Neyman-Pearson-Theorie statistischen Testens

Klaus Willmes

Häufig werden empirische Untersuchungen in der Psychologie ausgeführt, um psychologische Hypothesen empirisch zu prüfen. Nach dem Definitionsvorschlag von Hager (1992, S. 11) stellen psychologische Hypothesen, die sich auf psychisches Erleben und Verhalten beziehen, eine Annahme oder Behauptung über kausale oder nichtkausale Beziehungen und Zusammenhänge zwischen psychologischen Konstrukten, Variablen und Sachverhalten dar. Hussy und Möller (1994) geben einen Überblick über verschiedene Arten wissenschaftlicher Hypothesen. Obwohl sich psychologische Hypothesen (PHn) üblicherweise auf psychische Vorgänge bei einzelnen Individuen beziehen, werden sie sehr häufig unter Verwendung statistischer (Test-)Verfahren überprüft, die zur Prüfung statistischer Hypothesen (SHn) geeignet sind, welche sich auf über Beobachtungseinheiten aggregierte Daten, auf Aggregat- oder Populationsaussagen beziehen. Wie Bredenkamp (1980) sowie Erdfelder und Bredenkamp (1994) erläutert haben, können Aggregataussagen zwar notwendige, aber keine hinreichenden Bedingungen für die Gültigkeit von allgemeinpsychologischen PHn darstellen, die ja für *einzelne* Individuen gelten sollen. Es ist also stets zu analysieren, ob eine statistische Aggregathypothese aus der allgemeinpsychologischen Hypothese logisch folgt. Erdfelder und Bredenkamp (1994) weisen z.B. für eine aus der Theorie der dualen Kodierung abgeleitete PH: „Imaginale Verarbeitungsprozesse führen im Vergleich zu sprachlichen Verarbeitungsprozessen zu besseren Gedächtnisleistungen" nach, daß neben dem Konstanthalten von Störeinflüssen die experimentell randomisierte Zuordnung der Pbn zu den experimentellen Bedingungen erforderlich ist, um die Implikationsbeziehung PH $\Longrightarrow$ SH zu sichern. Enthält die PH, wie häufig und auch in dem Beispiel der Fall, (universelle) theoretische Begriffe, ist der Theorie-Empirie-Überbrückungsvorgang um einen Zwischenschritt zu ergänzen, indem in einem Schritt der Operationalisierung (Gadenne, 1994) unter Hinzunahme von Hilfshypothesen aus der PH eine empirische Hypothese (EH, Hager, 1987, 1992) abgeleitet wird (im Beispiel eine EH über empirische Indikatoren der Gedächtnisleistung), die dann die SH(n) impliziert. Eine weitere „Komplikation" tritt ein, wenn die EH nicht adäquat in eine einzelne SH umgesetzt werden kann, sondern in eine statistische Vorhersage (SV), die eine Menge (Familie) von testbaren SHn umfaßt (Hager, 1987, 1992; Westermann & Hager, 1986). Dann handelt es sich um ein sogenanntes Mehrentscheidungsproblem.

Nachfolgend werden einige Grundgedanken einer prominenten statistischen Testtheorie, der Testtheorie von J. Neyman und E. S. Pearson vorgestellt, die zur statistischen Entscheidung über eine PH verwendet werden kann. Sie ist geeignet, eine „strenge" und „faire" Prüfung von SHn zu ermöglichen, da die Wahrscheinlichkeit einer fälschlichen Bewährung sowie einer fälschlichen Nichtbewährung der in

https://doi.org/10.25521/HQM09

einer empirischen Untersuchung zu prüfenden PH kontrolliert werden kann. Diese Strategie entspricht einem methodologischen Falsifikationismus (Lakatos, 1974), der zuläßt, daß eine empirisch nicht bewährte Theorie dennoch wahr sein kann. Dazu werden im ersten Abschnitt statistische Grundbegriffe bereitgestellt, bevor im zweiten Abschnitt Grundzüge der Neyman-Pearson-Theorie (NPT) dargestellt werden. An dessen Ende steht eine kurze Analyse der Eignung der NPT für eine deduktive Methodologie empirischer Untersuchungen in der Psychologie. Den Abschluß bilden einige Empfehlungen zur weiterführenden Lektüre.

1 Statistische Grundlagen

1.1 Zufallsvariablen

Ziel einer mathematischen Theorie des Testens von SHn, wie sie in der Mathematischen Statistik betrachtet wird, ist es, Entscheidungen unter Unsicherheit, d.h. statistische Entscheidungen, die auf zufallsabhängigen Ergebnissen (Daten) beruhen, zum Gegenstand mathematischer Überlegungen zu machen. Grundlegend dafür ist die Wahrscheinlichkeitstheorie (sowie gewisse Maß- und integrationstheoretische Hilfsmittel; Literaturhinweise in Steyer, 1994).

Die formale Repräsentation der empirischen Phänomene in einem Zufallsexperiment (Werfen einer Münze; Lösen einer Testaufgabe durch einen Pb) stellt der Wahrscheinlichkeitsraum dar (Steyer, 1994). Er besteht aus einer Menge Ω der (möglichen) Ergebnisse, der Menge der (möglichen) Ereignisse $\mathcal{A}$ und einem Wahrscheinlichkeitsmaß P, das jedem (möglichen) Ereignis $A \in \mathcal{A}$ eine Wahrscheinlichkeit $P(A)$ mit einem Wert zwischen 0 und 1 zuordnet. Die Menge der (möglichen) Ereignisse kann die Menge aller Teilmengen von Ω (Potenzmenge von Ω) sein, aber auch jedes gegenüber abzählbarer Vereinigungs- und Durchschnittsbildung abgeschlossene Teilmengensystem von Ω (sog. σ-Algebra). Um eine einfachere Beschreibung eines Zufallsexperimentes zu erreichen, werden Zufallsvariablen (stochastische Variablen) definiert. Eine Zufallsvariable (ZV) X bildet die Ergebnisse $\omega \in \Omega$ im allgemeinen auf reelle Zahlenwerte ab; $x = X(\omega)$ heißt Realisierung (Ausprägung) von X. Es gibt diskrete ZVn mit endlich vielen Realisierungen x_i und den durch die ZV induzierten Wahrscheinlichkeiten $p_i = P(\omega_i) = P(X(\omega_i))$. Bei stetigen ZVn ist der Vorgang in mathematisch exakter Form komplizierter darzustellen. In diesem Fall ist für den Wertebereich (Bildraum) von X ein Wahrscheinlichkeitsmaß dafür definiert, daß X Zahlenwerte in einem Intervall annimmt, inbesondere $P(\{(-\infty, x)\})$. Bei statistischen Entscheidungsproblemen interessiert man sich nicht für X selbst, sondern für seine Verteilung, die eindeutig charakterisierbar ist durch seine Verteilungsfunktion

$$F^X(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\}). \tag{1}$$

Für eine stetige (z.B. normalverteilte) ZV kann zusätzlich zu F^X in der Regel die Dichtefunktion f^X angegeben werden mit $F^X(x) = \int_{-\infty}^{x} f^X(y)dy$. Mit dem Übergang vom ursprünglichen Wahrscheinlichkeitsraum zu den reellen Zahlen (genauer: zu einem Meßraum Ξ) hat man die Unabhängigkeit des Modells vom speziellen Anwendungsfall erreicht. Deshalb geht man im allgemeinen von der ZV X als der primär gegebenen Größe aus mit einer Verteilung P^X, charakterisiert

durch ihre Verteilungsfunktion $F^X(x)$. Die ZV X muß nicht eindimensional sein. Das Ergebnis eines Zufallsexperiments kann nun durch einen p-dimensionalen Zahlenvektor $\mathbf{x} = (x_1, \ldots, x_p)$ charakterisiert sein. Auch in diesem Fall sind für die p-dimensionale ZV $\mathbf{X}$ deren p-dimensionale Verteilungsfunktion und u.U. die zugehörige p-dimensionale Dichte definiert (im Fall unabhängiger Beobachtungen das „Produkt" der eindimensionalen Verteilungen der X_i).

