**Dirk Büsch**

# Sequ(T)est: *Ein einfaches Statistikprogramm zum sequenziellen Testen in sportwissenschaftlichen Untersuchungen*

*SEQU(T)EST: A SIMPLE STATISTICS PROGRAM FOR SEQUENTIAL TESTING IN SPORTS SCIENCE EXAMINATIONS*

## Zusammenfassung

*Bei vielen quasi-experimentellen Untersuchungen im (Leistungs-)Sport stellt sich das Problem, dass oftmals nur wenige Sportler(innen) zur Verfügung stehen, aber zur gruppenstatistischen Absicherung von Trainingseffekten viele Probanden benötigt werden. Das sequenzielle Testen (ST) nach Wald (1947) ist eine wenig beachtete Alternative zu herkömmlichen nichtsequenziellen Tests, bei der nur so viele Daten zu erfassen sind, bis eine statistische Entscheidung gefällt werden kann. Ein entscheidender Vorteil des ST manifestiert sich in der damit verbundenen Reduktion des Stichprobenumfangs. Das Statistikprogramm Sequ(T)est bietet die Möglichkeit, die Vorteile des ST direkt zu nutzen. Mit einfachen statistischen Vorüberlegungen zum α- und β-Fehler sowie der Größe des zu erwartenden Effekts kann die/der Untersuchungsleiter(in) die Tendenz der Ergebnisse von Beginn an sowohl grafisch als auch nummerisch verfolgen sowie eine Untersuchung zum frühestmöglichen Zeitpunkt beenden. Mit Hilfe des selbst entwickelten Statistikprogramms wird für den sequenziellen Binomialtest, d.h. alternativ verteilte Merkmale (z.B. „erfolgreich" vs. „nicht erfolgreich") gezeigt, mit welchen systematischen Vorüberlegungen sowohl für in der Sportpraxis als auch in den Sportwissenschaften Tätige inhaltlich und statistisch relevante Ergebnisse zu erreichen sind. Die Möglichkeiten des ST können Schritt für Schritt mit dem selbst entwickelten und kostenfreien Statistikprogramm Sequ(T)est nachvollzogen werden.*

## Abstract

*Quasi-experimental investigations among (competitive) athletes have to deal with the inherent problem that only few subjects are available most of the time. However, many subjects are necessary for the statistical evaluation of training effects. The sequential testing according to Wald (1947) is a hardly considered alternative to conventional non-sequential tests, with the former requiring the collection of only so much data that a statistical decision can be made. A decisive advantage of ST is the according reduction of the sample. The statistical program Sequ(T)est allows to benefit directly from the advantages of sequential testing. By means of simple statistical considerations in view of α- and β-errors as well as of the expected size of the effect the tendency of the results can be observed both graphically and numerically. Additionally, the investigation can be finished as early as possible. By applying the software we developed to the binominal test, which is used with alternatively distributed features (e.g. 'successful' vs. 'unsuccessful'), we will show how systematic pre-considerations can lead to results that are relevant for people dealing with sport practice as well as for sport scientists, both with regards to the contents and to statistical procedures. The opportunities provided by ST can be explored step by step with our freeware Sequ(T)est.*

## 1 Einleitung

Eine Aufgabe der Sportwissenschaft besteht darin, Methoden zu entwickeln resp. zur Verfügung zu stellen, die sowohl in der sportlichen Praxis als auch in quasi-experimentellen Untersuchungen einfach anzuwenden sind und zu wissenschaftlich fundierten sowie validen Ergebnissen führen. Insbesondere das statistische Denken und Schließen hat sich zu einem Bereich entwickelt, der einerseits mit Argwohn betrachtet, aber andererseits als notwendige Bedingung für richtige Schlussfolgerungen aus empirischen Beobachtungen angesehen wird (Gigerenzer, 2002). Im Gegensatz dazu basiert z.B. die Entwicklung neuer Trainingsmethoden sehr häufig auf einer subjektiven Versuchs-Irrtums-Heuristik, die wissenschaftlichen Ansprüchen nicht genügt. Im schlimmsten Fall kann dabei sogar eine effiziente Trainingsmethode fälschlicherweise abgelehnt werden, obwohl Entwicklungspotenzial vorhanden ist. Für die Sportwissenschaft erscheint es daher äußerst lohnenswert, unter anderem nach statistischen Methoden Ausschau zu halten, die sowohl den Besonderheiten des Sports gerecht werden als auch einfach anzuwenden sind.

