SFU Forschungsbulletin SFU Research Bulletin 9/1 (2021), S. 68-86 DOI: 10.15135/2021.9.1.68-86

Die einfaktorielle Varianzanalyse für unabhängige Stichproben und der Kruskal-Wallis-Test im psychotherapiewissenschaftlichen Kontext

-Empfehlungen für Anwendung und Interpretation-

The one-way analysis of variance for independent samples and the Kruskal-Wallis test in psychotherapy science

-Recommendations for application and interpretation-

Aden, J., Bunina, A. & Vavrik, C.

Kurzzusammenfassung

In diesem dritten Beitrag der Serie Statistik in der Psychotherapiewissenschaft wird die Anwendung von Verfahren zum Vergleich von k- unabhängigen Stichproben (einfaktorielle unabhängige ANOVA und der Kruskal-Wallis-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 Varianzanalyse, Kruskal-Wallis-Test, Reportkonventionen, Effektstärken, post-hoc Test

© SFU Forschungsbulletin SFU Research Bulletin 2021 (1)

Die einfaktorielle Varianzanalyse für unabhängige Stichproben und der Kruskal-Wallis-Test im psychotherapiewissenschaftlichen Kontext
-Empfehlungen für Anwendung und Interpretation-

Abstract

In this third contribution in the series Statistics in Psychotherapy Science, the application use of methods for comparing k-independent samples (one-way independent ANOVA and the Kruskal-Wallis 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 ANOVA, Kruskal-Wallis test, report conventions, effect-sizes, post-hoc tests

Einsatzfeld und Background

Einsatzbereiche und Fragestellungen

Die empirische Untersuchung psychotherapeutischer Interventionsmethoden sowie der sie interessierenden psychologischen Merkmale bildet die Grundlage für eine evidenzbasierte Psychotherapie und ermöglichen die stetige Weiterentwicklung sowie die optimierte Anpassung von Therapieangeboten an verschiedene Klient*innengruppen. In diesem Kontext stellen Verfahren zur Überprüfung statistisch relevanter Unterschiede die am häufigsten angewandten Auswertungsformen in der quantitativen Psychotherapieforschung dar. Diese Verfahrensgruppe ermöglicht die Evaluation von Differenzen zwischen verschiedenen Gruppen/Stichproben/Messzeitpunkten und dadurch den Nachweis sowie die Steigerung der Effektivität psychotherapeutischer Behandlungsmethoden. In diesem Sinne dient die Anwendung von Unterschiedstestungen und deren Generalisierung nicht zuletzt der allgemeinen Qualitätssicherung sowie einer Förderung der Akzeptanz psychotherapeutischer Interventionsmaßnahmen.

In den ersten beiden Artikeln der Serie Statistik des Forschungsbulletins wurde die optimale Verfahrenswahl sowie eine korrekte Anwendung der Verfahren für die Unterschiedstestung bei zwei unverbundenen (Seistock, Bunina & Aden, 2020) bzw. zwei verbundene (Gattermeyer, Vladarski & Aden, 2020) Gruppen vorgestellt. Häufig ist es im Kontext psychotherapiewissenschaftlicher Fragestellungen jedoch notwendig oder inhaltlich ergiebiger, mehr als zwei Gruppen (k-Gruppen/ -Stichproben/ -Messzeitpunkte) bezüglich möglicher Unterschiede in einem interessierten Merkmal zu vergleichen. Dies kann beispielsweise der Fall sein, wenn untersucht werden soll, ob sich über den Therapieverlauf hinweg intraindividuelle Schwankungen in der Ausprägung eines bestimmten Störungsbildes zeigen (z.B. Depression, Angststörung). Ebenso könnte von Interesse sein, ob sich Personen verschiedener Therapieformen (z.B. Therapieform A, B & C) in Hinblick auf die wahrgenommene Qualität der Therapeut*innen-Klient*innen-Beziehung unterscheiden. Darüber hinaus ist die Messung ausgewählter Merkmale, wie beispielsweise der Psychotherapiemotivation, in Abhängigkeit verschiedener Alters- oder Diagnosegruppen im Rahmen der Bedarfsplanung, der Interventionsforschung sowie ebenfalls in der psychotherapeutischen Praxis maßgeblich für die Bereitstellung eines optimalen Behandlungsangebotes und nicht zuletzt für die Vorhersage des Behandlungserfolges (vgl. Schulz, Lang, Nübling & Koch, 2003; Nübling, Schulz, Schulz, Koch & Wittmann, 2005). Der Artikel in dieser Ausgabe des Bulletins befasst sich mit den statistischen Verfahren zur Überprüfung eben solcher Studiendesigns. Speziell werden in vorliegender Ausgabe statistische Testverfahren zur Überprüfung von Unterschieden zwischen k-unabhängigen Stichproben vorgestellt, welche als Erweiterung der Unterschiedstestungen für zwei unverbundene Gruppen verstanden werden können.

Unterschiedstestungen für k-Stichproben kommen in der psychotherapeutischen Forschungspraxis häufig in Hinblick auf die Wirksamkeit und Effektivität verschiedener Therapiemethoden zum Einsatz. Diesbezüglich liegt der Fokus auf dem Vergleich von mehr als zwei unterschiedlichen Stichproben/Gruppen/Messzeitpunkten. Ebenso wie bei den Verfahren zur Unterschiedstestung von zwei Stichproben (siehe Seistock, Bunina & Aden, 2020; Gattermeyer, Vladarski & Aden, 2020) wird auch bei der statistischen Untersuchung von k-Stichproben zwischen abhängigen und unabhängigen

Bestimmung der Stichproben-art (verbunden vs. unverbunden)

Studiendesigns differenziert. Verfahren zur Überprüfung von Differenzen zwischen k-verbundenen Stichproben fokussieren häufig die Evaluation der Wirksamkeit von Therapiemethoden im Sinne einer Veränderung bestimmter Merkmalsausprägungen (z.B. Angst). In diesem Kontext wird den Klient*innen in der Regel ein störungsspezifisches Messinstrument zu mehr als zwei Zeitpunkten, beispielsweise vor Beginn einer Therapie, direkt nach Beendigung der Therapie und drei Monate nach Abschluss der Therapie, vorgegeben (z.B. das Beck Angst-Inventar (Beck & Steer, 1990)) und überprüft, ob sich über die verschiedenen Zeitpunkte hinweg signifikante Schwankungen in der Symptomausprägung beobachten lassen. In solch einem abhängigen Studiendesign wird somit dieselbe Personengruppe zu mehr als zwei Zeitpunkten (zu Beginn, direkt nach Beendigung und drei Monate nach Abschluss einer Therapie) getestet. Dadurch lassen sich die erhobenen Messwerte einander paarweise zuordnen und intraindividuelle Unterschiede in der Merkmalsausprägung evaluieren (siehe Tabelle 1).