1.2 Stichprobenraum

Grundlegend für die Modellierung eines (psychologischen) Zufallsexperiments ist der Stichprobenraum (von R. A. Fisher eingeführt). Üblicherweise werden in einer empirischen Untersuchung mehrere ($n > 1$) Beobachtungen gewonnen; man hat somit n Daten $x_1, \ldots, x_n$. Im einfachsten Fall sind alle Daten voneinander unabhängige Realisierungen derselben ZV X. Der Datenvektor $\mathbf{x} = (x_1, \ldots, x_n)$ heißt Stichprobe vom Umfang n als Realisierung des Zufallsvektors $\mathbf{X} = (X_1, \ldots, X_n)$ und ist damit ein Punkt in einem n-dimensionalen (Euklidischen) Raum. Ist die Verteilung von X bekannt, so läßt sich die (gemeinsame) Verteilung von $(X_1, \ldots, X_n)$ bestimmen. Die n ZVn müssen nicht identisch verteilt sein. Beobachtet man verschiedene Pbn unter k verschiedenen experimentellen Bedingungen (k-Stichprobenproblem), ist $\mathbf{X} = (X_{11}, \ldots, X_{1n_1}, \ldots, X_{k1}, \ldots, X_{kn_k})$ mit voneinander unabhängigen X_{ij} je experimenteller Bedingung i verteilt nach $F_i(x)$. Stammen alle oder ein Teil der Beobachtungen von einem Pbn, sind Abhängigkeiten zwischen den ZVn anzunehmen.

1.3 Parameterraum

Falls die Verteilung der ZV X vollständig bekannt ist, kann man auch die Verteilung von $\mathbf{X}$ vollständig bestimmen. Häufig hat man aber nur die Möglichkeit anzunehmen, daß die Verteilung von X Element einer bestimmten Klasse von Verteilungen über dem Stichprobenraum ist, welche durch einen ein- oder m-dimensionalen Parameter θ charakterisiert ist. Die Gesamtheit Θ aller (im Modell) möglichen Parameterwerte $\theta \in \Theta$, der Parameterraum, sei eine Teilmenge des m-dimensionalen Euklidischen Raums. Eine Verteilungsklasse ist die der Normalverteilungen mit unbekannten Parameterwerten aus dem zweidimensionalen Parameterraum (μ, σ^2) mit $\sigma^2 > 0$ oder die Klasse der (diskreten) Binomialverteilungen mit Parameterraum (n, p), wobei n eine (feste) natürliche Zahl und $0 < p < 1$ die unbekannte Wahrscheinlichkeit ist. Man nennt die Spezifizierung einer (m-parametrigen) Verteilungsklasse auch statistische Oberhypothese. Im obigen Beispiel mit k Bedingungen könnte der Parameterraum $(\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2)$ sein mit der zusätzlichen Einschränkung durch die Annahme identischer Varianzen $\sigma_1^2 = \ldots = \sigma_k^2$ (also $m = k + 1$). Für viele Anwendungsfälle in der Psychologie ist die Annahme einer parametrisierten Klasse von Wahrscheinlichkeitsverteilungen nicht sinnvoll, sondern lediglich die Festlegung, daß es sich um eine stetige Verteilung aus der (sehr großen) Funktionenklasse eindimensionaler stetiger Verteilungen (mit Dichtefunktion) handelt. In diesem Fall spricht man von einer nichtparametrischen Verteilungsannahme.

1.4 Statistische Hypothesen

Unter einer statistischen Hypothese H (oder auch H_0) versteht man bei einer parametrisierten Verteilungsklasse eine spezielle Annahme über die Verteilung einer ZV und über den Parameter θ, welcher vollständig oder (meistens nur) teilweise spezifiziert wird. Diese (nichtleere) Teilmenge des Parameterraumes Θ wird häufig mit Θ_0 (oder auch mit H) bezeichnet. Beispielsweise könnte es in einem Einstichprobenproblem mit Normalverteilungen als Verteilungsklasse die Menge aller Parameterpaare $\theta_0 = (\mu_0, \sigma^2)$ sein, mit einem festen Wert μ_0 und nicht näher spezifiziertem $\sigma^2 > 0$. Für die Klasse der Binomialverteilungen könnte es $\theta_0 = (n, p_0)$ mit $p_0 = 0.5$ sein. In einer nichtparametrischen Verteilungsannahme könnte es die Einschränkung auf um Null symmetrische Verteilungsfunktionen sein. Ganz allgemein kann man jede Annahme über die Verteilung von X als statistische Hypothese bezeichnen. Eine Hypothese heißt „einfach", wenn durch sie die unbekannte Verteilung von X eindeutig festgelegt wird (Θ_0 enthält dann nur ein Element). Alle anderen Hypothesen nennt man „zusammengesetzt". Im obigen Fall $\theta_0 = (\mu_0, \sigma^2)$ handelt es sich also um eine zusammengesetzte Hypothese, die durch die Spezifizierung $\sigma^2 = \sigma_0^2$ zu einer einfachen Hypothese wird. Auch bei $\theta_0 = (\mu \leq 0, \sigma_0^2)$ handelt es sich um eine zusammengesetzte Hypothese.

1.5 Statistischer Test

Eine PH kann so beschaffen sein, daß aus ihr per Implikation als SH in einem bestimmten Wahrscheinlichkeitsmodell eine spezielle Annahme über den Parameterraum einer parametrisierten Verteilungsklasse folgt, deren Zutreffen anhand der Daten aus einem Zufallsexperiment beurteilt werden soll. Für diese Beurteilung sollte zudem festgelegt sein, mit welchen Unsicherheiten oder Urteilsfehlern zu rechnen ist und wie diese eventuell zu kontrollieren sind.

Mit einem statistischen Test soll also eine Entscheidung zwischen verschiedenen Aussagen, d.h. zwischen verschiedenen Hypothesen über den Parameter θ getroffen werden, von denen für jeden (möglichen) Wert von θ genau eine zutrifft. Die Menge der Entscheidungen Δ für eine der Hypothesen (zusammen mit einer über Δ definierten σ-Algebra) wird Entscheidungsraum genannt. Bei einem (nichtrandomisierten) statistischen Test handelt es sich um eine (zweiwertige) Entscheidungsfunktion, die jedem Element des Stichprobenraums genau eine der beiden Entscheidungen für die Hypothese H_0 oder H_1 – mit dH_0 und dH_1 bezeichnet – zuordnet. Der Stichprobenraum wird damit ebenfalls in zwei sich nicht überschneidende Mengen A (mit Entscheidung dH_0 für $x \in A$) und C, dem Komplement von A ($C = A^c$, mit Entscheidung dH_1 für $x \in C$), aufgeteilt.