## 2 Das Problem des Testens in sportnahen Untersuchungen

Quasi-experimentelle Untersuchungen im Sport sind oftmals mit zwei Problemen konfrontiert: Erstens bereitet die geringe Anzahl von Probanden, z.B. innerhalb einer Trainings- oder Schülergruppe, auswertungstechnische Schwierigkeiten, da zur gruppenstatistischen Absicherung meistens mehr Probanden benötigt werden als Gruppenmitglieder vorhanden sind. Allerdings ist im (Leistungs-)Sport gar nicht von Interesse, ob sich die Gruppe im Mittelwert bedeutsam verbessert hat, sondern ob jede Sportlerin bzw. jeder Sportler ein vorab festgelegtes Trainingsziel erreicht hat. Als Konsequenz wird unter anderem die Verwendung von Einzelfallanalysen vorgeschlagen (Schlicht & Janssen, 1990; Lames, 1994; Hohmann & Wichmann, 2001). Die Übertragung der Einzelfallergebnisse auf (Trainings-)Gruppen ist jedoch nicht unproblematisch, so dass Alternativen zum „Mythos des Mittelwerts“ (Sixtl, 2000) in Betracht zu ziehen sind. Zweitens sind Trainer(innen) und Athlet(inn)en gegenüber der weiteren Durchführung einer neuen Trainingsmethode skeptisch, wenn schon die ersten Probanden Leistungsverschlechterungen zeigen, aber aus statistisch-methodischen Gründen das Trainingsprogramm bis zum „bitteren Ende“ durchgeführt werden muss.

Angesichts der kursorisch beschriebenen Probleme bietet das *sequenzielle Testen* (ST) eine Möglichkeit, wie die Anzahl der Probanden in Stichprobenuntersuchungen drastisch reduziert (Witte, 1977; Sachs, 2002) und bei konkreten Vorüberlegungen über die Größe des zu erwartenden und auch des zu vermeidenden (experimentellen) Trainingseffekts eine Untersuchung zum frühestmöglichen Zeitpunkt beendet werden kann. Um ein Training an dem erreichten oder nicht erreichten Effekt zu bemessen, wird hier der Vorzeichentest bzw. Binomialtest vorgeschlagen (Hartung, 1993; Sixtl, 2000). Ziel ist die Entscheidung darüber, ob ein Kriterium erreicht (+) oder nicht erreicht (-) wird. So könnte von einer Trainerin oder einem Trainer im Vorhinein formuliert werden, welche Sprunghöhenverbesserung ein Sprungkrafttraining für eine Trainingsgruppe haben muss, damit das Training als erfolgreich be-

trachtet werden kann. Im Folgenden soll unter pragmatischen Gesichtspunkten nicht der mathematisch-statistische Hintergrund des ST expliziert werden, sondern zuvorderst die untersuchungsmethodische Vorgehensweise beschrieben werden. Alle Verfahrensschritte können mit der selbst entwickelten Software *Sequ(T)est*[1] Schritt für Schritt nachvollzogen werden.

## 3 Sequenzielles Testen

Die theoretischen Grundlagen für das ST bzw. den sequenziellen Wahrscheinlichkeitsverhältnistest (*sequential probability ration test* oder SPR-Test) wurden von Wald (1945, 1947) publiziert. Sequenzielle Verfahren zeichnen sich im Wesentlichen dadurch aus, dass nach jeder sukzessiv und unabhängig erfassten Beobachtung aus einer interessierenden (endlichen) Grundgesamtheit drei statistische Entscheidungen möglich sind: (1) Annahme der Nullhypothese, (2) Annahme der Alternativhypothese und (3) Fortsetzung der Untersuchung durch Erfassung eines weiteren Datums (Indifferenzbereich). Im Gegensatz zu den „klassischen" statistischen Verfahren, bei denen der Stichprobenumfang „indirekt" über die Effektgröße definiert ist, wird der Stichprobenumfang bei sequenziellen Verfahren zu einer Variablen $n$. Der Stichprobenumfang wird durch die a priori festgelegte Effektgröße „erzwungen". Es ist jedoch möglich, den durchschnittlich maximal zu erwartenden Stichprobenumfang sowie den durchschnittlichen Stichprobenumfang für die Annahme der Null- oder der Alternativhypothese im Vorhinein abzuschätzen (Cohen, 1988; Bortz, Lienert & Boehnke, 2000). Sequenzialtests, die entweder bei der Annahme der Null- oder Alternativhypothese oder nach einer vorab festgelegten Anzahl von Beobachtungen beendet werden, bezeichnet man auch als *geschlossene Sequenzialtests* (Sixtl, 2000). Aus einer untersuchungsmethodischen Perspektive zeichnen sich sequenzielle Verfahren (auch Ergebnisfolgeverfahren oder Folgetests genannt) gegenüber den nicht-sequenziellen Tests durch eine Stichprobenersparnis von durchschnittlich 30 % und bisweilen sogar 70 % aus (Diepgen, 1996). Ein besonders auffälliges und damit auch für den Sport interessantes Merkmal des ST ist daher die Versuchsökonomie (Witte, 1977; Weber, 1980; Bortz et al., 2000).

