Zeitschrift: Bulletin de la Société Fribourgeoise des Sciences Naturelles = Bulletin der Naturforschenden Gesellschaft Freiburg

Herausgeber: Société Fribourgeoise des Sciences Naturelles

Band: 52 (1962)

Artikel: Mathematische Testtheorie : Grundlagen und neuere Probleme

Autor: Kres, Heinz

DOI: https://doi.org/10.5169/seals-308388

Nutzungsbedingungen

Die ETH-Bibliothek ist die Anbieterin der digitalisierten Zeitschriften auf E-Periodica. Sie besitzt keine Urheberrechte an den Zeitschriften und ist nicht verantwortlich für deren Inhalte. Die Rechte liegen in der Regel bei den Herausgebern beziehungsweise den externen Rechteinhabern. Das Veröffentlichen von Bildern in Print- und Online-Publikationen sowie auf Social Media-Kanälen oder Webseiten ist nur mit vorheriger Genehmigung der Rechteinhaber erlaubt. Mehr erfahren

Conditions d'utilisation

L'ETH Library est le fournisseur des revues numérisées. Elle ne détient aucun droit d'auteur sur les revues et n'est pas responsable de leur contenu. En règle générale, les droits sont détenus par les éditeurs ou les détenteurs de droits externes. La reproduction d'images dans des publications imprimées ou en ligne ainsi que sur des canaux de médias sociaux ou des sites web n'est autorisée qu'avec l'accord préalable des détenteurs des droits. En savoir plus

Terms of use

The ETH Library is the provider of the digitised journals. It does not own any copyrights to the journals and is not responsible for their content. The rights usually lie with the publishers or the external rights holders. Publishing images in print and online publications, as well as on social media channels or websites, is only permitted with the prior consent of the rights holders. Find out more

Download PDF: 28.01.2026

ETH-Bibliothek Zürich, E-Periodica, https://www.e-periodica.ch

Mathematische Testtheorie
Grundlagen und neuere Probleme

Teil I

Grundbegriffe der Testtheorie mit Anwendungen

Von Heinz Kres
Mathematisches Institut der Universität Freiburg im Üchtland (Schweiz)

INHALTSVERZEICHNIS

Einleitung 73

A. Die Grundzüge der Neyman-Pearsonschen Testtheorie

1. Hypothese, Test und kritische Region 75
2. Die möglichen Fehler und die Güte eines Tests 78
3. Einseitiges und zweiseitiges Testen 80
4. Einfache und zusammengesetzte Hypothesen 81
5. Einige weitere Forderungen an Testverfahren 82
6. Das Zwei- (bzw. k-)Stichprobenproblem 83
7. Parametrische und parameterfreie Methoden in der Mathematischen Statistik 85

B. Einige wichtige Tests für die Behandlung des Zweistichprobenproblems.

8. Der Studentsche t-Test 86
9. Der Zeichentest 88
10. Der Iterationen-Test von Wald und Wolfowitz 90
11. Der Wilcoxon-Test 91
12. Der X-Test von van der Waerden 96

C. Anwendungsbeispiele zu den behandelten Testverfahren

13. Die Auswertung psychologischer Untersuchungen mit dem Iterationen-Test 98
14. Die Untersuchung von Wachstumseinwirkungen zweier Vitamine bei Pilzen. (Mit dem Student-Test) 99
15. Die Untersuchung der Brenndauer von Glühlampen mit dem X-Test 100
16. Die Untersuchung von Titrationen mit dem Zeichentest 101
17. Vergleich zweier Produktionsverfahren mittels X-Test, Wilcoxon-Test und Student-Test 102

Anhang : **Auswahlgesichtspunkte für die verschiedenen Testverfahren** . . . 104

Literaturverzeichnis 106

Einleitung

Die Mathematische Testtheorie gehört neben der Theorie der Parameterschätzungen und der Theorie der Vertrauens- und Toleranzbereiche zu den bedeutendsten Teilgebieten der statistischen Erfahrungsbildung.

In der Praxis der experimentellen Forschung geht es oftmals weniger darum, einen statistischen Parameter – etwa den Mittelwert einer Grundgesamtheit – genau zu schätzen, als vielmehr um die Behandlung von Problemen der folgenden Art :

1. Ein Physiker hat eine neue Methode zum Auftragen der Thoriumschicht eines Neutronenzählrohres unter folgenden Gesichtspunkten zu beurteilen :

a) Ist die gewünschte Verkürzung der Löschzeit eingetreten ?

b) Hat sich dabei die durchschnittliche Lebensdauer des Rohres in unerwünschter Weise verkürzt ?

2. Ein Fabrikant möchte wissen, ob eine neue Art der Nahrungsmittelkonservierung durch radioaktive Präparate die Vitamine schonender behandelt als die bisherige.

3. Ein Mediziner möchte testen, ob ein neu entwickeltes Antibiotikum in der Behandlung von Infektionskrankheiten erfolgreicher ist als das bisher verwendete Standardpräparat.

4. In der Züchtungsforschung hat ein Botaniker die Erträge einer neuen Population mit denen der bisherigen zu vergleichen.

5. In der vergleichenden Anatomie – speziell etwa in der Primatologie – soll die « Echtheit » zweier Arten nachgewiesen werden.

Weitere Probleme finden sich in den Anwendungsbeispielen des Abschnittes C.

All diesen Aufgabenstellungen liegt ein gemeinsames Problem zugrunde :

Hat sich im Vergleich mit einer zweiten Grundgesamtheit – oder durch Abänderung der Versuchsbedingungen – der durchschnittliche Wert einer zu prüfenden Eigenschaft wesentlich verändert, sei es vergrößert oder verkleinert ? Zur mathematischen Behandlung solcher Probleme werden wir im Abschnitt A eine *Nullhypothese* H_0 aufstellen, welche besagt, daß beide « Stichproben » (d. h. diejenige aus der ersten und diejenige aus der zweiten zum Vergleich herangezogenen Population) zu Grundgesamtheiten mit der gleichen Verteilung gehören.

Zusätzlich werden wir die *Alternativhypothese* H_a betrachten müssen, wonach die beiden Stichproben als zu verschieden verteilten Grundgesamtheiten gehörig anzusehen sind. Die Nullhypothese, die auf Gleichheit der beiden Verteilungen lautet, ist in diesem Falle zu verwerfen.

In der Testtheorie dient dann auf der Basis der erhobenen Stichproben eine besondere *Stichprobenfunktion* mit einer zugehörigen *Testvorschrift* zur Entscheidung über Annahme oder Verwerfung der aufgestellten Nullhypothese. Die Grundzüge dieser Testtheorie, deren Entwicklung vornehmlich mit den Namen NEYMAN und PEARSON verbunden ist, werden im folgenden Abschnitt A dargelegt. Dabei soll bezüglich der Definitionen und der Sprechweise stets das sogenannte *Problem der zwei Stichproben*, wie es in den weiter oben skizzierten Aufgabenstellungen bereits angedeutet wurde und im § 6 noch genauer formuliert wird, im Vordergrund des Interesses stehen.

Der Abschnitt B bringt dann die wichtigsten und gebräuchlichsten Testverfahren zur Behandlung des Zwei-Stichproben-Problems.

Zur Veranschaulichung der Theorie werden im Abschnitt C einige typische Anwendungsbeispiele aus verschiedenen Gebieten dargestellt.

Der Anhang soll schließlich noch einige Hinweise geben auf die Frage, welchen der zahlreichen Tests man im Einzelfall günstigerweise anwendet.

Abschnitt A

Die Grundzüge der Neyman-Pearsonschen Testtheorie

1. Hypothese, Test und kritische Region

In der Mathematischen Statistik nennt man eine eindeutige reellwertige Funktion X, die ihre möglichen Werte gemäß Zufall annimmt, eine *zufällige Größe* oder *stochastische Variable.*

Hiervon ausgehend definiert man durch

$$(1;1)\quad F(t) = P(X \leqq t)$$

eine neue Funktion, die *Verteilungsfunktion* der zufälligen Veränderlichen X. Dabei bedeutet $P(X \leqq t)$ die Wahrscheinlichkeit dafür, daß die zufällige Veränderliche X einen Wert kleiner oder gleich t annimmt.

Die Ableitung der Verteilungsfunktion heißt *Dichtefunktion.*

Auf die vielfältigen mathematischen Eigenschaften dieser beiden Funktionen soll hier nicht weiter eingegangen werden. Es sei lediglich an die wichtigen Forderungen

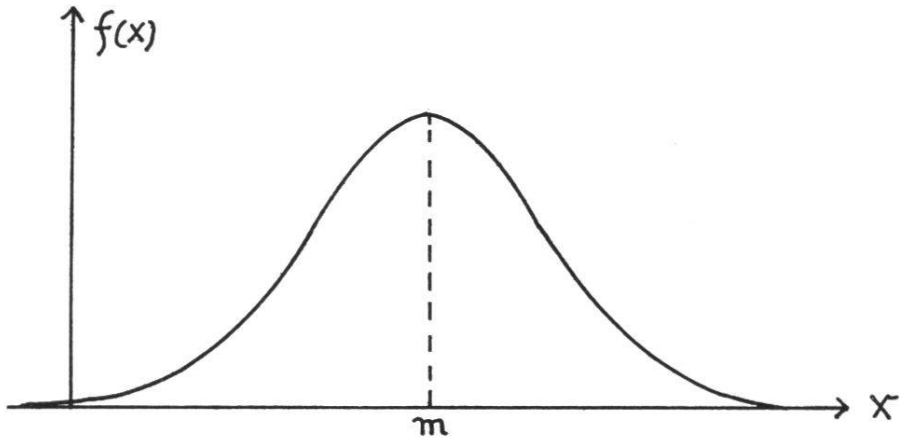
$$(1;2)\quad F(-\infty) = 0 \text{ und } F(+\infty) = 1$$

erinnert. Zur Veranschaulichung verweisen wir auf das bekannteste Beispiel, die « normale » Dichtefunktion f(x) – auch Gaußsche Glockenkurve genannt – mit ihrem Integral, der « normalen » Verteilungsfunktion F(t) :

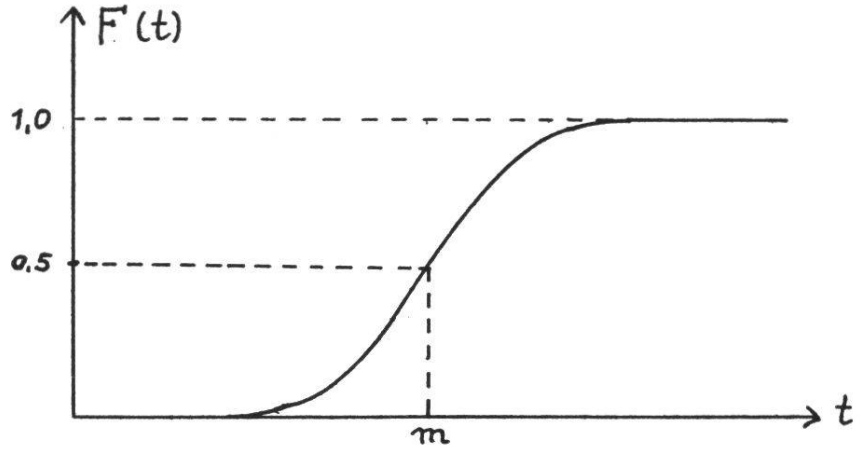
$$(1;3)\quad f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(x-m)^2}{2 \cdot \sigma^2}\right) \quad \text{mit}$$

$$(1;4)\quad F(t) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{t} \exp\left(-\frac{(x-m)^2}{2 \cdot \sigma^2}\right) dx\,;$$

und in graphischer Darstellung :



a)



b)

Fig. 1

Der Begriff der zufälligen Größe ist für die folgenden Ausführungen von Bedeutung, weil alle statistischen Untersuchungen über die Merkmale von Grundgesamtheiten auf die Betrachtung von zufälligen Größen hinauslaufen.