Häufig ist aufgrund der psychologischen Fragestellung eine der Hypothesen ausgezeichnet; diese wird oft als Nullhypothese bezeichnet. C heißt dann kritischer Bereich (Ablehnungsbereich), sodaß die Beobachtung von $x \in C$ zum „Verwerfen der Nullhypothese" führt. A wird als Annahmebereich des Tests bezeichnet. Einfache Beispiele für solche Hypothesenpaare H_0 und H_1 sind die einfache $H_0 = \{\theta_0\}$ und die zusammengesetzte $H_1 = \{\theta : \theta \neq \theta_0\}$ in einem zweiseitigen Test oder ein einseitiger Test mit den beiden zusammengesetzten Hypothesen $H_0 = \{\theta : \theta \geq \theta_0\}$ und $H_1 = \{\theta : \theta < \theta_0\}$. Neben der Zerlegung des Parameterraums in zwei komplementäre

Mengen ist auch der Fall zweier einfacher Hypothesen $H_0 = \{\theta_0\}$ und $H_1 = \{\theta_1\}$ von Interesse.

Für die Entscheidung darüber, ob eine Stichprobe $\mathbf{x}$ im Annahme- oder Ablehnungsbereich liegt, muß man aus ihr Informationen über die Verteilung von $\mathbf{X}$ und deren unbekannte(n) Parameter gewinnen. Dazu verwendet man Stichprobenfunktionen, die auch Statistik(en) T genannt werden. Bei ihnen handelt es sich um (meßbare) Abbildungen des Stichprobenraumes in einen anderen ein- oder mehrdimensionalen (meßbaren Euklidischen) Raum. Eine Statistik T ist als Funktion der ZV $\mathbf{X} = (X_1, \ldots, X_n)$ selbst wieder eine ZV $T = T(\mathbf{X}(\omega))$ mit Werten $t(\mathbf{x})$. Wenn die Verteilung(sklasse) von $\mathbf{X}$ (vollständig) spezifiziert ist, ist das auch für die durch T induzierte Verteilung(sklasse) der Fall, und die Verteilung von T ist im parametrischen Fall im allgemeinen auch wieder von θ abhängig. Zwei einfache Beispiele sollen das erläutern:

1. Seien $X_1, \ldots, X_n$ unabhängig voneinander und jeweils binomialverteilt mit $n = 1$ und Wahrscheinlichkeitsparameter p. Der Stichprobenraum kann repräsentiert werden durch die Eckpunkte eines n-dimensionalen Einheitswürfels. Mögliche Ereignisse sind alle Teilmengen von Eckpunkten. Als Statistik kann man, um Informationen über die Stichprobenverteilung von $\mathbf{X}$ und damit über den Parameter $\theta = p$ zu gewinnen, $T(\mathbf{X}) = \sum X_i$ betrachten. Als Parameterraum wählt man im allgemeinen $\Theta = \{p : 0 < p < 1\}$. $T(\mathbf{X})$ folgt (als induzierter Klasse von Wahrscheinlichkeitsverteilungen) einer Binomialverteilung mit den Parametern n und p (abgekürzt $B(n,p)$). Wie für andere ZVn auch, kann man Erwartungswert und Varianz von T bestimmen: $E(T) = np$ und $Var(T) = np(1-p)$.

2. Als Modell für die unabhängig und identisch verteilten $X_1, \ldots, X_n$ werde jeweils eine Normalverteilung mit unbekanntem μ und σ^2 angenommen, abgekürzt $N(\mu, \sigma^2)$. Die Information in den Daten wird erfaßt durch die zweidimensionale Statistik $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X}))$ mit $T_1(\mathbf{X}) = 1/n \sum X_i = \bar{X}$ und $T_2(\mathbf{X}) = S^2 = \sum (X_i - \bar{X})^2/(n-1)$. T_1 folgt einer $N(\mu, \sigma^2/n)$-Verteilung und T_2 einer von T_1 stochastisch unabhängigen χ^2-Verteilung mit $n-1$ Freiheitsgraden ($df = n-1$).

Solche Statistiken sind dann besonders hilfreich, wenn man zur Festlegung der Entscheidungsfunktion nicht auf die n-dimensionale ZV $\mathbf{X}$ zurückgreifen muß, sondern nur auf $T(\mathbf{X})$. Das ist der Fall, wenn T dieselbe Information über $\theta \in \Theta$ enthält, wenn T also suffizient für θ ist (vgl. hierzu und zu weiteren wünschenswerten Eigenschaften von Statistiken Klauer, in diesem Band). Im ersten Beispiel hängt die gemeinsame Verteilung von $\mathbf{X}$ als n-faches Produkt von $B(1,p)$-Verteilungen nur von $T(\mathbf{X}) = \sum X_i$ ab: die Wahrscheinlichkeitsfunktion ist hier $P(\mathbf{X} = \mathbf{x}) = p^{(\sum x_i)} \cdot (1-p)^{(n - \sum x_i)}$. Im zweiten Beispiel ist die gemeinsame n-dimensionale Dichtefunktion als n-faches Produkt von $N(\mu, \sigma^2)$-Verteilungen nur von $(\bar{X}, S^2)$ abhängig:

$$P(\mathbf{X} = \mathbf{x}) = (1/\sigma\sqrt{2\pi})^n \cdot exp((-\sum (x_i - \bar{x})^2/2\sigma^2) - n(\bar{x} - \mu)^2/2\sigma^2) \ , \qquad (2)$$

da $\sum (X_i - \bar{X})^2 = (n-1)S^2$ ist. Also kann man für eine Entscheidung den Stichprobenraum so partitionieren, daß alle $\mathbf{x}$ zusammengefaßt werden, die zum selben Wert $T = t$ führen.

Bei der Wahl einer Statistik zur Entscheidung bezüglich einer statistischen Hypothese (über einen Parameter θ) wurde häufig intuitiv vorgegangen. Es wurde eine (reellwertige) Teststatistik ausgewählt, deren Verteilung bei Gültigkeit der Nullhypothese spezifiziert ist und deren Verteilungsfunktion in tabellierter Form vorliegt. Als kritischen Bereich C wählt man die möglichen Werte der Teststatistik aus, die in „stärkstem Widerspruch" zu den Gegebenheiten bei Gültigkeit der Nullhypothese stehen. Zum Testen der einfachen Hypothese $\mu = \mu_0$ gegen die zusammengesetzte $H_1 : \mu \neq \mu_0$ bei n unabhängig $N(\mu, \sigma^2)$-verteilten ZVn hat beispielsweise R. A. Fisher nach heuristischen Überlegungen die Teststatistik $|T(X)| = \sqrt{n}(\bar{X} - \mu_0)/S$ vorgeschlagen. Der kritische Bereich wird so festgelegt, daß er „große" Werte einer t-verteilten ZV mit $df = n - 1$ umfaßt (siehe Abschnitt 1.6).