Des Weiteren ergibt sich ein zusätzlicher Vorteil daraus, dass das Hypothesentesten beim ST dem testtheoretisch bevorzugten Signifikanztestkonzept von Neyman und Pearson (1933) folgt, und nicht, wie in der Forschungspraxis leider weit verbreitet, als „forschungsmethodischer Zwitter" oder „Hybrid- oder Mixtur-Statistik" realisiert werden kann (Diepgen, 1987; Sedlmeier, 1998; Conzelmann, 1999 u.v.m.). Beim ST können daher entsprechend nicht nur Entscheidungen zu Gunsten der Alternativhypothese mit einer zuvor festgelegten Irrtumswahrscheinlichkeit $\alpha$, sondern auch Entscheidungen zu Gunsten der Nullhypothese mit einer zuvor festgelegten Irrtumswahrscheinlichkeit $\beta$ abgesichert werden. Dafür ist es zwingend erforderlich, dass die erwartete Effektgröße aus inhaltlichen und messmethodischen Überlegun-

[1] Das Statistikprogramm *Sequ(T)est* für den ein- und zweiseitigen sequenziellen Binomialtest ist unter dem URL www.sport.uni-bremen.de/buesch als Freeware verfügbar (Büsch & Schmidt, 2003). Die Beispiele in diesem Beitrag wurden mit diesem Programm berechnet.

gen a priori festgelegt wird (für eine ausführliche Darstellung unterschiedlicher Effektgrößen vgl. Rosenthal, Rosnow & Rubin, 2000; Bortz & Döring, 2002). Die a priori Bestimmung der erwarteten Effektgröße ist zwar nicht unproblematisch, aber kommt z.B. einem systematischen Training oder einem Untersuchungsdesign mit spezifischen Hypothesen entgegen, da die Trainer(innen) und Forscher(innen) in vielen Fällen konkrete Vorstellungen über den Mindesteffekt, d.h. das konkrete Ziel eines Treatments formulieren können. Sowohl für die Sportpraxis als auch für die Sportwissenschaft ist damit eine größere Aufrichtigkeit beim statistischen Hypothesentesten – sowohl bei sequenziellen als auch bei nicht-sequenziellen Verfahren – verbunden, da die theoretisch-inhaltlich relevanten Entscheidungskriterien immer im Vorhinein bestimmt werden und nicht in fast schon ritualisierter Form im Nachhinein erfolgen (Diepgen, 1987).

Die möglichen Entscheidungen beim einseitigen ST resultieren aus zwei Gleichungen, wobei eine Gleichung über die Annahme der Nullhypothese und eine Gleichung über die Annahme der Alternativhypothese entscheidet. Dadurch kann nach jeder Beobachtung entschieden werden, ob die Null- oder Alternativhypothese angenommen wird oder eine weitere Beobachtung notwendig ist. Bei diesem Vorgehen müssen beide Gleichungen bzw. das Wahrscheinlichkeitsverhältnis nach jeder Beobachtung neu berechnet werden (Bortz et al., 2000; Sixtl, 2000; Sachs, 2002). Da es sich bei den Gleichungen aufgrund der identischen Steigungskoeffizienten um parallel verlaufende Geraden handelt, können die Gleichungen der Annahmegeraden in ein rechtwinkliges Koordinatensystem mit $n$ als Abszisse und $m$ als Ordinate eingezeichnet werden. Dabei entspricht $n$ der Anzahl der Probanden, die bisher an der Untersuchung teilgenommen haben, und $m$ der Anzahl der Personen, die das vorher festgelegte Kriterium erreicht haben. Dadurch kann die Entscheidung auf einfache Art und Weise auch grafisch erfolgen. Für jede Beobachtung wird ein Punkt mit den Koordinaten $n$ und $m$ eingetragen, so dass sich eine *Stichprobenspur* ergibt. Wenn die Stichprobenspur eine der beiden Annahmegeraden schneidet, dann kann entweder die Null- oder die Alternativhypothese angenommen werden. So lange die Stichprobenspur keine der beiden Geraden schneidet, wird die Datenerfassung fortgesetzt. Diepgen (1987) bezeichnet diese Verfahrensweise aufgrund seiner Einfachheit auch als „Entscheidungsspaziergang". Abbildung 1 veranschaulicht das grafische Entscheidungsverfahren für das einseitige ST alternativ verteilter Merkmale.