In einem Experiment bezeichnet man die n beobachteten Werte der zu untersuchenden Größe X mit x_i $(i=1,2,\ldots,n)$. Die Gesamtheit der möglichen n-Tupel bildet den *Stichprobenraum X*. Seine Dimension ist also gleich der Anzahl der entnommenen Stichprobenelemente.

Die Menge der in Frage kommenden Verteilungsfunktionen der zu untersuchenden zufälligen Größe bezeichnen wir mit $\langle P_\theta \rangle$, wobei der Parameter θ alle Werte eines Parameterraumes Ω durchlaufen kann. Die bei der Untersuchung zugelassenen Verteilungsfunktionen bilden also die Menge

$$(1;5) \quad \langle P_\theta \,;\, \theta \varepsilon \Omega \rangle$$

Das Problem des Prüfens einer Hypothese durch einen Test stellt sich dann in folgender Form :

Nach der Durchführung des Experiments hat man auf Grund der Daten mittels einer Testvorschrift eine der beiden folgenden Entscheidungen zu treffen :

d_1 : $\theta \varepsilon \omega \subset \Omega$ d. h. die den Ausgang des Experiments herbeiführende Verteilung liegt in dem Unterraum ω des Hypothesenraumes Ω.

d_2 : $\theta \varepsilon \Omega - \omega$ d. h. die Verteilung der zufälligen Größe liegt in dem zu ω komplementären Unterraum $\Omega - \omega$ des Hypothesenraumes Ω.

Bei der Planung des Experiments stellt man eine Annahme über seinen Ausgang auf und legt den Bereich ω so fest, daß die zu prüfende Annahme – die dann durch den Test entweder bestätigt oder widerlegt wird – durch $\theta \varepsilon \omega$ charakterisiert ist.

Diese Annahme nennen wir *Nullhypothese* H_0.

Die zu ω komplementäre Hypothesenmenge $\Omega - \omega$ trägt den Namen *Alternativhypothese* H_a oder kurz Alternative. Eine statistische Hypothese schlechthin spezifiziert eine Untermenge ω von Ω und stellt fest, daß die Verteilungsfunktion der zu untersuchenden zufälligen Größe X eine Funktion P_θ mit $\theta \varepsilon \omega$ ist.

Für jede Stichprobe $(x_1, x_2, \ldots, x_n)$ wird nun durch eine *Testvorschrift* anhand einer *Testgröße T*, die ihrerseits ebenfalls eine zufällige Größe ist, genau vorgeschrieben, welche der beiden Entscheidungen d_1 oder d_2 zu treffen ist. Diesen beiden möglichen Entscheidungen entsprechen zwei zueinander komplementäre Punktmengen des Stichprobenraumes *X*.

Diejenige Punktmenge des Raumes *X* (und ebenfalls die ihr zugeordnete Punktmenge im Wertebereich W der Testgröße T), die die Entscheidung d_2 nach sich zieht, heißt *kritische Region* oder *Verwerfungsbereich V*, da sie das Verwerfen der zu prüfenden Nyllhypothese H_0 bewirkt.

Die Entscheidung trifft man im allgemeinen nicht primär anhand der Punkte des Stichprobenraumes, sondern mittels der ihnen zugeordneten Werte t der Testgröße T. Analog dazu hat man dann die kritische Region V nicht primär im Stichprobenraum abzustecken, sondern als eine Punktmenge w im Wertebereich W der Testgröße T.

Faßt man alle Punkte des Stichprobenraumes, die durch T in w abgebildet werden, zur Menge s zusammen, so besteht ein Test für die statistische Hypothese H_0 auch in der Aufteilung des Stichproben-

raumes X in die beiden komplementären Teilmengen s und $X - s$ mit der Maßgabe, daß H_0 zugunsten von H_a zu verwerfen ist, sofern das beobachtete n-Tupel $(x_1, x_2, \ldots, x_n)$ in s liegt.

Nullhypothese und Alternative können wir also kurz wie folgt beschreiben :

$$H_0 : \theta \varepsilon \omega, \quad \text{falls } t \varepsilon W - w \text{ bzw. } (x_1, x_2, \ldots, x_n) \varepsilon X - s$$
$$H_a : \theta \varepsilon \Omega - \omega, \text{ falls } t \varepsilon w \quad \text{bzw. } (x_1, x_2, \ldots, x_n) \varepsilon s.$$

2. Die möglichen Fehler und die Güte eines Tests

Die Wahl des Verwerfungsbereiches $V = w$ wird unter folgenden Gesichtspunkten vorgenommen :

Man gibt zunächst das Testniveau, die sogenannte *Irrtumswahrscheinlichkeit* α, mit $0 < \alpha < 1$, vor und wählt dazu im Wertebereich W der Testgröße T eine zugehörige kritische Region $V = w$ derart aus, daß H_0 aufgrund einer Stichprobe höchstens mit einer Wahrscheinlichkeit α abgelehnt wird, vorausgesetzt, daß H_0 in der Grundgesamtheit wirklich gilt ; d. h., daß $\theta \varepsilon \omega$ ist. α gibt also die Wahrscheinlichkeit des Verwerfens der Nullhypothese an, und zwar unter der Voraussetzung, daß sie richtig ist. α bezeichnet man auch als die *Wahrscheinlichkeit eines Fehlers 1. Art.* Wir haben also

$$(2;1) \quad P(w\,;\, \theta \varepsilon \omega) \leqq \alpha.$$

Andererseits besteht aber die Möglichkeit, daß H_0 falsch ist. Um auch in diesem Falle die richtige Entscheidung, d. h. die Verwerfung von H_0, mit möglichst großer Wahrscheinlichkeit zu treffen, stellt man an die kritische Region $V = w$ noch die Forderung, daß sie unter der Voraussetzung der Richtigkeit von H_a eine möglichst große Wahrscheinlichkeit auf sich vereinigt.

Man fordert also

$$(2;2) \quad P(w\,;\, \theta \varepsilon \Omega - \omega) \text{ maximal.}$$

Der im Falle der Richtigkeit von H_a mögliche Fehler ist die Wahrscheinlichkeit des Nichtverwerfens (d. h. des Annehmens) von H_0, obwohl es falsch ist. Sie wird mit β bezeichnet und trägt den Namen *Wahrscheinlichkeit eines Fehlers 2. Art.* Man wird bestrebt sein, diesen

Fehler möglichst klein zu halten ; das ist gleichwertig mit der Forderung (2;2).

Das Problem, eine gute kritische Region und damit einen guten Test zu bekommen, liegt in der Schwierigkeit, bei vorgegebenem Fehler 1. Art einen möglichst kleinen Fehler 2. Art zuzulassen.

Statt der Forderung, den Fehler 2. Art möglichst klein zu halten, kann man auch verlangen, daß die komplementäre Wahrscheinlichkeit nach (2;2), nämlich diejenige des Verwerfens von H_0, wenn es falsch ist, möglichst groß wird. Diese Wahrscheinlichkeit

$$(2;3)\quad M(w\,;\,\theta) = 1 - \beta = P(w\,;\,\theta\varepsilon\Omega\text{-}\omega)$$

trägt den Namen *Güte- oder Machtfunktion* des zu V = w gehörenden Tests.

Von großer theoretischer wie auch praktischer Bedeutung für die Testtheorie ist auch die sogenannte *Operations- oder Testcharakteristik*

$$(2;4)\quad L(w\,;\theta) = 1 - P(w\,;\,\theta\varepsilon\Omega\text{-}\omega).$$

Sie gibt die Annahmewahrscheinlichkeit für H_0 in Abhängigkeit von θ an. Zwischen Gütefunktion und Operationscharakteristik besteht folgende Beziehung :

$$(2;5)\quad L(w\,;\,\theta) = 1 - M(w\,;\,\theta).$$

Insbesondere gilt $L(w\,;\,\theta\varepsilon\omega) = 1 - \alpha$

und $L(w\,;\,\theta\varepsilon\Omega\text{-}\omega) = \beta.$

Da die Punkte des Stichprobenraumes im allgemeinen sowohl unter H_0 wie auch unter H_a eine von Null verschiedene Wahrscheinlichkeit tragen, läßt sich ein Test mit der Irrtumswahrscheinlichkeit $\alpha = 0$ und der Güte 1, d. h.

$$(2;6)\quad L(w\,;\,\theta) = \begin{cases} 1, \text{ falls } \theta\varepsilon\omega \\ 0, \text{ falls } \theta\varepsilon\Omega\text{-}\omega, \end{cases}$$

nicht verwirklichen.

Die Berechnung der Güte- oder Machtfunktion (engl. : power-function), die von NEYMAN und PEARSON in die Testtheorie eingeführt wurde, bereitet besonders bei den noch zu besprechenden verteilungsfreien Methoden große Schwierigkeiten, weil sie die Kenntnis der Verteilungsfunktion der zufälligen Größe unter H_a voraussetzt. Im allgemeinen beschränkt man sich bei den verteilungsfreien Verfahren

auf einen Gütevergleich mit dem verteilungsbehafteten Student-Test und nimmt dazu normal verteilte Grundgesamtheiten an.

3. Einseitiges und zweiseitiges Testen

Die Wahl der kritischen Region als Teil des Wertebereiches W der Testgröße T ist unter Berücksichtigung der im vorigen Paragraphen dargestellten Gesichtspunkte vorzunehmen.

