$) oder Huynh-Feldt (wenn $\epsilon > .75$) im Anschluss durchgeführt werden (Bortz & Schuster, 2010; Rasch, Frieze, Hofmann & Naumann, 2014). Detaillierte Erklärungen zur Bestimmung der Skalenniveaus, Prüfung der Normalverteilung und Signifikanzbewertung eines Ergebnisses finden sich im ersten Artikel der Serie Statistik im Forschungsbulletin (Seistock, Bunina & Aden, 2020). Können die Voraussetzungen als gegeben angesehen werden, so kann die Berechnung des Innersubjekteffekts durchgeführt und somit die statistische Relevanz von beobachteten Unterschieden (p-Wert) „over-all“ und zwischen je zwei Gruppen/Messzeitpunkten, entlang der Post-Hoc-Testungen, bestimmt werden. Unter Berücksichtigung von Stichprobeneffekten sollten zur anschließenden inhaltlichen Diskussion der Ergebnisrelevanz die deskriptiven Kennwerte (Mittelwerte und Standardabweichungen der einzelnen Gruppen/Messzeitpunkten) und die Effektstärke interpretiert werden. Im Rahmen einfaktorieller varianzanalytischer Verfahren wird zumeist das partielle Eta Quadrat (η_p^2) als Effektstärke berechnet. Die Bewertungsstufen nach Cohen (1988) bieten einen Orientierungsrahmen zur Interpretation der Größe des eigentlichen Effekts und ermöglichen studienübergreifende Vergleiche (Rasch, Frieze, Hofmann & Naumann, 2014).

Report- konventionen

Innersubjekt- effekt

In Form eines statistischen Reports im Rahmen des Hauptverfahrens einer abhängigen einfaktoriellen Varianzanalyse (Innersubjekteffekt) empfiehlt es sich folgende Angaben anzuführen, welche anhand eines Beispiels hier illustriert sind:

$$(F(2,198) = 8.63, p < .001, \eta_p^2 = 0.87)$$

Die Resultate der Post-Hoc-Testungen, die im Anschluss an einen signifikanten Innersubjekteffekt durchgeführt werden, lassen sich z.B. wie folgt reportieren:

Im Rahmen der paarweisen Vergleiche der Messzeitpunkte zeigt sich, dass sich Messzeitpunkt Prä ($M_{Prä} = 29$ ($SD = 3.00$)) signifikant von Messzeitpunkt Post ($M_{Post} = 35$ ($SD = 7.50$)) sowie von Messzeitpunkt Follow-Up ($M_{FollowUp} = 24$ ($SD = 8.25$)) unterscheidet (jeweils $p < .001$). Weiters konnte ein signifikanter Unterschied zwischen Messzeitpunkt Post und Messzeitpunkt Follow-Up festgestellt werden ($p < .001$).

Zwischensubjekteffekt

Bei der Berechnung des Zwischensubjekteffekts handelt es sich um eine einfaktorielle Varianzanalyse für unabhängige Stichproben, welche bereits im Zuge des dritten Beitrags der Serie Statistik im Forschungsbulletin abgehandelt wurde (sh. Aden, Bunina & Vavrik, 2021) und daher im vorliegenden Beitrag ebenfalls keine erneute Explikation erfahren werden.

Auf die Voraussetzungen sei jedoch auch an dieser Stelle kurz verwiesen: Analog zur unabhängigen Varianzanalyse muss das Skalenniveau des zu untersuchenden Merkmals metrisch, mindestens intervallskaliert, sein (z.B. Depression gemessen mit BDI-II). Darüber hinaus sollte die Normalverteilung der Merkmalsausprägung in den einzelnen Stichproben (k-unabhängigen Gruppen) mit den Anpassungstests (Goodness-of-Fit-Tests nach Kolmogorov-Smirnov und Shapiro Wilk) und die Varianzhomogenität mittels des Levene-Tests (erfüllt, wenn $p > .05$) überprüft werden. Gelten die Voraussetzungen als gegeben, kann die Signifikanzbestimmung des Zwischensubjekteffekts („over-all“) folgen. Resultiert ein signifikantes Ergebnis im Hauptverfahren der unabhängigen Varianzanalyse, so empfiehlt es sich im Anschluss die Post-Hoc-Testungen, für eine differenzierte Betrachtung der paarweisen Mittelwertsunterschiede, durchzuführen. Wird wie im hier beschriebenen Beispiel der Unterschied zwischen zwei Gruppen (Intervention- und Kontrollgruppe) im Rahmen des Zwischensubjekteffekts untersucht, so kann bei einem signifikanten „over-all“ Ergebnis von einer nachfolgenden Post-hoc-Testung abgesehen werden, da in diesem Fall offensichtlich wäre, zwischen welchen Gruppen ein beobachteter Unterschied vorliegt. Die Berechnung der statistischen Relevanz von beobachteten Unterschieden (p-Wert) sollte anschließend durch die inhaltliche Interpretation der deskriptiven Kennwerte (Mittelwerte und Standardabweichung der einzelnen Gruppen) und der Bestimmung der Effektstärke, partielles Eta Quadrat (η_p^2), unter Berücksichtigung von Stichprobeneffekten, ergänzt werden, um Aussagen über die Ergebnisrelevanz treffen zu können.

