Empirische Sonderpädagogik, 2012, Nr. 3/4, S. 265–274

Inferenzstatistischer Nachweis intraindividueller Unterschiede im Rahmen von Einzelfallanalysen

Thomas Köhler

Universität Hamburg

Zusammenfassung
Obwohl einzelfallanalytische Studien sich sowohl in der pädagogischen wie der psychologischen Forschung anbieten würden, sind diese ausgesprochen selten. Dies ist sicher nicht zuletzt darin begründet, dass vergleichsweise komplizierte statistische Verfahren erforderlich sind, um individuelle Veränderungen in seriell abhängigen Daten gegen Zufälligkeit abzusichern. In diesem Artikel wird angeregt, einfache statistische Verfahren wie t-Test oder Varianzanalysen zur Lösung dieses Problems heranzuziehen. Es wird vorgeschlagen, einen großen Datensatz zu erheben, um danach hinreichend viele Messwerte eliminieren zu können, sodass serielle Unabhängigkeit vorliegt. Alternativ kann die Methode des „prewhitening" angewandt werden, um die serielle Abhängigkeit zu eliminieren.

Schlüsselwörter: Einzelfallanalytische Untersuchungen, seriell abhängige Daten, konventionelle inferenzstatistische Methoden, Prewhitening

The use of inferential statistics in the examination of within-subject changes

Abstract
Although single subject designs would be most appropriate in pedagogical and psychological research they are rather uncommon. This is mainly due to the fact that the application of fairly complicated statistical methods is necessary in order to demonstrate the significance of individual changes in serially dependent data. In this article some simple conventional statistics (such as t-test or analysis of variance) are proposed to overcome this problem. It is suggested either to obtain a large data set and eliminate a sufficient number of these to assure the lack of serial dependency as required for application. Alternatively, methods of "pre-whitening" can be used to eliminate serial dependency.

Keywords: single subject analysis, serial dependency, conventional methods of inferential statistics, prewhitening

Einleitung

Einzelfallanalysen – zumindest solche quantitativer Art – haben sich in den Sozialwissenschaften nie wirklich durchsetzen können. Dies ist insofern nur schwer nachvollziehbar, als gerade in diesen Disziplinen, sei es in der Psychologie, sei es in der Pädagogik, oft die Voraussetzungen für die Anwendung aggregatstatischer Verfahren nicht wirklich gegeben sind. Dazu gehören bekanntlich gewisse Mindestumfänge der Untersuchungsstichproben, des Weiteren vergleichsweise einheitliches Verhalten der einzelnen Stichprobenelemente. Nimmt man etwa den keineswegs seltenen Fall an, dass ein Teil der Stichprobe sich auf eine Intervention hin im erwarteten und gewünschten Sinne ändert, ein anderer Teil dies aber nicht tut, sondern vielleicht sogar genau in entgegengesetzter Richtung reagiert, so ließe sich mit den üblichen gruppenstatistischen Prozeduren (etwa *t*-Test für abhängige Stichproben, Varianzanalyse mit Messwertwiederholungen) in diesen Fällen meist keine Signifikanz nachweisen; entweder ist die durchschnittliche Veränderung klein oder aber – selbst wenn diese substantiell ist – die Varianz der Veränderungswerte so groß, dass der Nachweis der Überzufälligkeit nicht gelingt.

Hinzu kommt das erwähnte Problem der häufig ausgesprochen kleinen Stichprobenumfänge, sodass – selbst wenn Interventionseffekte evident sind und sogar bei sämtlichen Probanden in die gleiche Richtung zielen – der Nachweis der Signifikanz grundsätzlich nicht gelingen kann. Schließlich ist natürlich der Fall sehr häufig, dass es sich überhaupt nur um eine Studie an einer einzigen Person handelt, womit das übliche aggregatstatistische Methodenarsenal zum Nachweis der nicht zufälligen Veränderung sich überhaupt nicht anwenden lässt.

In allen genannten Fällen: 1) große Stichproben, jedoch bei inhomogenen Interventionseffekten, 2) für die Anwendung aggregatstatistischer Verfahren zu geringen Stichprobenumfängen 3) Studien mit $n = 1$ bietet sich eine einzelstatistische Auswertung an – es sei denn, man begnügt sich mit den Verweis auf die Evidenz der Daten oder schiebt die Frage der Absicherung gegen Zufälligkeit beiseite. Allerdings ist zu konzedieren, dass die statistische Behandlung von Einzelfalldaten teilweise ausgesprochen diffizil ist und die entsprechenden Modelle oft nur bedingt nachvollziehbar und auch nicht immer unumstritten sind.

