Wahrscheinlichkeitstheoretische Kriterien für die Beurteilung der Güte der Ausgleichung einer Sterbetafel

Autor(en): **Ammeter, Hans**

Objekttyp: **Article**

Zeitschrift: **Mitteilungen / Vereinigung Schweizerischer Versicherungsmathematiker = Bulletin / Association des Actuaires Suisses = Bulletin / Association of Swiss Actuaries**

Band (Jahr): **52 (1952)**

PDF erstellt am: **18.09.2024**

Persistenter Link: https://doi.org/10.5169/seals-550729

B

Wissenschaftliche Mitteilungen

Wahrscheinlichkeitstheoretische Kriterien für die Beurteilung der Güte der Ausgleichung einer Sterbetafel [1])

Von *Hans Ammeter*, Zürich

Einleitung

Unmittelbar aus Beobachtungen abgeleitete Sterbetafeln verlaufen — entgegen der intuitiven Erwartung — erfahrungsgemäss mehr oder weniger unregelmässig und sind daher für die versicherungstechnische Praxis nicht ohne weiteres verwendbar. Die auftretenden Unregelmässigkeiten erscheinen gewissermassen als Beobachtungsfehler, die den wahren Verlauf der Sterbetafel entstellen. Man versucht deshalb, diese Unregelmässigkeiten durch einen geeigneten Prozess — eben die Ausgleichung — zu eliminieren, um die «wahre» Sterbetafel zu erhalten. Im Verlaufe der Zeit sind eine ganze Reihe von Methoden entwickelt worden, die mit mehr oder weniger Erfolg zu einer Ausgleichung der Sterbetafel führen.

Von einer guten Ausgleichung verlangt man, dass die ausgeglichenen Werte eine möglichst glatte Kurve bilden, und ferner, dass sie möglichst getreu die Beobachtungen wiedergibt und nicht etwa charakteristische Eigentümlichkeiten verwischt. Ob und inwieweit diese letztere Forderung im konkreten Fall erfüllt werden konnte, lässt sich genau genommen nicht feststellen, weil die «wahre» Tafel selbst stets unbekannt bleibt. Man musste sich deshalb lange Zeit mit einer mehr oder weniger gefühlsmässigen Überprüfung der Ausgleichung begnügen.

[1]) *Anm. der Red.:* Die vorliegende Arbeit wurde auf ein Preisausschreiben der «Vereinigung schweizerischer Versicherungsmathematiker» hin eingereicht und mit dem höchstmöglichen Preis ausgezeichnet.

Die Entwicklung der mathematischen Statistik im laufenden Jahrhundert hat hier zu einem wesentlichen Fortschritt geführt. Die im folgenden zu erörternden Kriterien der mathematischen Statistik berechtigen allerdings im konkreten Fall auch nicht zu zwingenden Schlüssen. Gegenüber einer rein gefühlsmässigen Beurteilung weisen die von der mathematischen Statistik entwickelten Verfahren jedoch den grossen Vorteil auf, dass bei ihrer Anwendung die Wahrscheinlichkeit eines richtigen oder falschen Urteils stets in einem bestimmten, dem verfolgten Zweck angepassten Rahmen gehalten werden kann. Die Festlegung dieses Rahmens bleibt allerdings dem subjektiven Ermessen vorbehalten.

Bei der Anwendung der Kriterien der mathematischen Statistik ist stets der Umstand wichtig, welche Ausgleichungsmethode benützt wurde. Im folgenden werden daher zuerst die wahrscheinlichkeitstheoretische Bedeutung der verschiedenen Ausgleichungsmethoden und erst anschliessend die Kriterien für die Güte der Ausgleichung behandelt. Diesen Ausführungen wird eine kurze Darstellung der wahrscheinlichkeitstheoretischen Grundlagen dieser Verfahren vorausgeschickt.

I. Wahrscheinlichkeitstheoretische Grundlagen

Die Erfahrung lehrt, dass die bei unausgeglichenen Sterbetafeln auftretenden Unregelmässigkeiten um so mehr ins Gewicht fallen, je kleiner das zugrunde liegende Beobachtungsmaterial ist. Die Unregelmässigkeiten in der unausgeglichenen Sterbetafel scheinen somit irgendwie eine Folge des begrenzten Umfanges des Beobachtungsmaterials zu sein. Diese Erklärung lässt sich noch näher präzisieren und schliesslich in ein wahrscheinlichkeitstheoretisches Modell ausbauen, das die in der Wirklichkeit auftretenden Vorgänge hinreichend genau beschreibt.

A. Das grundlegende Stichprobenmodell

Könnte man das verfügbare Beobachtungsmaterial in allen Altersklassen gleichmässig und beliebig vermehren, so würde man schliesslich zu einem hypothetischen Beobachtungsbestand gelangen, der Grundgesamtheit genannt werden soll. Aus der Grundgesamtheit

könnte ohne Ausgleichung die wahre Sterbetafel abgeleitet werden. Von der wahren Sterbetafel wird vorausgesetzt, dass sie keinerlei Unregelmässigkeiten aufweist und daher durch glatte Kurven darstellbar ist. Die wirklich vorhandenen Beobachtungen fasst man demgegenüber als eine blindlings ausgewählte Stichprobe aus der Grundgesamtheit auf. Die Unregelmässigkeiten in der Stichprobe erklären sich dann zwanglos durch den begrenzten Umfang der Stichprobe.

Im Rahmen dieses Stichprobenmodells entspricht der Ausgleichung die Aufgabe, aus den Daten der Stichprobe die Sterblichkeit der Grundgesamtheit zu ermitteln. Der Überprüfung der Ausgleichung entspricht anderseits die Frage, ob die gegebenen Beobachtungen eine Stichprobe aus der Grundgesamtheit sein könnten.

B. Das Verteilungsgesetz der beobachteten Sterbefälle

Aus der Grundgesamtheit könnte man nicht nur die gegebene, sondern unendlich viele andere, analoge Stichproben entnehmen. Jede dieser Stichproben würde zu einer etwas anderen Sterbetafel führen. Fasst man die gleichartigen Stichproben zusammen, und bestimmt man für jede Konstellation der Sterbefälle $T_1, T_2 \dots T_x \dots T_n$ den relativen Anteil in der Gesamtheit aller Stichproben, so gelangt man zur Frequenzfunktion der Sterbefälle $f(T_1, T_2 \dots T_x \dots T_n)$. Diese n-dimensionale Frequenzfunktion der n zufälligen Variablen $T_1, T_2 \dots T_x \dots T_n$ ist für die nachstehenden Betrachtungen von grundlegender Bedeutung. Sie lässt sich mit den Hilfsmitteln der Wahrscheinlichkeitsrechnung ermitteln, wenn die Grundgesamtheit und die Methoden der Stichprobenauswahl eindeutig gegeben sind. Diese Einzelheiten im Stichprobenmodell sollen möglichst im Einklang mit den wirklichen Verhältnissen festgelegt werden, soweit dies bei dem rein hypothetischen Charakter der ganzen Konstruktion überhaupt möglich ist.

Zunächst sei angenommen, in der Grundgesamtheit gelte in jeder Altersklasse eine bestimmte und feste Sterbenswahrscheinlichkeit $\bar{q}_x$. Ferner erfolge die Auswahl der Stichproben getrennt nach Altersklassen so, dass jedes Element die gleiche Chance hat, in die Stichprobe aufgenommen zu werden. Für eine einzelne Altersklasse x gilt dann bei R_x unter Risiko stehenden Personen für die Anzahl der beobachteten

Sterbefälle T_x die aus dem klassischen Urnenschema folgende Binomialverteilung

$$f(T_x) = \binom{R_x}{T_x} \bar{q}_x^{T_x} (1 - \bar{q}_x)^{R_x - T_x}. \tag{1}$$

Die gewählten Festsetzungen tragen dem Umstand nicht Rechnung, dass die Sterblichkeit als ein in der Zeit ablaufender Vorgang zu betrachten ist und dass die einjährige Sterbenswahrscheinlichkeit $\bar{q}_x$ sich ausdrücklich auf das Jahr als Beobachtungszeit bezieht, während beim Urnenschema die Ziehungsdauer gar keine Rolle spielt. Den Verhältnissen in der Wirklichkeit kommt man näher, wenn man nicht eine einzige, sondern eine kontinuierliche Folge von Ziehungen in jedem Zeitelement annimmt, wobei gleichzeitig an Stelle der einjährigen Sterbenswahrscheinlichkeit $\bar{q}_x$ die entsprechende Wahrscheinlichkeit pro Zeitelement

$$\frac{\bar{q}_x}{m}$$

$(m \to \infty)$ in Rechnung zu stellen ist. Unter diesen etwas mehr an die wirklichen Verhältnisse angepassten Annahmen geht die Frequenzfunktion (1) in die Poisson-Verteilung

$$f(T_x) = \frac{e^{-\bar{T}_x}\, \bar{T}_x^{T_x}}{(T_x)!} \tag{2}$$

über, in der die Sterbenswahrscheinlichkeit $\bar{q}_x$ nicht mehr direkt, sondern nur noch indirekt im Erwartungswert $\bar{T}_x = R_x \bar{q}_x$ auftritt. Für grosse Werte von $\bar{T}_x$, wie sie bei den Anwendungen gewöhnlich auftreten, darf die diskontinuierliche Poissonverteilung (2) durch die stetige Normalverteilung

$$f(T_x) = (2\pi\, \bar{T}_x)^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{(\bar{T}_x - T_x)^2}{\bar{T}_x}} \tag{3}$$

ersetzt werden, bei der Mittelwert und Streuung, wie bei der Verteilung (2), gleich dem Erwartungswert $\bar{T}_x$ sind.

Die Formeln (2) und (3) gelten zunächst für den Fall, dass die Sterbenswahrscheinlichkeit pro Zeitelement $\bar{q}_x/m$ fest bleibt. Es lässt sich jedoch zeigen, dass sie auch unter weit allgemeineren Voraussetzungen gültig bleibt, nämlich auch, wenn die Wahrscheinlichkeit $\bar{q}_x/m$ während der Beobachtungsperiode sich stetig oder sogar sprunghaft verändert.

Der Erwartungswert $\overline{T}_x$ ist in diesem allgemeineren Fall gleich der Summe aller in den einzelnen Zeitelementen aufgetretenen Erwartungswerte. Die sogenannte «übernormale Dispersion» bei den Sterblichkeitsschwankungen, welche sich aus den sprunghaften Veränderungen der Sterblichkeit ergibt, spielt somit im vorliegenden Problem gar keine Rolle, während sie umgekehrt bei risikotheoretischen Fragestellungen die Anwendung der Formel (2) oft nicht zulässt. Diese unterschiedliche Bedeutung der «übernormalen Dispersion» bei sterblichkeitsstatistischen Untersuchungen einerseits und bei risikotheoretischen Fragestellungen anderseits rührt davon her, dass bei sterblichkeitsstatistischen Untersuchungen nur die *zufälligen* Abweichungen von der mittleren Sterblichkeit mit Einschluss von wesentlichen Schwankungen der Beobachtungsperiode von Interesse sind, während in der Risikotheorie stets nach den Abweichungen von einer erwarteten, normalen Sterblichkeit unter Ausschluss von wesentlichen Schwankungen gefragt wird.

Die unter recht wirklichkeitsnahen Voraussetzungen abgeleitete Verteilung (3) darf somit als Verteilungsgesetz der Sterbefälle für eine einzelne Altersklasse gelten. Nimmt man an, dass die Sterbefälle in den verschiedenen Altersklassen untereinander stochastisch unabhängig sind, so lässt sich das gesuchte n-dimensionale Verteilungsgesetz der Sterbefälle ohne weiteres als Produkt der Verteilungen vom Typus (3) darstellen, d. h. man hat dann

$$f(T_1, T_2 \dots T_x \dots T_n) = (2\pi\, \overline{T}_1\, \overline{T}_2 \dots \overline{T}_x \dots \overline{T}_n)^{-\frac{1}{2}}\, e^{-\frac{1}{2}\sum_{x=1}^{n} \frac{(\overline{T}_x - T_x)^2}{\overline{T}_x}} . \quad (4)$$

Die Annahme der Unabhängigkeit erleichtert die Rechnung wesentlich; sie ist gerechtfertigt durch statistische Untersuchungen (siehe z. B. die Arbeit [9]) und durch die plausible Überlegung, dass eine allfällige Abhängigkeit der Sterbefälle untereinander innerhalb jeder einzelnen Altersklasse noch stärker in Erscheinung treten müsste als in voneinander verschiedenen Altersklassen. Die Verteilungen (2) und (3) stützen sich jedoch wesentlich auf die Annahme der Unabhängigkeit der einzelnen Sterbefälle. Es wäre daher geradezu abwegig, Unabhängigkeit der Sterbefälle innerhalb einer einzelnen Altersklasse vorauszusetzen, nicht aber innerhalb verschiedener Altersklassen.

Die Verteilungen (3) und (4) gelten nur für Personensterblichkeit. Für Sterbetafeln, die sich auf Policen- oder gar Summensterblichkeit stützen, sind sie mit Rücksicht auf das ungleiche Gewicht der einzelnen Sterbefälle nicht anwendbar. Die Entwicklung von geeigneten Methoden,

welche auch in diesen allgemeineren Fällen anwendbar wären, bleibt weiteren Forschungen vorbehalten. Als Ausgangspunkt könnte dabei das in der Arbeit [23] hergeleitete Verteilungsgesetz der Policensterbefälle dienen.

II. Die Methoden für die Ausgleichung der Sterbetafeln

Die verschiedenen Methoden für die Ausgleichung von Sterbetafeln lassen sich in drei Hauptgruppen einteilen, nämlich in

A. analytische Methoden,
B. mechanische Methoden,
C. graphische Methoden.

Die letztgenannten Methoden überlassen dem subjektiven Ermessen des Ausgleichers einen zu weiten Spielraum und können daher nicht Anspruch darauf erheben, als wissenschaftlich begründete Methoden zu gelten, obschon sie für manche praktische Zwecke durchaus genügen. Die nachstehenden Erörterungen berücksichtigen daher nur die analytischen und mechanischen Methoden, wobei nur die wahrscheinlichkeitstheoretischen Eigenschaften dieser Methoden und nicht das praktische Vorgehen behandelt wird.

A. Die analytischen Methoden

Bei analytischen Ausgleichungen geht man von der Annahme aus, die wahren Werte der Sterbenswahrscheinlichkeiten $\bar{q}_x$ lassen sich durch eine analytische Funktion von der Form

$$\bar{q}_x = \psi(x, \bar{\theta}_1, \bar{\theta}_2 \ldots \bar{\theta}_k) \tag{5}$$

darstellen, worin $\bar{\theta}_1, \bar{\theta}_2 \ldots \bar{\theta}_k$ k Parameter sind, die in der Grundgesamtheit ganz bestimmte Werte annehmen. Hauptaufgabe einer analytischen Ausgleichung ist es, aus den Daten der vorhandenen Stichprobe möglichst plausible Näherungswerte $\theta'_1, \theta'_2 \ldots \theta'_k$ zu finden. Die Ausgleichung führt dann auf die Funktion

$$q'_x = \psi(x, \theta'_1, \theta'_2 \ldots \theta'_k), \tag{5a}$$

welche je nach der gewählten Ausgleichungsmethode von der durch (5) gegebenen wahren Tafel abweicht.