Neben dem beschriebenen einseitigen ST besteht die Möglichkeit, auch zweiseitig zu testen. In diesem Fall wird nicht nur darüber entschieden, ob ein Kriterium erreicht oder nicht erreicht wurde, sondern ob von einer konkurrierenden Alternativhypothese mit negativer Effektgröße auszugehen ist. Für das Beispiel „Sprungkrafttraining" bedeutet das zweiseitige ST, dass nicht nur darüber entschieden wird, ob sich die Sportler(innen) in der vorab definierten Größenordnung bedeutsam verbessert haben oder nicht, sondern auch, ob sich die Sportler(innen) um die vorab definierte Größenordnung verschlechtert haben. Im letzten Fall hätte das Sprungkrafttraining zu einer bedeutsamen Reduktion der Sprunghöhe geführt. Allerdings ist darauf hinzuweisen, dass aufgrund der geringeren Teststärke eines zweiseitigen Tests immer ein größerer Stichprobenumfang notwendig ist. Die Auswahl eines

zweiseitigen Tests ist jedoch immer eine theoretisch und praktisch zu begründende Entscheidung. Eine inhaltlich begründete einseitige Fragestellung ist in den meisten Fällen zu bevorzugen und entspricht auch eher einem sportpraktischen und -wissenschaftlichen Anliegen.

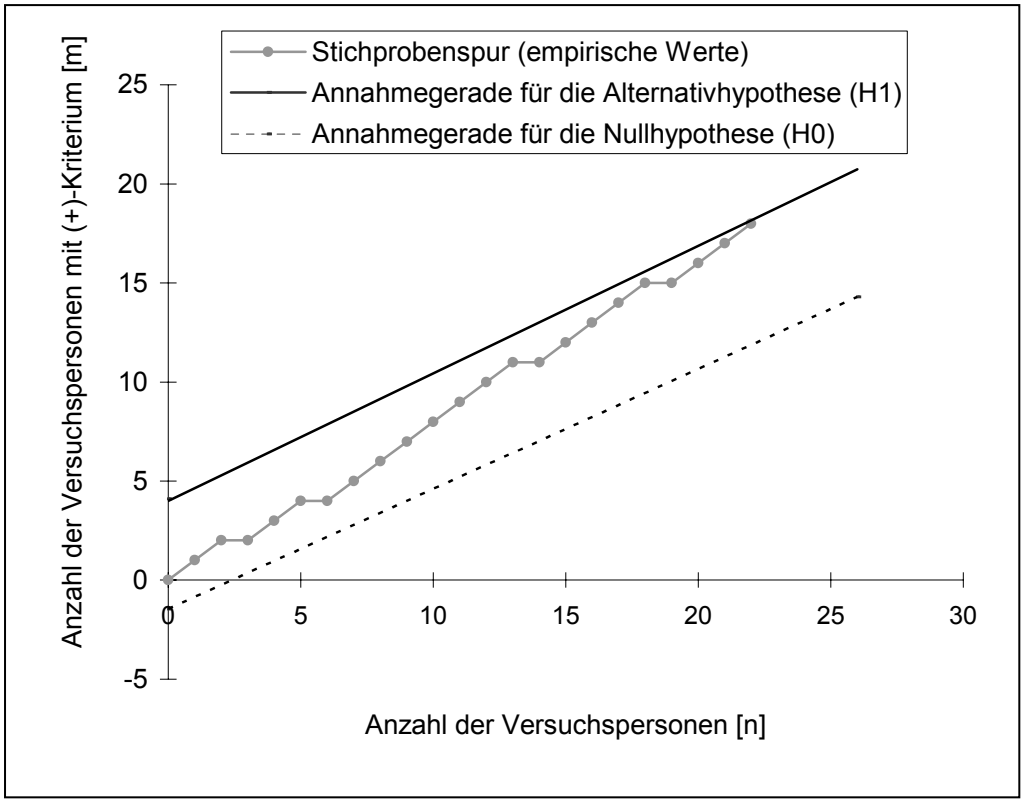


**Abb. 1: Beispiel für eine Stichprobenspur beim einseitigen Binomialtest**

## *Exkurs: Sequenzielle Paarvergleiche*

Die beschriebene zweiseitige Prüfung kann auch als sequenzieller Paarvergleich angewendet werden. Dabei gilt ein Paar als Versuchseinheit, bei der eine Athletin oder ein Athlet sukzessiv zwei unterschiedliche Trainingsmethoden ausprobiert oder zwei Athlet(inn)en nach unterschiedlichen Trainingsmethoden trainieren. Die Paarbildung erfolgt jeweils nach inhaltlichen Gesichtspunkten, wobei im zweiten Fall die Athlet(inn)en in wesentlichen Leistungskriterien zu parallelisieren sind. Für Paarvergleiche können zwei prinzipielle Testsituationen unterschieden werden (Bauer, Scheiber & Wohlzogen, 1986):

(1) Die Ergebnisse zweier Trainingsmethoden A und B können nur kategorial bewertet werden („erfolgreich" versus „nicht erfolgreich"). In diesem Fall können diskordante Paare („+,-: Plus-Einheit" oder „-,+: Minus-Einheit") und konkordante Paare, so genannte *ties* („+,+" oder „-,-") auftreten. Für eine sequenzielle Auswertung genügt es, nur die diskordanten Paare zu berücksichtigen.