Wir beschäftigen uns hier nur mit der für die Forschungspraxis sicher relevantesten Aufgabe der Absicherung von Unterschieden innerhalb einer Person – beispielsweise im Rahmen der Untersuchung von pretreatment- und posttreatment-Werten (etwa den Leistungen vor und nach einer Fördermaßnahme) oder beim Vergleich von Verhaltensdaten in unterschiedlichen Situationen (beispielsweise Aggressivität in der Schule und außerhalb dieses Rahmens). Für die zwar interessante und methodisch nachvollziehbarere, jedoch in ihrer praktischen Bedeutung häufig überschätzte Zeitreihenanalyse sei auf Köhler (2008) verwiesen. Der Tatsache Rechnung tragend, dass die LeserInnen dieses Beitrags sicher nicht unbedingt spezielles Methodeninteresse aufweisen, sollen der Didaktik zuliebe bei der Darstellung gewisse Ungenauigkeiten in Kauf genommen werden.

Die Anwendung gruppenstatistischer Verfahren zum Vergleich von Daten innerhalb ein und derselben Person: Möglichkeiten und Probleme

Es sei also beispielsweise die Situation gegeben, dass von einem einzigen Schüler, der einen bestimmten Förderunterricht erhält, über mehrere Monate im wöchentlichen Abstand erhobene Leistungsdaten vorlie-

gen, und zwar sowohl vor Einsetzen der Fördermaßnahme als auch danach; damit stellt sich die Frage, wie weit eine eventuell gefundene Verbesserung nicht auf zufällige intraindividuelle Schwankungen zurückzuführen ist, sondern eine systematische Veränderung darstellt. (Dass natürlich eine solche systematische Veränderung nicht kausal als Interventionseffekt erklärt werden kann, weil die unerlässliche Kontrollbedingung hier fehlt, sei nur erwähnt; dies ist aber eine allgemeine Problematik von Einzelfallstudien, die auch durch noch so ausgeklügelte Designs letztlich nicht wirklich gelöst werden kann; s. dazu die Diskussion in Köhler, 2008). Wir wollen weiter der Einfachheit halber annehmen, dass eine eindeutige Zuordnung entsprechender Zeitpunkte vor und nach Intervention nicht gelingt, beispielsweise weil unterschiedlich viele Messungen unter beiden Bedingungen vorliegen.

Dann ist formal jene Situation gegeben, bei der etwa ein *t*-Test für unabhängige (eindeutiger: nichtkorrelierende) Stichproben vorliegt. Rufen wir uns also das aus der Statistikausbildung bekannte Problem in Erinnerung, dass beispielsweise in einer Stichprobe von Frauen und einer Stichprobe von Männern die Werte in einem Konzentrationstest erhoben wurden und nun diese Gruppen bezüglich ihres mittleren Abschneidens verglichen werden sollen. Mit dem *t*-Test für unabhängige Stichproben wird überprüft, ob der mittlere Unterschied im Vergleich zu den Varianzen in beiden Stichproben so groß ist, dass diese Differenz nicht mehr durch Zufall erklärt werden kann, mit anderen Worten: signifikant ist. Der anhand der Daten berechnete *t*-Wert wird dazu mit einem in Tabellen aufgelisteten kritischen *t*-Wert verglichen, welcher unter anderem von der Zahl der so genannten Freiheitsgrade abhängt (in diesem Fall der Summe der beiden Stichprobenumfänge, vermindert um 2). Je größer die Zahl der Freiheitsgrade ist, je mehr als voneinander unabhängig vorausgesetzte Probandenwerte somit in die Berechnung eingeben, umso eher erreicht der ermittelte *t*-Wert Signifikanz.

Es ist nun nahe liegend, diese Logik auf die Messwerte eines Probanden (z. B. vor und nach Therapie oder innerhalb und außerhalb der Schule) zu übertragen und zu bestimmen, ob der Unterschied zwischen diesen situativen Gegebenheiten groß genug im Verhältnis zur Varianz in den beiden Bedingungen ist, um nicht mehr durch Zufall erklärt werden zu können.