1. Übersicht über die klassischen Methoden der Parameterbestimmung

a) Die Methode der ausgewählten Punkte

Am naheliegendsten und einfachsten ist es, die Parameter θ_i' so zu bestimmen, dass die durch die Gleichung (5a) gegebene Kurve durch k geeignet gewählte Punkte hindurch geht. Diese Methode lässt den grössten Teil des verfügbaren Beobachtungsmaterials ausser acht und überlässt die Wahl der gegebenen Punkte dem subjektiven Ermessen des Ausgleichers. Die Methode der ausgewählten Punkte ist daher wie die nahe mit ihr verwandte Methode der graphischen Ausgleichung für die wahrscheinlichkeitstheoretische Behandlung ungeeignet.

b) Die Methode der Momente

Nach der Methode der Momente werden die k Parameter θ_i' so bestimmt, dass k Gleichungen von der Form

$$\sum_x [q_x'(\theta_i') - q_x]\, x^r = 0 \tag{6a}$$

oder

$$\sum_x [T_x'(\theta_i') - T_x]\, x^r = 0 \tag{6b}$$

erfüllt sind. Die notwendigen k Gleichungen werden erhalten, indem man r nacheinander die Werte von 0 bis $k-1$ annehmen lässt, oder auch, indem man r nur die Werte von 0 bis $k' < k-1$ annehmen lässt und gleichzeitig die Summen (6) in mehrere Teilsummen zerlegt, so dass wiederum k Bestimmungsgleichungen entstehen. Nach diesem letzteren Prinzip verfährt z. B. die Methode von King-Hardy für Ausgleichungen nach der Makehamschen Formel.

c) Die Methode der kleinsten Quadrate

Die Methode der kleinsten Quadrate geht aus von der Bedingung

$$\sum_{x=1}^{n} [T_x'(\theta_i') - T_x]^2 = \text{Minimum}; \tag{7a}$$

die gesuchten Parameter erhält man dann durch Differenzieren der in (7a) links stehenden Funktion nach den k Parametern θ_i', d. h. aus den Bestimmungsgleichungen

$$\sum_{x=1}^{n} \frac{\partial [T_x'(\theta_i') - T_x]^2}{\partial \theta_i'} = 0 \quad (i = 1, 2 \ldots k). \tag{7a'}$$

Aus praktischen Gründen wird oft an Stelle der Bedingung (7a) mit dem Ansatz

$$\sum_{x=1}^{n} [q'_x(\theta'_i) - q_x]^2 = \text{Minimum}, \tag{7b}$$

d. h. ohne Berücksichtigung des Gewichts der Beobachtungen, gerechnet. Ferner werden an Stelle der absoluten Abweichungen zwischen Erwartung und Beobachtung gelegentlich auch die sogenannten standardisierten Abweichungen

$$\chi_x = \frac{T'_x(\theta'_i) - T_x}{\sqrt{T'_x(\theta'_i)}}$$

in Rechnung gestellt. Man gelangt dann zur χ^2-Minimum-Methode mit dem Ansatz

$$\sum_{x=1}^{n} \chi_x^2 = \sum_{x=1}^{n} \frac{[T'_x(\theta'_i) - T_x]^2}{T'_x(\theta'_i)} = \text{Minimum} \tag{7c}$$

oder etwas vereinfacht

$$\sum_{x=1}^{n} \tilde{\chi}_x^2 = \sum_{x=1}^{n} \frac{[T'_x(\theta'_i) - T_x]^2}{T_x} = \text{Minimum}. \tag{7d}$$

Für hinreichend grosse Stichproben führen alle vier Ansätze (7) zum gleichen Resultat. Praktisch ergeben sich jedoch stets gewisse Unterschiede. Die Bedingung (7a) führt zur stärksten Anpassung der Ausgleichung an die Beobachtungen in den Altern mit den grössten Anzahlen an beobachteten Sterbefällen. Anderseits führt der Ansatz (7b), der von den Sterbenswahrscheinlichkeiten ausgeht, zur stärksten Anpassung bei den hohen Altern, wo die grössten Sterbenswahrscheinlichkeiten auftreten. Die Formeln (7c) und (7d) der χ^2-Minimum-Methode weisen den Vorzug auf, dass die Genauigkeit der Ausgleichung in allen Altern gleichmässig wird.

2. Die Fisherschen Kriterien

Die oben angegebenen Bedingungen (6) und (7), welche zu den für die Parameterberechnung notwendigen k Gleichungen führen, erscheinen zunächst ziemlich willkürlich. Es stellt sich die Frage, welcher der verschiedenen Ansätze als der beste zu gelten hat, oder ob gar irgendwelche weiteren Bedingungen noch besser wären. Diese Fragestellung führt auf die von R. A. Fisher herrührende statistische Schätzungstheorie [15].

Nimmt man an, man könnte an jeder der unendlich vielen Stichproben, welche blindlings aus der Grundgesamtheit entnommen werden könnten, die k Parameter θ_i' nach einer bestimmten Methode — z. B. der Methode der Momente — berechnen, so würden sich bei jeder Stichprobe etwas andere Parameterwerte ergeben. Die Parameter θ_i' haben somit als Funktionen der n zufälligen Variablen T_x ihrerseits den Charakter von zufälligen Variablen, die je nach der gewählten Ausgleichungsmethode einem bestimmten Verteilungsgesetz $f(\theta_i')$ folgen. Die Eigenschaften dieser Verteilungsgesetze erlauben es, die verschiedenen Ausgleichungsmethoden gegeneinander abzuwägen, wobei die nachstehenden Kriterien massgebend sind.

a) Eine Ausgleichungsmethode heisst *folgerichtig* (consistent), wenn die aus dem Verteilungsgesetz $f(\theta_i')$ des Parameters θ_i' zu entnehmende Wahrscheinlichkeit, dass der berechnete Wert θ_i' um mehr als einen beliebig kleinen Betrag ε vom wahren Wert $\bar{\theta}_i$ abweicht, bei wachsendem Stichprobenumfang beliebig nahe gegen Null sinkt. Etwas weniger präzis, aber einfacher ausgedrückt bedeutet dies, dass bei einer folgerichtigen Ausgleichungsmethode die berechneten Parameterwerte θ_i' mit wachsendem Stichprobenumfang gegen die wahren Werte $\bar{\theta}_i$ streben. Die Folgerichtigkeit ist somit eine für grosse Stichproben geltende Grenzwerteigenschaft.

b) Eine folgerichtige Ausgleichungsmethode ist *frei von systematischen Fehlern* (unbiased), wenn auch bei endlichem Stichprobenumfang der Erwartungswert

$$E(\theta_i') = \int_{-\infty}^{\infty} f(\theta_i')\, \theta_i'\, d\theta_i'$$

identisch ist mit dem wahren Wert $\bar{\theta}_i$.

c) Die *Wirksamkeit* (efficiency) einer Ausgleichungsmethode lässt sich durch die Streuung der Parameterverteilungen

$$\sigma^2(\theta_i') = \int_{-\infty}^{\infty} f(\theta_i')\, [\theta_i'^2 - E^2(\theta_i')]\, d\theta_i'$$

messen. Von zwei konkurrierenden Ausgleichungsmethoden ist diejenige wirksamer, bei der die Parameterstreuung $\sigma^2(\theta_i')$ kleiner ist.

d) Eine Ausgleichungsmethode heisst *hinreichend* (sufficient), wenn keine weitere Methode existiert, die eine zusätzliche Information über

den wahren Wert $\bar{\theta}_i$ des Parameters θ'_i liefern könnte. Ist z. B. θ'_i nach einer hinreichenden Methode und θ''_i nach irgendeiner anderen, nicht funktional von der hinreichenden Methode abhängigen Methode bestimmt worden, so lässt sich das simultane Verteilungsgesetz von θ'_i und θ''_i in der Form

$$f(\theta'_i, \theta''_i) = f_1(\theta'_i, \bar{\theta}_i)\, f_2(\theta'_i, \theta''_i)$$

darstellen, d. h. als Produkt von zwei Funktionen, von denen die erste vom hinreichend bestimmten Parameterwert θ'_i und vom wahren Wert $\bar{\theta}_i$ abhängt, während die zweite Funktion den wahren Wert nicht enthält und daher auch keine zusätzlichen Informationsquellen über den wahren Wert bieten kann.

Hinreichende Methoden sind stets am wirksamsten, sofern überhaupt eine wirksamste Methode existiert. Leider lassen sich hinreichende Methoden nur unter ganz bestimmten Voraussetzungen angeben; im allgemeinen existieren keine hinreichenden Methoden.

3. Die Likelihoodmethode

Die Fisherschen Kriterien erlauben einen objektiven Vergleich der verschiedenen Ausgleichungsmethoden und führen darüber hinaus zu einer optimalen Methode, die unter dem Namen *Likelihoodmethode* bekannt ist. Nach dieser Methode werden die Parameter θ'_i so bestimmt, dass die Wahrscheinlichkeit für das Auftreten der beobachteten Konstellation der Sterbefälle ein Maximum erreicht. Dies führt zur Bedingung

$$f(T_1, T_2 \ldots T_x \ldots T_n) = \prod_{x}^{x} \left[2\pi\, T'_x(\theta'_i)\right]^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{[T_x - T'_x(\theta'_i)]^2}{T'_x(\theta'_i)}} = \text{Max.}, \quad (8)$$

die durch Logarithmieren in die handlichere Form

$$-\tfrac{1}{2}\sum_{x=1}^{n} \ln\left[2\pi\, T'_x(\theta'_i)\right] - \tfrac{1}{2}\sum_{x=1}^{n} \frac{[T_x - T'_x(\theta'_1)]^2}{T'_x(\theta'_i)} = \text{Maximum} \quad (8a)$$

übergeht. Die Likelihoodmethode ist *folgerichtig*, wenigstens für grosse Stichproben *am wirksamsten* und *hinreichend*, wenn überhaupt eine hinreichende Methode existiert. Die Verteilungsgesetze der Parameter θ'_i streben für grosse Stichproben gegen Normalverteilungen.

Für grosse Stichproben verliert das erste Glied links in Formel (8a) immer mehr an Bedeutung gegenüber dem zweiten Glied. Berücksichtigt man nur dieses, für grosse Stichproben allein wesentliche Glied, so geht die Bedingung (8a) in die Bedingung (7c), die der χ^2-Minimum-Methode entspricht, über. Die χ^2-Minimum-Methode darf daher im Sinne der Fisherschen Kriterien als beste Ausgleichungsmethode gelten. Die nahe mit der χ^2-Minimum-Methode verwandten, aus dem Prinzip der kleinsten Quadrate folgenden Bedingungen (7a), (7b) und (7d) weisen für grosse Stichproben nahezu die gleichen Vorzüge auf wie die Bedingung (7c). Die durch die Bedingungen (6) charakterisierte Methode der Momente hingegen ist im allgemeinen weniger wirksam als die χ^2-Minimum-Methode. In gewissen Spezialfällen können beide Methoden zum gleichen Ergebnis führen.

Abschliessend sei noch festgestellt, dass die Fisherschen Kriterien nur Aussagen über die Stichprobenverteilungen der Parameter θ_i' geben. Daraus folgt, dass diese Kriterien über den Erfolg einer Ausgleichung im konkreten Fall keine Anhaltspunkte liefern. Die χ^2-Minimum-Methode führt demnach nicht immer zur besten Ausgleichung, sondern nur bei häufiger Anwendung im Durchschnitt zu besseren Resultaten als andere Methoden.

B. Die mechanischen Methoden

Die analytischen Verfahren bewähren sich trotz ihrer wahrscheinlichkeitstheoretischen Vorzüge in der Praxis oft nicht, weil keine geeigneten analytischen Sterbegesetze gefunden werden können, die für die ganze Sterbetafel gelten und ausserdem nur wenige Parameter enthalten. Unter diesen Umständen wird oft einem der zahlreichen mechanischen Verfahren der Vorzug gegeben, die stets zu ziemlich befriedigenden, wenn auch nicht erstklassigen Ausgleichungen führen.

Nach den mechanischen Methoden werden die ausgeglichenen Werte der Sterbenswahrscheinlichkeiten q_x' stets mit Hilfe von Ausdrücken von der Form

$$q_x' = \sum_{\nu=-k}^{+k} a_\nu \, q_{x+\nu}, \tag{9}$$

d. h. als gewogenes Mittel von $2k+1$ Werten berechnet. Die verschiedenen, wohlbekannten mechanischen Formeln unterscheiden sich

untereinander nur durch die Anzahl und den Verlauf der Gewichtskoeffizienten a_ν, die ihrerseits stets der Bedingung

$$\sum_{\nu=-k}^{+k} a_\nu = 1 \tag{9'}$$

genügen.

Die beobachteten Werte q_x kann man aufteilen in den wahren Wert $\bar{q}_x$ und in die Abweichung Δ_x, d. h. man hat

$$q_x = \bar{q}_x + \Delta_x .$$

Für den ausgeglichenen Wert q'_x erhält man somit

$$q'_x = \sum_{\nu=-k}^{+k} a_\nu \bar{q}_{x+\nu} + \sum_{\nu=-k}^{+k} a_\nu \Delta_{x+\nu} . \tag{9a}$$

Führt die Ausgleichungsformel (9), auf die wahren Werte $\bar{q}_x$ angewendet, zu richtigen Ergebnissen, d. h. ist

$$\bar{q}_x = \sum_{\nu=-k}^{+k} a_\nu \bar{q}_{x+\nu} , \tag{10}$$

so liegen die wahren Werte $\bar{q}_x$ auf einer analytischen Kurve, die aus der Differenzengleichung $(2k+1)$-ter Ordnung (10) bestimmt werden kann. Unter diesen Voraussetzungen stellt das erste Glied rechts in Formel (9a) den wahren Wert $\bar{q}_x$ dar und das zweite Glied den Fehler des ausgeglichenen Wertes. Gilt die Differenzengleichung (10) nicht, so führt die mechanische Ausgleichsformel (9) zu gewöhnlich wellenartig verlaufenden, systematischen Abweichungen zwischen den wahren Werten $\bar{q}_x$ und den ausgeglichenen Werten q'_x.

Wenn die Differenzengleichung (10) erfüllt ist, so ist die Abweichung Δ_x eine zufällige Variable, die um den Mittelwert Null mit der Streuung

$$\sigma^2(\Delta_x) = \frac{\bar{q}_x}{R_x}$$

normal verteilt ist. Der durch das zweite Glied rechts in Formel (9a) gegebene zufällige Fehler des ausgeglichenen Wertes $\Delta q'_x$ ist somit ebenfalls normal verteilt um den Mittelwert Null. Für die Streuung gilt die Formel

$$\sigma^2(\Delta q'_x) = \sum_{\nu=-k}^{+k} a_\nu^2 \frac{\bar{q}_{x+\nu}}{R_{x+\nu}} .$$

Zu beachten ist ferner, dass die Fehler der ausgeglichenen Werte innerhalb eines gewissen Bereichs von Nachbarwerten teilweise aus den gleichen Elementen aufgebaut sind. Die Fehler der ausgeglichenen Werte sind daher untereinander stochastisch abhängig. Dieser Umstand ist bei der Überprüfung von Ausgleichungen bedeutsam.

Die mechanischen Ausgleichungen führen im allgemeinen nicht zu wirklich glatten Kurven; das zweite Glied in Formel (9a) rechts bewirkt stets einen unregelmässigen Verlauf. Dieser Nachteil lässt sich vermeiden, wenn man, wie beispielsweise nach der Methode von King, die Ausgleichsformel nur auf äquidistante, sogenannte Kardinalpunkte anwendet und die fehlenden Werte durch oskulatorische Interpolation ergänzt. Dieses Vorgehen weist anderseits den Nachteil auf, dass die Genauigkeit der ausgeglichenen Werte untereinander verschieden wird, und dass das Ausgleichungsergebnis von der willkürlichen Wahl der Kardinalpunktfolge abhängt.