Nun interessiert man sich bei praktischen Untersuchungen häufig lediglich dafür, ob sich die zu untersuchende Eigenschaft in einer ganz bestimmten Richtung (z. B. nur Vergrößerung oder nur Verkleinerung) verändert hat. In diesem Falle spricht man von *einseitigem* (entweder *rechtsseitigem* oder *linksseitigem*) Testen. Dieses läuft auf eine Einschränkung des Hypothesenraumes hinaus.

Testet man beispielsweise in einer Grundgesamtheit die Nullhypothese H_0: Mittelwert $E(X) = \mu = 0$ gegenüber der Alternative H_a: $E(X) = \mu > 0$, so handelt es sich um ein rechtsseitiges Testen mit dem Raum der zugelassenen Hypothesen $\Omega = \langle \mu \geqq 0 \rangle$ und speziell H_0: $\omega = \langle \mu = 0 \rangle$ sowie H_a: $\Omega - \omega = \langle \mu > 0 \rangle$. Entsprechendes gilt für linksseitiges Testen.

Interessiert man sich dagegen für Abweichungen in beiden Richtungen, so testet man *zweiseitig* gegenüber H_a: $\langle \mu \neq 0 \rangle$.

Da die Testgrößen im allgemeinen eindimensionale symmetrische Verteilungen mit dem Mittelwert 0 besitzen, kann man die kritischen Regionen durch Schranken wie folgt angeben:

I. $T > T_\alpha$ (rechtsseitig) mit der Irrtumswahrscheinlichkeit α,
II. $T < -T_\alpha$ (linksseitig) ebenfalls auf dem Niveau α,
III. $|T| > T_\alpha$ (zweiseitig) auf dem Niveau 2α.

Dabei bestimmt man T_α aus der Beziehung

$$(3;1) \quad P(T > T_\alpha) \leqq \alpha.$$

4. Einfache und zusammengesetzte Hypothesen

Man wird an einen Test die Forderung stellen, daß er der mächtigste (d. h. derjenige mit der größten Güte) für das gegebene Problem ist. Leider ist man aber in der Theorie noch weit davon entfernt, ein Verfahren angeben zu können, das zu jedem Verwerfungsbereich die Konstruktion eines *mächtigsten Testes* gestattet. Lediglich für den Spezialfall, daß H_0 und H_a in einem sogleich zu definierenden Sinne « *einfach* » sind, ist dieses Problem durch ein Lemma von Neyman und Pearson gelöst.

Wir hatten die Hypothesen durch H_0: $\theta \varepsilon \omega$ und H_a: $\theta \varepsilon \Omega - \omega$ definiert. Besteht nun ω lediglich aus einem einzigen θ-Wert, dann heißt H_0 eine *einfache* Hypothese. Entsprechend heißt H_a einfach, wenn $\Omega - \omega$ nur aus einem einzigen θ-Wert besteht. *Zusammengesetzt* heißt eine Hypothese immer dann, wenn sie mehrere Parameterwerte umfaßt.

Die testtheoretischen Grundbegriffe lassen sich an folgendem Beispiel mit einfachen Hypothesen verdeutlichen :

Es sei bekannt, daß die Grundgesamtheit entweder die Dichte $f_0(x)$ oder $f_1(x)$ besitzt. Wir setzen also :

$$H_0 : f(x) = f_0(x) \text{ und } H_a : f(x) = f_1(x)$$

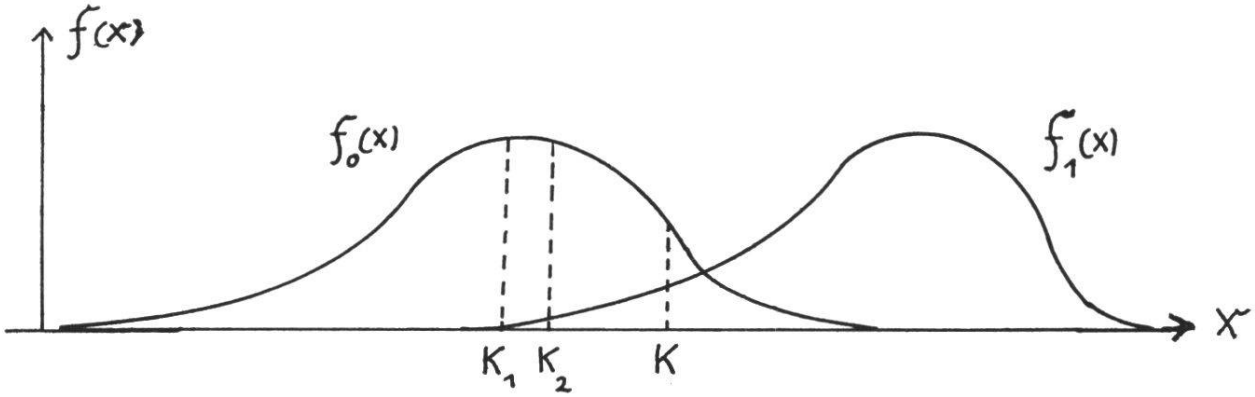


Fig. 2

Ist das Testniveau – d. h. die Wahrscheinlichkeit eines Fehlers 1. Art – mit dem Wert α vorgegeben, so hat man als zugehörigen Verwerfungsbereich einen Teilbereich der x-Achse auszuwählen, beispiels-

weise $V = \langle K_1 \leqq x \leqq K_2 \rangle$. Dazu müssen K_1 und K_2 so gewählt werden, daß

$$(4;1) \quad \int_{K_1}^{K_2} f_0(x)dx = \alpha$$

ist. Der zugehörige Fehler 2. Art beträgt dann

$$(4;2) \quad \int_{-\infty}^{K_1} f_1(x)dx + \int_{K_2}^{+\infty} f_1(x)dx.$$

Je nachdem, ob nun das Stichprobenelement x_1 in den Bereich V fällt oder nicht, wird H_0 verworfen oder angenommen.

Die angeschriebenen Integrale geben die Flächeninhalte unter den Dichtefunktionen f_0 bzw. f_1 in den angegebenen Intervallen an. Daher sieht man sofort, daß bei dieser Wahl des Verwerfungsbereiches der Fehler 2. Art mehr als 90 % beträgt.

Um einen mächtigsten Test, d. h. einen solchen mit möglichst kleinem Fehler 2. Art zu bekommen, wird man stattdessen als Verwerfungsbereich $V = \langle x \geqq K \rangle$ mit

$$(4;3) \quad \int_{K}^{+\infty} f_0(x)dx = \alpha$$

wählen. Als Fehler 2. Art bekommt man dann nur noch

$$(4;4) \quad \int_{-\infty}^{K} f_1(x)dx.$$

5. Einige weitere Forderungen an Testverfahren

Man wird stets bemüht sein, Tests mit in gewisser Hinsicht optimalen Eigenschaften zu verwenden. Zu diesem Zwecke hat man eine Reihe von Forderungen aufgestellt, die man sinnvollerweise von einem Test verlangen kann. Die wichtigsten sollen hier angeführt werden.

a) Unverfälschte oder biasfreie Tests: Ein Verwerfungsbereich w mit der Eigenschaft

$$(5;1) \quad P(x \varepsilon w\,;\, H_0) > P(x \varepsilon w\,;\, H_a)$$

würde dazu führen, daß die Nullhypothese häufiger verworfen würde unter der Voraussetzung, daß sie richtig ist, als unter der Voraussetzung,

daß sie falsch ist. Ein Test mit einem solchen Verwerfungsbereich heißt nach Neyman und Pearson « biased » oder mit Bias (d. h. mit einem systematischen Fehler) behaftet. Die wünschenswerte Biasfreiheit eines durch einen bestimmten Verwerfungsbereich festgelegten Testverfahrens ist mathematisch oft sehr schwer zu verwirklichen.

b) *Gleichmäßig mächtigere Tests:* Wenn zwei kritische Regionen w und w′ mit

$$(5;2)\quad P(x \varepsilon w\,; H_o) = P(x \varepsilon w'\,; H_o)$$

beide biasfreie Tests ergeben, dann ist w eine bessere kritische Region als w′, wenn die Ungleichung

$$(5;3)\quad P(x \varepsilon w\,; H_a) > P(x \varepsilon w'\,; H_a)$$

für alle zulässigen Alternativhypothesen gilt. Der auf dem Verwerfungsbereich w beruhende Test heißt dann *gleichmäßig mächtiger* als der auf w′ beruhende. Man wird also bemüht sein, nach Möglichkeit *gleichmäßig mächtigste Tests* zu konstruieren.

c) *Die Konsistenz eines Tests:* Die im Stichprobenraum *X* abgesteckte kritische Region w hängt von dessen Dimension ab. Wir drücken das kurz durch $W = w_n$ aus. Nach Wald und Wolfowitz ist ein Test genau dann *konsistent,* wenn die Wahrscheinlichkeit für das Verwerfen von H_o, sofern H_a gilt, gegen eins strebt, vorausgesetzt, daß der Stichprobenumfang (d. h. die Dimension des Stichprobenraumes) gegen unendlich geht. Man fordert also

$$(5;4)\quad \lim_{n \to \infty} P(x \varepsilon w_n\,; H_a) = 1.$$

6. Das Zwei- (bzw. k-) Stichprobenproblem

Ein Experiment möge bei g-facher Wiederholung die Werte $x_1, x_2, \ldots, x_g$ geliefert haben und dann unter abgeänderten Verhältnissen bei h-facher Wiederholung die Werte $y_1, y_2, \ldots, y_h$. Man möchte nun wissen, ob die g+h Werte als aus ein und derselben Grundgesamtheit stammend betrachtet werden können.

Wir betrachten die beiden Stichproben $\langle x_i \rangle$ und $\langle y_k \rangle$ als Realisationen (d. h. zufällig angenommene Werte) zweier Mengen von unabhängigen zufälligen Größen, nämlich der $X_1, X_2, \ldots, X_g$ und der Y_1,

$Y_2, \ldots, Y_h$. In geometrischer Sprechweise stellen alle möglichen Realisationen der einen (g+h) dimensionalen zufälligen Größe $(X_1, X_2, \ldots, X_g, Y_1, Y_2, \ldots, Y_h)$ eine Punktmenge in einem (g+h=n)-dimensionalen Raume dar. Jede Stichprobe repräsentiert einen Punkt dieses Raumes. Deshalb nennt man die Menge aller Stichproben vom Umfange n den n-dimensionalen Stichprobenraum.