Ein solches Vorgehen, bei dem die Methoden der üblichen Gruppeninferenzstatistik auf individuelle Daten angewandt werden (die damit formal den Probanden entsprechen), wurde von Gentile, Roden und Klein (1972) vorgeschlagen, jedoch mit guten Argumenten von Hartmann (1972) als unzulässig abgewiesen. Anders als im gruppenstatistischen Fall kann man hier nämlich nicht von zufällig gezogenen, voneinander unabhängigen Probandenwerten ausgehen, sondern hat bei wiederholt an einem Individuum erhobenen Daten deren mögliche serielle Abhängigkeit anzunehmen (zur genaueren Definition siehe unten). Es kann sicher nicht problemlos davon ausgegangen werden, dass die Aggressivität eines Kindes in einer Schulstunde auf die in den folgenden Stunden keinen Einfluss hat, und selbst wenn man die Aggressivität nur einmal pro Tag erhebt, dass diese Werte nicht in irgendeiner Weise miteinander zu tun haben. Die gefundene Varianz unterschätzt also möglicherweise die tatsächlich gegebene, und mit Sicherheit setzt man die Zahl der Freiheitsgrade, also der angenommenen unabhängigen Beobachtungen, zu niedrig an, geht also damit leichter von einer nicht gegebenen Überzufälligkeit aus. So interessant diese seriellen Abhängigkeiten individueller Daten im Rahmen von Zeitreihenanalysen sein mögen, bei inferenzstatistischen Prä-Post-Vergleichen auf

Einzelfallniveau stellen sie somit unzweifelhaft eine Komplikation dar.

Berücksichtigung serieller Abhängigkeiten

Ein eher zynischer Ratschlag wäre es, sich einfach von diesen Einwänden nicht stören zu lassen und trotzdem den *t*-Test zu „rechnen". Es dürfte sicher nicht ganz falsch sein, dies zu tun, um sich selbst zunächst einen Eindruck von den Interventionseffekten zu verschaffen, und wahrscheinlich ist dieser Eindruck erheblich sicherer als der auf der schieren optischen Inspektion der Daten basierend. Will man dieses Problem systematischer angehen, so bieten sich prinzipiell zwei Möglichkeiten an, die auch ohne allzu tiefes Eindringen in die Einzelfallstatistik angewendet werden können. Es ist zu betonen, dass die vorgeschlagenen Vorgehensweisen nicht im ganz strengen Sinne exakt sind, und möglicherweise bei methodologischen Puristen nicht auf uneingeschränkte Zustimmung stoßen. Allerdings besteht bekanntlich eine gewisse Diskrepanz zwischen dem, was in der empirischen Arbeit möglich und sinnvoll ist und jenen Maximalforderungen, die von Außenstehenden herangetragen werden. Man sollte sich doch zwischendrin vor Augen halten, dass stillschweigend nicht selten gewisse Regelverletzungen in Kauf genommen werden, so etwa die immer wieder vorausgesetzte Intervallskalierung der Daten keineswegs sicher gegeben ist.

Vermeidung serieller Abhängigkeiten durch geeignete Erhebungsmethoden

Im Vorfeld, im Rahmen der Versuchsplanung, lässt sich das Problem insofern schon weitgehend entschärfen, als man bei genügend großer Anzahl von Erhebungen in der Prä- wie der Post-Phase (allgemeiner: in den unterschiedlichen Untersuchungsbedingungen) bei teilweiser Nicht-Berücksichtigung von Daten immer noch genügend viele erhält, um mit diesen legitim konventionelle *t*-Tests (oder Varianzanalysen oder nonparametrische Tests) anwenden zu können. Hat man beispielsweise 30 (möglichst in gleichen Zeitabständen erfolgende) Erhebungen in der Phase vor der Intervention durchgeführt und danach immerhin noch einmal 25 Messungen, lassen sich in der Regel problemlos jeweils mindestens 5 oder 6 Werte auswählen, die in genügend großen und dabei unregelmäßigen Abständen liegen; es ist sicher vertretbar anzunehmen – auch wenn man es nicht streng beweisen kann –, dass diese Daten unsystematisch um die Prä- und pPst-Mittelwerte schwanken und so die Voraussetzungen für die Anwendung konventioneller prä-post-Vergleiche mit inferenzstatistischen Verfahren gegeben sind. Erfahrungsgemäß reicht – weil die störenden interindividuellen Unterschiede wegfallen – ein so kleiner Datensatz aus, um pädagogisch oder psychologisch relevante Effekte im Einzelfall auch statistisch als solche abzusichern. Geheimnis ist also lediglich, genügend viele Daten zu erheben, um schadlos bei der Auswertung einige opfern zu können.