(2) Die Ergebnisse zweier Trainingsmethoden A und B können quantitativ, z.B. in Zentimetern bewertet werden. Die im Vergleich bessere Trainingsmethode A wird als Plus-Einheit und die im Vergleich bessere Trainingsmethode B als Minus-Einheit bezeichnet. Konkordante Paare sind bei intervallskalierten Daten sehr unwahrscheinlich (1/∞) und können ebenfalls unberücksichtigt bleiben.

Abschließend ist anzumerken, dass sequenzielle Verfahren nicht zwingend eine Entscheidung nach jedem Datum erfordern, sondern auch Entscheidungen nach gruppierten Daten zulassen [Gruppierungspläne oder gruppensequenzielle Pläne, Wassmer, 2001; Diepgen, 1996; Bauer, 1986; Wald, 1947].

## 4 Festlegung der statistischen Parameter im Programm Sequ(T)est

Für das untersuchungspraktische Vorgehen beim einseitigen sequenziellen Binomialtest – und auch bei allen anderen Sequenzialtests – müssen vor dem Untersuchungsbeginn vier Entscheidungen getroffen werden:

(1) Wie groß soll das Risiko sein, sich fälschlicherweise für das neue Trainingsprogramm zu entscheiden, obwohl es gar nicht besser ist (*Alpha-* oder $\alpha$*-Risiko*)?

Üblicherweise werden für $\alpha$ konventionelle Schranken, d.h. Werte von 0.01, 0.05 oder 0.10 angegeben. In einigen Fällen findet man auch die Angabe 0.001. Ein kleiner Wert für $\alpha$ (0.001 oder 0.01) bedeutet eine strenge Prüfung für die Annahme der Alternativhypothese (Westermann, 2000). Damit erhöht sich die Wahrscheinlichkeit, eher die Gültigkeit der Nullhypothese anzunehmen. Beispielsweise könnte ein neu entwickeltes Krafttrainingsprogramm mit hohen finanziellen Aufwendungen, z.B. durch neue computergestützte Krafttrainingsmaschinen verbunden sein, die nur dann zu rechtfertigen sind, wenn das neue Krafttrainingsprogramm der alten Methode deutlich überlegen ist.

(2) Wie groß soll das Risiko sein, sich fälschlicherweise gegen das neue Trainingsprogramm zu entscheiden, obwohl es besser ist (*Beta-* oder $\beta$*-Risiko*)?

Für $\beta$ werden vorzugsweise Werte von 0.01, 0.05, 0.10 oder 0.20 vorgeschlagen. Ein kleiner Wert für $\beta$ (0.01 oder 0.05) entspricht einer strengen Prüfung für die Annahme der Nullhypothese. Zieht die fälschliche Annahme der Nullhypothese keine gravierenden Konsequenzen nach sich, dann kann man sich auch mit einem größeren Wert (0.10 oder 0.20) zufrieden geben (Bortz et al., 2000). Beispielsweise kann die weitere Verwendung vorhandener Krafttrainingsmaschinen zu Leistungssteigerungen führen, die etwas geringer ausfallen, aber ein kostengünstigeres Verfahren darstellen.

Aus Gründen einer anzustrebenden Test-Fairness, d.h., dass beide Arten von Fehlentscheidungen gleich gravierend erscheinen, sollten $\alpha$ und $\beta$ gleich groß

gewählt werden (Hussy & Jain, 2002). Allerdings können inhaltliche Überlegungen dazu führen, die Konsequenzen des $\alpha$- und $\beta$-Risikos unterschiedlich zu gewichten, wobei $\alpha \leq \beta$ zu wählen ist. Insgesamt entspricht $\alpha + \beta < 1$ dem Gesamtrisiko einer falschen Entscheidung durch den Sequenzialtest. Zusätzlich ist aus untersuchungsökonomischen Gesichtspunkten zu beachten, dass bei einem kleineren $\alpha$ und $\beta$ ein durchschnittlich größerer Stichprobenumfang notwendig ist. Entsprechende Überlegungen zu $\alpha$ und $\beta$ sind daher nicht nur inhaltlich notwendig, sondern auch ökonomisch sinnvoll.

(3) Wie groß soll die Wahrscheinlichkeit sein, mit der *kein* Unterschied innerhalb der zu untersuchenden Stichprobe angenommen wird (Wahrscheinlichkeit für die Nullhypothese, $P_0$)?