Man macht nun zunächst die Annahme, daß die zufälligen Größen X_i alle der Verteilungsfunktion F und die zufälligen Größen Y_k alle der Verteilungsfunktion G gehorchen.

Das Problem der zwei Stichproben besteht dann in der Feststellung, ob man *beide* Stichproben als aus ein und derselben Grundgesamtheit entnommen ansehen kann. Mit anderen Worten : Es ist festzustellen, ob ein gewisser Effekt vorhanden ist oder nicht.

Zu diesem Zwecke stellt man als zu prüfende Nullhypothese die Forderung $H_0: F = G$ (manchmal auch nur : $E(X) = E(Y)$) auf. Das besagt, daß beide Stichproben derselben Grundgesamtheit angehören. Die Alternativhypothese, gegenüber der man den Test aufbaut, soll besagen, daß beide Stichproben zu verschiedenen Grundgesamtheiten gehören. Das heißt, man hat als Alternativhypothese $H_a: F \neq G$ (manchmal fordert man auch $H_a: E(X) \neq E(Y)$).

Eine Verallgemeinerung des Problems der zwei Stichproben ergibt das *Problem der k Stichproben.* Es besteht in der Untersuchung, ob k zufällige Stichproben derselben Grundgesamtheit entnommen sein können, d. h. ob sie zu Populationen mit identischen Verteilungen gehören.

Wenn die Verteilungen der zu untersuchenden k Grundgesamtheiten mit F_i bezeichnet werden, so wird man als Nullhypothese $H_0: F_1 = F_2 = \ldots = F_k$ festsetzen. Die Klasse der zuzulassenden Alternativen wird man in diesem Falle entsprechend dem größeren Schwierigkeitsgrade der mathematischen Behandlung des Problems kaum als $H_a: F_i \neq F_k$ für $i \neq k$ festsetzen, sondern durch Zusatzforderungen einschränken.

Beim *Einstichprobenproblem* liegt insofern ein etwas anderer Sachverhalt vor, als hier nicht die Gleichheit der zu mehreren Stichproben gehörenden Verteilungen untersucht wird, sondern – da nur eine Stichprobe vorliegt – die zu einer einzigen Stichprobe gehörende Verteilung auf Übereinstimmung mit einer theoretisch angenommenen geprüft wird.

Das Ein- und das k-Stichprobenproblem interessieren hier nur am Rande, und zwar insofern, als manche der im Abschnitt B behandelten Testverfahren Erweiterungen für die Behandlung des Ein- und des k-Stichprobenproblems besitzen.

7. Parametrische und parameterfreie Methoden in der Mathematischen Statistik

Man kann die Verteilungsfunktion einer zufälligen Größe X im allgemeinen durch die Angabe von endlich vielen Parametern charakterisieren. Beispielsweise ist die eindimensionale Normalverteilung – Vergl. (1;3) und (1;4) – durch die Angabe von zwei Parametern, nämlich des Mittelwertes m und der Streuung σ, eindeutig festgelegt.

Aufgrund dieser Tatsache nennt man alle statistischen Methoden, die für das zu untersuchende Material (= Grundgesamtheit) das Gelten einer gewissen Verteilungsfunktion voraussetzen, *parametrische oder verteilungsbehaftete Methoden*. Bekanntlich wird bei biologischen und medizinischen Untersuchungen häufig – mit mehr oder weniger großer Berechtigung – eine Normalverteilung angenommen.

Demgegenüber nennt man alle statistischen Methoden, die keinerlei Voraussetzung über die Art der in der Grundgesamtheit vorliegenden Verteilung machen, *verteilungsfreie* oder *parameterfreie* bzw. *nicht-parametrische* Verfahren. Schwache Forderungen, wie etwa die Stetigkeit der Verteilungsdichte, werden allerdings auch hier aus theoretischen Erwägungen oft erhoben.

Da man sich bei statistischen Untersuchungen häufig Verhältnissen gegenübergestellt sieht, bei denen nicht die geringste Aussage über den Typ der zugrunde liegenden Verteilung gemacht werden kann, hat man vornehmlich in den letzten drei Jahrzehnten eine ganze Reihe von nicht-parametrischen Verfahren ersonnen.

Dementsprechend soll im nächsten Abschnitt B bei der Darstellung von Testmethoden für das Zweistichprobenproblem das Schwergewicht auf die verteilungsfreien Testverfahren gelegt werden.

Abschnitt B

Einige wichtige Tests für die Behandlung des Zweistichprobenproblems

In diesem Abschnitt soll eine Reihe von häufig verwendeten Prüfverfahren für das im § 6 beschriebene Problem der zwei Stichproben dargestellt werden. Im Rahmen dieser Arbeit kann es sich nur darum handeln, die für die Anwendung notwendige Kenntnis der jeweiligen Testvorschrift im engeren Sinne zu vermitteln. Die sonstigen theoretischen Eigenschaften des Tests und der Verteilungen der Testgröße können hingegen nur angedeutet werden.

Zur weiteren Vertiefung in die Probleme der Testtheorie kann von der am Schluß zitierten Literatur ausgegangen werden.

Während im folgenden § 8 der bekannte und häufig verwendete parametrische Student-Test beschrieben wird, werden sich die dann folgenden Paragraphen ausschließlich mit nichtparametrischen Testverfahren befassen, das heißt mit solchen, bei denen man nicht von vornherein genötigt ist, dem zu untersuchenden Charakteristikum eine ganz bestimmte Verteilung zu unterschieben.

8. Der Studentsche t-Test

Der t-Test geht auf eine Arbeit des englischen Statistikers Gosset zurück, die dieser 1908 unter dem Pseudonym « STUDENT » veröffentlicht hat.

Man nimmt an, daß die beiden zu vergleichenden Grundgesamtheiten – aus denen die beiden Stichproben stammen – unabhängig normal verteilt sind mit gleichen (oder zumindestens in etwa gleichen) Streuungen und möglicherweise verschiedenen Mittelwerten.

Aus den $n = g+h$ Meßwerten der beiden Stichproben $x_1, x_2, \ldots, x_g$ und $y_1, y_2, \ldots, y_h$ hat man zunächst folgende Ausdrücke zu berechnen:

$$(8;1)\quad D = \bar{x} - \bar{y} \text{ mit } \bar{x} = \frac{1}{g}\sum_{i=1}^{g} x_i \text{ und } \bar{y} = \frac{1}{h}\sum_{j=1}^{h} y_j$$

$$(8;2)\quad s^2 = \frac{1}{n-2}\left\{\sum_{i=1}^{g}(x_i-\bar{x})^2+\sum_{j=1}^{h}(y_j-\bar{y})^2\right\}$$

$$(8;3)\quad S^2 = \left(\frac{1}{g}+\frac{1}{h}\right)\cdot s^2.$$

Mit diesen Ausdrücken bildet man dann die Testgröße

$$(8;4)\quad t = \frac{D}{S}.$$

Diese Testgröße t ist eine zufällige Größe. Ihre Dichtefunktion hat man berechnet :

$$(8;5)\quad h(\tau) = \frac{\gamma}{\sqrt{f}}\cdot\left(1+\frac{\tau^2}{f}\right)^{\frac{-(f+1)}{2}}$$

$$\text{mit } \gamma = \pi^{-1/2}\cdot\Gamma\left(\frac{f+1}{2}\right)\cdot\Gamma\left(\frac{f}{2}\right)^{-1}.$$

Darin heißt $f = n-2$ die Zahl der Freiheitsgrade. Γ ist das Funktionszeichen der Gammafunktion.
$h(\tau)$ hat die Gestalt einer Glockenkurve und nähert sich für $f \to \infty$ der normalen Dichtefunktion (1;3).

Die zur Dichte $h(\tau)$ und damit zu t gehörende Verteilungsfunktion

$$(8;6)\quad H(a) = \int_{-\infty}^{a} h(\tau)d\tau$$

heißt *Student-Verteilung*.

Die Testvorschrift lautet nun wie folgt :

Im Falle zweiseitigen Testens stellt man fest, ob der absolute Betrag der Testgröße $t = D/S$ eine positive Schranke t_α überschreitet, die vom vorgegebenen Testniveau α (=Wahrscheinlichkeit eines Fehlers 1. Art) abhängt. t_α ist so berechnet, daß die Wahrscheinlichkeit des Ereignisses $|t|>t_\alpha$ gleich α ist. Die Werte t_α liegen als Tabulierung der t-Verteilung vor ; beispielsweise bei VAN DER WAERDEN (1957). Die Nullhypothese H_0 : $E(X) = E(Y)$ (d. h. gleiche Mittelwerte für beide Verteilungen) ist zu verwerfen, falls $|t|>t_\alpha$ ausfällt. Die Alternativhypothese würde in diesem Falle unterschiedliche Mittelwerte für die Verteilungen der beiden zufälligen Größen X und Y fordern.

Will man den Test nur einseitig anwenden, so nimmt man als Alternative :

$$H_a: E(X) > E(Y)$$
$$\text{oder } H_a: E(X) < E(Y).$$

In diesen Fällen wird die Nullhypothese verworfen, sofern $t > t_\alpha$ beziehungsweise $t < -t_\alpha$ ausfällt. Die Irrtumswahrscheinlichkeit beträgt dann nur jeweils $\alpha/2$.

Beispiele zum t-Test wie auch zu den anderen noch zu behandelnden Testverfahren folgen im Abschnitt C.

9. Der Zeichentest

Der Zeichentest trägt seinen Namen wegen der Tatsache, daß er als Ausgangsdaten nur die Vorzeichen der Differenzen von Wertepaaren benutzt. Man hat also die Stichprobenwerte aus den beiden Grundgesamtheiten paarweise zu entnehmen. Dieser Test kann auch dann angewendet werden, wenn ein quantitatives Messen der Werte nicht möglich ist, sondern sich nur Größenvergleiche anstellen lassen. Das kommt besonders im Bereich der psychologischen Forschung häufig vor. Hinsichtlich der Anwendbarkeit des Zeichentests ist jedoch stets zu beachten, daß die Werte beider Stichproben in Paaren (x_i, y_i) vorliegen müssen, wobei die x_i der einen und die y_i der anderen Grundgesamtheit zu entnehmen sind. Der Test eignet sich also speziell zum Vergleich zweier Behandlungsverfahren, die auf dieselben Versuchsobjekte nacheinander angewendet werden.

Es seien nun n Paare unabhängiger Beobachtungen (x_i, y_i) gegeben. Man kann dann die *Anzahl der positiven Differenzen* unter den zu bildenden n Differenzen $(x_i - y_i)$ als Testgröße verwenden.