Tabelle 1: Bsp. für k-verbundene Stichproben

	Messzeitpunkt 1	Messzeitpunkt 2	Messzeitpunkt 3
Klient*in A	65	45	50
Klient*in B	70	53	57
Klient*in C	62	53	51
...	...	...	...

Im Gegensatz dazu kommen Unterschiedstestungen für k-unverbundene Stichproben -wie sie in diesem Artikel dargelegt werden- überall dort zum Einsatz, wo zwar ein Unterschied zwischen mehr als zwei Gruppen überprüft werden soll, die jeweiligen Gruppen jedoch nur einmalig zu einem bestimmten Zeitpunkt getestet werden. Beispielsweise kann untersucht werden, ob sich signifikante Unterschiede in der Psychotherapiemotivation in Abhängigkeit verschiedener Diagnosegruppen konstatieren lassen. Hierfür könnte zu Beginn der Therapie ein Fragebogen zur Evaluation der Psychotherapiemotivation (z.B. FPTM (Schulz, Nübling & Rüddel, 1995; Nübling et al., 2005)) sowohl Personen mit Diagnose A, als auch Personen, die eine Diagnose B erhalten haben, sowie darüber hinaus ebenfalls Personen mit einer Diagnose C vorgegeben werden. Im Vergleich zum erläuterten Beispiel bei k-verbundenen Stichproben (Messung von Angst mittels BAI-Inventar: vor, direkt nach Beendigung und drei Monate nach Abschluss der Therapie) werden hier die Motivationseinschätzungen von drei verschiedenen Personengruppen (Klient*innen der Diagnosegruppe A vs. Klient*innen der Diagnosegruppe B vs. Klient*innen der Diagnosegruppe C) zu nur einem Zeitpunkt erhoben. Die Messwerte können einander somit nicht paarweise zugeordnet und verglichen werden, vielmehr liegt der Fokus auf dem Vergleich der interindividuellen Wertausprägungen der Personen aus den Diagnosegruppen A, B und C (siehe Tabelle 2).

Tabelle 2: Bsp. für k-unverbundene Stichproben

Diagnosegruppe A		Diagnosegruppe B		Diagnosegruppe C	
Klient*in A	70	Klient*in D	60	Klient*in G	25
Klient*in B	50	Klient*in E	22	Klient*in H	75
Klient*in C	64	Klient*in F	45	Klient*in I	36
...	...	...	...	...	...

Um zu überprüfen, ob sich zwischen k-unverbundenen Stichproben statistisch signifikante Unterschiede in Hinblick auf ein mindestens rangskaliertes (ordinales) Merkmal konstatieren lassen, wird in der Regel, in Abhängigkeit bestimmter Voraussetzungen, eines der folgenden Auswertungsverfahren herangezogen: die einfaktorielle unabhängige Varianzanalyse (engl. *analysis of variance, ANOVA)* oder der Kruskal-Wallis-Test (*H-Test*). Hierbei handelt es sich um sogenannte inferenzstatistische Verfahren, welche einen Schluss der in den Stichproben gefundenen Effekte auf die Gesamtpopulation zulassen sollen. Die angeführten Tests setzen somit die Formulierung eines Hypothesenpaares (Null- und Alternativhypothese) voraus, welche für die Population verallgemeinert werden. Da sich die einfaktorielle Varianzanalyse bei ihrer Berechnung auf den Mittelwertvergleich der k-unverbundenen Gruppen stützt, wird die Hypothese überprüft, ob sich die Mittelwerte von mindestens zwei Gruppen statistisch signifikant unterscheiden. Das bereits erläuterte Beispiel für k-unverbundene Gruppen hinsichtlich der Untersuchung von Unterschieden in der Psychotherapiemotivation würde diesbezüglich u.a. folgende Hypothesenformulierung zulassen:

Hypothesen-Formulierung

H0: Klient*innen unterscheiden sich nicht signifikant in Abhängigkeit der Diagnosegruppe (A, B, C) hinsichtlich der Psychotherapiemotivation.

H1: Klient*innen unterscheiden sich signifikant in Abhängigkeit der Diagnosegruppe (A, B, C) hinsichtlich der Psychotherapiemotivation.

Bei Verfahren für den Vergleich von zwei Stichproben wird zwischen ein- und zweiseitigen Fragestellungen unterschieden (siehe z.B. Seistock, Bunina & Aden, 2020).

Im Rahmen der Verfahren für k-unabhängige Stichproben entfällt die Differenzierung zwischen ein- und zweiseitigen Fragestellungen sowie Testungen. Eine gerichtete Spezifizierung von

Fragestellungen kann bei einfaktoriellen ANOVAs aber z.B. im Rahmen von *Polynomialen Kontrasten* (diese werden in einem späteren Artikel der Serie Statistik behandelt) formuliert werden. Für eine konventionelle Durchführung einfaktorieller Varianzanalysen sowie Kruskal-Wallis-Tests mit anschließenden post-hoc-Testungen wird jedoch keine Unterscheidung zwischen ein- und zweiseitiger Testung vorgenommen.

Verfahren für k-unabhängige Stichproben als „Over-all-Tests"