Einen interessanten Weg, um im konkreten Fall eine möglichst glatte Kurve und gleichzeitig einen engen Anschluss an die Beobachtungen zu erreichen, schlägt Whittaker vor [8]. Von der Überlegung ausgehend, dass bei einer gut ausgeglichenen Reihe die Differenzen höherer Ordnung $\Delta^m q'_x$ klein sind, berechnet er die Gewichtskoeffizienten a_ν in Formel (9) so, dass der Ausdruck

$$\sum_{x=1}^{n} (q'_x - q_x)^2 + g_m \sum_{x=1}^{n} (\Delta^m q'_x)^2$$

ein Minimum wird; g_m bedeutet dabei einen willkürlichen Gewichtsfaktor. Zu einer wirklich glatten Kurve führt die Whittakersche Methode allerdings auch nicht; immerhin kann man die Glätte durch eine geeignete Wahl des Gewichtsfaktors g_m innerhalb eines gewissen Rahmens steigern. Vom wahrscheinlichkeitstheoretischen Standpunkt aus nimmt die Whittakersche Methode eine Mittelstellung zwischen den analytischen und mechanischen Verfahren ein. Ihre wahrscheinlichkeitstheoretischen Eigenschaften sind noch nicht genügend abgeklärt.

III. Die Überprüfung der Güte von Ausgleichungen

Von einer guten Ausgleichung verlangt man

a) einen glatten Kurvenverlauf;

b) gute Übereinstimmung mit den Beobachtungen, d. h. möglichst kleine Abweichungen zwischen Beobachtung und Ausgleichung;

c) einen nicht systematischen, zufallsartigen Verlauf dieser Abweichungen.

Im folgenden sei versucht, die Methoden zur Überprüfung dieser drei Eigenschaften zusammenzustellen und ihre Vor- und Nachteile gegeneinander abzuwägen.

A. Die Glätte der Ausgleichung

Bei den analytischen Ausgleichungen und bei mechanischen Ausgleichungen in der Art der Methode von King steht zum vorneherein fest, dass die Ausgleichung zu einer glatten Kurve führt. In diesen Fällen ist die Überprüfung der Glätte der Ausgleichung nicht notwendig, weil es wohl keinen Sinn hat, den Unterschied in der Glätte zu untersuchen, welcher zwischen verschiedenen analytischen Kurven, z. B. Parabeln und Exponentialkurven, besteht. Anders verhält es sich bei den meisten mechanischen Verfahren, weil bei diesen keine wirklich glatten Kurven entstehen, sondern günstigstenfalls scheinbar glatte Kurven, bei denen nur noch Unregelmässigkeiten im Rahmen von Rundungsfehlern auftreten.

Bei einer glatten Kurve sind die Differenzen höherer Ordnung, z. B. $\Delta^m q'_x$, gewöhnlich klein. Die Quadratsumme der $m$ten Differenzen der ausgeglichenen Reihe eignet sich deshalb als Mass für die Glätte einer Kurve. Für die unteren Alter genügt es in der Regel, auf die dritten Differenzen abzustellen; für die höheren Alter mit ihren progressiv wachsenden Sterbenswahrscheinlichkeiten empfiehlt es sich, eine Differenz $m > 3$ in Betracht zu ziehen. Je kleiner die Quadratsumme der $m$ten Differenzen ausfällt, desto besser ist die Glätte der untersuchten Kurve. Als befriedigend darf die Glätte immer dann gelten, wenn die Quadratsumme kleiner ausfällt als

$$\sum_{x}^{x+n-1} [\Delta^m q'_x]^2 = n\, 2^{2m-2}, \tag{11}$$

gemessen in Einheiten der letzten Dezimale von q'_x. Der Grenzwert (11) ergibt sich unter der Annahme, dass die m-te Differenz $\Delta^m q'_x$ an sich verschwindet, aber den maximalen Rundungsfehler aufweist. Dieser tritt auf, wenn die letzte Dezimale der ausgeglichenen Werte alternierend den Rundungsfehler $\pm \frac{1}{2}$ aufweist.

Beispiel: Die nach der Methode von Woolhouse mechanisch ausgeglichene Sterbetafel SM 1901/10 führt in dem 30 Alter umfassenden Intervall $10 \leqslant x \leqslant 39$ zu einer Quadratsumme der dritten Differenzen von 2·293 gemessen in Einheiten der letzten (fünften) Dezimale. Nach Formel (11) wäre bei einer glatten Kurve nur ein Betrag von $30 \cdot 2^4 = 480$ zulässig. Die Ausgleichung nach Woolhouse führt demnach nicht zu einer genügend glatten Kurve.

B. Grundsätzliche Bemerkungen über Testverfahren

1. Allgemeine Festsetzungen

Bei der Überprüfung einer Ausgleichung hinsichtlich der Grösse der Abweichungen zwischen Beobachtung und Ausgleichung in den einzelnen Altern und des unsystematischen Verlaufs dieser Abweichungen bedient man sich mit Vorteil eines sogenannten Testverfahrens. Diese Verfahren stützen sich auf das im Kapitel I geschilderte Stichprobenmodell und auf die Annahme, dass die Ausgleichung auf die wahre Sterbetafel geführt habe. Die gegebenen Beobachtungen werden somit als eine blindlings ausgewählte Stichprobe aus der Grundgesamtheit aufgefasst, in der die Sterblichkeit durch die gefundene Ausgleichung gegeben ist. Die Frage nach der Übereinstimmung zwischen Ausgleichung und Beobachtung geht dann in die andere Frage über, ob die vorhandenen Beobachtungen eine Stichprobe aus der angenommenen Grundgesamtheit sein könnten. Um dies abzuklären, teilt man die Gesamtheit aller Stichproben, deren Struktur durch das n-dimensionale Verteilungsgesetz (4) gegeben ist, in zwei Untergesamtheiten auf, nämlich in die Untergesamtheit der praktisch vorkommenden Stichproben und in die Untergesamtheit der theoretisch an sich möglichen, praktisch aber nicht auftretenden Stichproben. Diese beiden Untergesamtheiten sollen so abgegrenzt werden, dass auf die letztere ein praktisch zu vernachlässigender Anteil P — z. B. 5% — und auf die erstere ein Anteil $1 - P$ — z. B.

95 % — aus der Gesamtheit aller Stichproben entfällt. Man betrachtet dann die Übereinstimmung zwischen Ausgleichung und Beobachtung als genügend oder ungenügend, je nachdem die gegebenen Beobachtungen zur einen oder andern Untergesamtheit gehören. Wie hoch der Anteil P, die sogenannte Wesentlichkeitsschranke, zu bemessen ist, hängt vom subjektiven Ermessen des Ausgleichers und von den Folgen eines allfälligen Fehlurteils ab. Bei Sterbetafeln empfiehlt es sich in der Regel, P nicht allzu klein zu wählen, weil sonst auch recht schlechte Ausgleichungen noch als zulässig angesehen werden müssten. Im allgemeinen wird P im Bereich zwischen 5 % und 1 % als angemessen betrachtet.

Im Interesse einer möglichst klaren Darstellung ist es zweckmässig, die Gesamtheit aller Stichproben geometrisch zu veranschaulichen. Die n beobachteten Anzahlen der Sterbefälle $T_1, T_2 \ldots T_n$ lassen sich als rechtwinklige Koordinaten eines Punktes in einem n-dimensionalen euklidischen Raum deuten. Die Punkte der beiden oben erwähnten Untergesamtheiten der Stichproben erfüllen dann jede einen gewissen Teil dieses Raumes. Die Untergesamtheit der praktisch vorkommenden Stichproben führt auf Punkte in einem Raumbereich, welchen man den Annahmebereich (region of acceptance) nennt. Die Punkte der Untergesamtheit der praktisch nicht vorkommenden Stichproben liegen im restlichen Raum; den Bereich dieser Punkte nennt man den kritischen Bereich (critical region).

Durch die Wahl der Wesentlichkeitsschranke P sind die beiden genannten Raumbereiche, nämlich der Annahmebereich und der kritische Bereich, keineswegs eindeutig gegeben. Es liessen sich vielmehr unendlich viele derartige Bereiche angeben, die sich alle auf die gleiche Wesentlichkeitsschranke P stützen. Jeder einzelnen derartigen Abgrenzung entspricht dabei ein bestimmtes Testverfahren. Unter diesen unendlich vielen kritischen Bereichen sind jedoch nicht alle in gleicher Weise geeignet. Bei der Auswahl von geeigneten kritischen Bereichen sind zwei Gesichtspunkte massgebend, nämlich

a) der kritische Bereich muss, wenn immer möglich, so gewählt werden, dass die Feststellung, ob eine konkrete Stichprobe in den kritischen Bereich fällt oder nicht, nach einer einfachen, möglichst universal gültigen Regel erfolgen kann.

b) der kritische Bereich muss so gewählt werden, dass die Chance, eine ungenügende Ausgleichung als solche zu entdecken, möglichst gross wird.

Diese beiden Gesichtspunkte führen dazu, eine im Sinne von *b)* möglichst leistungsfähige Masszahl M zu bilden, die aus den beobachteten und erwarteten Sterbefällen zu berechnen ist. Im konkreten Fall liegt dann eine gegebene Stichprobe im kritischen Bereich oder nicht, je nachdem die Masszahl einen gewissen kritischen Wert M^* überschreitet oder nicht.

2. Das Verteilungsgesetz der Masszahl M

Die Masszahl M ist, da sie aus den zufälligen Variablen $T_1, T_2 \ldots T_n$ aufgebaut sein soll, ihrerseits eine zufällige Variable, die einem Verteilungsgesetz mit der Frequenzfunktion $f(M)$ folgt. Der gesuchte kritische Wert M^* ergibt sich dann als Funktion der Wesentlichkeitsschranke P aus der Beziehung

$$P = \int_{M^*}^{\infty} f(M)\, dM \tag{12}$$

und ist somit gegeben, sobald das Verteilungsgesetz $f(M)$ bekannt ist. Dieses Verteilungsgesetz lässt sich grundsätzlich aus dem n-dimensionalen Verteilungsgesetz der Sterbefälle (4) durch Summation der Wahrscheinlichkeiten für alle Wertkonstellationen der Sterbezahlen $T_1, T_2 \ldots T_n$ bilden, welche auf die gleiche Masszahl M führen. Analytisch ist somit $f(M)$ durch das n-fache Integral

$$f(M) = \int\int \ldots \int \prod_{x}^{x} (2\pi\, \overline{T}_x)^{-\frac{1}{2}}\, e^{-\frac{1}{2}\frac{(\overline{T}_x - T_x)^2}{\overline{T}_x}}\, dT_x \tag{13}$$

gegeben, wobei über ein Gebiet zu integrieren ist, so dass die Masszahl M als Funktion der beobachteten und erwarteten Sterbefälle stets den gegebenen Wert M annimmt. Die Auflösung des n-fachen Integrals (13) führt fast immer auf grosse, wenn nicht unüberwindliche Schwierigkeiten. Diese lassen sich unter Umständen vermeiden oder doch verkleinern, wenn man an Stelle der Frequenzfunktion zuerst die zugehörige charakteristische Funktion

$$\varphi_M(t) = \int_{-\infty}^{\infty} e^{itM} f(M)\, dM \tag{14a}$$

bestimmt. Die Frequenzfunktion selbst lässt sich anschliessend aus der charakteristischen Funktion zu

$$f(M) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itM} \varphi_M(t)\, dt \tag{14b}$$

bestimmen oder noch einfacher aus einer Transformationstabelle entnehmen, in der wie in einem Wörterbuch charakteristische Funktionen zu gegebenen Frequenzfunktionen und umgekehrt nachgeschlagen werden können.

Die charakteristische Funktion (14a) kann direkt aus der Verteilung (4) berechnet werden, nämlich als Erwartungswert der Funktion e^{itM} bezüglich der n-dimensionalen Verteilung der Sterbefälle (4). Es ist

$$\varphi_M(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{itM(T_1, T_2 \dots T_n)} f(T_1, T_2 \dots T_n) \overset{x}{\Pi} dT_x . \tag{15}$$

Die Formeln (13) und (15) tragen dem Umstand noch nicht Rechnung, dass die Ausgleichung aus den gegebenen Beobachtungen berechnet wurde. Die Berücksichtigung der Ausgleichungsmethode erfolgt in verschiedener Weise, je nachdem ob eine mechanische oder eine analytische Methode vorliegt.

Bei mechanischen Ausgleichungen sind in der Masszahl M durch Substitution der mechanischen Ausgleichsformel (9) die nach der Ausgleichung erwarteten Toten zu ersetzen durch Ausdrücke, die ausschliesslich von den beobachteten Toten abhängen. Die charakteristische Funktion kann dann ohne weiteres berechnet werden. Für die näheren Einzelheiten der Methode sei auf die Arbeit [10] verwiesen.

Bei analytischen Ausgleichungen ist in folgender Weise vorzugehen: Die k Parameter θ_r' des analytischen Sterbegesetzes $\psi(x, \theta_r')$ sind als Funktionen der beobachteten Sterbefälle $T_1, T_2 \dots T_n$ ihrerseits zufällige Variable und folgen zusammen mit der Masszahl M einem $(k+1)$ dimensionalen Verteilungsgesetz, dem die $(k+1)$ dimensionale charakteristische Funktion

$$\varphi(t_M, t_{\theta_1'} \dots t_{\theta_k'}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{it_M M + \sum_{r=1}^{k} it_{\theta_r'} \theta_r'} f(T_1, T_2 \dots T_n) \overset{x}{\Pi} dT_x \tag{14a'}$$

zugeordnet ist, die analog wie in Formel (15) als Erwartungswert von

$$e^{it_M M + \sum_{r=1}^{k} it_{\theta_r'} \theta_r'}$$

bezüglich der Verteilung (4) zu berechnen ist. Durch Inversion ergibt sich anschliessend das zugehörige simultane Verteilungsgesetz der $k+1$ zufälligen Variablen $M, \theta_1' \ldots \theta_k'$ und schliesslich — nachdem man über die k Parameter θ_r' integriert hat — das allein verlangte Verteilungsgesetz der Masszahl M zu

$$f(M) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(M, \theta_1', \theta_2' \ldots \theta_k') \overset{r}{\Pi} d\theta_r'.$$

Aus diesen Darlegungen dürfte zur Genüge hervorgehen, dass die Herleitung des Verteilungsgesetzes einer beliebigen Masszahl M im allgemeinen recht schwierig ist. Es ist daher nicht verwunderlich, dass derartige Verteilungsgesetze bis jetzt nur für verhältnismässig wenige Masszahlen gefunden werden konnten. Man ist daher gezwungen, sich auf die wenigen Masszahlen zu beschränken, bei denen das zugehörige Verteilungsgesetz bekannt ist.

Oft wird man sich auch mit dem Verteilungsgesetz begnügen, das die verwendete Ausgleichungsmethode ausser acht lässt. Dieses vereinfachte Vorgehen ist allerdings nur dann statthaft, wenn die Anzahl der Parameter im analytischen Sterbegesetz im Vergleich zur Anzahl der berücksichtigten Alter in der Sterbetafel klein ist.

Im allgemeinen hängt das Verteilungsgesetz der Masszahl M bei analytischen Ausgleichungen nicht nur von der Anzahl der aus den Beobachtungen bestimmten Parameter ab, sondern auch von der Methode der Parameterberechnung. Beispielsweise ist das Verteilungsgesetz verschieden, wenn bei der Makehamschen Formel die drei Parameter nach der Methode von King-Hardy oder nach der χ^2-Minimum-Methode von Cramér-Wold bestimmt werden, und zwar ist die Streuung von M bei Anwendung der Methode von King-Hardy grösser als bei der Methode von Cramér-Wold. Dies bedeutet nichts anderes, als dass die Ausgleichung nach King-Hardy mit Rücksicht auf die weniger leistungsfähige Ausgleichungsmethode nicht so streng beurteilt wird, wie die Ausgleichung nach Cramér-Wold. Eine derartige Beurteilung mit ungleichen Massstäben, die dazu führen könnte, dass die schlechtere Ausgleichung als die bessere erscheint, befriedigt nicht. Sinnvoller ist es, stets auf das Verteilungsgesetz abzustellen, das für die leistungsfähigste Methode der Parameterbestimmung, d. h. für die χ^2-Minimum-Methode gilt.