Die Möglichkeit von « Bindungen » – d. h. von Fällen, in denen $x_i = y_i$ ist und man als Differenz Null erhält – kann theoretisch wegen der Stetigkeit der Verteilungen ausgeschlossen werden. Sofern solche Bindungen in der Praxis infolge Meßungenauigkeit vorkommen, kann man sie durch einen einfachen Zufallsprozeß lösen oder aber einfach fortlassen.

Hat man die Differenzen

$$(9;1) \quad z_i = x_i - y_i \quad \text{mit } i = 1,2,3,\ldots,n$$

beobachtet, so mögen k von ihnen positiv und die restlichen n-k negativ ausgefallen sein. Die z_i kann man dann als Realisationen von un-

abhängigen zufälligen Größen Z_i auffassen. Unter H_0 (gleiche Verteilungen für die beiden zu untersuchenden Grundgesamtheiten) werden die z_i mit gleicher Wahrscheinlichkeit – d. h. mit $^1/_2$ – positiv wie negativ ausfallen :

$$(9;2)\quad P(Z_i > 0) = P(Z_i < 0).$$

Man hat also unter H_0 gleichgroße Anzahlen von positiven und negativen z-Werten zu erwarten.

Im Falle einseitigen Testens, d. h. gegenüber $H_a : X > Y$, wird die kritische Region dadurch festgelegt, daß man H_0 verwirft, sobald die Anzahl N der positiven z_i eine kritische Schranke N_α überschreitet.

Die Bestimmung von N_α zu vorgegebenem α geschieht wie folgt : Wenn die Anzahl der positiven Differenzen gleich k und damit diejenige der negativen Differenzen gleich n-k ist, dann kann man mit Hilfe der *Binomialverteilung für* $p = q = {}^1/_2$ Vertrauensgrenzen bilden, zwischen denen nach Vorgabe des Testniveaus die Zahlen k und n-k vermutlich liegen.

Unter H_0 ist die Wahrscheinlichkeit dafür, daß mehr als r von den n Werten z_i positiv ausfallen, gleich

$$(9;3)\quad P(N > r\,;\,H_0) = \left\{ \binom{n}{r+1} + \binom{n}{r+2} + \ldots + \binom{n}{n} \right\} \cdot \left(\frac{1}{2}\right)^n.$$

Nun hat man r so zu bestimmen, daß es die kleinste Zahl ist, für die nach (9;3) noch

$$(9;4)\quad P(N > r\,;\,H_0) = \ldots = \leqq \alpha \text{ gilt.}$$

Dieses r nennen wir r_α. Es übernimmt die Rolle der kritischen Schranke N_α.

H_0 wird verworfen, sobald für die vorliegenden Stichproben N größer als N_α ausfällt.

Bei zweiseitiger Anwendung wird H_0 nicht nur verworfen, wenn die Zahl der positiven Differenzen N größer als N_α ist, sondern auch dann, wenn die komplementäre Zahl n – N der negativen Differenzen dieselbe Schranke N_α übersteigt. Das Testniveau beträgt in diesem Falle wieder 2α.

Tafeln der N_α-Werte für die gängigsten Stichprobenumfänge und die üblichen Testniveaus findet man unter anderem bei VAN DER WAERDEN und in der Monographie von VAN DER WAERDEN – NIEVERGELT.

10. Der Iterationen-Test von Wald und Wolfowitz

Dieser Test geht davon aus, daß man die Werte der beiden Stichproben in einer einzigen Rangordnung der Größe nach anordnen kann. (Vgl. WALD und WOLFOWITZ). Als Testgröße benutzt man dann die Häufigkeit des Aufeinanderfolgens (d. h. Iterierens) von Beobachtungswerten *einer* Stichprobe innerhalb der gemeinsamen Rangordnung:

Sei $z_1, z_2, \ldots, z_{g+h=n}$ die größenmäßig geordnete Vereinigung der beiden Stichproben und $V = (v_1, v_2, \ldots, v_n)$ eine folgendermaßen definierte Folge:

(10;1) $v_i = a$, falls z_i aus der Stichprobe $\langle x_i \rangle$ stammt
$v_i = b$, falls z_i aus der Stichprobe $\langle y_k \rangle$ stammt.

Eine Teilfolge $v_{s+1}, v_{s+2}, \ldots, v_{s+r}$ nennt man *eine Iteration*, falls

(10;2) $v_{s+1} = v_{s+2} = \ldots = v_{s+r}$ gilt und dabei sowohl
(10;3) $v_s \neq v_{s+1}$, falls $s > 0$, wie auch
(10;4) $v_{s+r} \neq v_{s+r+1}$, falls $s+r < g+h$, ist.

Die Folge a a a b a b b b a a b b hat demnach die folgenden sechs Iterationen:

aaa ; b ; a ; bbb ; aa ; bb.

Als Testgröße definiert man daraufhin:

(10;5) R = Anzahl der Iterationen in V.

Für den Fall, daß H_o gilt, werden die x_i und die y_k in der gemeinsamen Rangordnung ziemlich gleichmäßig verteilt sein, so daß für R ein maximaler Wert zu erwarten ist.

Unter H_a (verschiedene Verteilungen für die beiden Grundgesamtheiten) wird sich der Wert von R vermindern, da ein « Entmischen » der x_i und der y_k zu erwarten ist.

Der Verwerfungsbereich zum Prüfen der Nullhypothese auf dem Niveau α wird wie üblich durch

(10;6) $R < R_\alpha$

definiert, wobei R_α zu vorgegebenem α aus der Beziehung

(10;7) $P(R < R_\alpha ; H_o) \leqq \alpha$

zu bestimmen ist.

Es läßt sich zeigen, daß unter gewissen Zusatzbedingungen die Testgröße R asymptotisch normal verteilt ist mit dem Mittelwert

$$(10;8)\quad E(R) = \frac{2g}{1+c}$$

und dem Streuungsquadrat

$$(10;9)\quad \sigma^2\,(R) = \frac{4gc}{(1+c)^3},$$

wobei $c = g/h$ ist.

11. Der Wilcoxon-Test

Der ursprünglich von WILCOXON 1945 vorgeschlagene Test beschränkt sich auf den Fall, daß die beiden Stichproben gleichen Umfang haben und die Ergebnisse paarweise gewonnen werden, wie uns das bereits beim Zeichentest begegnete. Diese Einschränkungen wurden dann 1947 von MANN und WHITNEY durch eine Verallgemeinerung des Testverfahrens beseitigt. Demzufolge wird der Wilcoxon-Test in der angelsächsischen Literatur häufig als U-Test von MANN und WHITNEY bezeichnet.

X und Y seien wieder zwei unabhängige zufällige Größen mit den unbekannten Verteilungsfunktionen $F(t)$ beziehungsweise $G(t)$. Man nennt dann X stochastisch kleiner bzw. größer als Y, wenn für jeden endlichen Wert von t die Ungleichung

$$(11;1)\quad F(t) > G(t) \text{ bzw. } F(t) < G(t)$$

gilt. Die zu prüfende Nullhypothese soll auf Gleichheit der beiden Verteilungen lauten : $H_o : F = G$. Getestet wird gegenüber der Alternativhypothese $H_a : F(t) > G(t)$.

Nun habe man als Realisationen der beiden zufälligen Größen die beiden Stichproben $x_1, x_2, \ldots, x_g$ und $y_1, y_2, \ldots, y_h$ erhalten. Wie im vorigen Paragraphen beim Iterationentest stellt man auch hier die Rangordnung der Werte beider Stichproben auf. (Parameterfreie Testverfahren, die die Rangordnung der Meßergebnisse zum Ausgangspunkt haben, pflegt man als *Rangtests* zu bezeichnen !)

Für ein Beispiel mit $g = 3$ und $h = 6$ könnte man etwa folgende Rangordnung erhalten :

$$(11;2)\quad y_2, x_3, y_3, x_1, x_2, y_6, y_5, y_4, y_1.$$

Im Falle stetiger Verteilungen ist diese Anordnung immer herstellbar, da gleiche Meßwerte nur mit der Wahrscheinlichkeit Null vor-

kommen, d. h. unmöglich sind. Falls in der Praxis – etwa aus Gründen der Meßungenauigkeit – gleiche Werte, d. h. Bindungen, vorkommen, löst man sie durch einen Zufallsprozeß auf oder läßt sie einfach unberücksichtigt.

Kommt nun in dieser Rangordnung ein y_k an einer früheren Stelle als ein x_i, so spricht man von einer *Inversion*. In dem Beispiel (11;2) bildet der erste x-Wert, nämlich x_3, genau eine Inversion, und zwar mit dem vorausgehenden y_2. Der zweite und der dritte x-Wert, x_1 und x_2, bilden je zwei Inversionen, und zwar beide mit den zwei vorausgehenden y-Werten y_2 und y_3. Und so fort.

Die *Gesamtzahl U* derartiger Inversionen in der Rangordnung ist nun unsere *Testgröße*.

Liegen aus den beiden Stichproben insgesamt $g+h=n$ Elemente vor, so sind unter H_0 alle $n!$ möglichen Anordnungen oder Permutationen gleich wahrscheinlich; und zwar hat jede die Wahrscheinlichkeit

$$(11;3)\quad P = 1/n!$$

Wir haben nunmehr den Verwerfungsbereich V aus einer Teilmenge dieser $n!$ möglichen Anordnungen zu bilden. An V haben wir die Forderung gestellt, bei vorgegebenem Fehler 1. Art einen möglichst kleinen Fehler 2. Art – d. h. eine möglichst große Güte – zu liefern. Wenn wir als Wahrscheinlichkeit eines Fehlers 1. Art α vorgeben, so darf V höchstens $\alpha \cdot n!$ von den $n!$ möglichen Anordnungen umfassen, damit die Forderung

$$(11;4)\quad P(V\,;\,H_0) \leqq \alpha$$

gewahrt bleibt. Hinzu kommt die Forderung, daß

$$(11;5)\quad P(V\,;\,H_a) = \text{maximal}$$

angestrebt werden soll. Wir wählen hier diejenigen $\alpha \cdot n!$ Anordnungen als Elemente von V aus, die die wenigsten Inversionen enthalten; unter anderem insbesondere auch den Extremfall mit *keiner* Inversion: $x_1, x_2, \ldots, x_g, y_1, y_2, \ldots, y_h$.

Der Wilcoxon-Test schreibt nunmehr vor, die Nullhypothese zu verwerfen, sobald die Anzahl der Inversionen U eine festgesetzte Schranke U_α unterschreitet. U_α hat die Forderung

$$(11;6)\quad P(U < U_\alpha\,;\,H_0) \leqq \alpha$$

zu erfüllen, die besagt, daß unter H_0 die Anzahl der Anordnungen mit

$U<U_\alpha$ höchstens α n! beträgt. Damit haben wir den Wilcoxon-Test für die einseitige – und zwar linksseitige – Anwendung definiert.