Bei der einfaktoriellen unabhängigen Varianzanalyse sowie dem H-Test handelt es sich um sogenannte „Over-all-Verfahren" / Hauptverfahren. Diese zeichnen sich dadurch aus, dass ein signifikantes Ergebnis lediglich Auskunft darüber gibt, dass mindestens ein signifikanter Unterschied zwischen den untersuchten Gruppen in Hinblick auf das interessierte Merkmal besteht. Es kann jedoch noch keine Aussage darüber getroffen werden, (1) zwischen wie vielen und (2) zwischen welchen der k-unabhängigen Gruppen Unterschiede zu finden sind. Um das den angeführten Hypothesen zugrundeliegende Beispiel erneut aufzugreifen, könnte anhand eines signifikanten Ergebnisses im Rahmen des Hauptverfahrens (je nach Voraussetzungen einfaktorielle unabhängige ANOVA oder Kruskal-Wallis-Test) geschlussfolgert werden, dass „over-all", also über alle Gruppen hinweg, ein statistisch signifikanter Unterschied zwischen den Personen in Abhängigkeit der Diagnosegruppe (A, B, C) hinsichtlich der Psychotherapiemotivation besteht und die H0 würde verworfen werden. Eine Aussage darüber, **zwischen wie vielen** und **zwischen welchen** der untersuchten Diagnosegruppen (A vs. B, A vs. C, B vs. C) jedoch paarweise Unterschiede bestehen, könnte anhand des signifikanten Ergebnisses **noch nicht getroffen** werden. Um zu identifizieren, zwischen wie vielen und zwischen welchen der k-unabhängigen Gruppen signifikante Differenzen zu finden sind, können im Anschluss an das Hauptverfahren sogenannte *post-hoc-Tests* durchgeführt werden. Im Rahmen der einfaktoriellen Varianzanalyse können nach einem signifikanten Ergebnis im Hauptverfahren z.B. die paarweisen post-hoc-Tests nach Bonferroni zum Einsatz kommen (zum spezifizierten Einsatz unterschiedlicher post-hoc-Tests in Abhängigkeit der empirischen Ausgangslage siehe z.B. Bortz (2006)).

Post-hoc Tests

U.a., da der Kruskal-Wallis-Test mit sogenannten Rängen operiert (siehe Seistock, Bunina & Aden, 2020), bietet es sich an, im Anschluss an ein signifikantes Ergebnis bei diesem Verfahren optional paarweise post-hoc U-Tests durchzuführen, wobei zur Orientierung bereits deskriptive Kennwerte ausreichen.

Durch den paarweisen Vergleich der untersuchten Gruppen im Rahmen der post-hoc-Testungen (z.B. Diagnosegruppe A vs. Diagnosegruppe B, Diagnosegruppe A vs. Diagnosegruppe C, Diagnosegruppe B vs. Diagnosegruppe C) kann der durch das Hauptverfahren identifizierte signifikante Unterschied zwischen den k-unabhängigen Gruppen differenzierter aufgeschlüsselt werden.

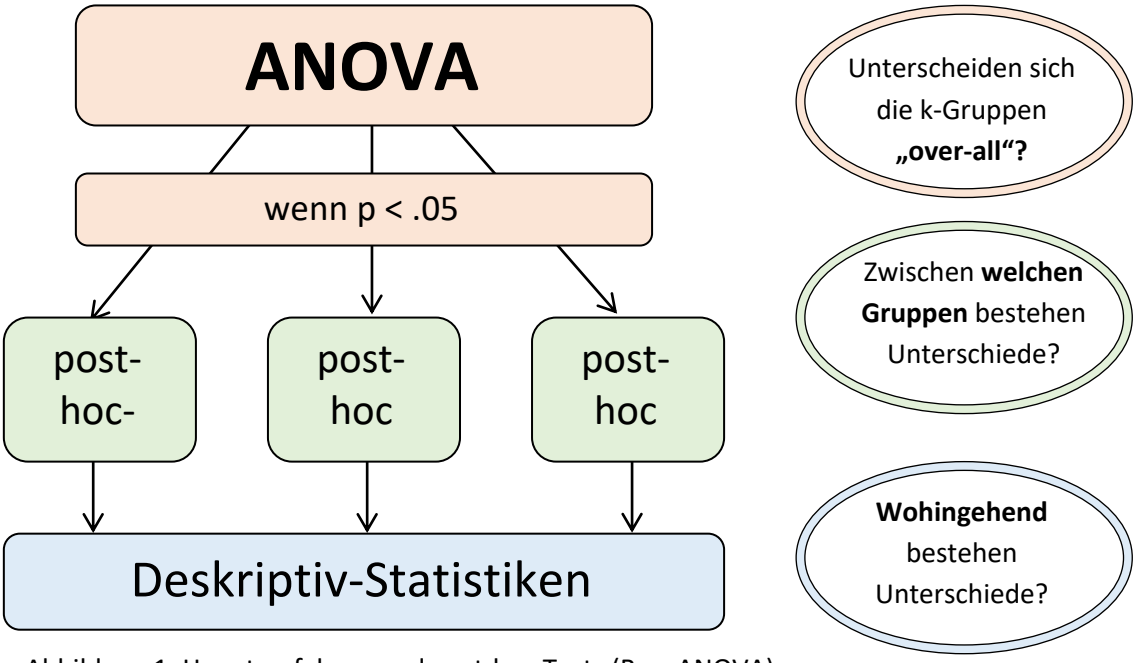


Abbildung 1: Hauptverfahren und post-hoc Tests (Bsp. ANOVA)

Multiples Testen

Obwohl sich der Einsatz von „Over-all" Verfahren positiv auf die Anzahl von Tests auswirken kann, kommt es dennoch im Rahmen der post-hoc-Testungen zu einer größeren Zahl an paarweisen Vergleichen. Der mit dem multiplen Testen verbundene Anstieg des Alpha-Fehlers muss dabei berücksichtigt werden. Verfahren, wie die post-hoc Bonferroni-Tests, berücksichtigen diesen Umstand mit einer integrierten Adjustierung der p-Werte. Zum Probelm der Alpha-Kumulierung beim multiplen Testen und wie diesem begegnet werden kann, siehe z.B. Bender, Lange & Ziegler (2007).

Wie bereits in den ersten beiden Artikeln der Serie Statistik beschrieben, erfolgt die Signifikanzbestimmung und die Bewertung auch bei den zwei angeführten Verfahren für k-unverbundene Stichproben (einfaktorielle unabhängige ANOVA und Kruskal-Wallis-Test) auf der Grundlage von Prüfverteilungen. Im Rahmen einer bestimmten Prüfverteilung (z.B. F-Verteilung bei der ANOVA) wird diesbezüglich ein numerischer Verteilungswert berechnet, welcher die Größe des Unterschieds zwischen den k-unverbundenen Gruppen in der Ausprägung des untersuchten Merkmals ausdrückt. Die Verfahren folgen somit einer einheitlichen Struktur, welche sich anhand der Frage, **wen** (welche unabhängigen Gruppen: z.B. Diagnosegruppe A, B und C) vergleiche ich **hinsichtlich wessen** (untersuchtes abhängiges Merkmal: z.B. Psychotherapiemotivation), veranschaulichen lässt. Die kategoriale unabhängige Variable (z.B. Diagnosegruppe), welche hinsichtlich der abhängigen Variable verglichen wird, wird im Rahmen der einfaktoriellen Varianzanalyse auch als „Faktor" bezeichnet, die Ausprägungen dieser Variable, also die spezifischen k-unabhängigen Gruppen (A, B, C), als „Faktorstufen". Um zu bestimmen, ob der berechnete Unterschied als überzufällig (statistisch signifikant/ bedeutsam) oder zufällig (statistisch nicht signifikant / unbedeutend) zu bewerten ist, muss der numerische Verteilungswert wiederum in ein einen Wahrscheinlichkeitswert (p-Wert) überführt werden. Durch den Vergleich dieses p-Wertes mit der festgelegten Signifikanzgrenze (in den Sozialwissenschaften üblicherweise 5%) ist ein Schluss über die statistische Signifikanz des Unterschiedes zwischen den k-unverbundenen Gruppen zulässig. Für genauere Ausführungen zur Bewertung der Signifikanz siehe Seistock, Bunina und Aden (2020).