In der Reportlegung des Over-All-Verfahrens der unabhängigen einfaktoriellen Varianzanalyse (Zwischensubjekteffekt) empfiehlt es sich vorliegende Kennwerte wie folgt anzugeben:

$$(F(1,98) = 6.16, p < .001, \eta_p^2 = 0.71)$$

Für die im Anschluss an ein signifikantes Ergebnis durchgeführten Post-Hoc-Testungen (hier nur exemplarisch dargestellt), können die Kennwerte beispielsweise folgendermaßen angeführt werden:

Im Rahmen der paarweisen Vergleiche zeigt sich eine signifikante Differenz ($p < .001$) zwischen der Kontrollgruppe ($M_{Kontroll.} = 30$ ($SD = 4.66$)) und der Interventionsgruppe ($M_{Intervention.} = 22$ ($SD = 5.84$)).

Interaktion

Im folgenden Abschnitt werden die wesentlichen empirischen Voraussetzungen im Rahmen des Interaktionseffekts sowie die Vorgehensweise zur Bestimmung der statischen Relevanz der Wechselwirkung und deren inhaltlichen Interpretation erläutert. Um zu untersuchen, ob zwischen Innersubjektfaktor und Zwischensubjektfaktor eine signifikante Interaktion besteht, müssen zunächst die Voraussetzungen der abhängigen und unabhängigen einfaktoriellen Varianzanalysen überprüft werden. Die Berechnung des Interaktionseffekts setzt neben dem metrischen, mindestens

intervallskalierten Skalenniveau der untersuchten abhängigen Variable, die Normalverteilung der Merkmalsausprägung in allen k-unabhängigen und k-abhängigen Gruppen sowie die Varianzhomogenität und Sphärizitätsbedingung voraus. Lassen sich die Voraussetzungen als erfüllt beurteilen, kann das Signifikanzniveau des Hauptverfahrens der Interaktion bestimmt werden. Gilt ausschließlich die Sphärizitätsbedingung als nicht erfüllt (Mauchly-Test signifikant), so muss auch im Rahmen des Interaktionseffekts eine Korrektur nach Greenhouse-Geiser (wenn $Epsilon \leq .75$) oder Huynh-Feldt (wenn $Epsilon > .75$) vorgenommen werden. Bei einem signifikanten Interaktionseffekt empfiehlt es sich in weiterer Folge die statistische Relevanz durch die Berechnung der Effektstärke und die inhaltliche Interpretation des Interaktionsdiagramms inklusive der Beurteilung deskriptiver Kennwerte zu ergänzen.

Die Wechselwirkung wird zumeist in einem Liniendiagramm grafisch dargestellt, in welchem die y-Achse die abhängige metrische Variable und die x-Achse sowie die separaten Linien je einen Gruppenfaktor abbilden. Bezugnehmend auf das bereits angeführte Beispiel empfiehlt es sich bei Wirksamkeitsforschungen die intraindividuellen Verläufe der Merkmalausprägung (z.B. Depression) zu den verschiedenen Messzeitpunkten (z.B. Prä, Post und Follow-Up) entlang der x-Achse und die interindividuellen Unterschiede zwischen den unabhängigen Gruppen (z.B. Kontrollgruppe und Interventionsgruppe) mittels separater Linien darzustellen. Durch den direkten Vergleich der Merkmalsausprägung der Kontrollgruppe und der Interventionsgruppe zu den einzelnen Messzeitpunkten, zeigen sich in der Grafik die Differenzen und die Veränderungen, die sich schließlich auf die Intervention (Treatment) zurückführen lassen (sh. Abb. 4). Auf Basis der deskriptiven Kennwerte (Mittelwerte und Standardabweichungen z.B. der Kontroll- und Interventionsgruppe zu allen drei Messzeitpunkten) können die im Diagramm dargestellten Unterschiede und Verläufe in der Merkmalsausprägung inhaltlich hinsichtlich Größe und Richtung beschrieben werden, wie beispielsweise welche Gruppen zu welchen Messzeitpunkten eine höhere oder niedrigere Merkmalsausprägung aufweisen und zu welchen Messzeitpunkten sich die Merkmalsausprägungen der unabhängigen Gruppen verändert (Bortz, 2006; Janczky, 2020).