3. Übersicht über die gebräuchlichsten Masszahlen

Die gebräuchlichsten Masszahlen lassen sich in drei Hauptgruppen einordnen, nämlich in

a) Masszahlen, welche die Güte der Übereinstimmung zwischen Ausgleichung und Beobachtung in den einzelnen Altern ohne Berücksichtigung der Reihenfolge dieser Alter messen;

b) Masszahlen, welche den unsystematischen, regellosen Verlauf dieser Abweichungen messen;

c) kombinierte Masszahlen, welche sowohl die Abweichungen in den einzelnen Altern als auch deren regellose Folge erfassen.

Mit diesen Masszahlen lässt sich nicht nur die Frage beantworten, ob eine bestimmte Ausgleichung genügt oder nicht, sondern auch, wie verschiedene Ausgleichungen nach ihrer Güte zu klassieren sind, und schliesslich, welche unter den vorliegenden Ausgleichungen die beste ist. Als ein objektives Mass für eine derartige Klassierung darf die Wahrscheinlichkeit

$$P(M) = \int_{M}^{\infty} f(M)\, dM \tag{16}$$

angesehen werden, in der M die bei der einzelnen Ausgleichung aufgetretene Masszahl bedeutet. Je grösser die Wahrscheinlichkeit $P(M)$ ausfällt, desto besser wird die Ausgleichung beurteilt. Die Ausgleichung mit dem grössten Wert von $P(M)$ wird als die beste angesehen. Diese Schlussweise ist allerdings im konkreten Fall nicht unbedingt stichhaltig; bei häufiger Anwendung trifft man jedoch in der Regel damit das richtige.

Im folgenden sollen die Masszahlen, die den wichtigsten Testverfahren zugrunde liegen, kurz erörtert werden. Auf eine Ableitung der Formeln wird im allgemeinen verzichtet; für diese Ableitungen sei auf die im Literaturverzeichnis angeführten Werke und Arbeiten verwiesen.

C. Testverfahren für die Güte der Übereinstimmung zwischen Ausgleichung und Beobachtung in den einzelnen Altern

1. Die Quadratsumme der Abweichungen

Es liegt nahe, die Güte der Übereinstimmung zwischen Ausgleichung und Beobachtung durch die Quadratsummen

$$\Delta T^2 = \sum_{x=1}^{n} (T'_x - T_x)^2 \qquad (17a)$$

oder

$$\Delta q^2 = \sum_{x=1}^{n} (q'_x - q_x)^2 \qquad (17b)$$

zu messen. Diese Masszahlen verschwinden, wenn Beobachtung und Ausgleichung zusammenfallen, und werden um so grösser, je mehr die Ausgleichung von den Beobachtungen divergiert. Die Verteilungsgesetze der Masszahlen (17) könnten mit Hilfe der im Abschnitt B, 2, skizzierten Methoden verhältnismässig einfach berechnet werden. Diese Rechnung würde zeigen, dass die Verteilungsgesetze der Masszahlen (17) zwei für die Anwendungen nachteilige Eigenschaften aufweisen, nämlich dass

a) die unbekannten n Erwartungswerte $\overline{T}_x$ der wahren Sterbetafel als Parameter auftreten, und dass

b) die Verteilungsgesetze der Masszahlen (17) von den besonderen Daten der Sterbetafel und des Beobachtungsmaterials abhängen und daher bei jeder Anwendung wieder neu berechnet werden müssten.

Die praktische Anwendung der Masszahlen (17) stösst somit auf beträchtliche Schwierigkeiten. Für grosse n (n = Anzahl der Altersklassen in der Sterbetafel) lassen sich die Verteilungsgesetze der Masszahlen (17) immerhin näherungsweise ermitteln, wenn man berücksichtigt, dass diese Verteilungsgesetze gegen Normalverteilungen mit den Mittelwerten

$$E(\Delta T^2) = \sum_{x=1}^{n} \overline{T}_x \qquad \text{und} \qquad E(\Delta q^2) = \sum_{x=1}^{n} \frac{\overline{q}_x}{R_x} \qquad (17')$$

und den Streuungen

$$\sigma^2(\Delta T^2) = 2\sum_{x=1}^{n} \overline{T}_x^2 + \sum_{x=1}^{n} \overline{T}_x \quad \text{und} \quad \sigma^2(\Delta q^2) = \sum_{x=1}^{n}\left(2\frac{\overline{q}_x^2}{R_x^2} + \frac{\overline{q}_x}{R_x^3}\right) \quad (17'')$$

streben.

Ersetzt man die wahren Werte $\overline{T}_x$ und $\overline{q}_x$ durch die aus der Ausgleichung hervorgegangenen Werte T'_x und q'_x, so kann man die Verteilungsgesetze der Masszahlen (17) wenigstens in erster Näherung berechnen. Die Güte der Näherung im konkreten Fall bleibt allerdings ziemlich ungewiss.

2. Der χ^2-Test

Die unter Abschnitt 1 behandelten Kriterien ΔT^2 und Δq^2 lassen sich durch eine einfache lineare Transformation, die sogenannte Standardisierung, so umgestalten, dass ihr Verteilungsgesetz von den «lästigen» Parametern $\overline{T}_1, \overline{T}_2 \ldots \overline{T}_n$ (nuisance parameters) befreit wird. Man ersetzt die absoluten Abweichungen $\overline{T}_x - T_x$ durch die standardisierten Abweichungen

$$\chi_x = \frac{\overline{T}_x - T_x}{\sqrt{\overline{T}_x}}, \quad (18)$$

die für alle Alter einheitlich normal verteilt sind um den Mittelwert Null mit der Streuung Eins. Man erhält dann die Quadratsumme

$$\chi^2 = \sum_{x=1}^{n} \chi_x^2 = \sum_{x=1}^{n} \frac{(\overline{T}_x - T_x)^2}{\overline{T}_x}, \quad (18a)$$

welche nach Helmert (1876) [16] und K. Pearson (1900) [20] dem Verteilungsgesetz

$$f(\chi^2) = \frac{e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\, \Gamma\left(\frac{n}{2}\right)}, \quad (18b)$$

der sogenannten χ^2-Verteilung mit n Freiheitsgraden folgt. Dank der Standardisierung hängt die χ^2-Verteilung (18b) nur von der Anzahl n der Altersklassen ab. Die Anzahlen der unter Risiko stehenden Personen, die Sterbetafel und andere von Fall zu Fall ändernde Daten spielen dagegen keine Rolle mehr. Die Masszahl χ^2 erlaubt somit eine Beurteilung von Sterbetafeln auf einer universal gültigen Grundlage.

a) Anwendung auf analytische Ausgleichungen

Die Formeln (18) gelten zunächst nur für Sterbetafeln, die nicht aus den gegebenen Beobachtungen hergeleitet worden sind. Wie R. A. Fisher (1924) [15] gezeigt hat, darf jedoch die χ^2-Methode ohne weiteres auch auf analytische Ausgleichungen angewendet werden, wenn folgende Voraussetzungen erfüllt sind:

α) Die Anzahl der Beobachtungen ist in allen Altersklassen gross.

β) Die Parameter des analytischen Sterbegesetzes sind nach der Likelihoodmethode oder der nahe mit diesem Verfahren verbundenen χ^2-Minimum-Methode oder auch der Methode der kleinsten Quadrate bestimmt worden.

γ) die Anzahl der Freiheitsgrade n wird für jeden aus den Beobachtungen ermittelten Parameter um je eine Einheit reduziert.

Wird somit eine Sterbetafel nach einem Sterbegesetz mit k Parametern ausgeglichen, so gilt für die aus den beobachteten (T_x) und erwarteten (T'_x) Anzahlen der Sterbefälle gebildete Masszahl

$$\chi'^2 = \sum_{x=1}^{n} \chi_x'^2 = \sum_{x=1}^{n} \frac{(T'_x - T_x)^2}{T'_x} \tag{18a'}$$

das Verteilungsgesetz

$$f(\chi'^2) = \frac{e^{-\frac{\chi'^2}{2}} (\chi'^2)^{\frac{n-k}{2}-1}}{2^{\frac{n-k}{2}} \Gamma\left(\frac{n-k}{2}\right)}, \tag{18b'}$$

die χ^2-Verteilung mit $(n-k)$ Freiheitsgraden.

Im Abschnitt II, A, wurde gezeigt, dass die χ^2-Minimum-Methode die leistungsfähigste Methode für die Parameterbestimmung ist. Nach den Ausführungen am Schlusse des Abschnittes III, B, 2, darf daher die Verteilung (18b') für alle analytischen Ausgleichungen schlechthin angewendet werden, selbst wenn die Parameter nicht nach der leistungsfähigsten Methode bestimmt wurden.

b) Anwendung auf mechanische Ausgleichungen

Entgegen einer weit verbreiteten Übung gilt die Verteilung (18b') nicht, wenn eine mechanische Ausgleichung vorliegt. Wie in der

Arbeit [10] gezeigt wird, folgt die Masszahl χ'^2 [Formel (18a')] bei mechanischen Ausgleichungen einem allgemeineren Verteilungsgesetz, dessen charakteristische Funktion in der Form

$$\varphi_{\chi'^2}(t) = [1 + c_1(-2it) + c_2(-2it)^2 \ldots c_n(-2it)^n]^{-\frac{1}{2}} \qquad (19)$$

darstellbar ist.

Die charakteristische Funktion der gewöhnlichen χ^2-Verteilung (18b)

$$\varphi_{\chi^2}(t) = [(1-2it)^n]^{-\frac{1}{2}} \qquad (20)$$

ist in der allgemeineren Verteilung (19) als Spezialfall enthalten. Die Frequenzfunktion der verallgemeinerten χ^2-Verteilung (19) scheint nicht in expliziter Form darstellbar zu sein, hingegen lassen sich ihre Momente direkt aus der charakteristischen Funktion (19) berechnen. Für den Mittelwert gilt z. B. die Formel

$$E(\chi'^2) = n \sum_{\nu=-k}^{+k} \alpha_\nu^2 = n\lambda, \qquad (21)$$

$$\text{mit } \alpha_\nu = \begin{cases} a_\nu & \text{für } \nu \neq 0 \\ a_\nu - 1 & \text{für } \nu = 0, \end{cases}$$

d. h. der Mittelwert der χ^2-Verteilung (19) ist gleich der n-fachen Quadratsumme der Koeffizienten α_ν, die ihrerseits durch die Gewichtskoeffizienten a_ν der mechanischen Ausgleichsformel (9) gegeben sind. Der Mittelwert (21) entspricht der Anzahl der Freiheitsgrade $n-k$ bei den analytischen Ausgleichungen. Bemerkenswert ist der Umstand, dass die Anzahl der Freiheitsgrade bei mechanischen Ausgleichungen durch proportionale Kürzung der Anzahl n und bei analytischen Ausgleichungen durch Subtraktion der Parameterzahl k aus der Anzahl n, hervorgeht.

In erster Näherung darf für mechanische Ausgleichungen mit der gewöhnlichen χ^2-Verteilung (18b') gerechnet werden, wobei der Mittelwert (21) für die Anzahl der Freiheitsgrade in Rechnung zu stellen ist. In der nachstehenden Tabelle sind die Mittelwerte von χ'^2 für einige bekannte mechanische Formeln zusammengestellt.

Ausgleichsformel (Koeffizientenfolge α_ν)	Mittelwert $E(\chi'^2)$ Formel (21)
5-Punkte-Formel von Wittstein $\frac{1}{5}(1, 1, -4 \ldots)$	0,800 n
9-Punkte-Formel von Finlaison $\frac{1}{25}(1, 2, 3, 4, -20 \ldots)$	0,736 n
15-Punkte-Formel von Woolhouse $\frac{1}{125}(-3, 0, -2, 3, 7, 21, 24, -100 \ldots)$	0,779 n
19-Punkte-Formel von Karup $\frac{1}{1250}(-4, -12, -18, -16, 0, 42, 106, 174, 228, -1000 \ldots)$	0,763 n
15-Punkte-Formel von Spencer $\frac{1}{320}(-3, -6, -5, 3, 21, 46, 67, -246 \ldots)$	0,730 n
21-Punkte-Formel von Spencer $\frac{1}{350}(-1, -3, -5, -5, -2, 6, 18, 33, 47, 57, -290 \ldots)$	0,800 n
30-Punkte-Formel von King für Ausgleichungen vom 2-ten Kardinalpunkt an	0,825 n

c) Bemerkungen

Die χ^2-Methode nimmt unter den verschiedenen bereits bekannten Testverfahren eine dominierende Stellung ein. Diese Stellung verdankt sie folgenden Vorzügen:

α) Die Masszahl χ^2 ist leicht und mit elementaren Hilfsmitteln berechenbar.

β) Die χ^2-Verteilung (18b') ist analytisch verhältnismässig einfach aufgebaut und lässt sich auch leicht anwenden.

γ) Der χ^2-Test ist eigentlich das einzige Testverfahren, bei dem die Auswirkungen der angewendeten Ausgleichungsmethode auf das Verteilungsgesetz der Masszahl genau bekannt sind. Bei den analytischen Ausgleichungen ist überdies die dabei zur Anwendung gelangende,

von R. A. Fisher herrührende Freiheitsgradregel so einfach, dass sie selbst von Personen angewendet werden kann, welche die mathematischen Grundlagen des Verfahrens nicht restlos beherrschen.

δ) Die numerischen Untersuchungen im IV. Kapitel zeigen, dass das χ^2-Verfahren unter allen Umständen zu einem einigermassen brauchbaren Resultat führt; Fälle, bei denen die χ^2-Methode gänzlich versagt, wie sie bei andern Verfahren auftreten, kommen nicht vor.

Diesen Vorzügen stehen anderseits gewisse Nachteile gegenüber; insbesondere ist es eine Schwäche des χ^2-Tests, dass er das Vorzeichen und die Reihenfolge der einzelnen Abweichungen unberücksichtigt lässt. Unter besonderen Voraussetzungen ist es daher oft möglich, schärfere Kriterien anzugeben als die Masszahl χ^2.

3. Der χ-Test

Die gegenüber der wahren Tafel berechneten standardisierten Abweichungen χ_x [Formel (18)] wären untereinander stochastisch unabhängige, zufällige Variable, die alle ein und demselben normalen Verteilungsgesetz

$$f(\chi) = (2\pi)^{-\frac{1}{2}} e^{-\frac{\chi^2}{2}} \tag{22}$$

folgen würden. Werden die standardisierten Abweichungen gegenüber einer aus den gegebenen Beobachtungen abgeleiteten Ausgleichung berechnet, so sind sie untereinander nicht mehr unabhängig; bei langen Beobachtungsreihen und wenn bei analytischen Ausgleichungen das Sterbegesetz nur verhältnismässig wenige Parameter aufweist, fällt diese Abhängigkeit jedoch nicht stark ins Gewicht. Man darf dann die einzelnen Werte von χ'_x als untereinander unabhängig betrachten; die Anzahl der Freiheitsgrade der Ausgleichung lässt sich überdies näherungsweise berücksichtigen, indem man an Stelle der Verteilung (22) mit der Verteilung

$$f(\chi') = \left(2\pi \frac{n-k}{n}\right)^{-\frac{1}{2}} e^{-\frac{1}{2}\frac{n}{n-k}\chi'^2} \tag{22'}$$

rechnet. Dieser theoretischen Verteilung kann man die Verteilung der in den einzelnen Altern wirklich aufgetretenen Werte von χ'_x gegenüberstellen. Ergibt dieser Vergleich eine genügende Übereinstimmung, so darf die zu prüfende Ausgleichung als befriedigend gelten.