Effektstärken

Da der p-Wert nicht nur von der Größe des beobachteten Unterschiedes, sondern unter anderem auch von der Stichprobengröße abhängig ist, könnte eine signifikante Differenz im Rahmen von Verfahren für k unabhängige Stichproben nur deshalb beobachtet worden sein, weil für die Berechnungen nicht die optimale, sondern eine zu große Personenanzahl (Oversampling) herangezogen wurde (vgl. Jones, Carley & Harrison, 2003; Krzywinski & Altman, 2013).

Die tatsächliche Größe sowie inhaltliche Relevanz eines beobachteten, signifikanten Unterschiedes kann erst durch die Interpretation von Effektstärken beurteilt werden (genauere Ausführungen zur Relevanz von Effektstärken für die Beurteilung statistischer Ergebnisse siehe z.B. Sullivan & Feinn, 2012; Fritz, Morris & Richler, 2012; Seistsock, Bunina & Aden, 2020).

Im Rahmen der unabhängigen einfaktoriellen Varianzanalyse (ANOVA) sind das *klassische Eta-Quadrat* (η^2) und das *partielle Eta-Quadrat* (η_p^2) die am häufigsten verwendeten Effektstärken (Fritz, Morris & Richler, 2012).

Neben dem η^2 und η_p^2 existieren noch weitere Effektmaße, die im Rahmen der einfaktoriellen ANOVA berechnet und interpretiert werden können (z.B. ω^2 oder ω_p^2). Im vorliegenden Beitrag sollen sich die Ausführungen jedoch ausschließlich auf das *klassische Eta-Quadrat* (η^2) und das *partielle Eta-Quadrat* (η_p^2) beziehen, da diese im sozialwissenschaftlichen Forschungskontext die häufigste Anwendung finden. Eine Übersicht und Diskussion bezüglich weiterer Effektmaße im Rahmen varianzanalytischer Berechnungen sowie deren Auswahl siehe z.B. Lakens (2013).

Klassisches Eta-Quadrat

Die Effektstärke η² gibt an, wie hoch der Anteil an der Gesamtvariation ist, der auf den Gruppenfaktor zurückzuführen ist. Anders formuliert: η² gibt an, welcher Anteil an Variation in einem bestimmten Merkmal durch einen Gruppierungsfaktor erklärt werden kann. Bezogen auf das vorangegangene Beispiel zum Verhältnis von Diagnosegruppe und Therapiemotivation gäbe das η² wieder, wie stark Unterschiede in der Therapiemotivation auf das jeweilige Krankheitsbild „zurückzuführen“[1] sind. Der η²-Wert lässt sich dabei als Prozentsatz interpretieren. So bedeutete ein Eta-Quadrat von η²= .13, dass 13% der Gesamtvariation in der metrischen Zielvariable durch den Gruppenfaktor erklärt werden können. Das Eta-Quadrat bildet das Verhältnis von systematischer Variation *SSfactor* (durch Gruppenfaktor erklärbar) und Gesamtvariation *SStotal* des Merkmals (gruppenunabhängige Variation) ab (z.B. Pierce, Block & Aguinis, 2004). Das klassische η² ist der Quotient aus systematischer und gesamter Variation:

$$\eta^2 = \frac{SSfactor}{SStotal}$$

Das partielle Eta-Quadrat (η_p^2) ist vor allem bei der Berechnung mehrfaktorieller ANOVAs, d.h. mit mehr als einem Gruppenfaktor, relevant; z.B. wenn simultan in einem Modell bzw. einer Hypothese

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

geprüft würde, ob im Hinblick auf die Therapiemotivation Unterschiede in Abhängigkeit der Diagnosegruppe *und* des Geschlechts[2] bestünden.

Partielles Eta-Quadrat

Im Zuge der Berechnung des partiellen Eta-Quadrats für einen Gruppenfaktor (z.B. Diagnosegruppe) wird der Anteil anderer Faktoren (z.B. Geschlecht) an der Gesamtvariation (z.B. im Merkmal Therapiemotivation) aus selbiger „(heraus-) partialisiert", d.h. exkludiert (Pierce, Block & Aguinis, 2004). Die Variation im Merkmal Therapiemotivation, welche durch die Diagnose der Klient*innen erklärt werden kann (*SSfactor*), wird durch die „bereinigte" Gesamtvariation dividiert. Diese Gesamtvariation ergibt sich aus den Bestimmungsstücken *SSfactor* und *SSerror*. Für η_p^2 gilt folgende Formel:

$$\eta_p^2 = \frac{SSfactor}{(SSfactor + SSerror)}$$

Bei *einfaktoriellen ANOVAs*, d.h. mit nur einem Gruppenfaktor, sind das *klassische Eta-Quadrat* (η^2) und das *partielle Eta-Quadrat* (η_p^2) identisch. In mehrfaktoriellen Designs ist die Entscheidung für eines der beiden Effektmaße von entscheidender Bedeutung für die Schätzung des Effekts. Weiterführende Ausführungen zum Verhältnis dieser beiden Effektstärken siehe Pierce, Block und Aguinis (2004).

Beurteilung des Effekts

Insgesamt gilt es, die Wahl des Effektmaßes bewusst zu treffen und an die jeweiligen Untersuchungsdesigns angepasst auszuwählen.