Rechnerische Elimination von seriellen Abhängigkeiten (Prewhitening)

Die sicher deutlich häufigere Situation ist jedoch die, dass die Daten bereits erhoben sind und nun mit mathematischen Methoden versucht werden muss, eine eventuelle serielle Abhängigkeit zu berücksichtigen. Es handelt sich um die Verfahren des so genannten Prewhitenings, der rechnerischen Erzeugung eines „weißen Rauschens", also statistisch unabhängiger Schwankungen in den Datensätzen.

Dazu muss zunächst kurz der Begriff der Autokorrelation von Zeitreihendaten eingeführt werden (s. dazu ausführlich Köhler, 2008, S. 38 ff.). Korreliert man die Werte

einer Zeitreihe (also an einem Individuum in regelmäßigen Abständen erhobene Daten in einer Variable) mit den Werten der um eine Einheit nach rechts (oder auch – was am Ergebnis nichts ändert – nach links) verschobenen Zeitreihe derselben Variable, so erhält man den Autokorrelationskoeffizienten der Zeitreihe mit lag 1 (oder ihren lag 1-Autokorrelationskoeffizienten). Ein einfaches Beispiel: Eine (als sehr lang angenommene) Zeitreihe bestehe aus den alternierenden Werten 2 und –2, habe also die Gestalt 2, –2, 2, –2, 2, –2…. Zur Bestimmung der lag 1-Autokorrelation schreibt man in einer Doppelspalte jeweils den Wert und den ihm folgenden Wert auf, siehe Tabelle 1.

Tabelle 1: Schematische Darstellung der Zeitreihendaten zur Illustration der lag 1-Autokorrelation.

2	–2
–2	2
2	–2
–2	2
2	–2
..	..
..	..

Korreliert man die entsprechenden Spaltenwerte, ergibt sich eine lag 1-Autokorrelation von –1.

Ganz entsprechend ist die lag 2-Autokorrelation definiert: Man schreibt in die linke Spalte den 1. Wert der Zeitreihe, daneben den um zwei Einheiten verschobenen Wert (also den dritten der Zeitreihe), dann in die folgende Zeile der linken Spalte den zweiten Wert der Zeitreihe, ihm gegenüber den vierten Wert; es folgt nun in der linken Spalte der dritte Wert, ihm gegenüber der fünfte, usw. Man erhält also folgende Tabelle (Tabelle 2):

Tabelle 2: Schematische Darstellung der Zeitreihendaten zur Illustration der lag 2-Autokorrelation.

2	2
–2	–2
2	2
–2	–2
2	2
..	..
..	..

Es berechnet sich damit eine lag 2-Autokorrelation von +1 (entsprechend an dieser einfachst strukturierten Zeitreihe eine lag 3-Autokorrelation von –1 , eine lag 4-Autokorrelation von +1 ...; vorausgesetzt ist allerdings, dass die Zeitreihe so lang ist, dass durch die bei höher werdenden lags notwendige Verminderung der Werte sich Mittelwert und Varianz nicht wesentlich ändern.)

Sind diese Autokorrelationen diverser lags sämtlich niedrig, spricht man von serieller Unabhängigkeit der Zeitreihendaten. Liegt hingegen serielle Abhängigkeit vor, so lässt sich diese in mathematischen Modellen formulieren. Bei Kenntnis dieser Autokorrelationen ist es nämlich beispielsweise möglich, so genannte autoregressive Modelle (AR-Modelle) der Zeitreihe zu erstellen, d. h. eine Schätzformel zu erstellen, mittels welcher jedes Element der Zeitreihe

mit Hilfe eines oder mehrerer vorangehender Glieder vorhergesagt wird. Bei einem AR1-Modell benutzt man zur Vorhersage nur das unmittelbar vorangehende Glied, bei AR2-Modellen die beiden, welche vor dem jeweils vorherzusagenden Wert liegen. In der Zeitreihe des Beispiels würde ein AR1-Modell genügen: Jeder Wert lässt sich – ausnahmsweise sogar fehlerfrei – dadurch erhalten, dass man den unmittelbar vorangehenden Wert mit –1 multipliziert; der einzige Autoregressionskoeffizient in dem AR1-Modell wäre also –1. Wie man in weniger einfachen Fällen aus den Autokorrelationskoeffizienten die Autoregressionskoeffizienten erhält und welche Ordnung für ein solches Modell am zweckmäßigsten gewählt wird, ist in Köhler (2008, S. 67 ff.) ausgeführt und an Rechenbeispielen erläutert.