Bei der praktischen Durchführung empfiehlt es sich, die beobachteten Werte von χ_x' nach einigen zweckmässig abgegrenzten Klassen auszuzählen. Beispielsweise könnte folgende Klasseneinteilung in Frage kommen:

Klasse	Intervalle für χ_x'	Beobachtete Anzahl von standardisierten Abweichungen im gegebenen Intervall	Erwartete Anzahl von standardisierten Abweichungen im gegebenen Intervall
1	$\chi_x' < -1{,}0$	n_1	$\bar{n}_1$
2	$-1{,}0 \leqslant \chi_x' < -0{,}5$	n_2	$\bar{n}_2$
3	$-0{,}5 \leqslant \chi_x' < 0$	n_3	$\bar{n}_3$
4	$0 \leqslant \chi_x' < 0{,}5$	n_4	$\bar{n}_4$
5	$0{,}5 \leqslant \chi_x' < 1{,}0$	n_5	$\bar{n}_5$
6	$\chi_x' \geqslant 1{,}0$	n_6	$\bar{n}_6$
Total	$-\infty < \chi_x' < +\infty$	n	n

Die an sich willkürliche Klasseneinteilung muss so gewählt werden, dass stets $\bar{n}_r \geqslant 6$ ist.

Die Übereinstimmung der beobachteten und der erwarteten Verteilung von χ_x' kann schliesslich mit Hilfe der Masszahl

$$\chi^2 = \sum_{r=1}^{k} \frac{(n - \bar{n}_r)^2}{\bar{n}_r} \tag{23}$$

geprüft werden, die — wenn k Klassen gebildet werden (im Beispiel oben sechs Klassen) — einer χ^2-Verteilung mit $k-1$ Freiheitsgraden folgt. Die Güte der Ausgleichung beurteilt sich so schliesslich nach der Wahrscheinlichkeit $P(\chi^2)$, mit der ein grösserer Wert für χ^2 als der nach Formel (23) berechnete, zu erwarten ist.

Grundsätzlich bedeutet das im vorliegenden Abschnitt geschilderte Verfahren einen Fortschritt gegenüber der im vorigen Abschnitt

geschilderten gewöhnlichen χ^2-Methode, weil das Verteilungsgesetz der standardisierten Abweichungen selbst und nicht nur sein zweites Moment überprüft wird. Praktisch erhält man aber kaum ein zuverlässigeres Resultat, weil die Klasseneinteilung bei höchstens 100 Altern zu grob gewählt werden muss. Nicht befriedigend ist ferner der Umstand, dass die für das Ergebnis nicht unwesentliche Klasseneinteilung ziemlich willkürlich vorgenommen werden muss. Ob und wie der χ-Test auch auf mechanische Ausgleichungen angewendet werden darf, wäre noch abzuklären.

4. Der ω^2-Test

Der unter Abschnitt 3 erläuterte χ-Test gipfelt im Vergleich der theoretischen und beobachteten Frequenzfunktionen $f(\chi_x')$ der standardisierten Abweichungen. Dieser Vergleich kann nach einer von Cramér (1928) [12] und v. Mises (1931) [7] unabhängig voneinander entwickelten Methode noch besser an Hand der Verteilungsfunktion

$$F(\chi) = \int_{-\infty}^{\chi} f(\chi)\, d\chi$$

erfolgen; dabei kann insbesondere auf eine willkürliche Bildung von Klassen wie beim χ-Test verzichtet werden. Die ω^2-Methode stützt sich auf die Masszahl

$$\omega^2 = \frac{1}{c} \int_{-\infty}^{\infty} [F(\chi) - F(\chi')]^2\, d\chi, \tag{24}$$

die aus den theoretischen $F(\chi)$ und den beobachteten $F(\chi')$ Verteilungsfunktionen der standardisierten Abweichungen berechnet wird. Nach v. Mises gelten, wenn die theoretische Verteilung eine Normalverteilung ist, für Erwartungswert und Streuung der Verteilung von ω^2 die Formeln

$$E(\omega^2) = 1 \qquad \text{und} \qquad \sigma^2(\omega^2) \sim 0{,}63 - \frac{0{,}12}{n}. \tag{25}$$

Das Verteilungsgesetz von ω^2 selbst konnte bisher noch nicht gefunden werden.

Smirnoff (1936) [25] hat eine interessante Modifikation der ω^2-Methode vorgeschlagen. An Stelle der Masszahl (24) definiert er

$$\omega_n^2 = \int_{-\infty}^{\infty} [F(\chi) - F(\chi')]^2 \, dF(\chi) \tag{26}$$

und erreicht damit, dass das Verteilungsgesetz von ω_n^2 unabhängig wird von der theoretischen Verteilung $F(\chi)$. Nach Cramér [2] gelten für Mittelwert und Streuung von ω_n^2 die Formeln

$$E(\omega_n^2) = \frac{1}{6n} \qquad \text{und} \qquad \sigma^2(\omega_n^2) = \frac{4n-3}{180\,n^3}. \tag{27}$$

Für grosse n strebt die Verteilung von ω_n^2 gegen eine Grenzverteilung, deren P-Funktion durch den Ausdruck

$$\lim_{n\to\infty} P(\omega_n^2) = \frac{1}{\pi} \sum_{k=1}^{\infty} \int_{(2k-1)n}^{2kn} \frac{e^{-\frac{1}{2}z^2\omega_n^2}\, dz}{\sqrt{-z \sin z}} \tag{28}$$

dargestellt werden kann. Formel (28) sieht für die praktischen Anwendungen nicht gerade verlockend aus und ist scheinbar noch nicht numerisch ausgewertet worden.

Die ω^2-Tests sind zweifellos dem χ-Test überlegen. Für die praktischen Anwendungen ist das Verfahren aber noch zu wenig entwickelt.

5. Die $P(\lambda)$-Tests

a) Die Verteilungsfunktionstransformation

Mit Hilfe der Standardisierung konnte erreicht werden, dass die Abweichungen zwischen Beobachtung und Erwartung in allen Altern einer einheitlichen Normalverteilung folgen. Neben der Standardisierung gibt es eine weitere Methode, nach der sich eine Reihe von zufälligen Variablen mit untereinander verschiedenem, aber stetigem Verteilungsgesetz so transformieren lässt, dass eine einheitliche Verteilung entsteht. Diese, Verteilungsfunktionstransformation genannte, Operation ordnet der zufälligen Variablen mit der Frequenzfunktion $f(\chi)$ die neue Variable

$$y = \int_{-\infty}^{\chi} f(\chi)\, d\chi = F(\chi) \tag{29}$$

zu.

Die neue Variable y genügt stets der Rechtecksverteilung (oft auch Gleichverteilung genannt)

$$f(y) = \left\{ \begin{array}{l} 1 \text{ für } 0 \leqslant y \leqslant 1 \\ 0 \text{ für } y < 0 \text{ und } y > 1 \end{array} \right\} \tag{29'}$$

Substituiert man ferner

$$\lambda = -2 \ln y, \tag{30}$$

so folgt die Variable λ der Exponentialverteilung

$$f(\lambda) = \tfrac{1}{2} e^{-\frac{\lambda}{2}} \qquad \lambda \geqslant 0. \tag{30'}$$

Führt man die beiden Transformationen (29) und (30) an einer Folge von stochastisch unabhängigen zufälligen Variablen $\chi_1, \chi_2 \cdots \chi_n$ durch, und vereinigt man anschliessend die neuen Variablen λ_x zur Masszahl

$$\lambda_I = 2 \left| \ln \prod_x^x y_x \right| = \sum_x \lambda_x, \tag{31}$$

so erhält man eine zufällige Variable, die dem Verteilungsgesetz

$$f(\lambda) = \frac{e^{-\frac{\lambda}{2}} \lambda^{n-1}}{\Gamma(n)\, 2^n}, \tag{31'}$$

d. h. einer χ^2-Verteilung mit $2n$ Freiheitsgraden, folgt.

b) Anwendung auf Sterbetafeln

Bei der Anwendung auf Sterbetafeln ist folgende Rechnung durchzuführen:

1. Berechnung der standardisierten Abweichungen

$$\chi_x = \frac{T'_x - T_x}{\sqrt{T'_x}}; \tag{32}$$

2. Verteilungsfunktionstransformation von χ_x:
   Man bestimmt mit Hilfe einer Tabelle über das Gausssche Wahrscheinlichkeitsintegral die der Rechtecksverteilung (29') folgenden Variablen

$$y_x = \Phi(\chi_x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\chi x} e^{-\frac{\chi^2}{2}} d\chi; \tag{32a}$$

3. Man ermittelt die Hilfsgrössen

$$\lambda_x = -2 \ln y_x;$$

und bildet die Masszahl

$$\lambda_I = \sum_x \lambda_x; \tag{32b}$$

4. Man berechnet die Wahrscheinlichkeit $P(\lambda_I)$ auf Grund der χ^2-Verteilung mit $2n$ Freiheitsgraden und zieht aus dem berechneten Wert die üblichen Schlüsse über die Güte der Ausgleichung.

Zu beachten ist, dass man an Stelle der Transformation (32a) ebensogut mit der Transformation

$$y_x = 1 - \Phi(\chi_x) = (2\pi)^{-\frac{1}{2}} \int_{\chi_x}^{\infty} e^{-\frac{\chi^2}{2}} d\chi \tag{32a'}$$

rechnen könnte, die ebenfalls auf die Rechtecksverteilung (29′) führen würde. Geht man an Stelle von (32a) von (32a′) aus, so erhält man eine andere Masszahl λ_{II}, die dem gleichen Verteilungsgesetz (31′) folgt wie die Masszahl λ_I. Die beiden Masszahlen λ_I und λ_{II} haben jede für sich eine besondere Bedeutung. Die Masszahl λ_I ist nur wirksam für schlechte Ausgleichungen, bei denen die zu prüfende Tafel zu tief verläuft und umgekehrt die Masszahl λ_{II}, wenn die zur Prüfung vorgelegte Tafel zu hoch verläuft. Diese Eigenschaft der beiden Tests ist eine Folge der logarithmischen Transformation (29), die für gegen Null strebende Werte von y_x zu progressiv wachsenden Beträgen für λ_x führt, während umgekehrt λ_x für gegen Eins strebende Werte von y_x nur schwach reagiert. Weiss man nicht zum voraus, in welcher Richtung die zu prüfende Tafel eventuell von der wahren Tafel abweichen könnte, so ist es ratsam, beide Masszahlen λ_I und λ_{II} nebeneinander anzuwenden.

Die Formeln (32) tragen dem Umstand nicht Rechnung, dass die standardisierten Abweichungen dank der Ausgleichung nicht vollständig unabhängige Variable sind. Bei langen Beobachtungsreihen und wenn nur wenige Parameter des analytischen Sterbegesetzes aus den Beobachtungen bestimmt werden, kann man näherungsweise an Stelle der standardisierten Abweichungen gemäss Formel (32) die Werte

$$\chi_x' = \sqrt{\frac{n}{n-k}} \frac{T_x' - T_x}{\sqrt{T_x'}} \tag{32'}$$

in Rechnung stellen.

c) Die Verbindung von mehreren unabhängigen Tests in einen einzigen Test [19]

Die $P(\lambda)$-Tests können nicht nur zur direkten Überprüfung von Sterbetafeln benützt werden, sondern auch, um mehrere untereinander unabhängige Tests in einen einzigen Test zu kombinieren. Bei Sterbetafeln können z. B. einzelne einander nicht überschneidende Abschnitte nach verschiedenen Methoden ausgeglichen werden. Für jeden Abschnitt kann ferner ein besonderer Test auf Grund einer nicht notwendig einheitlichen Masszahl angewendet werden. Im $i$ten Abschnitt sei z. B. auf Grund einer Masszahl M_i eine Wahrscheinlichkeit $P(M_i)$ berechnet worden. Aus allen Wahrscheinlichkeiten $P(M_i)$ zusammen lässt sich dann die kombinierte Masszahl

$$\lambda_{II} = \sum_{i=1}^{r} 2 \left| \ln P(M_i) \right| \tag{33}$$

aufbauen, die einer χ^2-Verteilung mit $2r$ Freiheitsgraden folgt. Die Wahrscheinlichkeit $P(\lambda_{II})$ beurteilt dann die Ausgleichung über die ganze Sterbetafel.

Die $P(\lambda)$-Tests setzen bei ihrer Anwendung bedeutende wahrscheinlichkeitstheoretische Kenntnisse voraus. Sie werden deshalb leider nur verhältnismässig selten praktisch angewendet, obschon sie — wie im nächsten Abschnitt gezeigt werden soll — in gewisser Hinsicht das schärfste nur denkbare Kriterium darstellen. Unbefriedigend ist es, dass der Test für mechanische Ausgleichungen mit Rücksicht auf die starke Abhängigkeit der einzelnen Abweichungen untereinander nicht anwendbar ist. Auch für analytische Ausgleichungen ist die Anwendung der Ausgleichungsmethode auf das Verteilungsgesetz der Masszahlen λ_I und λ_{II} eigentlich noch nicht einwandfrei abgeklärt.

6. Die Likelihood-Kriterien von Neyman und Pearson [5]

a) Grundsätzliche Erwägungen

Die bisher geschilderten Tests wurden im wesentlichen auf intuitiver Grundlage gefunden. Es stellt sich die Frage, welcher dieser Tests am leistungsfähigsten ist, oder ob gar irgendwelche weiteren Kriterien noch leistungsfähiger wären. Um diese Frage beantworten zu können, muss zuerst abgeklärt werden, ob und allenfalls wie die Leistungsfähigkeit eines Tests beurteilt werden kann.

Bei der Anwendung eines Testverfahrens bestehen zwei Möglichkeiten, ein Fehlurteil zu fällen, nämlich

1. das *Fehlurteil erster Art,* bei dem man eine richtige Hypothese H_0 über die Sterbetafel verwirft, weil zufällig eine Masszahl über dem kritischen Wert aufgetreten ist. Geht man bei allen Tests stets von der gleichen Wesentlichkeitsschranke P aus, so ist ein Fehlurteil erster Art bei allen Tests gleich wahrscheinlich.

2. das *Fehlurteil zweiter Art,* bei dem man eine falsche Hypothese H_0 über die Sterbetafel annimmt, weil zufällig eine Masszahl unter dem kritischen Wert aufgetreten ist. Die Wahrscheinlichkeit für ein Fehlurteil zweiter Art ist je nach dem gewählten Testverfahren verschieden. Das Komplement dieser Wahrscheinlichkeit, d. h. die Wahrscheinlichkeit, die Hypothese H_0 als falsch zu entdecken, ist daher ein Mass für die Leistungsfähigkeit der verschiedenen Verfahren. Je grösser diese Wahrscheinlichkeit ausfällt, um so leistungsfähiger ist der betreffende Test.

Geht man von einer bestimmten Annahme über die zu prüfende falsche Hypothese H_0 und die richtige Gegenhypothese H_1 aus, so besteht die Möglichkeit, eine geeignete Masszahl so zu wählen, dass

1. die Wahrscheinlichkeit eines Fehlurteils erster Art einen bestimmten, durch die gewählte Wesentlichkeitsschranke gegebenen Wert P annimmt, und dass
2. die Wahrscheinlichkeit eines Fehlurteils zweiter Art gleichzeitig ein Minimum erreicht.