Die η^2-Maße können Werte zwischen 0 und +1 annehmen. Zur Beurteilung des Effekts im Hinblick auf dessen Stärke sind beispielsweise von Cohen (1988) folgende Bewertungsstufen vorgeschlagen:

0.01 kleiner Effekt

0.06 mittlerer Effekt

0.14 großer Effekt

Cohen (1988) selbst sah globale Bewertungskonventionen kritisch und plädierte dafür, diese Stufen lediglich als Orientierungspunkte zu begreifen. Eine kritische Auseinandersetzung mit der Anwendung von Bewertungskonventionen in der psychologischen Forschung ist beispielsweis bei Schäfer und Schwarz (2019) zu finden.

Im Rahmen des Kruskal-Wallis-Tests (H-Tests) besteht zumindest die Möglichkeit die Effektstärke η_H^2 zu berechnen, die auf der H-Statistik basiert:

[2] An dieser Stelle würde eine zwei-faktorielle ANOVA berechnet, die neben den Haupteffekten A und B auch die Wechselwirkung zwischen A und B (A x B) mit berücksichtigte.

$$\eta_H^2 = \frac{H - k + 1}{n - k}$$

In der Praxis wird jedoch häufig auf die Berechnung und Angabe einer eigenen Effektstärke im Rahmen des Kruskal-Wallis-Test (H-Test) verzichtet.

Voraussetzungen

Skalen-niveau

Die Entscheidung bezüglich der Verfahrenswahl (einfaktorielle unabhängige ANOVA oder H-Test) ist, wie auch bei den Verfahren für zwei Stichproben, an bestimmte Voraussetzungen gebunden, welche in Abhängigkeit vom jeweiligen Test erfüllt sein müssen, um eine korrekte Anwendung und Interpretation der erzielten Ergebnisse gewährleisten zu können. Die Wahl des Verfahrens ist dabei stark von der jeweiligen Datenlage abhängig und vor allem durch das Skalenniveau des zu untersuchenden Merkmals (abhängige Variable, hinsichtlich wessen? z.B. Psychotherapiemotivation) bestimmt. Die einfaktorielle Varianzanalyse kommt ausschließlich bei (normalverteilten) metrischen (verhältnis- oder intervallskalierten) Variablen zum Einsatz, der Kruskal-Wallis-Test kann bei nicht normalverteilten metrischen und ordinalen Variablen angewandt werden. Metrische Variablen ergeben sich dabei aus all jenen Werten, welche mittels standardisierter Testinstrumente oder Fragebögen erhoben wurden (z.B. Angstwerte mittels Beck Angst-Inventar (Beck & Steer, 1990); Psychotherapiemotivation mittels FPTM (Nübling et al., 2005)). Ordinale Variablen speisen sich hingegen aus Daten, welche beispielsweise auf subjektiven Einschätzungen (z.B. Zufriedenheitseinschätzungen) beruhen und mittels unstandardisierter Fragebögen erhoben wurden. Zunächst muss im Rahmen der Verfahrenswahl somit die Frage geklärt werden, ob es sich bei dem zu untersuchenden abhängigen Merkmal, hinsichtlich derer die k-unabhängigen Gruppen verglichen werden, um eine metrische Variable (einfaktorielle Varianzanalyse) oder um eine ordinale Variable (Kruskal-Wallis-Test) handelt.

Neben dem Messniveau der untersuchten Variable, hinsichtlich derer k-unverbundene Stichproben verglichen werden, ist die korrekte Verfahrenswahl jedoch ebenso abhängig von unterschiedlichen empirischen Voraussetzungen.

Normal-verteilung in jeder Gruppe

Grundsätzlich sollte von den zwei angeführten Verfahren, wenn möglich, auf die einfaktorielle unabhängige Varianzanalyse zurückgegriffen werden, da diese die größte Testmacht aufweist, d.h. über eine höhere Wahrscheinlichkeit verfügt, eine vorhandene Systematik (Unterschied) korrekterweise zu identifizieren. Allerdings verlangt dieses Verfahren auch die Einhaltung der meisten empirischen Voraussetzungen. Diesbezüglich können die Voraussetzungen des t-Tests für zwei unabhängige Gruppen (Normalverteilung des untersuchten Merkmals und Varianzhomogenität, siehe Seistock, Bunina & Aden, 2020) auf k-unabhängige Gruppen im Rahmen der einfaktoriellen Varianzanalyse verallgemeinert werden. Das bedeutet, dass die Anwendung der einfaktoriellen unabhängigen Varianzanalyse neben dem metrischen Skalenniveau der untersuchten Variable (mindestens intervallskaliert) ebenfalls deren Normalverteilung voraussetzt. Eine detaillierte Beschreibung zur Voraussetzung und Überprüfung der Normalverteilung ist im ersten Artikel der

Serie Statistik des Forschungsbulletins nachzulesen (Seistock, Bunina & Aden, 2020). Ist diese Voraussetzung verletzt, das zu untersuchende metrische Merkmal somit nicht in jeder der k Gruppen normalverteilt, kann auf das Ausweichverfahren des Kruskal-Wallis-Test zurückgegriffen werden, welches in weiterer Folge noch genauer beschrieben wird.

Homogenität der Varianzen

Analog zum t-Test für zwei unverbundene Gruppen muss darüber hinaus auch bei der Anwendung der einfaktoriellen Varianzanalyse bei mehr als zwei unabhängigen Gruppen die empirische Voraussetzung der Varianzhomogenität erfüllt sein. Die k-unabhängigen Stichproben sollten im Sinne dieser Voraussetzung im untersuchten Merkmal somit eine ähnliche Streuung aufweisen, d.h. die Variation des Merkmals innerhalb der Gruppen in einem vergleichbaren Ausmaß vorliegen. Die Homogenität der Varianzen kann beispielsweise mit dem sogenannten Levene-Test überprüft werden. Dieser sollte nicht signifikant ausfallen, d.h. einen p-Wert von >.05 aufweisen, damit die Voraussetzung der Varianzhomogenität als erfüllt angesehen werden kann. Ist die Voraussetzung verletzt, d.h. die Varianzen innerhalb der k-unabhängigen Gruppen weisen Heterogenität auf, kann beispielsweise auf den sogenannten Brown-Forsythe-Test der die die Welch-ANOVA[3] zurückgegriffen werden.