Hat man ein solches Modell für eine gegebene Zeitreihe erstellt, lässt sich für jeden Wert – mit der eventuellen Ausnahme so und so vieler Werte am Anfang, die noch zu wenige für die Prognose erforderliche Vorgänger besitzen – ein Schätzwert liefern, der auf der seriellen Abhängigkeit der Daten basiert. Subtrahiert man von diesem Schätzwert (z. B. $\hat{z}(10)$ dem Schätzer für das 10. Glied der Zeitreihe) den tatsächlichen Wert $z(10)$, erhält man einen Fehlerwert an der Stelle 10, üblicherweise bezeichnet mit $u(10)$. Diese Fehlerwerte schwanken dann unsystematisch, und ihre Varianz kann deshalb zur Anwendung des *t*-Tests in die bekannten Formeln eingesetzt werden. Es ist zu ergänzen, dass sich durch eine solche Rechenprozedur nichts an den Mittelwerten der Zeitreihen ändert; der Mittelwert der Zeitreihe wird nämlich von den einzelnen Zeitreihenwerten vor Anwendung des Eliminationsverfahrens abgezogen, sodass die Prozedur nun an einer Zeitreihe mit Mittelwert 0 vollzogen wird, der gegenüber den durchgeführten Veränderungen invariant bleibt. In die Formel für den *t*-Test muss natürlich der ursprüngliche Mittelwert eingesetzt werden.

Eine Komplikation kann im Rahmen dieser kurzen einführenden Darstellung nicht behandelt werden, nämlich das mögliche Vorliegen eines deterministischen Trends, etwa eines linearen Anstiegs der Messwerte in der pretreatment-Phase. Zwar gibt es einfache Verfahren, solche deterministischen Trends zu eliminieren, wobei allerdings wertvolle Information verschenkt wird; dass die Messwerte (z. B. in der Variable Aggressivität) in der pretreatment-Phase, also spontan, anstiegen, ist natürlich ein Sachverhalt, der beim Betrachten des Treatment-Effekts zu berücksichtigen ist.

Wir zeigen nun das Vorgehen des „prewhitening“ an zwei einfachen Beispielen (in Anlehnung an Köhler, 2008, S. 138 ff.).

Fehlende serielle Abhängigkeit in beiden zu vergleichenden Zeitreihen

Angenommen, ein Proband wurde 14-mal in wöchentlichem Abstand vor Therapie nach seinen sozialen Kontakten befragt (operationalisiert über die Zahl der gemeinschaftlich verbrachten Zeit im Laufe der jeweils vergangenen Woche). Dann sei ein vierwöchiges soziales Kompetenztraining erfolgt (ohne Messungen in diesem Zeitraum); anschließend wurde 10-mal, wieder in wöchentlichen Abständen, erneut diese Variable erhoben. Es ergaben sich die Werte in Tabelle 3.

Bezeichnet man als $k_1 = k_{\text{vorTh}}$ die Anzahl der Messwerte vor Therapie, entsprechend $k_2 = k_{\text{nachTh}}$ die Zahl der nach Abschluss der Behandlung erhobenen Messwerte, mit $\overline{x}_{\text{vorTh}}$; $s_{x\ \text{vorTh}}$; $\overline{x}_{\text{nachTh}}$; $s_{x\ \text{nachTh}}$; die Mittelwerte und Standardabweichungen in den Phasen, so ergibt sich: $k_1 = k_{\text{vorTh}} = 14$; $\overline{x}_{\text{vorTh}} = 7{,}44$; $s_{x\ \text{vorTh}} = 0{,}53$; $k_2 = k_{\text{nachTh}} = 10$; $\overline{x}_{\text{nachTh}} = 8{,}02$; $s_{x\ \text{nachTh}} = 0{,}14$; also zeigen sich nicht geringe intraindividuelle Unterschiede, deren Zufälligkeit ausgeschlossen werden sollte. Es stellt sich daher die Fra-

Tabelle 3: Soziale Kontakte, erhoben im wöchentlichen Abstand (fiktive Daten).

Wochen vor Therapie	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$x(t)$ (Score für Sozialkontakte)	6,5	7,1	7,3	7,1	7,7	8,3	6,7	7,4	7,4	7,3	8,4	7,6	7,9	7,5

Wochen nach Therapie	1	2	3	4	5	6	7	8	9	10
$x(t)$ (Score für Sozialkontakte)	8,1	7,9	8,2	7,9	8,1	8,0	7,8	7,9	8,1	8,2

ge: Hat sich die durchschnittliche Zahl an Stunden sozialer Kontakte nach Therapie signifikant erhöht?