1.6 Fehlerwahrscheinlichkeiten und Gütefunktion (Teststärke)

Bei der Festlegung der Entscheidungsfunktion ist es möglich, daß Fehlentscheidungen vorkommen. Ein Fehler erster Art liegt vor, wenn man sich bei beobachtetem $\mathbf{x}$ oder $T(\mathbf{x})$ für H_1 (d.h. $d(\mathbf{x}) = dH_1$) entscheidet, obwohl H_0 (d.h. $\theta \in \Theta_0$) zutrifft. Von einem Fehler zweiter Art spricht man, wenn man sich für H_0 (d.h. $d(\mathbf{x}) = dH_0$) entscheidet, obwohl H_1 (d.h. $\theta \in \Theta_0^c$, Komplement von Θ_0) zutrifft. Die zugehörigen Wahrscheinlichkeiten sind für einen Fehler erster Art $P_{\theta \in \Theta_0}(\mathbf{x} \in C)$ und für einen Fehler zweiter Art $P_{\theta \in \Theta_0^c}(\mathbf{x} \in A) = 1 - P_{\theta \in \Theta_0^c}(\mathbf{x} \in C)$. Beide Arten von Fehlerwahrscheinlichkeiten für einen Test und damit auch die Wahrscheinlichkeiten für richtige Entscheidungen in Abhängigkeit von $\theta \in \Theta$ lassen sich einheitlich durch die Gütefunktion (*power function*), d.h. die Güte der Entscheidungsfunktion d darstellen:

$$\beta(\theta) = P_\theta(d(\mathbf{x}) = dH_1) = P_\theta(\mathbf{x} \in C) \; . \tag{3}$$

Wichtig ist, daß die Gütefunktion auch vom Stichprobenumfang n abhängig ist. Ideal wäre die Gütefunktion, falls sie den Wert Null für alle $\theta \in \Theta_0$ und ansonsten den Wert 1 annähme. Bis auf triviale Fälle ist das aber nicht zu erreichen. Für eine feste Stichprobengröße ist es auch unmöglich, beide Fehlerarten beliebig klein zu machen. Auf der Suche nach einem „guten" Test geht man allgemein so vor: Üblicherweise legt man fest, daß die Testgüte in Θ_0 nicht größer als α sein soll:

$$sup_{\theta \in \Theta_0}(\beta(\theta)) \leq \alpha \quad \text{für alle} \quad \theta \in \Theta_0 \tag{4}$$

und α eine (kleine) Wahrscheinlichkeit mit $0 < \alpha < 1$ ist. Man nennt dann den Test einen Test zum Niveau α. Somit ist der Umfang der kritischen Region C festgelegt. Unter allen Tests mit dieser Eigenschaft versucht man den Test zu wählen, mit dem man einen möglichst kleinen Fehler zweiter Art begeht bzw. die Testgüte maximiert, d.h. es soll gelten:

$$\begin{aligned} & P_\theta(d(\mathbf{x}) = dH_0) \quad \text{minimal für} \quad \theta \in H_1 \\ \text{oder äquivalent} \quad & \beta(\theta) = \quad \text{maximal für alle} \quad \theta \in H_1. \end{aligned} \tag{5}$$

Die Wahrscheinlichkeit für die Entscheidung dH_1 soll also maximal sein, wenn die Alternative zutrifft. Zu beachten ist, daß in obiger Festlegung von α eine unsymmetrische Betrachtung des eigentlich symmetrischen Entscheidungsproblems zwischen

zwei Alternativen vorliegt. Allerdings ist es oft nur so möglich, einen Test explizit anzugeben. Es läßt sich zeigen, daß für einseitige Hypothesen mit eindimensionalem θ ein optimaler Test eine Entscheidungsfunktion folgender Form hat:

$$d(\mathbf{x}) = dH_1, \quad \text{falls} \quad T(\mathbf{x}) > c, \quad \text{und} \quad d(\mathbf{x}) = dH_0, \quad \text{falls} \quad T(\mathbf{x}) \leq c \qquad (6)$$

für die Teststatistik $T(\mathbf{X})$. Der kritische Wert c wird gemäß der Forderungen (4) und (5) so bestimmt, daß

$$\begin{aligned} &P_\theta(T(\mathbf{X}) > c) \leq \alpha \quad \text{für alle} \quad \theta \in H_0 \\ &\text{und} \quad P_\theta(T(\mathbf{X})) \quad \text{maximal für} \quad \theta \in H_1. \end{aligned} \qquad (7)$$

Der kritische Wert muß also möglichst klein gewählt werden, ohne daß α überschritten wird. An den zwei Beispielen sollen diese Überlegungen kurz erläutert werden:

1. Die ZV und Teststatistik $T(\mathbf{X}) = \sum X_i$ sei binomialverteilt $B(5, \theta)$. Getestet werden soll $H_0 : \theta \leq 1/2$ gegen $H_1 : \theta > 1/2$. Legt man die kritische Region so fest, daß H_0 nur verworfen wird, falls $T = 5$ (d.h. nur „Erfolge"), hat man: $\beta(\theta) = P_\theta(\mathbf{X} \in C) = P_\theta(T(\mathbf{X}) = 5) = \theta^5$. Die Wahrscheinlichkeit für einen Fehler erster Art ist klein: $\beta(\theta) \leq (1/2)^5 = 0.0312$ für alle $\theta \leq 1/2$; aber der Fehler zweiter Art ist zu groß, denn $1 - \beta(\theta)$ ist für viele $\theta > 1/2$ sehr groß. Erst ab $\theta > (1/2)^{1/5} = 0.87$ ist $1 - \beta(\theta) < 1/2$. Durch Vergrößerung des kritischen Bereiches durch Hinzunahme von $T(\mathbf{X}) = 4$ wird der Fehler zweiter Art gesenkt, aber auch der Fehler erster Art erhöht.

2. Bei n unabhängig $N(\mu, \sigma_0^2)$-verteilten ZVn mit bekanntem σ_0^2 und einseitiger $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$ hat man:

$$\begin{aligned} \beta(\mu) = P((\bar{X} - \mu_0)/\sigma_0 > c) &= P(\sqrt{n}(\bar{X} - \mu)/\sigma_0 > c + \sqrt{n}(\mu_0 - \mu)/\sigma_0) \\ &= P(Z > c + \sqrt{n}(\mu_0 - \mu)/\sigma_0) \end{aligned} \qquad (8)$$

   für eine ZV Z mit $N(0,1)$-Verteilung. Die Gütefunktion $\beta(\mu)$ wächst monoton mit wachsendem μ, und es gilt $\beta(\mu_0) = \alpha$, falls $P(Z > c) = \alpha$. Auch in diesem Beispiel ist die Gütefunktion in (8) vom Stichprobenumfang n abhängig.

1.7 Die Theorie des Signifikanztests von R. A. Fisher

Ohne auf die historischen Aspekte der Entwicklung statistischer Testverfahren (z.B. Cowles, 1989) eingehen zu können, läßt sich nach R. A. Fisher die Aufgabe der mathematischen Statistik folgendermaßen charakterisieren: Ein statistischer Test wird als das Prüfen einer Hypothese H_0 aufgefaßt, welche die Identität der Verteilung einer im Experiment vorliegenden ZV mit einer hypothetischen Verteilung behauptet. Im obigen Beispiel 2 bei einer parametrischen Verteilungsklasse und eindimensionalem Parameterraum wäre das also die Nullhypothese $H_0 : \theta = \theta_0$. Deshalb handelt es sich um ein zweiseitiges Testproblem. Die Abweichung der tatsächlich vorliegenden von der hypothetischen Verteilung ist dann von Interesse. Dazu wurde eine oft aufgrund heuristischer Überlegungen gewonnene Teststatistik $T(\mathbf{X})$ herangezogen. Daß diese sich dann u.U. später als in noch zu präzisierendem Sinn optimal herausstellte, sei