Die einfaktorielle unabhängige Varianzanalyse

einfakt. unabhäng. ANOVA

Die einfaktorielle Varianzanalyse basiert auf der Erklärung von Variation bzw. Varianz eines Merkmals. Dabei werden verschiedene Formen der Variation berechnet und zu einander in Verhältnis gesetzt. Entscheidend dabei ist die systematische Variation, die denjenigen Anteil der Gesamtvariation umfasst, der durch den Gruppenfaktor erklärt werden kann. Im Rahmen der Varianzzerlegung wird dieser Anteil als *SSfactor* (Sum of Squares Factor)[4] bezeichnet. Diese Varianz wird auf Basis der Variation *zwischen den Gruppen* errechnet (*SSfactor*$= \sum\sum(\bar{x}_{mi} - \bar{x}_{Gesamt})^2$) und der unsystematischen Variation *SSerror* (Sum of Squares Error) gegenübergestellt. Diese - auch als Quadratsumme Innerhalb bezeichnete - Variation basiert im Wesentlichen auf der Variation *innerhalb der Gruppen* (*SSerror*$= \sum\sum(x_{mi} - \bar{x}_{mi})^2$). Alltagssprachlich formuliert werden also zwei Formen von „Verschiedenheit", also Variation berechnet. Eine, welche die Verschiedenheit **ZWISCHEN** den Gruppen abbildet (SSfactor) und eine, welche die Verschiedenheit **INNERHALB** der Gruppen abbildet (SSerror). Zur Überprüfung, ob sich die Gruppen signifikant voneinander unterscheiden, werden diese beiden Verschiedenheiten ins Verhältnis zueinander gesetzt. Dazu werden die beiden Quadratsummen - *SSfactor* und *SSerror* - in Varianzen umgerechnet, indem diese an den jeweiligen Freiheitsgraden relativiert werden:

$$\hat{\sigma}^2{}_{factor} = \frac{SS\_factor}{df_{factor}}$$

[3] Abhängig von der empirischen Ausgangslage bestehen unterschiedliche Möglichkeiten mit einer Verletzung der Varainzhomogenität umzugehen (siehe dazu auch Bortz (2006))
[4] Im Deutschen auch als *Quadratsumme Zwischen* bezeichnet.

$$\hat{\sigma}^2_{error} = \frac{SS\_error}{df_{error}}$$

Ins Verhältnis werden die beiden Varianzen dann im Rahmen des sogenannten F-Tests gesetzt:

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

Als Resultat des F-Tests steht der sogenannte *empirische F-Wert*, der einen Verteilungsparameter darstellt, auf Basis dessen eine Beurteilung der Signifikanz erfolgt.

Angenommen Sie erhalten beim Vergleich von drei unverbundenen Stichproben (A, B, C) hinsichtlich einer metrischen Variable mittels einfaktorieller unabhängiger ANOVA einen *F-Wert* von 110.16, bei df_{factor}= 2 und df_{error}= 90, mit einem *p-Wert* von p< .001, sowie einer Effektstärke von η_p^2= 0.71.

Beim statistischen Report des Hauptverfahrens ANOVA empfiehlt sich die Angabe folgender Kennwerte:

Reportkonventionen einfakt. unabhäng. ANOVA

(F(2,90)= 110.16, p< .001, η_p^2= 0.71)

Für die im Anschluss an eine signifikante ANOVA durchgeführten post-hoc-Tests empfiehlt sich beispielsweise folgende Reportlegung:

Im Zuge der paarweisen Vergleiche der Gruppen nach der signifikanten ANOVA erhalten Sie folgende Ergebnisse: Gruppe B $M_{GruppeB}$= 3.16 (SD= 2.68), unterscheidet sich signifikant von Gruppe A $M_{GruppeA}$= 12.25 (SD= 2.89) und Gruppe C $M_{GruppeC}$= 11.85 (SD= 2.63) (jeweils p< .001). Gruppe A und B unterscheiden sich jedoch nicht signifikant (p= .987) voneinander im metrischen Merkmal[5].

Der Kruskal-Wallis-Test

Beim H-Test nach Kruskal und Wallis (1952) werden über die Chi²-Quadrat-Verteilung ebenfalls k-unabhängige Gruppen auf Unterschiede geprüft. Dabei wird dieses Verfahren angewendet, wenn die k-unabhängigen Gruppen a) hinsichtlich eines ordinal-skalierten Merkmals oder b) hinsichtlich eines metrischen, aber nicht in jeder der k Gruppen normalverteilten, Merkmals verglichen werden.

Zur Berechnung des Kruskal-Wallis-Tests werden an Stelle der Originalwerte sogenannte Ränge verwendet. D.h. die ursprünglichen Beobachtungen werden in Ränge transformiert und fungieren in

Kruskal-Wallis-Test

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

weiterer Folge als Berechnungsgrundlage. Mit der Verwendung von Rängen gehen einige Vorteile einher, die beispielsweise Anforderungen an die Verteilung der Daten betrifft. So kann ein Kruskal-Wallis-Test - im Gegensatz zur ANOVA - auch dann problemlos berechnet werden, wenn die Normalverteilung der Daten nicht gewährleistet ist. Darüber hinaus ergibt sich aus der Verwendung von Rängen die Möglichkeit auch mit ordinalen Merkmalen zu rechnen bzw. diese zu untersuchen (Kruskal & Wallis, 1952).

$$H = \frac{12}{N*(N+1)}\left(\sum \frac{R^2{}_i}{n_i}\right) - 3 * (n+1)$$

Auf Basis des berechneten H-Wertes kann dann eine Beurteilung der Signifikanz der Gruppenunterschiede vorgenommen werden. Zusätzlich besteht die Möglichkeit noch die Effektstärke η_H^2 anzugeben. Im Anschluss an einen signifikanten Kruskal-Wallis-Test können analog zur einfaktoriellen Varianzanalyse paarweise post-hoc Tests durchgeführt werden. Im Rahmen des Kruskal-Wallis-Tests werden üblicherweise paarweise Mann-Whitney-U-Tests angewendet.

Angenommen Sie erhalten beim Vergleich von drei unverbundenen Stichproben (A, B, C) hinsichtlich einer ordinalen Variable mittels Kruskal-Wallis-Test einen *χ²-Wert* von 8.35, bei *df*= 2 und *n*= 93, mit einem *p-Wert* von p= .015.