Hierfür bietet sich der *t*-Test für unabhängige Stichproben an, vorausgesetzt, es kann von Unabhängigkeit der Fehlerwerte (hier: der Messwerte, bezogen auf ihren Durchschnitt in der jeweiligen Phase) ausgegangen werden. Der *t*-Test für korrelierende Stichproben ist nicht angezeigt, da keine sichere Zuordnung zwischen Paaren von Zeitpunkten vor und nach Therapie möglich ist; dies ist schon daraus zu ersehen, dass die Stichprobenumfänge unterschiedlich sind.

Zur Überprüfung der seriellen Abhängigkeit werden – praktischerweise nach Abzug des Zeitreihenmittelwerts von den einzelnen Zeitreihendaten – die Autokorrelationen mit verschiedenen lags bestimmt, und zwar für jede der beiden Zeitreihen getrennt. Dafür ergeben sich folgende Werte:

Autokorrelationen vor Therapie:
$r_1 = 0{,}02$; $r_2 = 0{,}06$;
Autokorrelationen nach Therapie:
$r_1 = -0{,}11$; $r_2 = -0{,}04$.

Diese Koeffizienten sind im Großen und Ganzen niedrig, sodass die Werte innerhalb der beiden Stichproben offenbar *nicht wesentlich voneinander abhängig* sind (*keine serielle Abhängigkeit zeigen*); es lassen sich deshalb die Stichprobenvarianzen zur Schätzung der Populationsvarianz heranziehen. Bekanntermaßen ist aus den beiden Stichprobenvarianzen zunächst eine gewichtete mittlere Standardabweichung zu berechnen, also

$$s = \sqrt{\frac{(k_{\text{vorTh}} - 1)\cdot s^2_{\text{vorTh}} + (k_{\text{nachTh}} - 1)\cdot s^2_{\text{nachTh}}}{k_{\text{vorTh}} + k_{\text{nachTh}} - 2}},$$

hier

$$s = \sqrt{\frac{(14-1)\cdot 0{,}28 + (10-1)\cdot 0{,}02}{14+10-2}} = 0{,}41.$$

Daraus bestimmt sich die Prüfgröße

$$t = \frac{\bar{x}_{\text{nachTh}} - \bar{x}_{\text{vorTh}}}{s} \sqrt{\frac{k_{\text{vorTh}}\cdot k_{\text{nachTh}}}{k_{\text{vorTh}} + k_{\text{nachTh}}}},$$

hier $\frac{7-8}{0{,}31} \sqrt{\frac{15\cdot 14}{15+14}} = 3{,}42,$

deren Betrag mit dem kritischen *t*-Wert bei $k_{vorTh} + k_{nachTh} - 2 = 22$ Freiheitsgraden für das gewählte Signifikanzniveau (bei hier einseitiger Fragestellung) zu vergleichen ist; statistischen Tafeln lässt sich dafür der Wert $k_{krit;\,einseitig;\,22;\,5\%} = 1{,}72$ entnehmen. Der empirische *t*-Wert überschreitet den kritischen, sodass bei dem untersuchten Probanden von einer überzufälligen Zunahme sozialer Kontakte nach Kompetenztraining auszugehen ist. Natürlich kann angesichts fehlender Kontrollbedingungen nicht sicher gefolgert werden, dass dies eine Folge des Trainings darstellt.

Zusammenfassend: Liegt in den beiden zu vergleichenden Zeitreihen *keine serielle Abhängigkeit* vor, kann zum Mittelwertvergleich der gewöhnliche *t*-Test herangezogen werden. Die Mittelwerte der Zeitreihen entsprechen den Stichprobenmittelwerten der Aggregatanalyse, die Varianzen der Zeitreihen denen der Stichproben, die Zeitreihenlängen den Stichprobenumfängen.