hier schon erwähnt. Diese Teststatistik wurde so gewählt, daß sie einen „extremen" Wert $T(\mathbf{x})$ annahm, wenn der Beobachtungsvektor $\mathbf{x}$ „sehr stark" im Widerspruch zu der Situation unter H_0 stand. Dennoch ist prinzipiell jeder Wert $\mathbf{x}$ und damit auch $T(\mathbf{x})$ bei Gültigkeit von H_0 beobachtbar, wenn auch mit sehr geringer Wahrscheinlichkeit. Fisher hat auf die explizite Vorgabe von α wie in (4) ganz verzichtet und den p-Wert

$$p(\mathbf{x}) = P(\mathbf{X} = \mathbf{y}, H_0 : T(\mathbf{y}) \geq T(\mathbf{x})), \tag{9}$$

also den Wert einer weiteren Statistik, betrachtet (evtl. wurde auch $|T(\mathbf{y})|$ bei zweiseitiger H_0 benutzt). Der so gewonnene (empirische) p-Wert wurde dann als „Maß der Evidenz gegen die Nullhypothese" angesehen. Häufig wurde bei $p(\mathbf{x}) \leq 0.05$ (Fisher's Empfehlung „*one in twenty*") die Nullhypothese als nicht plausibel verworfen. Allerdings hat Fisher wiederholt sinngemäß betont, daß es sich bei diesem Vorgehen nicht um eine mechanische Entscheidungsregel handeln sollte, sondern um eine dem jeweiligen Forschungsgegenstand angemessene Festlegung. Will man den Signifikanztest in ein Entscheidungsverfahren umwandeln, ist ein kritischer Wert c (üblicherweise als Quantil einer bekannten und vertafelten Verteilungsfunktion) für $T(\mathbf{X})$ so zu wählen, daß eine irrtümliche Ablehnung der H_0 höchstens mit Wahrscheinlichkeit α eintreten könnte. Wird ein $T(\mathbf{X}) > c$ beobachtet, nennt man die Abweichung von H_0 „signifikant zum Niveau α". Eine (7) entsprechende Aussage ist für die Entscheidung $d(\mathbf{x}) = dH_0$ nur möglich, wenn eine Alternative spezifiziert wird, wenn $sup_{\theta \in H_1} P_\theta(d(\mathbf{x}) = dH_1)$ berechnet wird. Ansonsten kann die Entscheidung dH_0 nur bedeuten, daß $\mathbf{x}$ bezüglich des für die Fehlerwahrscheinlichkeit erster Art festgelegten Wertes α nicht im Widerspruch zu H_0 steht. Mit einem Signifikanztest im Sinne von Fisher kann man die (vorläufige) Gültigkeit einer statistischen Hypothese statistisch nur sichern, wenn sie als H_1 formuliert wird. Falls $T(\mathbf{x})$ nicht in den kritischen Bereich fällt, ist eine Entscheidung für H_0 nicht gerechtfertigt. Diese Nichtberücksichtigung von zur H_0 alternativen Hypothesen ist eine der meistdiskutierten Entscheidungen von R. A. Fisher und Ursache vieler Auseinandersetzungen (Literatur in Hager, 1992). Einerseits erfolgte die Wahl einer Teststatistik nicht willkürlich, sondern sie war implizit auf bestimmte Abweichungen von der H_0 ausgerichtet (sensitiv). Andererseits hatte für Fisher das Überprüfen einer H_0 die Funktion, unter Kontrolle des Zufalls abzusichern, daß ein erklärungsbedürftiges Phänomen vorliegt. Seine Ablehnung eines Vorgehens, das eine Entscheidung zwischen zwei alternativ formulierten Hypothesen erlaubt, beruhte u.a. auf der „frequentistischen" Interpretation der Fehler erster und zweiter Art als relativer Häufigkeit von Fehlentscheidungen bei wiederholter Stichprobenziehung aus derselben Population zum Testen derselben Hypothesen, wie das etwa bei der Qualitätskontrolle in der Fertigung von Massengütern wie Schrauben (mit festem Durchmesser θ_0 gegenüber einer Abweichung der Maschineneinstellung mit Durchmesser θ_1) der Fall ist. Solch eine „mechanische" Anwendung einer fixen Entscheidungsprozedur hielt Fisher bei Fragestellungen der (Grundlagen-)Forschung für unangemessen.

1.8 Randomisierungs- und Permutationstests

R. A. Fisher hat auch exemplarisch eine große Klasse von statistischen Testverfahren vorgeschlagen, die verteilungsfrei sind. Es wird keine parametrische Verteilungsan-

nahme für $\mathbf{X}$ gemacht. Mit einem Permutationstest können Nullhypothesen H^P der Art getestet werden, daß die gemeinsame Verteilungsfunktion $F(\mathbf{x})$ bei Gültigkeit von H^P invariant (unverändert) bleibt unter einer bestimmten Menge Q_n von Vertauschungen (Permutationen) π der Beobachtungen, also H^P: $F(\pi\mathbf{x}) = F(\mathbf{x})$ für alle $\pi \in Q_n$. Unter H^P besitzen alle $\pi\mathbf{x}$ dieselbe Wahrscheinlichkeit 1/Anzahl Elemente in $Q_n = 1/card(Q_n)$. Wie zuvor kann man einen Annahme- und Ablehnungsbereich festlegen. Der Ablehnungsbereich C enthält für jedes in einem Experiment neu beobachtete $\mathbf{x}$ alle die Permutationen von $\mathbf{x}$, die im stärksten Widerspruch zu H^P stehen. Sei eine „geeignete" Teststatistik $T(\mathbf{X})$ zum Testen von H^P so beschaffen, daß sie große Werte annimmt, wenn $\pi\mathbf{x}$ aus C stammt. Ein Signifikanztest ist nun definiert über den p-Wert als Maß für die Konsistenz der Daten mit der Nullhypothese (evidentialistisches Stützungsmaß; vgl. Hager, 1992, S. 37):

$$p(\mathbf{x}) = P(H^P : T(\mathbf{X}) \geq T(\mathbf{x})) = card\left(\{\pi \in Q_n : T(\pi\mathbf{x}) \geq T(\mathbf{x})\}\right) / card(Q_n) \; . \quad (10)$$

Je kleiner p, umso größer ist die Evidenz gegen H^P. Wichtig ist, daß es sich um einen auf die gegebenen Daten $\mathbf{x}$ bedingten Test handelt. Allein die randomisierte Zuweisung der Beobachtungen zu den experimentellen Bedingungen reicht aus, die Validität dieses Signifikanztests zu sichern (Lehmann, 1959, S. 282). Als Beispiel stelle man sich vor, daß n Vpn per Zufall je einer von k experimentellen Bedingungen zugewiesen werden. Die übliche (globale) H_0 eines Randomisierungstests ist H^R: alle k experimentellen Bedingungen haben für alle n Vpn denselben Effekt. Einziges Zufallselement für die Ableitung eines statistischen Tests sind die Designvariablen δ_l^j, die die Zuweisung der l-ten Vp zur j-ten Bedingung indizieren. Wenn H^R gilt, können Unterschiede in den beobachteten Werten nur durch Variabilität der Vpn selbst (und evtl. Meßfehler) entstehen. Ohne Wirken der experimentellen Bedingungen würde das zu Beobachtungen $\mathbf{u} = (u_1, \ldots, u_n)$ führen. Die Teststatistik ist eine Funktion der n Design-ZV und der Einheitenwerte (*unit values*) $\mathbf{u}$. Analog zum Permutationstest betrachtet man:

$$p(\mathbf{x}) = P(H^R \text{ und } \mathbf{u} : T(\mathbf{x}, \{\delta_l^j\}) \geq T(\mathbf{x})) \; . \quad (11)$$

Sowohl bei Gültigkeit von H^P wie H^R gilt: $P(p(\mathbf{x}) \leq \alpha) = \alpha$ für jedes erreichbare Niveau α mit $0 < \alpha < 1$; es handelt sich also um einen Test zum Niveau α.