$P_0$ entspricht der Wahrscheinlichkeit, mit der kein Unterschied innerhalb der zu untersuchenden (endlichen) Grundgesamtheit bzw. die Gültigkeit der Nullhypothese angenommen wird. Mit anderen Worten: Die Häufigkeiten der Plus- und Minus-Kriterien (+ vs. -) unterscheiden sich in der untersuchten Stichprobe und damit in der (endlichen) Grundgesamtheit von Sportler(inne)n nicht ($H_0$) bzw. überschreiten nicht ein zuvor festgelegtes Kriterium. In der Vielzahl der Fälle gilt hierbei die Leibnizbedingung 0.5 (Stegmüller, 1983). D.h., dass der „Unterschied" in der Stichprobe zwischen zwei Kriterien nicht größer als eine Ratewahrscheinlichkeit von 50 % ist. Aufgrund theoretisch-inhaltlicher Überlegungen kann $P_0$ auch kleiner oder größer als 0.50 festgelegt werden.

(4) Wie groß soll die Wahrscheinlichkeit sein, mit der ein Unterschied innerhalb der zu untersuchenden Stichprobe angenommen wird (Wahrscheinlichkeit für die Alternativhypothese, $P_1$)?

$P_1$ entspricht der Wahrscheinlichkeit, mit der ein Unterschied innerhalb der (endlichen) Grundgesamtheit bzw. die Gültigkeit der Alternativhypothese angenommen wird. Die Festlegung von $P_1$ entspricht der Festlegung einer Effektgröße $g$, um die sich die angenommenen Parameter für die Null- und Alternativhypothese unterscheiden müssen, um von einem praktisch bedeutsamen Unterschied sprechen zu können. Dieser Wert muss grundsätzlich größer sein als $P_0$. Aus vorliegenden Untersuchungen oder konkreten Erwartungen an ein spezifisches Training sollte $g$ bestimmbar sein. Ansonsten empfiehlt sich unter der Bedingung $P_0 = 0.50$ ein Wert von $P_1 = 0.55$ (entspricht einer kleinen Effektgröße $g = 0.05$, Cohen, 1988), da somit zumindest die unspezifischen Erfahrungen zu Gunsten einer Ausprägung definiert sind. Des Weiteren können unter der Bedingung $P_0 = 0.50$ Werte für $P_1 = 0.65$ mit einer mittleren Effektgröße ($g = 0.15$) und $P_1 = 0.75$ mit einer großen Effektgröße ($g = 0.25$) festgelegt werden (Bortz, Österreich & Vogelbusch, 1979; Cohen, 1988; Bortz & Döring, 2002). Die unterschiedlichen Effektgrößen sind jedoch nicht als obligatorisch, sondern allenfalls als Hilfestellung oder bewährte Konvention zu verstehen. Die Entscheidung über die a priori festzulegende Effektgröße bleibt letztendlich vom Einzelfall abhängig. Unter ökonomischen Gesichtspunkten ist dabei zu berücksichtigen, dass eine größere Differenz zwischen $P_0$ und $P_1$ im Durchschnitt einen geringeren Stichprobenumfang erfordert.

Für das untersuchungsmethodische Vorgehen beim zweiseitigen sequenziellen Binomialtest müssen vor dem Untersuchungsbeginn fünf Entscheidungen getroffen werden. Zusätzlich zu den Entscheidungen (1) bis (4) des einseitigen Testens muss noch die Wahrscheinlichkeit für die konkurrierende Alternativhypothese ($P_1$-) festgelegt werden. Hierfür wird die Effektgröße von der Wahrscheinlichkeit für die Nullhypothese subtrahiert.

Nachdem die Vorentscheidungen für $\alpha$, $\beta$, $P_0$ und $P_1$ (beim zweiseitigen Testen noch $P_{1-}$) getroffen wurden, müssen die Beobachtungsdaten $n$ sowie die Daten mit dem Plus-Kriterium *m sequenziell* (!) eingegeben werden. Nach jeder Beobachtung kann jetzt grafisch entschieden werden, ob die Null- oder die Alternativhypothese angenommen werden kann oder weitere Beobachtungen notwendig sind. Das praktische Vorgehen wird abschließend an einem konkreten Beispiel erläutert. Im Programm *Sequ(T)est* ist der entsprechende Beispieldatensatz enthalten, so dass die Verfahrensschritte und Ergebnisse einfach nachvollzogen werden können.