Die Bestimmung der Inversionenzahl kann man sehr praktisch auch aus den Rangnummern vornehmen: Es sei $z_1, z_2, \ldots, z_{g+h=n}$ die aus den beiden Stichproben gebildete Rangordnung. Wenn darin das kleinste aus einem x-Wert hervorgegangene z die Rangnummer r_1 trägt, dann bildet es mit den r_1-1 vorangehenden y-Werten genau r_1-1 Inversionen. Das zweitkleinste aus einem x-Wert hervorgegangene z möge die Rangnummer r_2 tragen; es bildet dann mit den r_2-2 vorausgehenden y-Werten genau r_2-2 Inversionen. Und so fort.

Man erhält also als Summe der Rangnummern der x_i und damit als Anzahl der Inversionen

$$(11;7)\quad U = (r_1-1) + (r_2-2) + \ldots + (r_g-g) = \sum_{i=1}^{g} r_i - \sum_{i=1}^{g} i = \sum_{i=1}^{g} r_i - \frac{g(g+1}{2}.$$

Da der Ausdruck $g(g+1)/2$ bei vorgegebenen Stichprobenumfängen konstant ist, kann man ebensogut den Ausdruck $R_x = \sum_{i=1}^{g} r_i$ d.h. die Summe der Rangnummern der x_i, als Testgröße nehmen. Häufig verwendet man stattdessen auch die Summe der Rangnummern der y_k. Zwischen beiden besteht die Beziehung

$$(11;8)\quad R_z = R_x + R_y = \frac{(n+1)n}{2}.$$

Bei *linksseitiger* Anwendung wird H_o genau dann zugunsten von $H_a: F(t)>G(t)$ verworfen, wenn die Anzahl der Inversionen U unter der Schranke U_α liegt. Dann ist zu erwarten, daß im allgemeinen $X<Y$ gilt. Insbesondere ist das der Fall, wenn die beiden Stichproben die extreme Rangordnung

$$(11;9)\quad x,x,x,\ldots,x,y,y,\ldots,y$$

aufweisen und damit die Minimalzahl an Inversionen $U_{min} = 0$ vorliegt.

Bei *rechtsseitiger* Anwendung des Wilcoxon-Testes wird H_o zugunsten von $H_a: F(t) < G(t)$ verworfen, wenn die Anzahl der Inversionen oberhalb einer Schranke U'_α liegt. Insbesondere ist das der Fall, wenn die Rangordnung

(11;10) y,y,...,y,x,x,...,x

mit der Maximalzahl an Inversionen $U_{max} = g \cdot h$ vorliegt. U'_α wird ähnlich wie U_α bestimmt.

Sind nun im Experiment Unterschiede in beiden Richtungen von Interesse, so wird man den Test *zweiseitig*, d. h. gegenüber der Vereinigung der beiden obigen Alternativmengen, aufbauen. In diesem Falle wird H_0 nicht nur verworfen, wenn (linksseitig) $U < U_\alpha$ ist, sondern auch dann, wenn (rechtsseitig) $U > U'_\alpha$ ausfällt. Die Irrtumswahrscheinlichkeit beträgt dann wieder 2α.

In der Praxis arbeitet man allerdings auch im Falle zweiseitigen Testens nur mit einer einzigen Schranke :

Berücksichtigt man nämlich neben der Anzahl der Inversionen U auch noch die Anzahl der *Nichtinversionen* U^+, so gilt

(11;11) $U + U^+ = g.h.$

Die Grenzfälle ergeben sich aus den Anordnungen

(11;9) mit $U = 0$ und $U^+ = g.h$ sowie
(11;10) mit $U = g.h$ und $U^+ = 0$.

Genau dann, wenn die Anzahl U der Inversionen die Schranke U'_α übersteigt, dann unterschreitet die Zahl U^+ der Nichtinversionen die Schranke U_α.

Daraus folgt : Bei zweiseitiger Anwendung des Wilcoxon-Testes wird H_0 verworfen, wenn $U < U_\alpha$ oder $U^+ < U_\alpha$ ist.

Die Bestimmung der Schranken U_α hat für kleinere Stichprobenumfänge mittels kombinatorischer Überlegungen zu erfolgen. Für größere Stichprobenumfänge macht man von dem Ergebnis Gebrauch, daß die Testgröße U asymptotisch normal verteilt ist mit dem Mittelwert

(11;12) $E_{g,h}(U) = \frac{gh}{2}$

und dem Streuungsquadrat

(11;13) $\sigma^2_{g,h}(U) = \frac{gh\,(g+h+1)}{12}.$

Man kann dann die Tafeln der Normalverteilung benutzen. Die nötigen Tabellen zur praktischen Anwendung des Wilcoxon-Testes findet man beispielsweise bei VAN DER WAERDEN (1957).

Zum Gebrauch der Tafeln sei abschließend noch folgendes bemerkt :

Nach den bisherigen Betrachtungen könnte man die Tafeln so aufbauen, daß zu den gebräuchlichsten Testniveaus ($1/100 = 1\,\%$; $1/50 = 2\,\%$; $1/20 = 5\,\%$) und zu verschiedenen Stichprobenumfängen g, h die Schranken U_α angegeben würden, bei deren Unter- bzw. Überschreiten die beiden Stichproben in den Verwerfungsbereich fallen und somit H_0 zu verwerfen ist.

Im allgemeinen verwendet man dagegen unmittelbar die Verteilung der Testgröße U und definiert als sogenannte *Testwahrscheinlichkeit* die Wahrscheinlichkeit des Ereignisses ($U \leqq u$; H_0), also

$$(11;14)\quad p(u) = P(U \leqq u\,;\, H_0),$$

wobei u der im Einzelfalle von der zufälligen Größe U angenommene Wert sein soll.

Liefert also das Experiment die Inversionenzahl u, so wird H_0 auf dem Niveau α verworfen, sofern

$$(11;15)\quad p(u) = P(U \leqq u\,;\, H_0) \leqq \alpha$$

ausfällt.

Wegen $P(U \leqq u)$ ist der Test hier *linksseitig* aufgebaut. *Rechtsseitiges* Testen führt man aus, indem man u durch

$$(11;16)\quad u' = g \cdot h - u$$

ersetzt, ohne die Bezeichnungen der beiden Stichproben zu vertauschen. Ergibt sich ein $u > g \cdot h/2$, so unterläßt man die Transformation nach (11;16) und nimmt statt p(u) die komplementäre Wahrscheinlichkeit

$$(11;17)\quad p(u') = 1 - p(u).$$

Im Falle *zweiseitigen* Testens prüft man sowohlp (u) als auch p(u′) nach (11;15). Die Irrtumswahrscheinlichkeit ist dann wiederum 2α.

Aufgrund dieser Überlegungen hat man dann nicht mehr die kritischen Schranken U_α zu tabulieren, sondern die Testwahrscheinlichkeiten nach (11;14), d. h. die Verteilungsfunktion von U unter H_0 zu den verschiedenen Stichprobenumfängen g und h.

12. Der X-Test von van der Waerden

In einer Reihe von Arbeiten hat VAN DER WAERDEN die Güte parameterfreier Tests untersucht und mit der des parametrischen Student-Tests verglichen. In diesem Zusammenhang entwickelt er eine wesentliche Abänderung des Wilcoxon-Testes, die er als X-Test bezeichnet.

Die $n = g + h$ Elemente der beiden Stichproben $x_1, x_2, \ldots, x_g$ und $y_1, y_2, \ldots, y_h$ mögen ihrer Größe nach die Rangordnung $z_1, z_2, z_3, \ldots, z_n$ bilden.

Wir bezeichnen nun mit

$$(12;1) \quad z = \Psi(u)$$

die *Umkehrfunktion* der normalen Verteilungsfunktion

$$(12;2) \quad u = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-(1/2)z^2} dz.$$

Wie im vorigen Paragraphen beim Wilcoxon-Test bezeichne r wiederum die Rangnummern in der Rangordnung der Stichprobenwerte.

Als Testgröße für den X-Test dient dann die Summe

$$(12;3) \quad X = \sum_r \Psi\left(\frac{r}{n+1}\right),$$

wobei r lediglich diejenigen Rangnummern zu durchlaufen hat, deren z_i-Wert einen x-Wert verkörpert. X besteht also aus g Summanden.

Die *Testvorschrift* lautet dann: Sobald die Summe (12;3) einen gewissen kritischen Wert X_α übersteigt, wird die Nullhypothese verworfen zugunsten der Alternative, nach der die x_i im allgemeinen größer sind als die y_k. Damit hat man *rechtsseitig* getestet. Die Schranke X_α ist so zu bestimmen, daß unter der Annahme gleicher Wahrscheinlichkeiten für alle n ! möglichen Rangordnungen (d. h. unter H_0) die Wahrscheinlichkeit des Ereignisses $(X > X_\alpha)$ noch $\leq \alpha$ ist.

Nimmt man zur Summe (12;3) noch die Summe

$$(12;4) \quad Y = \sum_s \Psi\left(\frac{s}{n+1}\right),$$

worin s die Rangnummern der y_k durchläuft, hinzu, so geht bei Vertauschung der Rollen der beiden Stichproben die Testgröße X in $Y = -X$ über. Man kann mit derselben kritischen Schranke auch *linksseitig* testen.

Bei *zweiseitiger* Anwendung des X-Texts hat man schließlich H_0 zu verwerfen, wenn X oder Y die Schranke X_α übersteigt. Für $X > X_\alpha$ werden die x_i im allgemeinen größer sein als die y_k, während für $Y > X_\alpha$ die entgegengesetzte Aussage zu machen ist.

Eine ausführliche Gebrauchsanleitung befindet sich in der bereits erwähnten Monographie von VAN DER WAERDEN – NIEVERGELT sowie bei VAN DER WAERDEN (1957).

In weiteren Untersuchungen hat VAN DER WAERDEN gezeigt, daß die Testgröße X asymptotisch

$$(12;5)\quad N\left((0\,;\ \frac{g \cdot h}{n-1}\,Q\right)\text{-verteilt ist ;}$$

das heißt : sie strebt gegen eine Normalverteilung mit dem Mittelwert Null und dem Streuungsquadrat $\frac{gh}{n-1}\,Q$. Dabei berechnet sich die Größe Q gemäß

$$(12;6)\quad Q = \frac{1}{n} \sum_{i=1}^{n} \Psi^2\left(\frac{i}{n+1}\right)$$

Eventuell auftretende Bindungen unter den Stichprobenwerten brauchen hier nicht gelöst zu werden ; man bildet einfach die Summe der Ψ-Werte zu den umstrittenen Rangnummern und fügt für jeden Meßwert das arithmetische Mittel der Summanden als Anteil zur Testgröße hinzu.