Beim statistischen Report des Hauptverfahrens H-Test nach Kruskal und Wallis empfiehlt sich die Angabe folgender Kennwerte:

Reportkonventionen Kruskal-Wallis-Test

(χ²(df= 2, n= 93)= 8.35, p= .015)

Für die im Anschluss an einen signifikanten H-Test nach Kruskal und Wallis durchgeführten post-hoc U-Tests empfiehlt sich beispielsweise folgende Reportlegung:

Gruppe B $M_{GruppeB}$= 3.16 (SD= 2.68), unterscheidet sich signifikant von Gruppe A $M_{GruppeA}$= 12.25 (SD= 2.89) (U= 196.00, Z= -2.43, p= .015) und von Gruppe C $M_{GruppeC}$= 11.85 (SD= 2.63) (U= 333.00, Z= -2.59, p= .010). Gruppe A und B unterscheiden sich jedoch nicht signifikant (U= 547.00, Z= -.17, p= .866) voneinander im ordinalen Merkmal[6].

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

Konklusion

Forscher*innen auf dem Feld der quantitativ ausgerichteten Psychotherapiewissenschaft sehen sich häufig mit Untersuchungs- und damit Auswertungsdesigns konfrontiert, die auf den Vergleich von mehr als zwei Gruppen/Stichproben ausgerichtet sind.

Verfahren für k-unabhängige Stichproben ermöglichen die Auswertung von Querschnittuntersuchungen, die etwa zur Charakterisierung von Unterschieden zwischen Klient*innengruppen psychotherapeutischer Versorgungseinrichtungen (z.B. Unterschiede in den Belastungsscores zwischen Klient*innen, die Behandlungen in a) Ambulanzen oder b) Privatpraxen oder c) Kliniken in Anspruch nehmen). Ebenso eigenen sich diese Verfahren, um die Ausgangs- bzw. Startbedingungen vor dem Beginn einer Therapie zwischen verschiedenen Gruppen zu vergleichen. So kann etwa die Rolle vermittelnder Instanzen für die anfängliche Therapiemotivation mit Verfahren für k-unabhängige Stichproben untersucht werden (z.B. Vergleich Therapiemotivation zwischen Personen, die a) auf hausärtzlichen Rat/Vermittlung, b) auf Anraten oder Vermittlung von Familienangehörigen, c) gänzlich eigeninitiativ eine Psychotherapie begonnen haben). Die psychotherapiewissenschaftlich relevanten Fragestellungen, die mit Verfahren für k- unabhängige Stichproben untersucht werden können, sind inhaltlich äußerst vielfältig.

Beim Vergleich von mehr als zwei Gruppen bestehen jedoch einige Besonderheiten, die es bei der Anwendung der geeigneten statistischen Auswertungsverfahren zu beachten gilt. Wie im Artikel bereits skizziert, ist es beim Vergleich von k Gruppen zunächst von Vorteil von bloßen paarweisen Vergleichen der untersuchten Gruppen abzusehen. Der erste Hinweis besteht also zunächst ganz simpel darin, beim Vergleich von k Gruppen auch Verfahren für den Vergleich von mehr als zwei Gruppen einzusetzen. Dies hat praktische Gründe, die hier zusammenfassend noch einmal aufgezeigt werden:

→ *Alpha-Kumulierung*

Ausschließlich paarweise Vergleiche durchzuführen, kann zu einer unnötigen Kumulierung des Alpha-Fehlers führen und damit die statistische Entscheidungssicherheit beeinträchtigen. Dies ist insbesondere bei umfangreicheren Untersuchungen mit einer höheren Zahl an Hypothesen von Bedeutung. Außerdem tragen verschiedene post-hoc-Prozeduren dem Problem des multiplen Testens (z.B. Bonferroni) und der damit verbundenen Alpha-Kumulierung Rechnung (siehe vertiefend Bender, Lange & Ziegler, 2007).

→ *Forschungsökonomie*

Außerdem kann durch den Einsatz der vorgestellten Hauptverfahren bereits anhand einer einzelnen Berechnung identifiziert werden, ob sich überhaupt zumindest zwei Gruppen voneinander systematisch in einem Merkmal unterscheiden, ohne erst alle Gruppen paarweise miteinander vergleichen zu müssen. Bei nicht signifikanten Ergebnissen in den Hauptverfahren erübrigen sich die weiteren und potentiell zahlreichen paarweisen Vergleiche/Testungen. Hier bieten die Verfahren für

k-unabhängige Stichproben auch schlichtweg forschungsökonomische Vorteile und tragen dazu bei, die Zahl unnötiger paarweiser Vergleiche zu reduzieren (siehe Alpha-Kumulierung).

Allerdings ist bei der Verwendung von Verfahren für k- unabhängige Stichproben auf Unterschiede zur Anwendung und Interpretation von Verfahren für zwei unabhängige Stichproben zu verweisen. Diese sind in Form von Hinweisen für die Forschungspraxis nachfolgend angeführt:

→ *Post-Hoc-Testungen*

In der praktischen Anwendung von ANOVA´s oder Kruskal-Wallis-Tests sollte den Anwender*innen bewusst sein, dass mit einem signifikanten Ergebnis in dem jeweiligen Hauptverfahren noch nicht zwangsläufig auch die Forschungsfrage zu beantworten ist. Ein signifikantes Ergebnis im Rahmen einer ANOVA oder eines Kruskal-Wallis-Tests gibt zunächst lediglich Aufschluss darüber, dass grundsätzlich eine Systematik zu konstatieren ist. Allzu häufig wird ein signifikantes Ergebnis im Hauptverfahren fälschlicherweise dahingehend interpretiert, dass sich alle zur Untersuchung stehenden Gruppen signifikant voneinander unterscheiden. Welche Gruppen sich jedoch in welcher Weise unterscheiden, muss im Zuge anschließender post-hoc Testungen eruiert werden. Neben der Berechnung von paarweisen post-hoc Tests besteht im Rahmen varianzanalytischer Verfahren ebenso die Möglichkeit mit verschiedenen Formen sogenannter Kontraste eine Spezifizierung der Gruppenunterschiede zu eruieren. Die polynomialen sowie die orthogonalen Kontraste werden in einem späteren Artikel dieser Serie eingehender behandelt.

→ *Effektstärke*

Die Effektstärken η^2 und η_p^2 , sowie η_H^2 beziehen sich auf den Haupteffekt und geben nicht die Stärke der paarweisen Unterschiede wieder. Der übergeordnete Effekt sollte nicht auf die einzelnen, paarweisen Unterschiede übertragen werden. Es besteht die Möglichkeit gesonderte Effektstärken für die Paarvergleiche anzugeben (z.B. *d* oder *r*).