Neyman und Pearson haben gezeigt, dass optimale Masszahlen dieser Art stets durch das Verhältnis der Likelihoods

$$\lambda = \frac{f(H_0; T_1, T_2 \dots T_n)}{f(H_1; T_1, T_2 \dots T_n)}, \tag{34}$$

berechnet für die zu prüfende Hypothese H_0 und die Gegenhypothese H_1, gegeben sind. Testverfahren, die sich auf die Masszahl (34) stützen, nennt man leistungsfähigste (most powerful) Tests.

Zu beachten ist, dass die Masszahl (34) nur definiert ist, wenn die Hypothese H_0 und die Gegenhypothese H_1 vollständig gegeben sind. Genaue Aussagen über allfällige Gegenhypothesen sind bei Anwendungen in der Regel nicht möglich. Es stellt sich deshalb die Frage,

ob man nicht einen leistungsfähigsten Test bei beliebiger Gegenhypothese H_1 angeben könnte. Die diesbezüglichen Untersuchungen von Neyman und Pearson haben leider gezeigt, dass derartige Universal-Masszahlen nicht existieren. Hingegen ist es in vielen Fällen möglich, Masszahlen zu finden, die für eine ganze Klasse von Gegenhypothesen am wirksamsten sind; diese Tests nennt man gleichmässig leistungsfähigste (uniformly most powerful) Tests. E. S. Pearson [19] hat z. B. gezeigt, dass die beiden $P(\lambda)$-Tests gleichmässig leistungsfähigste Tests sind, wenn angenommen wird, die Hypothese H_0 sei durch die Rechtecksverteilung (29′) und die Gegenhypothese durch Verteilungen von der Form

$$f(y) = (m+1)\, y^m \tag{35a}$$

oder

$$f(y) = (m+1)\,(1-y)^m \tag{35b}$$

$$\text{mit } -1 < m < 0$$

gegeben, wobei (35a) für den $P(\lambda_I)$- und (35b) für den $P(\lambda_{II})$-Test gilt.

b) Anwendung auf Sterbetafeln

Um die Theorie von Neyman und Pearson auf Sterbetafeln anwenden zu können, muss man zuerst prüfen, was für Gegenhypothesen bei Sterbetafeln in Frage kommen. Für eine einzelne Altersklasse betrachtet, ist stets damit zu rechnen, dass nicht der aus der Ausgleichung hervorgegangene Wert T_x' richtig ist, sondern irgend ein anderer Wert T_x''. Ist T_x'' tatsächlich richtig, wird aber bei den Verteilungsfunktionstransformationen (32a) und (32a′) mit T_x' gerechnet, so resultiert nicht mehr die Rechtecksverteilung (29′) als Verteilung von y, sondern eine Verteilung von der Form

$$f(y) = e^{-(a\chi+b)} \quad \text{mit} \quad y = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\chi} e^{-\frac{\chi^2}{2}}\, d\chi \tag{36a}$$

$$\text{oder} \quad y = \frac{1}{\sqrt{2\pi}} \int_{\chi}^{\infty} e^{-\frac{\chi^2}{2}}\, d\chi, \tag{36b}$$

wobei die Konstanten a und b aus den Werten T_x' und T_x'' berechnet werden können. In der Figur 4 ist der Verlauf der Frequenzfunktion (36a)

graphisch dargestellt, wenn $T'_x = 1000$ und $T''_x = 1010$ angenommen wird. Die auftretende Kurve lässt sich in der Tat — wenigstens in erster Näherung — durch Kurven des Typs (35) darstellen. Daraus folgt aber, dass die beiden $P(\lambda)$-Tests im Sinne der Theorie von Neyman und Pearson — jeder nach einer Seite — als schärfste Kriterien zu gelten haben, deren Leistungsfähigkeit von keinem anderen Kriterium übertroffen werden kann.

Besonders zu beachten ist der Umstand, dass jeder der beiden $P(\lambda)$-Tests nur nach einer Richtung hin wirksam ist, d. h. sie sind nur wirksam, wenn ausschliesslich Gegenhypothesen berücksichtigt werden, bei denen in allen Altern grössere oder kleinere Sterblichkeit vorausgesetzt wird als in der zu prüfenden Tafel. Für Gegenhypothesen, bei denen alterszonenweise beide Arten von Abweichungen auftreten, sind die beiden $P(\lambda)$-Tests keineswegs gleichmässig am leistungsfähigsten.

7. Die Smooth-Tests von Neyman [17]

Die beiden $P(\lambda)$-Tests sind nur bei Gegenhypothesen von der Form (35) gleichmässig am leistungsfähigsten. Neyman hat deshalb versucht, weitere Tests aufzustellen, welche unter allgemeineren Voraussetzungen am leistungsfähigsten sind. Als Gegenhypothesen zur Rechtecksverteilung (29′) zieht er ein System von Verteilungen in Betracht, das durch Frequenzfunktionen von der Form

$$f(y) = c\, e^{\sum_{t=1}^{k} \theta_t \pi_t(y)} \tag{37}$$

darstellbar ist. In Formel (37) sind die Grössen θ_t beliebige Parameter, die von Fall zu Fall geeignet gewählt werden können, und $\pi_t(y)$ ein System von im Intervall $0 \leqslant y \leqslant 1$ orthogonalen Polynomen. Die ersten dieser Polynome lauten

$$\left.\begin{aligned} \pi_1(y) &= \sqrt{12}\,(y - \tfrac{1}{2}) \\ \pi_2(y) &= \sqrt{5}\,\{6(y - \tfrac{1}{2})^2 - \tfrac{1}{2}\} \\ \pi_3(y) &= \sqrt{7}\,\{20(y - \tfrac{1}{2})^3 - 3(y - \tfrac{1}{2})\}. \end{aligned}\right\} \tag{37′}$$

Neyman postuliert für seine Smooth-Tests folgende Eigenschaften:

1. Die Wahrscheinlichkeit für ein Fehlurteil erster Art ist wie üblich gleich der Wesentlichkeitsschranke P.

2. Für grosse n ist bei allen Gegenhypothesen von der Form (37), bei denen der Ausdruck

$$\lambda = \left(\sum_{i=1}^{k} \theta_i^2\right)^2$$

einen kleinen, aber bestimmten Wert annimmt, die Wahrscheinlichkeit eines Fehlurteils zweiter Art gleich gross.

3. Die Wahrscheinlichkeit eines Fehlurteils zweiter Art ist für grosse n und kleine Werte von λ minimal.

Diese Forderungen führen auf die Masszahlen

$$\left.\begin{aligned} \psi_1^2 &= u_1^2 \\ \psi_2^2 &= u_1^2 + u_2^2 \\ \psi_3^2 &= u_1^2 + u_2^2 + u_3^2 \\ \psi_k^2 &= \sum_{i=1}^{k} u_i^2, \end{aligned}\right\} \qquad (38)$$

worin die Hilfsgrössen u_i^2 durch die Formeln

$$\left.\begin{aligned} u_1^2 &= 12n^{-1}\left\{\sum_{x=1}^{n} z_x\right\}^2 \\ u_2^2 &= 180n^{-1}\left\{\sum_{x=1}^{n} z_x^2 - \tfrac{1}{12}n\right\}^2 \\ u_3^2 &= 7n^{-1}\left\{20\sum_{x=1}^{n} z_x^3 - 3\sum_{x=1}^{n} z_x\right\}^2 \end{aligned}\right\} \qquad (38')$$

gegeben sind. In den Formeln (38′) ist $z = y - \frac{1}{2}$ und y die aus der Verteilungsfunktionstransformation (32a) hervorgegangene Hilfsvariable. Die Masszahlen $\psi_1^2, \psi_2^2 \ldots \psi_k^2$ genügen χ^2-Verteilungen mit k Freiheitsgraden und erlauben somit eine analoge Überprüfung der Sterbetafelausgleichung wie die Masszahlen χ^2, ω^2, λ_I, λ_{II} usw. Zu beachten ist, dass nicht alle ψ_k^2-Tests miteinander anzuwenden sind, sondern nur ein

einziger, nämlich derjenige, welcher den in Erwägung zu ziehenden Gegenhypothesen genügend Rechnung trägt. Lässt sich die Gegenhypothese (37) durch ein Polynom k-ter Ordnung im Exponenten von (37) genügend genau erfassen, so ist nur der Test k-ter Ordnung anzuwenden.

Bei Sterbetafeln sind Gegenhypothesen von der Form (36) zu berücksichtigen, für die ein Beispiel in Figur 4 graphisch dargestellt ist. In erster Annäherung darf hier mit dem Test erster Ordnung gerechnet werden. Numerische Untersuchungen bestätigen überdies, dass man mit den Tests zweiter und dritter Ordnung praktisch auf die gleichen Resultate kommt wie mit dem Test erster Ordnung. Wollte man wirklich einen engen Anschluss an die Verteilung (36) gewährleisten, so müsste ein Test von sehr hoher Ordnung gewählt werden.

Zu beachten ist ferner, dass der Smooth-Test k-ter Ordnung nur dann die oben postulierten Eigenschaften 1–3 aufweist, wenn für alle n Altersklassen eine beliebige, aber immer die gleiche Gegenhypothese k-ter Ordnung von der Form (37) auftritt. Dieser Fall kann bei Sterbetafeln nur vorkommen, wenn in allen Altern eine gleichartige Abweichung auftritt. Fälle, bei denen die zu prüfende Sterbetafel teils zu hoch und teils zu tief verläuft, eignen sich daher nicht zur Überprüfung durch die Neymanschen Smooth-Tests.

D. Testverfahren für den unsystematischen, regellosen Verlauf der Abweichungen

Alle unter C behandelten Tests nehmen keine Rücksicht auf die Reihenfolge der einzelnen Abweichungen innerhalb der Sterbetafel. Eine Ausgleichung, bei der z. B. alle Abweichungen zuerst negativ und später positiv sind, wird genau gleich beurteilt wie eine Ausgleichung, bei der an sich die gleichen Abweichungen auftreten, diese aber regellos über alle Alter verstreut sind, obschon offensichtlich die letztere den Vorzug verdient. Es ist daher notwendig, die im vorigen Abschnitt behandelten Verfahren zu ergänzen durch besondere Tests, welche die Regellosigkeitsfolge der Abweichungen überprüfen. Derartige Tests gibt es eine ganze Reihe. Die wichtigsten dieser Kriterien sollen im folgenden kurz behandelt werden.

1. Die Anzahl der Zeichenwechsel

Die Anzahl der Zeichenwechsel in der nach steigenden Altern geordneten Reihe der Abweichungen $q'_x - q_x$ stellt ein einfaches und naheliegendes Kriterium für die Regellosigkeitsfolge der Abweichungen dar. Bei schlechten Ausgleichungen ist die Anzahl der Zeichenwechsel gewöhnlich abnormal klein oder gross; bei guten Ausgleichungen bewegt sich diese Anzahl in einem mittleren Rahmen, der durch das Verteilungsgesetz der Zeichenwechsel abgegrenzt werden kann. Dieses Verteilungsgesetz lässt sich bestimmen, wenn man annimmt, dass die Wahrscheinlichkeit für eine positive oder negative Abweichung zwischen der beobachteten und der erwarteten Tafel in allen Altern gleich $\frac{1}{2}$ ist. Bei n Altersklassen beträgt dann die Wahrscheinlichkeit für das Auftreten von z Zeichenwechseln

$$f(z) = \binom{n-1}{z} 2^{-(n-1)}. \tag{39}$$

Für Mittelwert und Streuung der Verteilung (39) gelten ferner die Formeln

$$E(z) = \frac{n-1}{2} \quad \text{und} \quad \sigma^2(z) = \frac{n-1}{4}. \tag{39'}$$

Bei einer 100 Altersklassen umfassenden Sterbetafel sind demnach im Mittel 49,5 Zeichenwechsel zu erwarten. Einer Wesentlichkeitsschranke von beispielsweise 5% entspricht näherungsweise ein Schwankungsbereich von $\pm 2\sigma$. Demnach wären bei 100 Altersklassen alle Ausgleichungen mit $40 \leqslant z \leqslant 60$ als befriedigend zu betrachten.

2. Anzahl der Spitzen

In der nach steigenden Altern geordneten Reihe der standardisierten Abweichungen nennt man alle Abweichungen eine Spitze, bei denen die beiden Nachbarwerte entweder beide grösser oder beide kleiner sind als der Spitzenwert. Die Anzahl dieser Spitzen in der Sterbetafel ist ein Kriterium, das sich ebenfalls zur Beurteilung der Regellosigkeitsfolge der Abweichungen eignet. Für grosse n ist diese Anzahl s eine zufällige Variable, die um

$$\left.\begin{aligned} &\text{den Mittelwert} \quad E(s) = \tfrac{2}{3}(n-2) \\ &\text{mit der Streuung} \quad \sigma^2(s) = \frac{16n-29}{90} \end{aligned}\right\} \tag{40}$$

normal verteilt ist. Bei 100 Altersklassen wären demnach 65,3 Spitzen zu erwarten. Lässt man Abweichungen im Bereich $\pm 2\sigma$ zu, so wären alle Ausgleichungen mit $57 \leqslant s \leqslant 74$ Spitzen als befriedigend zu beurteilen.

3. Der Test von Stevens

Unter den Abweichungen zwischen Ausgleichung und Beobachtung weisen n_1 Werte das gleiche Vorzeichen auf wie die Abweichung im untersten Alter und $n_2 = n - n_1$ Werte das entgegengesetzte Vorzeichen. Die n_1 Abweichungen mit dem gleichen Vorzeichen zerfallen im ganzen in λ Folgen mit dem gleichen Vorzeichen, und die n_2 Abweichungen mit dem entgegengesetzten Vorzeichen in $\lambda - 1$ Teilfolgen. Man kann dann aus den Werten n, n_1, n_2 und λ die Vierfeldertafel

λ	$n_1 - \lambda$	n_1
$n_2 + 1 - \lambda$	$\lambda - 1$	n_2
$n_2 + 1$	$n_1 - 1$	$n = n_1 + n_2$

bilden, die für gegebene Werte von n_1 und n_2 nur einen Freiheitsgrad aufweist. Die Grösse λ ist eine zufällige Variable mit der Frequenzfunktion (Stevens)

$$f(\lambda) = \frac{\binom{n_1 - 1}{\lambda - 1}\binom{n_2 + 1}{\lambda}}{\binom{n_1 + n_2}{n_1}}. \tag{41}$$

Für Mittelwert und Streuung der Verteilung (41) gelten die Beziehungen

$$E(\lambda) = \frac{n_1(n_2 + 1)}{n} \quad \text{und} \quad \sigma^2(\lambda) = \frac{n_1(n_1 - 1)(n_2 + 1)\, n_2}{n^2(n - 1)}.$$

Vergleicht man den beobachteten Wert von λ mit seinem Erwartungswert, so kann man an Hand der Verteilung (41) wiederum die üblichen Schlüsse über die Güte der Ausgleichung ziehen. An Stelle dieser direkten Methode kann man auch von der oben angegebenen Vierfeldertafel ausgehen und für jede der vier auftretenden

Häufigkeiten einen erwarteten und beobachteten Wert und schliesslich die Grösse χ^2 berechnen, die für alle vier Fälle zusammen einer χ^2-Verteilung mit einem Freiheitsgrad folgt.