2 Die Theorie von J. Neyman und E. S. Pearson

In Publikationen der Jahre 1928 bis 1938 (Literatur s. Hager, 1992) entwickelten J. Neyman und E. S. Pearson die Grundzüge einer mathematischen Theorie des Testens statistischer Hypothesen, in welcher Tests als Lösungen von klar definierten und umschriebenen Optimierungsproblemen abgeleitet werden. Obwohl Neyman und Pearson wichtige Ansätze wie das Konzept des Stichprobenraumes und mathematische Methoden zur Bestimmung der Stichprobenverteilung von (Test-)Statistiken von R. A. Fisher aufnahmen, ging ihre Zielvorstellung darüber hinaus, statistische Tests aufgrund teilweise heuristischer Überlegungen aus eher praktisch motivierten Anforderungen heraus zu entwickeln. Wichtige Konzepte entstanden auch in kritischer Abgrenzung zu Fisher. Neyman und Pearson haben keine abgeschlossene

Theorie vorgelegt, vielmehr handelt es sich um ein von ihnen initiiertes statistisches Forschungsprogramm, zu dem mehrere andere Personen nachfolgend wesentliche Beiträge geleistet haben. Ein erstes wichtiges, zusammenfassendes Lehrbuch stammt von E. L. Lehmann (1959), einem Studenten von Neyman. Der bedeutsamste Unterschied zu Fisher liegt in der Einsicht, daß zum Testen einer Hypothese stets eine Alternativhypothese oder eine Klasse von alternativen Hypothesen zu berücksichtigen ist, zwischen denen eine Entscheidung zu treffen ist (Neyman, 1942).

2.1 Parametrische Hypothesen

Für den Fall einer einfachen H_0 und einer einfachen H_1 ist nach dem Fundamentallemma von Neyman und Pearson (Lehmann, 1959, S. 63ff.) ein bester Test stets konstruktiv angebbar; dieser Test ist ein Test zum Niveau α, der die Gütefunktion maximiert. Für diskrete ZVn ist das Vorgehen leicht darzustellen. Seien $P_0(\mathbf{X} = \mathbf{x})$ und $P_1(\mathbf{X} = \mathbf{x})$ die Wahrscheinlichkeitsfunktionen unter H_0 bzw. H_1, z.B. die Binomialverteilung $B(n, \theta = p)$ mit $H_0 = \{p_0\}$ und $H_1 = \{p_1\}$. Der optimale Test muß für Beobachtungen $\mathbf{x}$ in der kritischen Region C folgende Bedingung erfüllen: $\sum_{\mathbf{x}\in C} P_0(\mathbf{x}) \leq \alpha$ und $\sum_{\mathbf{x}\in C} P_1(\mathbf{x})$ maximal. C sollte also die $\mathbf{x}$ enthalten, die einen möglichst hohen Wert für den Quotienten $r(\mathbf{x}) = P_1(\mathbf{x})/P_0(\mathbf{x})$ aufweisen. Sukzessive nimmt man alle $\mathbf{x}$ mit $r(\mathbf{x}) > c$ auf; dabei ist c so festgelegt, daß $\sum_{\mathbf{x}} P_0(\mathbf{x}\text{: } r(\mathbf{x}) > c) \leq \alpha$ ist und die Hinzunahme eines weiteren $\mathbf{x}$ über α hinausgehen würde. Ein Beispiel findet man etwa in Casella und Berger (1990, S. 368). Auf sogenannte randomisierte Tests (die jedes α-Niveau bei diskreter Verteilung der ZV $\mathbf{X}$ exakt ausschöpfen) wird hier nicht eingegangen (siehe Lehmann, 1959). Für ZVn mit Dichtefunktion $p_0(\mathbf{x},\theta_0)$ und $p_1(\mathbf{x},\theta_1)$ geht man entsprechend vor. Man wählt den kritischen Bereich so, daß $p_1(\mathbf{x},\theta_1)/p_0(\mathbf{x},\theta_0) > c$, und c ist so zu wählen, daß $p_1(\mathbf{x},\theta_1) < cp_0(\mathbf{x},\theta_0)$ für $\mathbf{x}$ aus dem Annahmebereich und $P_{\theta_0}(\mathbf{X}\in C) = \alpha$. Statt $\mathbf{X}$ könnte man auch eine (suffiziente) Teststatistik $T(\mathbf{X})$ betrachten. Diese Aussage des Fundamentallemmas ist vorwiegend von theoretischer Bedeutung, denn in vielen Fällen sind auch bei eindimensionalem Parameterraum H_0 und H_1 zusammengesetzt, häufig in der Form $H_0 : \theta \leq \theta_0$ und $H_1 : \theta > \theta_0$.

Ein Test heißt gleichmäßig bester Test (*uniformly most powerful*, UMP) für H_0 gegen H_1 zum Niveau α, wenn für alle $\theta \in \Theta_1$ (also speziell für alle $\theta > \theta_0$) die Testgüte gleichmäßig maximiert wird; d.h. $\beta(\theta)$ ist nicht vom speziellen $\theta \in \Theta_1$ abhängig. Die Bedingung des Fundamentallemmas, die Existenz einer Konstante k, muß also für jedes $\theta \in \Theta_1$ gelten sowie $P_{\theta\in\Theta_0}(\mathbf{x}\in C) = \alpha$ für ein $\theta \in \Theta_0$. Im obigen Beispiel ist das θ_0 selbst. Schon in dem einfachen Fall eines eindimensionalen Parameters ist für zweiseitige Alternativen ($\theta = \theta_0$ vs. $\theta \neq \theta_0$) wie im Beispiel 2 ein UMP-Test nicht angebbar (Casella & Berger, 1990, S. 371). Um dennoch zu einem optimalen Test zu gelangen, schränkt man die Menge zugelassener Tests auf die Klasse der unverfälschten (*unbiased*) Tests ein, d.h. Tests mit der Eigenschaft $\beta(\theta) \geq \alpha$ für alle $\theta \in H_1$. Die Entscheidung dH_1 soll also unter H_1 mindestens so wahrscheinlich wie unter H_0 sein. Im Beispiel 2 kann man als Teststatistik $Z = \sqrt{n}(\bar{X}-\theta_0)/\sigma_0$ wählen. Der Test, der H_0 verwirft, falls $|Z| > z_{1-\alpha/2}$, dem $(1-\alpha/2)$-Quantil der $N(0,1)$, ist der unverfälschte UMP-Test zum Niveau α, d.h. unter den unverfälschten Tests ist er UMP (Casella & Berger, 1990, S. 374ff.).