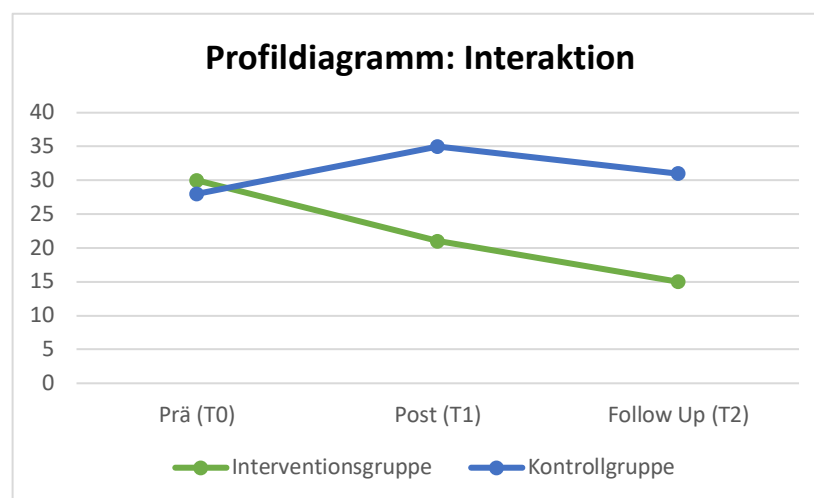


Abb. 4 beispielhafte Darstellung des Profildigramms einer Mixed ANOVA im Rahmen eines Prä-Post-FollowUp Designs für die erfolgreiche Behandlung einer Major Depression im Vergleich mit einer Kontrollgruppe

Durch die Prüfung der statistischen Relevanz des Interaktionseffekts und die anschließende inhaltliche Interpretation der Wechselwirkung, lässt sich die Wirksamkeit einer Intervention differenziert beschreiben. Darüber hinaus empfiehlt es sich, unter Berücksichtigung von Stichprobeneffekten, das partielle Eta Quadrat (η_p^2) als Effektstärke zu berechnen. Um die Effektstärke der signifikanten Interaktion vergleichbaren Wirksamkeitsforschungen gegenüberzustellen (wie beispielsweise in Metastudien) und somit auch die Legitimation der Intervention begründen zu können, sollte zudem die Größe des Effekts bestimmt werden. Die Bewertungsstufen nach Cohen (1988) dienen als Orientierungsrahmen zur Interpretation der Effektgröße (Bortz, 2006; Rasch, Frieze, Hofmann & Naumann, 2014).

Im Rahmen einer Mixed ANOVA sollten folgende Kennwerte des Interaktionseffekts, ergänzend zu denen des Innersubjekteffekts und Zwischensubjekteffekts, in der Reportlegung angeführt werden:

$$(F(2,198)= 7.34, p< .001, \eta_p^2= 0.80)$$

Im Anschluss an eine signifikante Interaktion kann die inhaltliche Interpretation des Profildigramms beispielweise wie folgt formuliert werden:

Zu Messzeitpunkt Prä zeigt sich eine marginale Differenz zwischen der Interventionsgruppe ($M_{Prä}= 30(SD= 2.25)$) und der Kontrollgruppe ($M_{Prä}= 28(SD= 2.75)$), die zu den Messzeitpunkten Post und Follow-Up an Größe zunimmt dahingehend, dass Personen der Interventionsgruppe ($M_{Post}= 21(SD= 2.55)$, $M_{FollowUp}= 15(SD= 2.30)$) eine deutlich niedrigere Merkmalsausprägung als Personen der Kontrollgruppe ($M_{Post}= 35(SD= 3.88)$; $M_{FollowUp}= 31(SD= 4.12)$) aufweisen. Während die Merkmalsausprägung der Interventionsgruppe kontinuierlich sinkt, erreicht die Merkmalsausprägung der Personen in der Kontrollgruppe zu Messzeitpunkt Post den Höhepunkt, bevor die Personen in der Kontrollgruppe zu Messzeitpunkt Follow-Up eine niedrigere Merkmalsausprägung als zu Zeitpunkt Post berichten (siehe Tab.2).