4. Allgemeine Bemerkungen über die Regellosigkeitstests

Die Nützlichkeit der Regellosigkeitstests wird im allgemeinen überschätzt. In der Regel sagen diese Tests nicht mehr aus, als was schon aus einer flüchtigen Durchsicht der Abweichungen zwischen Ausgleichung und Beobachtung erkennbar wäre. Eine Ausgleichung muss sehr deutlich von den Beobachtungen abweichen, bis ein Regellosigkeitstest das Ungenügen der Ausgleichung anzeigt.

Ein Beispiel: Das unter 1 behandelte Zeichenwechsel-Kriterium werde auf eine ungenügende Ausgleichung angewendet; das Auftreten von positiven oder negativen Abweichungen ist dann nicht mehr wie bei der guten Ausgleichung gleich wahrscheinlich, sondern mit voneinander verschiedenen Wahrscheinlichkeiten p und q ($p + q = 1$) zu erwarten. Die erwartungsmässige Anzahl der Zeichenwechsel beträgt dann $E(z) = 2(n-1)pq$ und ist somit tatsächlich etwas kleiner als bei einer guten Ausgleichung mit $p = q = \frac{1}{2}$. Bei einer 100 Altersklassen umfassenden Sterbetafel müssten jedoch die Wahrscheinlichkeiten um mehr als 0,225 vom Normalwert 0,5 abweichen, bis die erwartungsmässige Zahl der Zeichenwechsel unter die angegebene kritische Anzahl von 40 Zeichenwechseln fällt. Dieser Fall kann erst auftreten, wenn die Ausgleichung einseitig um wenigstens 60% der Streuung von der wahren Tafel abweichen würde, d. h. wenn die Ausgleichung derart offensichtlich von den Beobachtungen abweicht, dass jeder andere Test ebenfalls zur Verwerfung der Ausgleichung führt. Unter diesen Umständen ist es nur von geringem Nutzen, neben einem der üblichen Tests noch einen Regellosigkeitstest anzuwenden.

E. Kombinierte Tests

Die Prüfung der Abweichungen und der Regellosigkeitsfolge dieser Abweichungen durch gesonderte Tests kann nur dann zu einem befriedigenden Ergebnis führen, wenn die beiden Tests übereinstimmend zum gleichen Urteil führen. Oft ergeben sich aber entgegengesetzte

Schlüsse, z. B. kann der $P(\lambda)$-Test zur Verwerfung und gleichzeitig ein Regellosigkeitstest zur Annahme der Ausgleichung führen. In derartigen, ziemlich oft auftretenden Fällen, möchte man gerne die beiden Urteile in ein einziges Gesamturteil kombinieren. Im allgemeinen ist dies jedoch nicht ohne weiteres möglich, weil die beiden getrennten Urteile voneinander abhängig sein könnten. Es stellt sich deshalb die Frage, ob man nicht geeignetere Masszahlen aufstellen kann, die gleichzeitig die Grösse der Abweichungen und ihre Regellosigkeitsfolge messen.

1. Der χ^2-Smooth-Test von David [14]

F. N. David hat gezeigt, dass die im Abschnitt C, 2 eingeführte klassische Prüfgrösse χ^2 (18a) und irgendwelche andere Prüfgrössen, die ausschliesslich auf das Vorzeichen der einzelnen Abweichungen abstellen, als gegenseitig unabhängige zufällige Variable zu betrachten sind. Diese Eigenschaft erlaubt es, den klassischen χ^2-Test und beispielsweise den Test von Stevens in einen einzigen Test zu kombinieren, wobei nach dem im Abschnitt C, 5, *c)*, dargelegten Prinzip verfahren wird. Diese Methode gestaltet sich bei der praktischen Anwendung allerdings etwas mühsam, weil die beim Test von Stevens auftretende Prüfgrösse λ keine stetige Verteilung aufweist. Einen eleganten Weg, um diese Schwierigkeit zu überwinden, hat H. L. Seal [24] gewiesen. Er schlägt vor, von der unter D, 3, eingeführten Vierfeldertafel auszugehen und nach der üblichen Methode mit Hilfe der erwarteten und beobachteten Anzahl λ eine Prüfgrösse $\chi^2(\lambda)$ zu berechnen. Diese dermassen berechnete Grösse $\chi^2(\lambda)$ genügt einer χ^2-Verteilung mit einem Freiheitsgrad und ist unabhängig von der nach Formel (18a') berechneten Grösse $\chi^2(T)$, die aus den erwarteten und beobachteten Anzahlen der Gestorbenen gefunden wurde. Die Summe der beiden Prüfgrössen folgt somit einer χ^2-Verteilung mit $n - k + 1$ Freiheitsgraden und erlaubt so eine gleichzeitige Überprüfung der Abweichungen in den einzelnen Altern und ihrer regellosen Folge.

2. Das $(I\chi)^2$-Verfahren [11]

Ein weiteres Kriterium, das sowohl die Grösse der Abweichungen in den einzelnen Altern als auch ihre regellose Folge berücksichtigt, erhält man, wenn man die standardisierten Abweichungen von beiden Tafelenden her aufsummiert und die aufsummierten Werte quadriert.

Man gelangt so zur Masszahl

$$(I\chi)^2 = \frac{1}{n(n+1)} \left\{ \sum_{x=1}^{n} \left(\sum_{r=1}^{x} \chi_r \right)^2 + \sum_{x=n}^{1} \left(\sum_{r=n}^{x} \chi_r \right)^2 \right\}, \qquad (42a)$$

die sich in die übersichtliche Doppelsumme

$$(I\chi)^2 = \frac{1}{n(n+1)} \sum_{x=1}^{n} \sum_{y=1}^{n} (n+1-|x-y|)\, \chi_x \chi_y \qquad (42b)$$

überführen lässt. Der Nenner $n(n+1)$ wird eingeführt, damit die Masszahl den Erwartungswert Eins aufweist.

Die Masszahl $(I\chi)^2$ hängt im Gegensatz zur Grösse χ^2 wesentlich von der Reihenfolge der standardisierten Abweichungen ab. Systematisch verlaufende Abweichungen bewirken stets eine Vergrösserung der Masszahl. Bei hinreichend langen Beobachtungsreihen ist der $(I\chi)^2$-Test stets dem gewöhnlichen χ^2-Test überlegen. Besonders empfindlich ist die Masszahl $(I\chi)^2$ gegen einseitig abweichende Ausgleichungen, bei denen die ausgeglichene Tafel systematisch zu hoch oder zu tief verläuft.

Das Verteilungsgesetz von $(I\chi)^2$ selbst kann nicht in expliziter Form dargestellt werden. Die zugehörige charakteristische Funktion lässt sich hingegen angeben; es ist

$$\varphi_{(I\chi)^2}(t) = \{1 + a_1(2it) + a_2(2it)^2 \ldots a_n(2it)^n\}^{-\frac{1}{2}}. \qquad (43)$$

Für grosse n strebt der Ausdruck (43) gegen die Grenzfunktion

$$\lim_{n \to \infty} \varphi_{(I\chi)^2}(t) = \left\{ 1 + \sum_{r=1}^{\infty} \frac{r+1}{2} \frac{2^{2r}}{(2r)!} (-it)^r \right\}^{-\frac{1}{2}} \qquad (44a)$$

$$= \sqrt{2}\, \{\cos \sqrt{it} - \sqrt{it} \sin \sqrt{it}\}^{-\frac{1}{2}}, \qquad (44b)$$

die selbst für verhältnismässig bescheidene Werte von n schon recht gut mit der genauen Funktion (43) übereinstimmt.

Die Verteilungs- und Frequenzfunktion von $(I\chi)^2$ für grosse n lässt sich durch gewisse asymptotische Ausdrücke hinreichend genau darstellen. Für manche praktische Zwecke genügt es, wenn man von der Grösse

$$\xi = 0{,}4552\, \chi_1^2 + 0{,}1781\, \chi_2^2 + 0{,}1886 \qquad (45)$$

ausgeht, die näherungsweise dem gleichen Verteilungsgesetz genügt wie die Grösse $(I\chi)^2$ für grosse n. In Formel (45) folgen die Grössen χ_1^2 und χ_2^2 den χ^2-Verteilungen mit einem und zwei Freiheitsgraden.

Für die Anwendungen genügt es, wenn man die kritischen Werte der Masszahl für einige in Betracht fallende Werte der Wesentlichkeitsschranke P kennt. Diese können aus der nachstehenden Tabelle entnommen werden:

P	$(I\chi)^2$
10%	2,19
5%	2,96
1%	4,85
0,1%	7,60

Die angeführten Formeln über die $(I\chi)^2$-Verteilung beziehen sich eigentlich auf den Fall, wo die zur Prüfung vorgelegte Tafel nicht aus den vorhandenen Beobachtungen abgeleitet worden ist. Würde man die angewendete Ausgleichungsmethode berücksichtigen, so ergäbe sich eine gewisse Modifikation der $(I\chi)^2$-Verteilung. Die neue Verteilung würde dabei von den Daten des konkreten Falls abhängig und daher für die praktische Anwendung schwerfällig. Ob und wie die Masszahl $(I\chi)^2$ modifiziert werden muss, damit sie der angewendeten Ausgleichungsmethode Rechnung trägt, ist noch abzuklären.

3. Weitere Methoden

Die Masszahl $(I\chi)^2$ ist besonders empfindlich gegenüber einseitigen Abweichungen, bei denen die ausgeglichene Tafel durchwegs zu hoch oder zu tief verläuft. Divergiert jedoch die zur Prüfung vorgelegte Ausgleichung von der wahren Tafel so, dass ein oder gar mehrere Schnittpunkte zwischen der wahren und der ausgeglichenen Tafel auftreten, so heben sich bei der Aufsummierung die mit verschiedenen Vorzeichen auftretenden Abweichungen ganz oder teilweise auf und die Masszahl $(I\chi)^2$ wächst nicht über einen gewissen Rahmen hinaus. Es lässt sich leicht einsehen, dass dieser Nachteil des $(I\chi)^2$-Verfahrens vermieden werden könnte, wenn man an Stelle der einfachen Summierung der standardisierten Abweichungen mit der doppelten, dreifachen ... N-fachen Summe rechnen würde. Auf diese Weise würde man zu den $(II\chi)^2$, $(III\chi)^2$ ... $(N\chi)^2$-Tests gelangen, deren Verteilungsgesetze in ähnlicher Weise bestimmbar wären wie dasjenige der $(I\chi)^2$-Verteilung. In dieser Richtung kann die Theorie der kombinierten Tests, die erst am Anfang ihrer Entwicklung steht, noch beträchtlich ausgebaut werden.

F. Numerische Untersuchungen

Die Leistungsfähigkeit der verschiedenen Testverfahren lässt sich theoretisch mit Hilfe der im Abschnitt C, 6, skizzierten Theorie von Neyman und Pearson überprüfen. Numerische Untersuchungen auf dieser Grundlage sind aber bei Sterbetafeln mit einem unverhältnismässig grossen Zeitaufwand verbunden. Im folgenden werden daher die wichtigsten der oben behandelten Testverfahren auf einer etwas anderen Grundlage untersucht.

Diese Untersuchungen gehen aus vom Makehamschen Gesetz

$$\mu_x = a + b\,c^x \tag{46}$$

und vom Material, das der schweizerischen Volkssterbetafel SM 1939/44 in den Altersstufen von $40 \leqslant x \leqslant 89$ zugrunde liegt. Variiert man die drei in (46) auftretenden Parameter systematisch, so kann man ein ganzes System von Sterbetafeln erzeugen und anhand der geschilderten Testverfahren mit den Beobachtungen vergleichen. Geht man ferner von einer einheitlich gewählten Wesentlichkeitsschranke P aus, so ist für jedes Kriterium ein bestimmter Bereich von Parameterwerten a, b und c gegeben, der auf Sterbetafeln führt, die im Sinne des betreffenden Kriteriums als annehmbar zu betrachten sind. Die Länge dieser Parameterintervalle ist dann ein Mass für die Leistungsfähigkeit des betreffenden Kriteriums, wobei ein Test um so leistungsfähiger ist, je kürzer die Parameterintervalle ausfallen.

Aus der Fülle der beim Makehamschen Gesetz denkbaren Parametervariationen werden nur die nachstehenden fünf Typen in Betracht gezogen:

Typ I: Variationen von a allein.

Typ II: Variationen von b allein.

Typ III: Gleichgerichtete simultane Variationen von a und b: Variation $(a + b)$.

Typ IV: Entgegengesetzte simultane Variationen von a und b: Variation $(a - b)$.

Typ V: Simultane Variationen von a, b und c, wobei die Parameter a und b für gegebene Werte von c nach der Methode der Momente aus den Beobachtungen berechnet werden: Variation $[c - (a - b)]$.

Für die wahre Sterbetafel sind die Mittelwerte der standardisierten Abweichungen alle gleich Null. Für andere Sterbetafeln liegen diese Mittelwerte auf bestimmten Kurven, die man Regressionslinien der standardisierten Abweichungen nennen kann. In der beiliegenden Figur 1 sind je zwei derartige Regressionslinien für die Parametervariationen I bis V graphisch dargestellt. Zu beachten ist vor allem der Umstand, dass die Regressionslinien der Typen I, II und III die x-Achse nie schneiden, während bei Typ IV stets ein und bei Typ V stets zwei Schnittpunkte mit der x-Achse auftreten.

Wendet man die erläuterten Testverfahren auf das durch die Parametervariationen I bis V erzeugte System von Sterbetafeln an, so ist die dem gewählten Testverfahren zugrunde liegende Masszahl eine Funktion der Makeham-Parameter der zu prüfenden Sterbetafel. Diese Funktionen sind für die fünf Variationstypen und für die stetig verlaufenden Masszahlen χ^2, $(I\chi)^2$, λ_I, λ_{II} und ψ_1^2 in der Figur 2 graphisch dargestellt. Eine nähere Betrachtung dieser Masszahlfunktionen führt zu folgenden Feststellungen:

Die nach Formel (18a) berechnete klassische Masszahl χ^2 und die Masszahl $(I\chi)^2$ [Formel (42)] liegen stets auf nach oben geöffneten parabelähnlichen Gebilden. Für die Variationen I, II und III steigt die Masszahl $(I\chi)^2$ stets steiler an als die Masszahl χ^2; bei den Variationen IV und V wächst mit Rücksicht auf die Schnittpunkte zwischen den Regressionslinien und der x-Achse die Masszahl $(I\chi)^2$ langsamer als die Masszahl χ^2.

Die beiden Masszahlen λ_I und λ_{II} liegen für die Variationen I, II und III auf monoton von $+\infty$ bis 0 sinkenden, resp. von 0 bis $+\infty$ ansteigenden Kurven. Im Gegensatz zu den Masszahlen χ^2 und $(I\chi)^2$ ergibt sich bei den Masszahlen λ_I und λ_{II} immer nur je ein Parameterwert, bei dem die Masszahl einen gegebenen kritischen Wert erreicht. Die Masszahlen λ_I und λ_{II} sind daher nur nach einer Seite hin wirksam, und zwar ist, wie bereits gezeigt wurde, λ_I wirksam für Parametervariationen nach unten und λ_{II} für Parametervariationen nach oben. Wenn man nicht zum voraus sicher weiss, in welcher Richtung eine zu prüfende Tafel von der wahren Tafel abweichen könnte, muss man daher stets beide $P(\lambda)$-Tests nebeneinander anwenden.

Die Parametervariationen IV und V führen auf Sterbetafeln, die verglichen mit der wahren Tafel teils zu hoch, teils zu tief verlaufen.

In derartigen Fällen können die beiden, nur auf einseitige Abweichungen zugeschnittenen $P(\lambda)$-Tests versagen. Im vorliegenden Fall liegt z. B. auf der λ_{II}-Kurve der Variation IV überhaupt kein kritischer Wert der Masszahl λ_{II}.