Abschnitt C

Anwendungsbeispiele zu den behandelten Testverfahren

An einer Reihe von Beispielen aus verschiedenen Gebieten sollen nun die Anwendungsmöglichkeiten der im Abschnitt B beschriebenen Testverfahren für das *Problem der zwei Stichproben* aufgezeigt werden. Die meisten Beispiele sind so geartet, daß sie ohne weiteres mit mehreren der besprochenen Tests behandelt werden können. Zum Teil wer-

den wir das praktisch auch durchführen, an anderer Stelle jedoch nur darauf hinweisen.

13. Die Auswertung psychologischer Untersuchungen mit dem Iterationen-Test

Im Rahmen einer psychologischen Untersuchung – Vgl. SIEGEL – wurden 12 Jungen und 12 Mädchen im Alter von 4 Jahren während gewisser Spielzeiten beobachtet. Für jedes Kind wurden Anzahl und Intensität von « aktiven Handlungen » nach einem bestimmten Verfahren registriert. Man erwartete, daß zwischen Jungen und Mädchen dieses Alters ein Unterschied im Ausmaß der « Aktivität » vorhanden sein würde und stellte folgende Hypothesen auf :

H_o : Das Ausmaß der Aktivität ist bei Jungen und Mädchen dieses Alters gleich.

H_a : Jungen und Mädchen dieses Alters unterscheiden sich hinsichtlich der Aktivität.

Das Experiment lieferte folgende Ergebnisse :

Jungen: 86 ; 69 ; 72 ; 65 ; 113 ; 65 ; 118 ; 45 ; 141 ; 104 ; 41 ; 50 ;
Mädchen: 55 ; 40 ; 22 ; 58 ; 16 ; 7 ; 9 ; 16 ; 26 ; 36 ; 20 ; 15 ;

Da die Ergebnisse rangmäßig geordnet werden können und die Hypothesen sich auf Unterschiede der beiden unabhängigen Gruppen (Jungen und Mädchen) beziehen, kann man den Iterationen-Test von WALD und WOLFOWITZ (neben anderen Prüfverfahren !) anwenden.

Wir bezeichnen nun mit J beziehungsweise mit M die Meßwerte, die zu einem Jungen beziehungsweise zu einem Mädchen gehören, und stellen die Rangordnung der beiden Stichproben auf. Ohne die numerischen Werte noch einmal zu wiederholen, erhält man :

M M M M M M M M M M M M ;	J J J ;	M M ;	J J J J J J J J J J ;
1.	2.	3.	4. Iteration.

Für unsere beiden Stichproben vom Umfange $g = h = 12$ entnehmen wir dieser Rangordnung als Anzahl der Iterationen den Wert $R = 4$.

Wir führen den Test zweiseitig auf dem Niveau $2 \cdot \alpha = 0{,}05$ durch und entnehmen der Tabelle als zugehörige Schranke den Wert $R_\alpha = 7$.

Ergebnis: $R < R_\alpha$. Unser Wert $R = 4$ fällt in den Verwerfungsbereich; H_o ist also zugunsten von H_a zu verwerfen.

14. Die Untersuchung von Wachstumseinwirkungen zweier Vitamine bei Pilzen (Mit dem Student-Test)

Nach LINDER wurde der Einfluß der Vitamine B_1 und H auf das Wachstum des Pilzes «Trichophyton album» untersucht. Man erhielt folgende Meßserien :

1) *Ohne Zusatz :* x_i = 18 ; 14,5 ; 13,5 ; 12,5 ; 23 ; 24 ; 21 ; 17 ; 18,5 ; 9,5 ; 14 ;
2) *Mit Vitamin B_1 :* y_k = 27 ; 34 ; 20,5 ; 29,5 ; 20 ; 28 ; 20 ; 26,5 ; 22 ; 24,5 ; 34 ; 35,5 ; 19 ;
3) *Mit Vitamin H :* z_1 = 21,5 ; 20,5 ; 19 ; 24,5 ; 16 ; 13 ; 20 ; 16,5 ; 17,5 ; 19 ;

Zunächst untersuchen wir den Einfluß des Vitamines B_1. Dazu haben wir die beiden Stichproben $\langle x_i \rangle$ mit g = 11 und $\langle y_k \rangle$ mit h = 13 zu vergleichen.

Die Berechnung der erforderlichen Ausdrücke nach (8;1), (8;2) und (8;3) ergibt die folgenden Werte :

$$\bar{x} = 16{,}86 \; ; \; \bar{y} = 26{,}19 \; ; \; s^2 = 27{,}86 \; ; \; S^2 = 4{,}68.$$

Daraus ist jetzt die Testgröße (8;4) zu bilden :

$$t = D/S = (26{,}19 - 16{,}86)/2{,}16 = 4{,}32$$

In der Tafel findet man zur Zahl der Freiheitsgrade

$$f = n - 2 = g + h - 2 = 11 + 13 - 2 = 22$$

und zu den gebräuchlichen Testniveaus die folgenden Schranken für zweiseitiges Testen :

0,05 = 5 % mit $t_\alpha = 2{,}074$
0,02 = 2 % mit $t_\alpha = 2{,}508$
0,01 = 1 % mit $t_\alpha = 2{,}819$
0,001 = 0,1 % mit $t_\alpha = 3{,}792$

Ergebnis : Der berechnete Wert t = 4,32 überschreitet auf allen angeführten Testniveaus die zugehörigen kritischen Schranken. Der zu untersuchende Effekt (Einfluß des Vitamins B_1 auf das Wachstum) ist also sehr gut gesichert.

Anmerkung : Üblicherweise bezeichnet man einen Effekt als «*schwach*» gesichert, wenn man zweiseitig nur auf dem 5 %-Niveau (entsprechend

einseitig auf dem 2,5 $^0/_0$-Niveau) zum Verwerfen von H_0 gelangt. Bei Verwerfung auf dem 2 $^0/_0$-Niveau spricht man von «*guter*» Sicherung des Effektes und schließlich bei Verwerfung auf dem 0,1 $^0/_0$-Niveau von «*sehr guter*» Sicherung.

Will man nun noch den Einfluß des Vitamins H auf das Wachstum prüfen, so hat man die entsprechenden Rechnungen mit den beiden Stichproben $\langle x_i \rangle$ und $\langle z_1 \rangle$ durchzuführen. In diesem Falle erhält man zu den Mittelwerten $\bar{x} = 16{,}86$ und $\bar{z} = 18{,}75$ als Wert der Testgröße $t = 1{,}09$.

Für $f = 11 + 10 - 2 = 19$ liefert die Tafel schon auf dem schwachen zweiseitigen 5 $^0/_0$-Niveau eine kritische Schranke von $t_\alpha = 2{,}093$.

Ergebnis: Der Einfluß des Vitamines H kann also nicht einmal als schwach gesichert angesehen werden.

15. Die Untersuchung der Brenndauer von Glühlampen mit dem X-Test

Es seien zwei Sorten von Glühlampen miteinander zu vergleichen, um etwa die Wirksamkeit einer neuen Füllgasmischung, einer neuen Legierung für die Siprale oder eines abgeänderten Produktionsverfahrens zu testen. Unter H_0 wird man annehmen, daß die durchschnittlichen Lebenszeiten beider Sorten gleich sind. Von der ersten Sorte wurden $g = 10$ Exemplare bis zum Durchbrennen beheizt ; von der zweiten Sorte waren es $h = 12$ Stück. Die in Stunden gemessenen Lebenszeiten betrugen :

x_i : 625 ; 637 ; 710 ; 770 ; 820 ; 843 ; 856 ; 920 ; 1070 ; 1225 ;

y_k : 630 ; 683 ; 780 ; 830 ; 889 ; 970 ; 1028 ; 1150 ; 1210 ; 1470 ; 1 520 ; 2090 ;

Daraus entnimmt man folgende Rangordnung :

x y x y xx y x y x x y x y y x y y x y y y.

Die zugehörigen Summanden der Testgrößen (12;3) und (12;4) entnimmt man der Tabelle :

X_{r_i}	= −1,71	+0,16	Y_{s_k}	= −1,36	+0,05
	−1,12	+0,51		−0,94	+0,28
	−0,78	+0,94		−0,51	+0,39
	−0,64	+1,61		−0,28	+0,64
	−0,39			−3,09	+0,78
	−0,16				+1,12
	−0,05				+1,36
	−4,85				+1,71
	+1,61				+6,33
	X = −3,24				−3,09
					Y = +3,24

(Rechenkontrolle : X + Y = −3,24 + 3,24 = 0)

Die Tafel liefert für $\alpha = 2{,}5\ \%$ und einseitiges Testen als kritische Schranke den Wert $X_\alpha = 4{,}06$.

Ergebnis : Da weder X noch Y größer als 4,06 ist, kann die Annahme gleicher durchschnittlicher Lebensdauer für die beiden Lampenarten (= H_0) nicht verworfen werden. Der Effekt ist also nicht einmal schwach gesichert.

Zu demselben Ergebnis würde man auch unter Verwendung des Wilcoxon-Testes gelangen. (Vgl. HEMELRIJK und WABEKE)

Um eventuell noch eine ganz schwache Sicherung des Effektes zu bekommen, könnte man noch auf dem allerdings ungebräuchlichen Niveau von 10 % zweiseitig bzw. 5 % einseitig prüfen. Hier erhält man als zugehörige Schranke den Wert $X_\alpha = 3{,}45$. Da auch hier weder X noch Y größer als 3,45 ist, kann man H_0 noch immer nicht verwerfen.

16. Die Untersuchung von Titrationen mit dem Zeichentest

Ein Chemiker führte eine gewisse Anzahl von Titrationen doppelt aus. (Vgl. HEMELRIJK und WABEKE). Jede zu titrierende Lösung wurde nach ihrer Herstellung auf zwei Kolben verteilt. Während der Inhalt des ersten Kolbens sofort titriert wurde, ließ man bis zur Titration des zweiten eine gewisse Zeit verstreichen. Es war zu untersuchen, ob die Wartezeit einen Einfluß auf das Ergebnis der Titration hat. Man wird

also H_0 (kein Einfluß) testen gegenüber der Alternative H_a, nach der ein Einfluß vorhanden ist.