Im vorliegenden Artikel wurde die einfaktorielle Varianzanalyse, also jene mit nur einem Gruppierungsfaktor, behandelt. Im Rahmen dieser ANOVA spielt die Differenzierung zwischen dem klassischen und dem partiellen Eta-Quadrat keine wesentliche Rolle. Bei mehrfaktoriellen Auswertungsdesigns ist diese Differenzierung jedoch von Bedeutung und es gilt die Wahl der jeweiligen Effektstärke genau abzuwägen.

Die Angabe eines standardisierten Effektmaßes ist für eine angemessene Interpretation – insbesondere der Varianzanalyse- unerlässlich. Aufgrund von Effekten, die beispielsweise durch die Stichprobengröße bedingt sind, etwa aufgrund von Over- oder Under-Sampling), ist eine Interpretation der Ergebnisse einzig entlang des p-Wertes unzureichend.

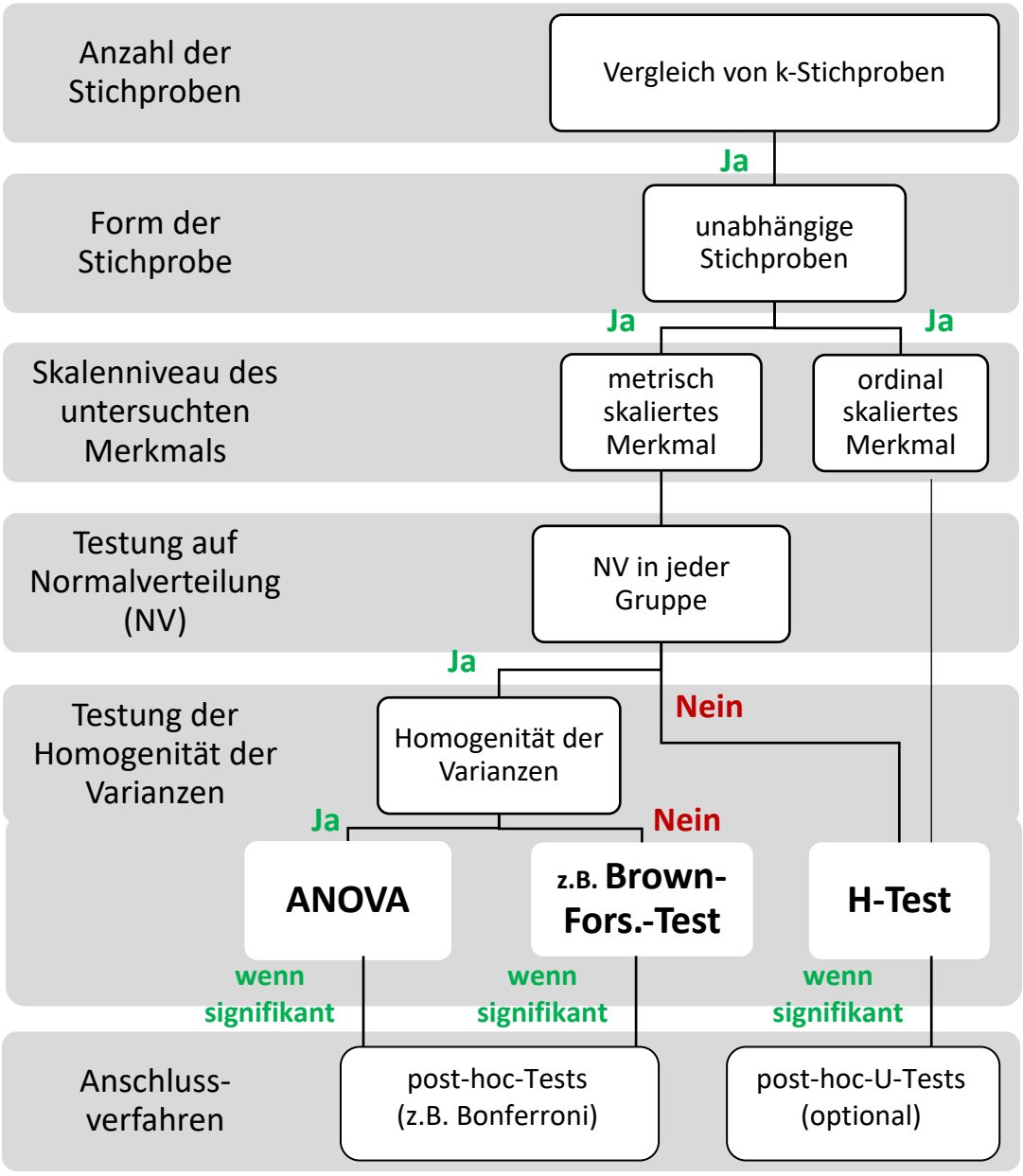
→ *Stichprobengröße- und Planung*

Um stichprobenbedingte Effekte dieser Art zu vermeiden und ein adäquates Verhältnis von Macht und Fehlerwahrscheinlichkeit sicherzustellen, ist eine Kalkulation des optimalen Stichprobenumfangs empfohlen (z.B. Chow, Shao, Wang & Lokhnygina, 2018).

→ *Wahl des korrekten Verfahrens*

Eine Verletzung der Voraussetzungen für eine einfaktorielle unabahängige ANOVA kann insbesondere bei kleinen Stichproben problematische Effekte aufweisen. Obgleich die ANOVA eher robust auf die Verletzung der Normalverteilung reagiert, ist insbesondere bei kleineren Stichprobenumfängen die Verwendung des non-parametrischen Verfahrens des Kruskal-Wallis-Tests zu empfehlen (siehe Entscheidungsbaum im Anhang).

Entscheidungsbaum



Literatur

Beck, A. T., Brown, G., Epstein, N. & Steer, R. A. (1988). An Inventory for Measuring Clinical Anxiety: Psychometric Properties. *Journal of Consulting and Clinical Psychology, 56* (6), 893–897.

Beck, A. T. & Steer, R. A. (1990). *Manual für the Beck Anxiety Inventory.* San Antonio, TX: 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 Verlag.

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 Edn. New York, NY: Academic Press.

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

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.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863, 1–12.

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.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and psychological measurement*, *64*(6), 916–924.

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).

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.

*Angaben zu den Autor*innen*

Jan Aden, Anastasiya Bunina, Caroline Vavrik
Institut für Statistik
Adresse: Freudplatz 1, 1020 Wien, Raum 6011
E-Mail: jan.aden@sfu.ac.at