## 5 Ein konkretes Beispiel für den einseitigen sequenziellen Binomialtest

In einer Jahrgangsstufe eines Gymnasiums soll untersucht werden, ob die Schülerinnen aufgrund eines sehr umfangreichen Sportangebots besser als der Durchschnitt sind. Der Normtabelle für den Jump-and-Reach-Test von Beck und Bös (1995) ist zu entnehmen, dass Schülerinnen im Alter von 14 Jahren mit ≥ 40 cm eine überdurchschnittliche und mit 32 bis 39 cm eine durchschnittliche Sprunghöhe haben. Es wird daher die Erwartung formuliert, dass mit einer Wahrscheinlichkeit von 75 % ($P_1 = 0.75$, $g = 0.25$) die Schülerinnen deutlich besser als der Durchschnitt ($P_0 = 0.5$) sind, d.h. mindestens 40 cm hoch springen. Insgesamt soll eine strenge Prüfung durchgeführt werden, da der hohe finanzielle Aufwand für den zusätzlichen Sportunterricht gegenüber der Schulbehörde sonst nicht mehr zu rechtfertigen ist. Aus diesen Überlegungen ergibt sich konsequenterweise ein Alpha-Risiko von $\alpha = .01$. Gleichzeitig ist davon auszugehen, dass ein umfangreicher und abwechslungsreicher Sportunterricht weitere positive Effekte für die Schülerinnen hat. So kann z.B. vermutet werden, dass ein zeitlich umfangreicherer Sportunterricht Gesundheit und Wohlbefinden der Schülerinnen verbessert. Daher erscheint ein Beta-Risiko von $\beta = .20$ angemessen.

Im Vorfeld der Untersuchung soll geklärt werden, wie viele Personen in Abhängigkeit vom wahren (aber unbekannten) Parameter $P$ erforderlich sein werden, um eine statistische Entscheidung zu treffen. Im günstigsten Fall wäre die Alternativhypothese unter der Bedingung, dass alle Probanden das Kriterium erreicht haben ($m = n$), bei $n = 11$ und die Nullhypothese unter der Bedingung, dass niemand das Kriterium erreicht hat ($m = 0$), bei $n = 3$ anzunehmen. Der im ungünstigsten Fall durchschnitt-

lich maximal zu erwartende Stichprobenumfang beträgt $n = 25$. Unter den beschriebenen Rahmenbedingungen erscheint die Untersuchung akzeptabel[2].

Die sequenzielle Erfassung der Jump-and-Reach-Leistungen von $n = 22$ Personen erbringt folgendes Resultat (1 = Kriterium erfüllt, d.h. mindestens 40 cm hoch gesprungen, 0 = Kriterium nicht erfüllt):

11011 01111 11101 11101 11

Die sequenziellen Daten werden in einen grafischen Testplan überführt, aus dem zu entnehmen ist, dass nach 22 Beobachtungen die Untersuchung abgebrochen werden kann und die Alternativhypothese anzunehmen ist (siehe Abbildung 1). Das Fazit lautet, dass die Schülerinnen mit einem umfangreicheren Sportangebot statistisch bedeutsam höher springen als der Durchschnitt bzw. eine überdurchschnittliche Sprunghöhe haben.

In einem nicht-sequenziellen Design wäre unter der Annahme, dass ein großer bedeutsamer Effekt ($g \geq 0.25$) zugrunde gelegt sowie ein identisches $\alpha$-Risiko von .01 und eine Teststärke von $1 - \beta = .80$ berücksichtigt wird, eine Gesamtstichprobe von mindestens $N = 37$ notwendig gewesen, um einen statistisch und praktisch bedeutsamen Effekt nachweisen zu können (Cohen, 1988; Bortz & Döring, 2002). Durch das sequenzielle Testen konnte der Stichprobenumfang für eine statistisch abgesicherte Aussage um 40 % reduziert werden.

## 6 Abschließende Bemerkungen

Auf die Bestimmung der minimalen und durchschnittlichen Anzahl von Beobachtungen (ASN-Funktion, „average sample number“) für eine Entscheidung zu Gunsten der Null- oder Alternativhypothese sowie den durchschnittlich maximal zu erwartenden Stichprobenumfang (Weber, 1980; Diepgen, 1996; Bortz et al., 2000) wurde in diesem Beitrag nicht eingegangen. Die Funktion ist jedoch in die Software *Sequ(T)est* implementiert und beschrieben, so dass auch die Berechnung für das Beispiel einfach nachvollzogen werden kann.

Wie die meisten statistischen Verfahren besitzt Sequ(T)est neben den beschriebenen Vorteilen auch gewisse Nachteile. Zum einen existiert keine allgemein verbindliche Regel für die Festlegung der Parameter. Anwender(innen) sind daher verpflichtet, die Kriterien für ihre theoretisch-inhaltlich begründeten Entscheidungen explizit zu dokumentieren. Zum anderen können mit Sequ(T)est bzw. dem implementierten Binomialtest keine Moderatorvariablen oder Interaktionen berücksichtigt werden. Für entsprechende Untersuchungsdesigns stehen die unterschiedlichsten sequenziellen statistischen Verfahren, z.B. auch F-Tests (für einen Überblick siehe Bauer et al., 1986; Whitehead, 1997) sowie kommerzielle Statistik-Programme zur Verfügung, die überwiegend in der klinischen Forschung verwendet werden. Am bekanntesten ist

[2] Die Berechnung der unterschiedlichen Stichprobenumfänge kann im Programm *Sequ(T)est* nach der Eingabe von $\alpha$, $\beta$, $P_0$ und $P_1$ nachvollzogen werden. Aus didaktischen Gründen wird an dieser Stelle auf die Darstellung der mathematischen Formeln verzichtet.

das Programm PEST (*Planning & Evaluation of Sequential Trials*), das federführend von John Whitehead seit 1985 in der *Medical and Pharmaceutical Statistics Research Unit* an der Universität Reading entwickelt wurde und mittlerweile in der Version 4.2 zur Verfügung steht. Des Weiteren können sequenzielle Designs mit dem Zusatzmodul S+SeqTrial im Statistik-Programmpaket S-PLUS (Version 6) analysiert werden.