Bereits für den t-Test im Ein- und Zweistichprobenproblem ist für den Nachweis der Optimalität die Verwendung recht komplizierter mathematischer Hilfsmittel erforderlich. Da σ^2 nicht fest ist, hat man einen zweidimensionalen Parameterraum (μ, σ^2) mit einer ein- oder zweiseitigen zusammengesetzten Hypothese lediglich bezüglich der ersten Komponente (Neyman, 1942, S. 310f.). Bekanntlich wird σ^2 durch die Statistik $S^2 = \sum(X_i - \bar{X})^2/(n-1)$ geschätzt, und die Teststatistik $T(\mathbf{X}) = \sqrt{n}(\bar{X} - \mu_0)/S$ hat eine zentrale t-Verteilung mit $df = n-1$. Für einen einseitigen Test verwirft man H_0, falls $t(\mathbf{x})$ größer ist als das $(1-\alpha)$-Quantil der zentralen t-Verteilung mit $df = n-1$. Die Gütefunktion $\beta(\mu, \sigma^2) = P(\mu, \sigma^2 : t(\mathbf{X}) > t_{n-1,1-\alpha})$ ist eine monoton wachsende Funktion des Nichtzentralitätsparameters (NZP) $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$ und von n. Für die ein- und zweiseitige Hypothese ist der t-Test ein unverfälschter UMP-Test zum Niveau α wie auch für das Zweistichprobenproblem mit gleichen Varianzen (Neyman, 1942; Lehmann, Kap. 6). Falls man ein k-Stichprobenproblem (einfaktorielle ANOVA mit festen Effekten) hat, existiert kein unverfälschter UMP-Test mehr. Um in diesem Fall (und anderen Modellen der ANOVA mit linearen Hypothesen) zu Optimalitätsaussagen zu gelangen, beschränkt man sich auf die Klasse der invarianten Tests. In diesem Fall wird der Stichprobenraum durch eine Transformation $\mathbf{y} = g(\mathbf{x})$ wie z.B. Lage- und Skalentransformationen so auf sich abgebildet, daß das Testproblem $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_0^c$ invariant bleibt, d.h. für die ZV X mit Dichtefunktion soll gelten: $\{f(\mathbf{x},\theta) : \theta \in \Theta_0\} = \{h(\mathbf{y},\theta) : \theta \in \Theta_0\}$ und $\{f(\mathbf{x},\theta) : \theta \in \Theta_0^c\} = \{h(\mathbf{y},\theta) : \theta \in \Theta_0^c\}$. Zum Testen verwendet man dann eine gegen die Transformation maximalinvariante Statistik $T(\mathbf{X})$. Die üblichen F-Tests der (einfaktoriellen) ANOVA besitzen die Eigenschaft, UMP-invariante Tests zu sein mit einer unter der jeweiligen H_0 zentralen F-Verteilung und nichtzentraler F-Verteilung unter H_1 mit NZP δ^2 (einfaktorielle ANOVA: $\delta^2 = 1/k \sum(\mu_i - \mu)^2/\sigma^2$). Es sei nur erwähnt, daß es in der MANOVA keinen UMP-invarianten Test gibt. Die verschiedenen vorgeschlagenen Teststatistiken von Pillai, Wilks, Roy, Hotelling-Lawley sind unterschiedlich teststark gegen verschiedene Abweichungen von der H_0 (vgl. das Kapitel zu multivariaten Verfahren von Andres, in diesem Band).

2.2 Nichtparametrische Hypothesen

Bei vielen stetig verteilten ZVn ist eine parametrische Verteilungsannahme nicht zu rechtfertigen. Insbesondere für ordinale Merkmale sind parametrische Verteilungsklassen nicht geeignet, da sie nicht invariant gegen stetige, streng monotone Transformationen sind. Bei der statistischen Entscheidung in solchen nichtparametrischen oder verteilungsfreien Testproblemen werden nicht die beobachteten Zahlenwerte selbst, sondern nur deren Rangordnung relevant. Bei intervallskalierten Merkmalen ist die Verwendung der Beobachtungswerte selbst zulässig. Bei Unkenntnis über die Verteilung der ZVn können Permutationstests (Randomisierungstests) angewendet werden. Eine typische nichtparametrische Hypothese, etwa für ein unabhängiges Zweistichprobenproblem ist: $H_0 : P(X_{1j} > x) \leq P(X_{2j} > x)$ für alle reellen x gegen $H_1 : P(X_{1j} > x) > P(X_{2j} > x)$ mit $j = 1, \ldots, n_1$ bzw. n_2, d.h. X_1 ist stochastisch größer als X_2. Für dieses Testproblem gibt es weder einen unverfälschten noch einen invarianten UMP-Test. Nur bei Einschränkung auf eine Teilklasse $K \subset H_1$ der Alternative ist bei einseitigen Testproblemen ein unverfälschter Test zum Niveau α,

der den Fehler zweiter Art gleichmäßig minimiert, möglich. Der Permutationstest von Pitman mit der Teststatistik $S(\mathbf{X}) = \sum X_{1j}$ hat diese Eigenschaft, wenn K die Klasse der Normalverteilungen ist. Die Menge der zulässigen Permutationen Q_n unter H_0 sind alle $n!/n_1!n_2!$ Vertauschungen von Beobachtungen zwischen den beiden Bedingungen, die zu unterschiedlichen Werten von $S(\mathbf{X})$ führen können. Der p-Wert wird wie in Abschnitt 1.8 (10) oder (11) bestimmt. Für Rangtests ist die Einschränkung auf Alternativen „in der Nähe der H_0" erforderlich. Man spricht dann von lokal gleichmäßig besten Rangtests. Schließlich ist es bei verteilungsunabhängigen Verfahren oft nur für unendlich großes n möglich, die asymptotische Optimalität von Rang- oder Permutationstests nachzuweisen. So strebt beispielsweise die Verteilung der geeignet standardisierten Teststatistik des Zweistichproben-Rangsummentests von Wilcoxon gegen die Standard-Normalverteilung und diejenige der Teststatistik des Zweistichproben-Permutationstests gegen die Verteilung des t-Tests (Pratt und Gibbons, 1981).

Zum Vergleich von nichtparametrischen Tests mit parametrischen Tests wird häufig die asymptotische, relative Effizienz (A.R.E) verwendet. Sie gibt an, welchen relativen Anteil der asymptotischen Teststärke des Vergleichstests ein bestimmter Test besitzt. Zum Beispiel ist die A.R.E. des k-Stichproben Kruskal-Wallis-Tests im Vergleich zum ANOVA-Test im Fall von normalverteilten Beobachtungen gegenüber Lagealternativen $3/\pi$. Die A.R.E. des Permutationstests relativ zum t-Test oder ANOVA-F-Test im k-Stichprobenproblem ist 1.