Die Ergebnisse der an 12 Lösungen vorgenommenen Titrationen lauteten :

Lösung Nr. :	Erste Titration :	Zweite Titration :	*Vorzeichen* der Differenz 2. – 1. Titration :
1	21,24 cm³	25,83 cm⁴	+
2	16,84	17,35	+
3	15,52	16,12	+
4	25,68	28,54	+
5	24,04	24,58	+
6	19,77	27,42	+
7	11,92	14,73	+
8	28,83	27,52	—
9	17,38	14,91	—
10	11,01	19,87	+
11	23,43	24,38	+
12	17,16	20,55	+

Die Anzahl der positiven Differenzen beträgt also $N = 10$. Der Tafel entnimmt man für $n = 12$ und zu einem Testniveau von $2 \cdot \alpha = 0{,}05$ die kritische Schranke $N_\alpha = 9$.

Ergebnis: Die beiden Stichproben liegen im Verwerfungsbereich, so daß man H_0 abzulehnen hat zugunsten der Annahme, daß die Wartezeiten das Ergebnis der Titrationen beeinflussen. Die zweite (= spätere) Titration liefert systematisch größere Titerwerte.

17. Vergleich zweier Produktionsverfahren mittels X-Test, Wilcoxon-Test und Student-Test

Bei der Entwicklung von geeigneten Verfahren zur Herstellung von Halbleitern war ein Qualitätsvergleich zwischen zwei vorläufigen Produktionsverfahren durchzuführen. Es lagen die folgenden beiden Reihen von Qualitätskoeffizienten vor :

$x_i = 27{,}1 ; 15{,}7 ; 34{,}3 ; 24{,}9 ; 19{,}4 ; 31{,}9 ;$
$y_k = 47{,}2 ; 34{,}8 ; 79{,}7 ; 35{,}0 ; 27{,}8 ;$

Diese $n = g+h = 6+5 = 11$ Stichprobenwerte ergeben die folgende Rangordnung :

$x_2\ x_5\ x_4\ x_1\ y_5\ x_6\ x_3\ y_2\ y_4\ y_1\ y_3,$

oder ohne Indizes :

x x x x y x x y y y y.

Zur Anwendung des Wilcoxon-Tests ermittelt man die Anzahl der Inversionen (= x nach y) dieser Rangordnung zu $u = 2$. Für die Stichprobenumfänge (g;h) = (6;5) bzw. (5;6) entnimmt man der Tabelle die folgende Testwahrscheinlichkeit :

$$p(u) = P(U \leqq 2 \; ; \; H_0) = 0{,}87 \; \%.$$

Ergebnis : Wegen $0{,}87 \; \% < 1 \; \%$ können wir H_0 verwerfen und den Effekt als gut gesichert betrachten. D. h. : Das Verfahren « Y » ist besser als das Verfahren « X ».

Zur Anwendung des X-Testes hat man zunächst wieder die Testgrößen X und Y zu bestimmen :

X_{r_i} =	−1,38	0,00	Y_{s_k} = −0,21	+0,43
	−0,97	+0,21		+0,67
	−0,67	+0,21		+0,97
	−0,43			+1,38
	−3,45			+3,45
	+0,21			−0,21
	X = −3,24			Y = +3,24

(Rechenkontrolle : X + Y = −3,24 + 3,24 = 0)

Der Tabelle entnimmt man folgende kritischen Schranken :

Einseitig auf dem 2,5 %-Niveau : $X_\alpha = 2{,}72$
Einseitig auf dem 1 %-Niveau : $X_\alpha = 3{,}20$
Einseitig auf dem 0,5 %-Niveau : $X_\alpha = 3{,}40$

Ergebnis : Es wird nicht nur die 2,5 %-Schranke, sondern auch die 1 %-Schranke überschritten. Man kann also den Unterschied als gut gesichert betrachten !

Zur Anwendung des Student-Tests ist zunächst folgendes zu bemerken : Bei der Betrachtung der beiden Stichproben kommt man zu der Vermutung, daß die zugehörigen Grundgesamtheiten möglicherweise nicht normal verteilt sind und ungleiche Streuungen besitzen. Es ist also sehr fraglich, ob die Voraussetzungen des Student-Tests überhaupt erfüllt sind.

Zur Durchführung der Rechnung hat man zunächst wieder die Werte nach (8;1), (8;2) und (8;3) zu bestimmen.

Die Ergebnisse sind :

$$D = \bar{y} - \bar{x} = 44{,}9 - 25{,}6 = 19{,}3$$
$$s^2 = 220{,}47 \; ; \; S^2 = 80{,}84 \text{ und } S = 8{,}99.$$

Daraus ergibt sich als Wert der Testgröße :

$$t = D/S = 2{,}147.$$

Ergebnis: Da die Tabelle für $f = n\text{-}2 = 9$ Freiheitsgrade zum einseitigen Testniveau $\alpha = 2{,}5\ \%$ eine kritische Schranke $t_\alpha = 2{,}262$ liefert, kann H_0 nicht verworfen werden. Der Unterschied in der Qualität der beiden Produktionsverfahren läßt sich mit Hilfe des Student-Tests also nicht einmal schwach sichern.

Dieses Beispiel zeigt deutlich, daß man bei Unsicherheit über das Vorliegen der Voraussetzungen des Studentschen t-Tests besser ein *verteilungsfreies* Verfahren anwendet.

Anhang

Auswahlgesichtspunkte für die verschiedenen Testverfahren

Im § 2 des Abschnittes A wurde bereits auf die Bedeutung der Güte- oder Machtfunktion eines Tests hingewiesen. Dazu ist zu bemerken, daß sich die Untersuchungen hierzu meist auf den Fall normal verteilter Grundgesamtheiten beschränken. Die Berechnung der Gütefunktion im Falle nicht normal verteilter oder gar unbekannt verteilter Grundgesamtheiten bereitet im allgemeinen erhebliche Schwierigkeiten.

Setzt man einmal – um einen angemessenen Vergleich mit dem parametrischen Student-Test zu erhalten – für die zu vergleichenden beiden Stichproben normal verteilte Grundgesamtheiten mit verschiedenen Mittelwerten und gleichen Streuungen voraus, so kann man die in dieser Arbeit behandelten Testverfahren (abgesehen vom Zeichentest) in folgender Rangordnung nach abnehmender Güte anführen :

1) Student-Test
2) X-Test von Van der Waerden
3) Wilcoxon-Test
4) Iterationen-Test von Wald und Wolfowitz.

Für den Zeichentest lassen sich aufgrund der Binomialverteilung besondere Überlegungen anstellen.

Läßt man nun aber die Voraussetzung der Normalverteilung fallen, um in das eigentliche Anwendungsgebiet der nicht-parametrischen oder verteilungsfreien Testverfahren zu kommen, so können die letzteren unter Umständen erheblich mächtiger sein als der parametrische Student-Test. Am Beispiel des § 17 wurde das sehr deutlich.

Abschließend soll daher folgendes festgestellt werden: Wenn die beiden Grundgesamtheiten annähernd normal verteilt sind und ziemlich gleiche Streuungen haben, dann ist der t-Test als der mächtigste von allen vorzuziehen.

In allen anderen Fällen wird man verteilungsfreie Testverfahren anwenden.

Für große Stichprobenumfänge ist der Wilcoxon-Test annähernd so mächtig wie der X-Test und mit erheblich weniger Rechenaufwand durchzuführen. – Das Argument des geringeren Rechenaufwandes trifft auch für die parameterfreien Methoden in ihrer Gesamtheit gegenüber dem Student-Test zu.

Für kleine Stichprobenumfänge ist der X-Test unbedingt dem Wilcoxon-Test vorzuziehen, weil er auch dort die genaue Abgrenzung des zu einem bestimmten Testniveau gehörenden Verwerfungsbereiches gestattet. Beim Wilcoxon-Test ist das mit den Permutationen kleinerer Rangordnungen nicht immer möglich.

Der Iterationen-Test ist trotz seiner etwas geringeren Güte besonders dann zu empfehlen, wenn sich beide Grundgesamtheiten nicht nur in *einem* der Charakteristika wie Mittelwert, Streuung, Symmetrie der Verteilung usw. unterscheiden, sondern gleichzeitig in *mehreren*.

Die Anwendung des Zeichentests empfiehlt sich besonders dann, wenn man zwei Behandlungsmethoden an denselben Objekten zu erproben hat. Bei häufiger Anwendung mag auch sein sehr geringer Rechenaufwand von Bedeutung sein.

Ein besonderer Vorteil der parameterfreien Rangtests liegt schließlich noch in der Tatsache, daß man sie im Unterschied zum Student-Test auch dann anwenden kann, wenn die zu untersuchenden Eigenschaften numerisch überhaupt nicht zu messen sind, sondern lediglich Vergleiche mit « größer als » und « kleiner als » gestatten. Im Bereich der Verhaltensforschung trifft man häufig auf derartige Probleme.

Literaturverzeichnis

HEMELRIJK, J. & WABEKE, D.: Elementaire statistische opgaven met uitgewerkte oplossingen. (Centrumreeks 2). Gorinchem 1957.

KRES, H.: Parameterfreie Testtheorie zur Behandlung des Zweistichprobenproblems. (Staatsarbeit). Münster 1958. (Unveröffentlicht).

LINDER, A.: Statistische Methoden für Naturwissenschafter, Mediziner und Ingenieure. 2. Aufl. Basel 1951.

MANN, H. B. & WHITNEY, D. R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18 (1947), 50-60.

NEYMAN, J. & PEARSON, E. S.: On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. Roy. Soc. London, Series A, 231 (1933), 289-337.

— — Contributions to the theory of testing statistical hypotheses. Stat. Res. Mem. 1 (1936) 1 ff. 2 (1938) 25 ff.

SIEGEL, S.: Nonparametric statistics for the behavioral sciences. New York - Toronto - London 1956. (Mc Graw-Hill Series in Psychology).

STUDENT (= W. S. Gosset): The probable error of the mean. Biometrica 6 (1908), 1 ff.

WAERDEN, B. L. VAN DER: Ein neuer Test für das Problem der zwei Stichproben. Mathematische Annalen 126 (1953), 93-107.

— — Mathematische Statistik. Berlin-Göttingen-Heidelberg 1957.

— — & NIEVERGELT, E.: Tafeln zum Vergleich zweier Stichproben mittels X-Test und Zeichentest. Berlin-Göttingen-Heidelberg 1956.

WALD, A. & WOLFOWITZ, J: On a test whether two samples are from the same population. Ann. Math. Stat. 11 (1940), 147-162.

WILCOXON, F.: Individual comparisons by ranking methods. Biometrics Bull. 1 (1945), 80-83.