## *Literatur*


Bauer, P., Scheiber, V. & Wohlzogen, F. X. (1986). *Sequentielle statistische Verfahren*. Stuttgart: Gustav Fischer.

Beck, J. & Bös, K. (1995). *Normwerte motorischer Leistungsfähigkeit*. Köln: Sport & Buch Strauss.

Bortz, J. & Döring, N. (2002). *Forschungsmethoden und Evaluation* (3. Aufl.). Berlin: Springer.

Bortz, J., Lienert, G. A. & Boehnke, K. (2000). *Verteilungsfreie Methoden in der Biostatistik* (2. Aufl.). Berlin: Springer.

Bortz, J., Österreich, R. & Vogelbusch, W. (1979). Die Ermittlung optimaler Stichprobenumfänge für die Durchführung von Binomial-Tests. *Archiv für Psychologie, 131*, 267-292.

Büsch, D. & Schmidt, R. (2003). *Sequ(T)est - ein sequenzieller Binomialtest für (sport-)wissenschaftliche Fragestellungen (Version 1.0) [Computer Software]. Bremen: Universität Bremen, Studiengang Sport.* Verfügbar unter: http://www.sport.uni-bremen.de/buesch

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hilldale, N.J.: Erlbaum.

Conzelmann, A. (1999). Grundlagen der Inferenzstatistik. In B. Strauß, H. Haag & M. Kolb (Hrsg.), *Datenanalyse in der Sportwissenschaft* (S. 213-276). Schorndorf: Hofmann.

Diepgen, R. (1987). Droje voor dropje. Oder: Sequentialstatistik, die ignorierte Alternative. *Zeitschrift für Sozialpsychologie, 18*, 19-27.

Diepgen, R. (1996). Sequentielles Testen. In G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 137-144). Weinheim: Beltz, PVU.

Gigerenzer, G. (2002). *Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken*. Berlin: Berlin Verlag.

Hartung, J. (1993). *Statistik: Lehr- und Handbuch der angewandten Statistik* (9. Aufl.). München: Oldenbourg.

Hohmann, A. & Wichmann, E. (2001). Die DEL-Analyse – eine Methode zur Trainingswirkungsanalyse. *Sportwissenschaft, 31* (2), 173-187.

Hussy, W. & Jain, A. (2002). *Experimentelle Hypothesenprüfung in der Psychologie*. Göttingen: Hogrefe.

Lames, M. (1994). Zeitreihenanalysen in der Trainingswissenschaft. *Spectrum der Sportwissenschaften, 6* (1), 27-50.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, 231*, 289-337.

Rosenthal, R., Rosnow, R. L. & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research*. Cambridge: University Press.

Sachs, L. (2002). *Statistik* (10. Aufl.). Berlin: Springer.

Schlicht, W. & Janssen, J.-P. (1990). Der Einzelfall in der empirischen Forschung der Sportwissenschaft: Begründung und Demonstration zeitreihenanalytischer Methoden. *Sportwissenschaft, 20* (3), 263-280.

Sedlmeier, P. (1998). Was sind die guten Gründe für Signifikanztests? *Methods of Psychological Research - Online, 3* (1), 39-42. Verfügbar unter: www.mpr-online.de.

Sixtl, F. (2000). *Der Mythos des Mittelwerts. Neue Methodenlehre der Statistik* (2. Aufl.). München: Oldenbourg.

Stegmüller, W. (1983). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie.* (2. Aufl.). Berlin: Springer.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annales of Mathematical Statistics, 16*, 117-186.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wassmer, G. (2001). *Statistische Testverfahren für gruppensequenzielle und adaptive Pläne in klinischen Studien* (2. Aufl.). München: Alexander Mönch.

Weber, E. (1980). *Grundriß der Biologischen Statistik* (8. Aufl.). Stuttgart: Fischer.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik*. Göttingen: Hogrefe.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials* (2nd ed.). Chichester: John Wiley & Sons.

Witte, E. H. (1977). Zur Logik und Anwendung der Inferenzstatistik. *Psychologische Beiträge, 19*, 290-303.

**Autorenhinweis:**
Ich bedanke mich bei Rüdiger Schmidt für die Programmierung der Software sowie bei einem anonymen Gutachter für einige hilfreiche Hinweise zu einer früheren Version dieses Beitrages.