2.3 Teststärke und Stichprobenumfang

Neyman (1942) weist darauf hin, daß die Analyse der Gütefunktion eines Tests nicht nur zum Vergleich verschiedener Tests sinnvoll ist. Vielmehr hat sie auch Bedeutung für die Planung eines Experiments. Der Stichprobenumfang, von dem $\beta(\theta)$ ebenfalls funktionell abhängig ist, sollte so gewählt werden, daß eine Abweichung des (der) interessierenden Parameter(s) von der Situation unter der H_0 mit großer Wahrscheinlichkeit durch den statistischen Test aufgedeckt wird, d.h. der Fehler zweiter Art sollte ebenfalls klein sein, wenn diese Abweichung, erfaßt in der Effektgröße, selbst nicht unbeträchtlich ist. Man hat also nach Auswahl eines bestimmten, möglichst optimalen statistischen Testverfahrens vier Bestimmungsgrößen: Fehler erster Art, Fehler zweiter Art = 1 minus Teststärke, Effektgröße und Stichprobenumfang. Nach Spezifizierung von drei dieser Größen ist die vierte festgelegt. Allerdings ist das exakt in aller Regel nur für parametrische Testprobleme möglich. Empfehlungen für approximative Lösungen im nichtparametrischen Fall unter Verwendung der A.R.E. sind etwa bei Bredenkamp (1980) zu finden. Für Randomisierungstests kann es, da es sich um auf die tatsächlichen Beobachtungen bedingte Tests handelt, keine a priori-Festlegung der Teststärke geben; auch hier müßte man approximative Lösungen verwenden. Mit seinem Buch hat J. Cohen (1988) für den parametrischen Fall die konsequente Beachtung der Teststärke bei der Planung von Experimenten für viele Anwender praktisch handhabbar gemacht. Insbesondere ermöglicht sein Buch die konsequente Wahl des Stichprobenumfangs so, daß ein a priori festzulegender Fehler zweiter Art relativ zu einem im Experiment aufzudeckenden Effekt der unabhängigen Variable(n) eingehalten werden kann. In der Folgezeit wurde deutlich,

daß viele Experimente oft wegen eines zu kleinen n mit dem Risiko eines zu großen Fehlers zweiter Art belastet waren. Von Bredenkamp (1980) und Hager (1987, 1992) wurden dann Vorschläge für verschiedene Strategien der Testplanung gemacht, die diese Probleme vermeiden.

2.4 Konsequenzen für die Überprüfung von PHn mit statistischen Testverfahren

Wie Erdfelder und Bredenkamp (1994) erläutern, ist zur Prüfung einer PH mittels einer SH die Wahrscheinlichkeit für eine Entscheidung über eine fälschliche Bewährung ebenso wie für eine fälschliche Nichtbewährung der PH möglichst klein zu halten. Dazu sind die Wahrscheinlichkeiten für fälschliche Ablehung und Annahme der implizierten SH zu kontrollieren. Je nachdem, ob die von der PH implizierte SH eine (einseitige) Alternativhypothese ist oder eine (exakte) Nullhypothese, ist der interessierende Effekt der experimentellen Bedingungen als minimaler aufzudeckender Effekt bzw. als (kleine) maximal zu tolerierende Abweichung von der H_0 zu konzipieren. Um beiden Alternativen der PH eine gleiche Chance der (vorläufigen) Bewährung zu geben, sollen Fehler erster und zweiter Art nach der Empfehlung von Bredenkamp (1980) gleich groß festgelegt werden. Die Tabellen in Cohen (1988) und eine Reihe inzwischen verfügbarer Computerprogramme (vgl. z.B. Buchner, Erdfelder & Faul, in diesem Band) erlauben für viele Testprobleme die Bestimmung eines (minimal) erforderlichen Stichprobenumfangs. Ein „nicht signifikantes" Ergebnis eines statistischen Tests läßt dann die Akzeptanz einer statistischen Nullhypothese zu, eine Entscheidung, die in der Konzeption des Signifikanztests nach Fisher nicht möglich ist. Eine weitere Eigenschaft dieses Vorgehens ist, daß, falls die PH eine Konjunktion von (einseitigen) statistischen Alternativhypothesen impliziert, u.U. keine Adjustierung des Fehlers erster Art für die einzelnen statistischen Tests erfolgen muß (Hager, 1987, 1992; Westermann & Hager, 1986).

3 Weiterführende Literatur

Für die wissenschaftstheoretische Reflektion des Einsatzes der von Neyman und Pearson begründeten Testtheorie sind die Überlegungen von Lakatos (1974) zum methodologischen Falsifikationismus und die Analysen von Stegmüller (1973) hilfreich. Die Originalarbeiten und Lehrbücher zur mathematischen Statistik, in denen die Neyman-Pearson-Theorie dargestellt ist, sind ohne gründliche mathematische Vorkenntnisse nicht im Detail zu verstehen. Dennoch lohnt sich die Durchsicht des Buches von E. L. Lehmann (1959) und von Pratt und Gibbons (1981) für nichtparametrische Methoden. Einfacher zu lesen ist das Buch von Casella und Berger (1990). Unter den Büchern in deutscher Sprache sind die Lehrbücher von Witting (1969) und Witting und Noelle (1970) zu nennen. Die Anwendung der Neyman-Pearson-Konzeption für eine Prüfung psychologischer Hypothesen findet man in dem Buch von Cohen (1988), bei Bredenkamp (1980), dem ausführlichen Buchbeitrag von Hager (1987) und der umfassenden Darstellung von Hager (1992), in der neben der Erläuterung verschiedener wissenschaftlicher Positionen und von Kernaussagen verschiedener Testtheorien eine detaillierte Anleitung zur statistischen Prüfung psycho-

logischer Hypothesen zu finden ist, welche dem Anspruch einer strengen und fairen Prüfung dieser wissenschaftlichen Hypothesen gerecht wird. Besonders zu beachten sind die Überlegungen für den Fall, in dem die PH nicht adäquat über eine einzige SH, sondern über eine Menge (Familie) von (gerichteten) SHn mit dem Problem der möglichen Kumulation von Fehlern erster oder zweiter Art zu prüfen ist. Band 1 der Forschungsmethoden der Psychologie in der Enzyklopädie der Psychologie, aus dem wiederholt Kapitel zitiert worden sind, kann ebenfalls empfohlen werden.

Literaturverzeichnis

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Casella, G. & Berger, R. L. (1990). *Statistical inference.* Belmont: Wadsworth.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition).* Hillsdale: Erlbaum.

Cowles, M. (1989). *Statistics in psychology: A historical perspective.* Hillsdale: Erlbaum.

Erdfelder, E. & Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann & W. H. Tack (Hrsg.), *Methodologische Grundlagen der Psychologie* (= Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 1, S. 604–648). Göttingen: Hogrefe.

Gadenne, V. (1994). Theorien. In T. Herrmann & W. H. Tack (Hrsg.), *Methodologische Grundlagen der Psychologie* (= Enzyklopädie der Psychologie, Serie Forschungsmethoden der Psychologie, Band 1, S. 295–427). Göttingen: Hogrefe.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Hrsg.) *Allgemeine Experimentelle Psychologie* (S. 43–264). Stuttgart: G. Fischer.

Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment.* Göttingen: Hogrefe.

Hussy, W. & Möller, H. (1994). Hypothesen. In T. Hermann & W. H. Tack (Hrsg.), *Methodologische Grundlagen der Psychologie* (= Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 1, S. 475–507). Göttingen: Hogrefe.

Lakatos, I. (1974). Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme. In I. Lakatos & A. Musgrave (Hrsg.) *Kritik und Erkenntnisfortschritt* (S. 89–189). Braunschweig: Vieweg.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* New York: Wiley.

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society, 105,* 292–327.

Pratt, J. W. & Gibbons, J. D. (1981). *Concepts of nonparametric theory.* New York: Springer.

Stegmüller, W. (1973). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band 4: Personelle und statistische Wahrscheinlichkeit. 2. Halbband: Statistisches Schließen.* Berlin: Springer.

Steyer, R. (1994). Stochastische Modelle. In T. Herrmann & W. H. Tack (Hrsg.), *Methodologische Grundlagen der Psychologie* (= Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 1, S. 649–693). Göttingen: Hogrefe.

Westermann, R. & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11,* 117–146.

Witting, H. (1969). *Mathematische Statistik.* Stuttgart: Teubner.

Witting, H. & Noelle, G. (1970). *Angewandte mathematische Statistik.* Stuttgart: Teubner.