Tab. 2 Deskriptive Kennwerte: Mittelwerte und Standardabweichungen zu allen drei Messzeitpunkten

		Mittelwerte (M)	Standardabweichungen (SD)
<i>Messzeitpunkt 1 (Prä = T0)</i>	Interventionsgruppe	30	2.25
	Kontrollgruppe	28	2.75
	Gesamt	29	3.00
<i>Messzeitpunkt 2 (Post = T1)</i>	Interventionsgruppe	21	2.55
	Kontrollgruppe	35	3.88
	Gesamt	28	7.50
<i>Messzeitpunkt 3 (Follow-Up = T2)</i>	Interventionsgruppe	15	2.30
	Kontrollgruppe	31	4.12
	Gesamt	23	8.25

EXKURS

Weitere Verfahren mit Interaktion – mehrfaktorielle Versuchspläne

Neben der Mixed ANOVA besteht ein weiteres varianzanalytisches Verfahren, in welchem Interaktionen bzw. Wechselwirkungen zwischen Gruppenvariablen hinsichtlich eines metrischen Merkmals verglichen werden können. Der Unterschied zur Mixed ANOVA besteht darin, dass es sich jedoch bei der sogenannten mehrfaktoriellen Varianzanalyse um mindestens 2 unverbundene Stichproben handelt, welche hinsichtlich eines metrischen Merkmals verglichen werden und nicht wie bei der Mixed ANOVA eine verbundene und eine unverbundene Stichprobe. Daraus ergibt sich ein anderer Einsatzbereich und nicht jener der Wirksamkeitsevaluation, weshalb nachfolgend nur die wichtigsten Aspekte der mehrfaktoriellen Varianzanalyse genannt werden.

Einsatzbereich

Die mehrfaktorielle ANOVA ermöglicht den zusätzlichen Einbezug eines weiteren unabhängigen Gruppenfaktors, was eine sinnvolle Erweiterung einer einfaktoriellen Varianzanalyse darstellen kann, da der Informationsgehalt und damit die Varianzaufklärung durch die Integration eines weiteren Gruppenfaktors zunimmt und dahingehend das Modell genauer wird. In diesem Punkt können Parallelen zur Mixed ANOVA aufgezeigt werden, da es in beiden Fällen auf einer Metaebene das Ziel ist, Unterschiede in einem psychologischen Merkmal hinreichend erklären zu können. Obgleich die Unterschiede bei einer mehrfaktoriellen Varianzanalyse nicht zugunsten einer Therapiewirksamkeit ausgelegt werden, so dient die Integration eines weiteren Gruppenfaktors ebenfalls dazu, den Einfluss möglicher beeinflussender Variablen sichtbar zu machen und beispielsweise Wechselwirkungen aufzudecken, die andernfalls latent bleiben würden.

Hypothesenerstellung

Ähnlich wie bei der Mixed ANOVA müssen bei der mehrfaktoriellen ANOVA mehrere Hypothesenpaare gebildet werden. Im Rahmen der Hypothesenbildung müssten nun für jeden Gruppenfaktor (= Haupteffekt) ein separates Hypothesenpaar aufgestellt werden, sowie ein Interaktions-Hypothesenpaar (AxB).

Voraussetzungen und Verfahren

Da die mehrfaktorielle ANOVA aus mindestens zwei einfaktoriellen Varianzanalysen für unverbundene Stichproben sowie einer Interaktion besteht, ergeben sich dieselben Voraussetzungen, welche für eine klassische einfaktorielle Varianzanalyse gelten und bei Aden, Bunina und Vavrik (2021) in einem bisherigen Methodenartikel nachzulesen sind.

Zeigt sich ein Haupteffekt als signifikant, so sind im Anschluss Post-hoc Testungen sowie deskriptive Statistiken zu interpretieren, um die Art und Richtung des Unterschiedes feststellen zu können. Dieses Vorgehen gilt selbstverständlich für alle berechneten signifikanten Haupteffekte. Genauere Informationen zur Durchführung und Interpretation von einfaktoriellen Varianzanalysen sind in einem weiteren Artikel der Serie Statistik zu finden (Aden, Bunina & Vavrik, 2021). Liegt eine signifikante

Interaktion vor, so ist analog zur Mixed ANOVA das Interaktionsdiagramm zu betrachten, um die Wechselwirkung zwischen den unverbundenen Stichproben adäquat interpretieren zu können.

Der Report gestaltet sich analog zu jenem der Mixed ANOVA, mit dem einzigen Unterschied, dass für die Haupteffekte die Richtlinien für den Zwischensubjekteffekt gelten.

Wie bei allen varianzanalytischen Verfahren ist es empfehlenswert neben Signifikanzen und Teststatistiken ebenso die Effektstärken, -sofern sinnvoll, zu berechnen und in die Interpretation miteinzubeziehen.