Serielle Abhängigkeit

Zur Abwechslung sei nun nicht eine einzige unterbrochene Zeitreihe betrachtet, sondern deren zwei (Fernsehkonsum eines Kindes geschiedener Eltern, welches die Wochenenden abwechselnd bei Mutter und Vater verbringt). Die Zeitreihen sind diesmal so konstruiert, dass sich in beiden eine serielle Abhängigkeit findet, in der ersten Zeitreihe im Sinne einer negativen lag 1-Autokorrelation, in der zweiten einer positiven. Wieder soll überprüft werden, ob unter diesen unterschiedlichen Bedingungen sich die durchschnittliche Zahl der Fernsehstunden an den Wochenenden signifikant unterscheidet. Diesmal ist, anders als im vorigen Beispiel, vor Anwendung des *t*-Tests zunächst die *serielle Abhängigkeit zu eliminieren*, um ohne wesentlichen Fehler von der Stichprobenvarianz auf die Varianz in der hypothetischen Grundgesamtheit (der prinzipiell Fernsehmöglichkeit bietenden Wochenenden in einem bestimmten Intervall) schließen zu können.

Man habe folgende fiktiven Daten erhalten (siehe Tabelle 4).

Für Mittelwert und Varianz der oberen Zeitreihe (Wochenenden bei der Mutter) berechnete sich: $\overline{x}_M = 7;\ s^2_{xM} = 0{,}13$; für die Autokorrelationen wurde bestimmt: $r_1 = 0{,}59$; $r_2 = 0{,}27$.

In der unteren Zeitreihe (Wochenende beim Vater) ergaben sich gerundet folgende Werte: $\overline{x}_V = 8;\ s^2_{xV} = 0{,}19$; die Autokorrelationen betragen: $r_1 = 0{,}56$; $r_2 = 0{,}08$.

Da die gefundenen Autokorrelationen substantiell sind, muss nun ein „prewhitening" durchgeführt werden und dazu ein geeignetes Zeitreihenmodell erstellt werden. Vorab ist die Zeitreihe so umzuschreiben, dass ihr Mittelwert 0 beträgt. Man subtrahiert also vom $x(t)$ den Mittelwert $\overline{x}_M$ und erhält einen von nun an mit $z(t)$ bezeichneten Wert.

Inspektion legt für die Zeitreihe $z_M(t) = x_M(t) - \overline{x}_M$ das AR1-Modell: $z_M(t) = -0{,}6 \cdot z_M(t-1) + u(t)$ nahe; die Schätzwerte $\hat{z}_M(t) = -0{,}6 \cdot z_M(t-1)$ und die Residuen $u(t) = z_M(t) - \hat{z}_M(t)$ finden sich in der oberen Hälfte von Tabelle 4 (Zeilen 4 und 5).

Für die Zeitreihe $z_V(t) = x_V(t) - \overline{x}_V$ scheint das AR1-Modell $z_V(t) = 0{,}56 \cdot z_V(t-1) + w(t)$ angemessen. Mit Hilfe der beiden Zeitreihenmodelle schätzen wir die Werte $\hat{z}_M(t) = -0{,}6 \cdot z_M(t-1)$ bzw. $\hat{z}_V(t) = 0{,}56 \cdot z_V(t-1)$ und berechnen daraus die Residuen $u(t) = z_M(t) - \hat{z}_M(t)$ bzw. $w(t) = z_V(t) - \hat{z}_V(t)$, welche gleichfalls in die Tabelle eingetragen werden. Als Stichprobenkennwerte dieser Residuen erhält man: $\overline{u} = 0{,}05$; $s^2_u = 0{,}07$; $\overline{w} = 0{,}03$; $s^2_w = 0{,}12$; dass die Mittelwerte der Residuen nicht den erwarteten Wert von exakt 0 ergeben, ist Folge von Rundungsfehlern. Im Sinne einer Residualanalyse bleibt nun zu prüfen, ob innerhalb der Residuen noch durch Anwendung des Modells nicht eliminierte Abhängigkeiten bestehen; dazu berechnet man die Autokorrelationen in den

Tabelle 4: Fernsehkonsum an Wochenenden bei der Mutter und beim Vater.

Wochenende bei der Mutter	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$x_M(t)$	7,20	6,76	7,36	6,74	6,58	7,29	6,50	7,65	7,14	6,80	6,94	6,64	7,35	6,65	7,44
$z_M(t) = x_M(t) - \bar{x}_M$	0,20	0,24	0,36	0,26	0,42	0,29	0,50	0,65	0,14	-0,2	0,06	0,36	0,35	0,35	0,44
$\hat{z}_M(t) = -0{,}6 \cdot z_M(t-1)$	–	0,12	0,14	0,22	0,16	0,25	0,17	0,03	0,39	0,08	0,12	0,04	0,22	0,21	0,21
$u(t) = z_M(t) - \hat{z}_M(t)$	–	0,12	0,22	0,04	0,58	0,04	0,33	0,35	0,25	0,12	0,18	0,40	0,13	0,14	0,23