Konklusion

Die Mixed ANOVA ist ein statistisches Verfahren, das im Zusammenhang unterschiedlicher Fragestellungen Anwendung findet bzw. das Auswertungsverfahren der Wahl darstellt. Ein prominentes Einsatzfeld der Mixed ANOVA stellt im psychotherapiewissenschaftlichen Kontext die Wirksamkeitsforschung dar. Exemplarisch für Untersuchungsdesigns, die der Zielsetzung eines Effektivitätsnachweises folgen, sind die im vorliegenden Beitrag angesprochenen RCTs.

Die Mixed ANOVA eröffnet die Möglichkeit, abhängige wie unabhängige Faktoren in einem Auswertungsmodell zu berücksichtigen. So wird es möglich Verlaufsmessungen auch zwischen unterschiedlichen Gruppen (z.B. Therapie- vs. Kontrollgruppe) zu vergleichen, was ein wesentliches Element randomisiert kontrollierter Studien darstellt. Entscheidend für Fragestellungen dieser Art ist dabei die Interpretation der Interaktion bzw. Wechselwirkung.

Wie bei fast allen statistischen Auswertungen gilt die Devise „statistische Signifikanz ist nicht gleich inhaltliche/klinische Relevanz“. Speziell im Bereich psychotherapiewissenschaftlicher Wirksamkeitsforschung ist es essenziell, neben der statistischen Bedeutsamkeit im Sinne einer Null-Hypothesen-Testung anhand des p-Wertes v.a. auch die inhaltliche Relevanz z.B. einer Symptomreduktion zu beurteilen. In diesem Zusammenhang sei neben Effektstärken ebenso auf das Über- bzw. Unterschreiten gewisser Cut-Off-Werte eingesetzter Messverfahren (z.B. BDI-II) zu verweisen.

Darüber hinaus ist eine a priori Stichprobenkalkulation geboten, um ein adäquates Verhältnis von Testmacht und Fehlerwahrscheinlichkeit sicher zu stellen sowie angenommene Effekte auch tatsächlich als signifikant nachweisen zu können.

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.
- 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 Verlag.
- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7., vollständig überarbeitete und erweiterte Auflage). Berlin, Heidelberg: Springer.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. New York, NY: Academic Press.
- Djulfegovic, B., & Guyatt, G. H. (2017). Progress in evidence-based medicine: a quarter century on. *The lancet*, 390(10092), 415–423.
- 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.
- Grossman, J. & Mackenzie, J. (2005). The Randomized controlled trial. Gold standard, or merely standard?. *Perspectives in Biology and Medicine*, volume 48, number 4 (autumn 2005), 516–34.
- Jadad, A. R., & Enkin, M. W. (2007). *Randomized controlled trials: Questions, answers and musings*. John Wiley & Sons, Incorporated.
- Janczyk, M., & Pfister, R. (2020). *Inferenzstatistik verstehen*. Heidelberg: Springer.
- Kuhner, C., Burger, C., Keller, F., & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II). Befunde aus deutschsprachigen Stichproben. *Der Nervenarzt*, 78, 651–656.
- Philips, B. & Falkenström, F. (2021). What Research Evidence Is Valid for Psychotherapy Research? *Frontiers in Psychiatry*, 11:625380. doi: 10.3389/fpsy.2020.625380.
- Pocock, S. J. & Elbourne, D. R. (2000). Randomized trials or observational tribulations? *N Engl J Med* 342:1907–09.
- Rasch, B., Fries, M., Hofmann, W., & Naumann, E. (2014). Varianzanalyse mit Messwiederholung. *Quantitative Methoden Band 2: Einführung in die Statistik* (S. 99–141). Heidelberg: Springer.
- Sanson-Fisher, R. W., Bonevski, B., Green, L. W., & D’Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *American journal of preventive medicine*, 33(2), 155–161.
- 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.
- Ware, J.H. & Hamel, M.B. (2011). Pragmatic trials—guides to better patient care? *N Engl J Med.*, 364:1685–7. doi: 10.1056/NEJMp1103502.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., & Mullen, D. (2008). Alternatives to the Randomized Controlled Trial. *American Journal of Public Health*, 98: 1359-1366.

Witte, K. (2019). *Angewandte Statistik in der Bewegungswissenschaft*. Heidelberg: Springer.

*Angaben zu den Autor*innen*

Institut für Statistik

Sophie Gattermeyer, Iris Enengl, Jan Aden

Adresse: Freudplatz 1, 1020 Wien, Raum 6011

E-Mail: jan.aden@sfu.ac.at, sophie.gattermeyer@sfu.ac.at