Wochenende beim Vater	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$x_V(t)$	8,50	8,20	8,16	8,03	7,52	7,31	7,71	8,07	8,67	8,34	8,57	7,76	7,71	7,66
$z_V(t) = x_V(t) - \bar{x}_V$	0,50	0,20	0,16	0,03	0,48	0,69	0,29	0,07	0,67	0,34	0,57	0,24	0,29	0,34
$\hat{z}_V(t) = 0{,}56 \cdot z_V(t-1)$	0,00	0,28	0,11	0,09	0,20	0,27	0,39	0,16	0,04	0,37	0,19	0,32	0,13	0,16
$w(t) = z_V(t) - \hat{z}_V(t)$	0,50	0,08	0,05	0,06	0,50	0,42	0,10	0,09	0,63	0,03	0,38	0,56	0,16	0,18

beiden Residuumsreihen $u(t)$ und $w(t)$, wofür sich folgende Koeffizienten ergaben: $r_1(u) = -0{,}3$; $r_2(u) = 0{,}03$; $r_1(w) = 0{,}07$; $r_2(w) = 0{,}19$.

Sie sind so gering, dass die beiden neuen Zeitreihen $x_M^*(t) = \bar{x}_M + u(t)$ sowie $x_V^*(t) = \bar{x}_V + w(t)$ als aus unabhängigen Werten bestehende Stichproben aufgefasst werden können. Somit scheint nun die Anwendung des *t*-Tests für nichtkorrelierende Stichproben gerechtfertigt; insbesondere sollten auch die Varianzschätzungen nicht mit einem systematischen Fehler behaftet sein.

Zunächst wird die gewichtete mittlere Standardabweichung bestimmt mittels der Gleichung

$$s = \sqrt{\frac{(k_M - 1) \cdot s_e^2 + (k_V - 1) \cdot s_w^2}{k_M + k_V - 2}}$$, hier

$$s = \sqrt{\frac{14 \cdot 0{,}07 + 13 \cdot 0{,}12}{15 + 14 - 2}} = 0{,}31.$$

Für die Prüfgröße *t* mit $k_M + k_V - 2$ Freiheitsgraden berechnet sich damit:

$$t = \frac{\bar{x}_M - \bar{x}_V}{s} \quad \sqrt{\frac{k_M \cdot k_V}{k_M + k_V}} =$$

$$\frac{7 - 8}{0{,}31} \quad \sqrt{\frac{15 \cdot 14}{15 + 14}} = -8{,}68.$$

Den Tabellen entnimmt man als kritischen Wert für $|t|$ bei 27 Freiheitsgraden und zweiseitiger Fragestellung für ein Sig-

nifikanzniveau von $\alpha = 0{,}01$ die Zahl 2,77; somit unterscheiden sich die durchschnittlichen Fernsehstunden des Kindes signifikant zwischen Wochenenden bei der Mutter und beim Vater.

Die Berechnung der Residuen würde nicht anders geschehen, wenn autoregressive Modelle höherer Ordnung den Daten eher gerecht würden; lediglich müsste man in diesen Fällen mehr als nur das erste Zeitreihenglied durch 0 schätzen.

Zusammenfassend: Im Falle stochastischer serieller Abhängigkeiten passt man am einfachsten den Zeitreihen autoregressive Modelle an. Danach sind die Residuen zu bestimmen und – im Falle von deren Unkorreliertheit – diese statt der Ursprungswerte in die Formel für den *t*-Test einzusetzen (so genanntes Prewhitening).

Literatur

Gentile, J.R., Roden, A.H. & Klein, R.D. (1972). An analysis of variance model for the intrasubject replication design.*Journal of Applied Behavior Analysis, 5*, 193–198.

Gottman, J.M. (1981). *Time series analysis*. Cambridge, UK: Cambridge University Press.

Hartmann, D.P. (1974). Forcing square pegs into round holes: Some comments on "an analysis-of-variance model for the intrasubject replication design". *Journal of Applied Behavior Analysis, 7*, 635–638.

Köhler, T. (2008). *Statistische Einzelfallanalyse: Eine Einführung mit Rechenbeispielen*. Weinheim: Beltz.

Anschrift des Autors

Prof. Dr. Thomas Köhler
Pädagogische Universität Hamburg
Psychologisches Institut III
Von-Melle-Park 5
20146 Hamburg
thomas.koehler@uni-hamburg.de