Der Smooth-Test erster Ordnung von Neyman stützt sich auf die Masszahl ψ_1^2, die durch die erste Formel (38) gegeben ist. Diese Masszahlen liegen stets auf nach oben geöffneten parabelähnlichen Kurven, die für einseitige Variationen (Typ I, II und III) verhältnismässig steil ansteigen. Bei den Typen IV und V dagegen verlaufen diese Kurven ziemlich flach, beim Typ V sogar nahezu horizontal, so dass der Test praktisch unbrauchbar wird. Die Smooth-Tests von Neyman sind somit wie die $P(\lambda)$-Tests nur wirksam, wenn einseitige Abweichungen von der wahren Tafel vorliegen.

Die Leistungsfähigkeit der verschiedenen Testverfahren untereinander bei bestimmten Variationstypen lässt sich in einfacher Weise an Hand der Sehnen beurteilen, welche einer bestimmten Wesentlichkeitsschranke P entsprechen. Die Länge dieser Sehnen (für $P = 5\%$ und $P = 1\%$ sind sie in der Figur 2 eingezeichnet) gibt das Parameterintervall an, das nach dem betreffenden Test auf zulässige Sterbetafeln führt. Je kürzer diese Sehne ausfällt, desto weniger läuft man Gefahr, irrtümlich eine falsche Tafel als richtig anzunehmen, und desto leistungsfähiger ist demzufolge der betreffende Test. In der beiliegenden Figur 3 sind die Längen dieser Sehnen, die den Parameterintervallen der annehmbaren Tafeln entsprechen, als Funktionen der Wesentlichkeitsschranke P graphisch dargestellt. Interessant in dieser Graphik ist weniger die absolute Höhe der einzelnen Kurven als deren relative Lage zueinander und insbesondere die Reihenfolge der Testkurven von unten nach oben innerhalb eines Variationstyps. Dieser Reihenfolge der Kurven entspricht nämlich die Rangfolge in der Leistungsfähigkeit der verschiedenen Tests. Beim Typ I ergibt sich folgende Rangfolge:

1. Die beiden $P(\lambda)$-Tests.
2. Der Smooth-Test von Neyman ψ_1^2.
3. Der $(I\chi)^2$-Test.
4. Der klassische χ^2-Test.

Bei den Typen II und III, bei denen wie bei I ebenfalls nur einseitige Variationen vorkommen, ergibt sich nahezu das gleiche Bild, nur dass die praktisch fast auf die gleichen Resultate führenden ψ_1^2- und $(I\chi)^2$-Tests ihre Ränge vertauschen. Charakteristisch ist es, dass die beiden $P(\lambda)$-Tests bei einseitigen Variationen stets das schärfste Kriterium abgeben. Dieses Resultat war zu erwarten, weil — wie im Abschnitt C, 6, ausgeführt wurde — nach der Testtheorie von Neyman und Pearson die $P(\lambda)$-Tests unter gewissen, praktisch erfüllten Voraussetzungen die absolut schärfsten Kriterien darstellen, die von keinem andern Verfahren übertroffen werden können.

Anders verhält es sich bei den Variationstypen IV und V. In diesen Fällen sind die besonderen Voraussetzungen des $P(\lambda)$-Tests und auch der Smooth-Tests von Neyman nicht mehr erfüllt. Der Smooth-Test von Neyman führt hier auf die schlechtesten Resultate, die praktisch nicht mehr brauchbar sind. Die $P(\lambda)$-Tests stehen im dritten resp. zweiten Rang. Diese verhältnismässig günstige Klassierung scheint mehr durch das benützte Beobachtungsmaterial als durch die Eigenschaften des Verfahrens bedingt zu sein.

Beim Typ IV liefert das $(I\chi)^2$-Verfahren noch nahezu gleich gute Resultate wie das klassische Kriterium χ^2, beim Typ V ist jedoch das letztere Verfahren mit Abstand dem $(I\chi)^2$-Verfahren überlegen. Dies erklärt sich ohne weiteres aus den besonderen Eigenschaften des $(I\chi)^2$-Verfahrens, das sich nicht mehr eignet, wenn — wie beim Typ V — zwei Schnittpunkte zwischen den Regressionslinien und der x-Achse auftreten.

Von besonderem Interesse ist schliesslich der Umstand, dass bei den Typen IV und V das klassische χ^2-Verfahren am besten abschneidet. Es zeigt sich somit, dass der χ^2-Test in komplizierter gelagerten Fällen, bei denen verschiedenartige Abweichungen innerhalb derselben Tafel auftreten, immer noch allen andern Verfahren überlegen ist. Der χ^2-Test darf somit gewissermassen als ein «Allround-Test» betrachtet werden, der allerdings im Einzelfall nicht das schärfste Kriterium abgibt, dafür aber in allen Fällen zu einem einigermassen brauchbaren Resultat führt und niemals gänzlich versagt.

Zusammenfassung

Es sei versucht, die wichtigsten Resultate der in dieser Arbeit dargestellten wahrscheinlichkeitstheoretischen und numerischen Untersuchungen zusammenzufassen und insbesondere auf die wichtigsten noch nicht oder nicht vollständig gelösten Fragen hinzuweisen.

1. Die Frage nach der Güte der verschiedenen Ausgleichungsmethoden ist durch die Fisherschen Kriterien — wenigstens für die analytischen Methoden — im wesentlichen abgeklärt. Es zeigt sich, dass bei analytischen Ausgleichungen die χ^2-Minimum-Methode als leistungsfähigste Methode zu gelten hat.

2. Im Laufe des zwanzigsten Jahrhunderts sind eine ganze Reihe von Testverfahren entwickelt worden, welche die wahrscheinlichkeitstheoretische Überprüfung von Ausgleichungen hinsichtlich

a) der in den einzelnen Altern auftretenden Abweichungen zwischen Ausgleichung und Beobachtung,

b) der Regellosigkeitsfolge dieser Abweichungen oder

c) beider Gesichtspunkte gleichzeitig

erlauben. Mit Hilfe dieser Tests lässt sich die relative Güte von verschiedenen Ausgleichungen in objektiver Weise überprüfen.

3. Von den bis heute bekannten Kriterien zur wahrscheinlichkeitstheoretischen Überprüfung von Ausgleichungen ist eigentlich nur das klassische χ^2-Verfahren von K. Pearson theoretisch genügend entwickelt, so dass es den bei analytischen und mechanischen Ausgleichungen auftretenden besonderen Verhältnissen Rechnung zu tragen vermag. Es ist wohl eine der wichtigsten Aufgaben für die weiteren Forschungen, die anderen Verfahren so auszubauen, dass auch sie die bei Ausgleichungen auftretenden Abhängigkeiten theoretisch einwandfrei berücksichtigen können.

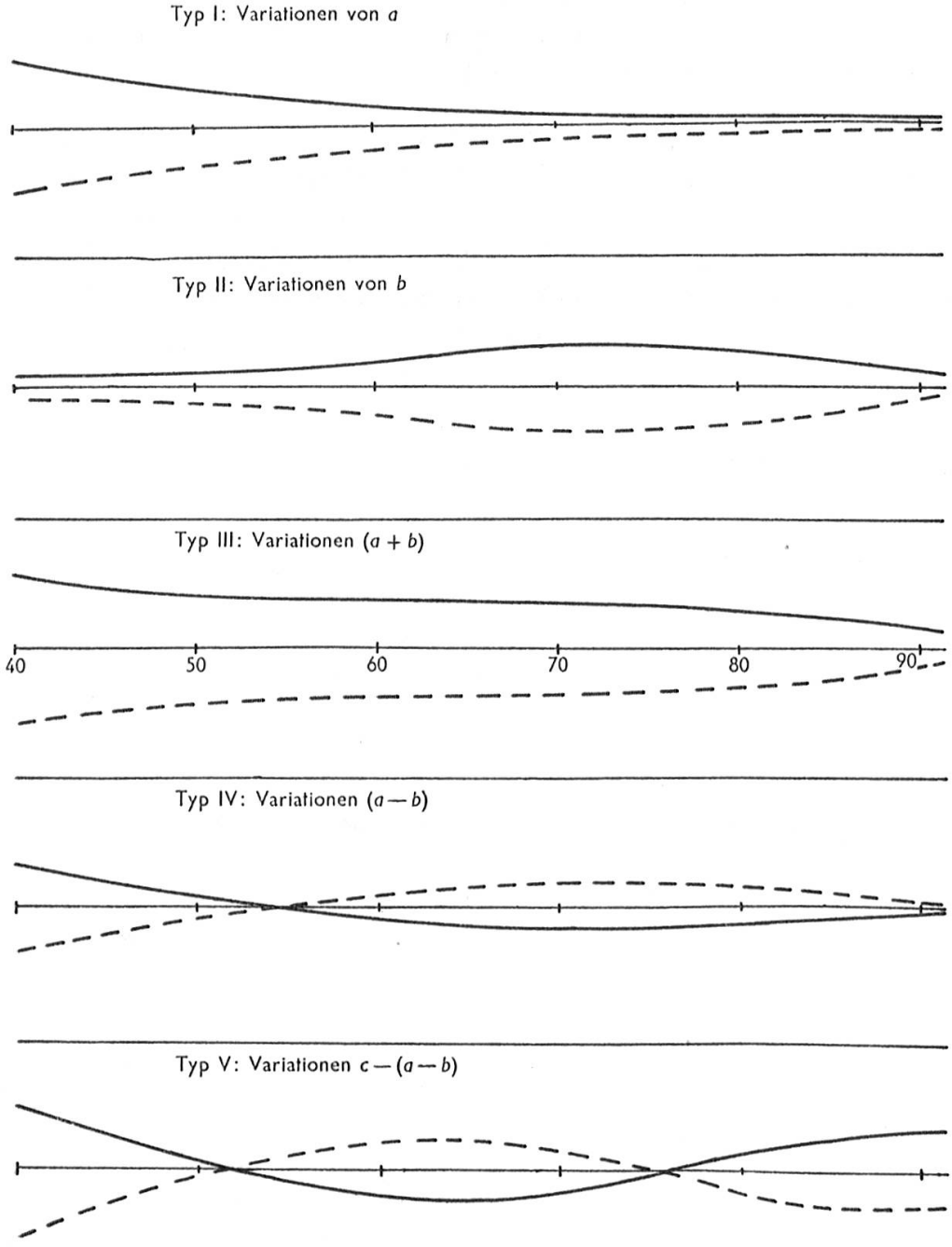
4. Die Theorie von Neyman und Pearson lehrt, dass es keinen «Universal-Test» gibt, der bei beliebiger Gegenhypothese das schärfste Kriterium darstellt. Dieses theoretische Ergebnis wird durch die im Abschnitt III, F, dargestellten numerischen Untersuchungen bestätigt.

5. Leistungsfähigste Tests lassen sich angeben, sobald ihre Anwendung auf ganz bestimmte Arten von Gegenhypothesen beschränkt wird. Für einseitige Abweichungen (durchwegs zu grosse oder zu kleine Sterblichkeit) stellen die beiden $P(\lambda)$-Tests die leistungsfähigsten Kriterien dar. Die Entwicklung von leistungsfähigsten Tests für allgemeinere Gegenhypothesen, z. B. für den Fall, wo zwischen der zu prüfenden und der wahren Sterbetafel einer oder mehrere Schnittpunkte auftreten, bleibt weiteren Untersuchungen vorbehalten.

6. Dem klassischen χ^2-Verfahren von K. Pearson kommt in dem Sinne der Charakter eines «Universal-Tests» zu, als es in allen Fällen zu einem einigermassen brauchbaren Ergebnis führt und niemals gänzlich unbrauchbar wird. Dieser universellen Anwendungsmöglichkeit steht der Nachteil gegenüber, dass für bestimmte Gegenhypothesen leistungsfähigere Spezialtests gefunden werden können.

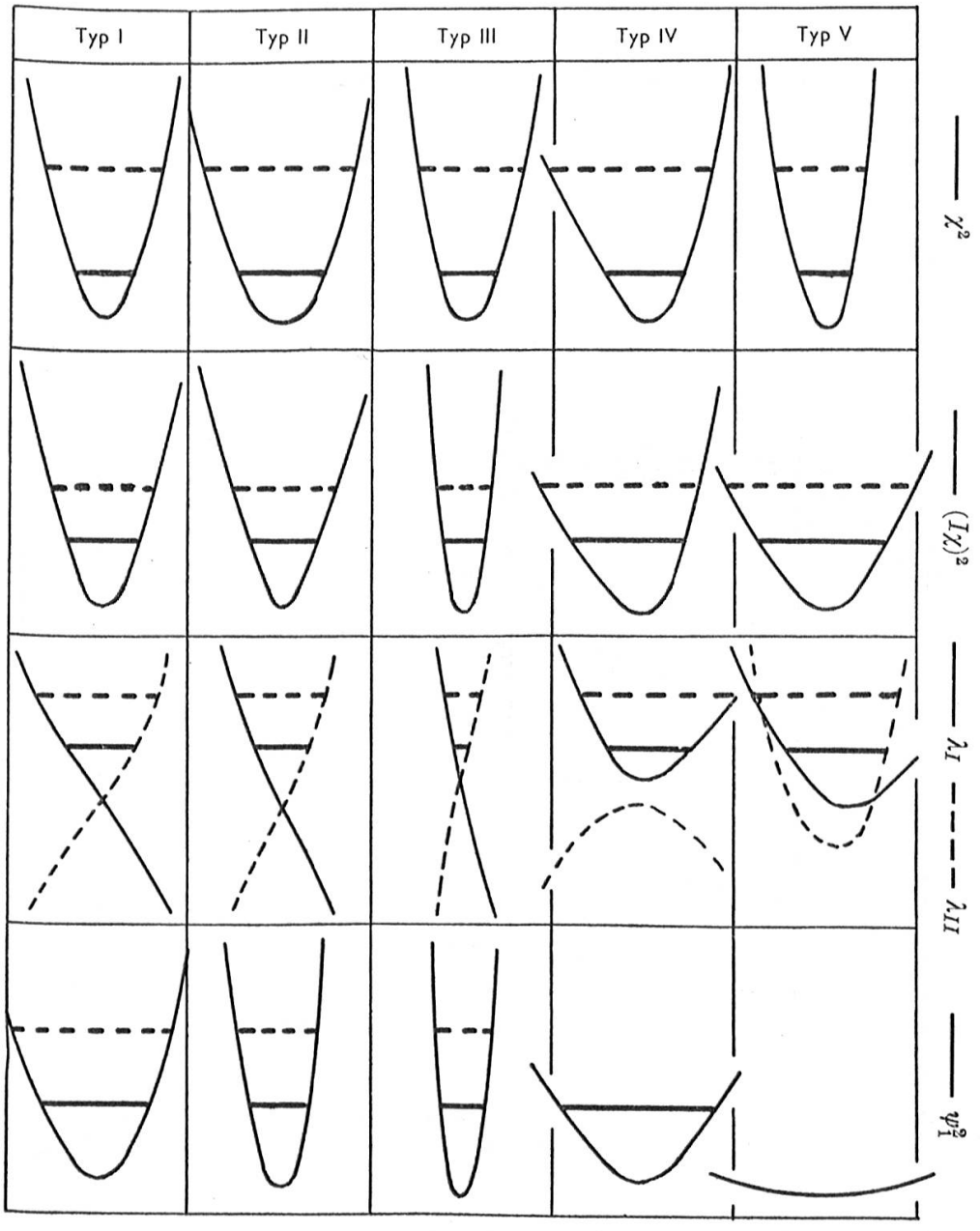
Figur 1

Regressionslinien der standardisierten Abweichungen für Parametervariationen bei Makehamschen Sterbetafeln



Figur 2

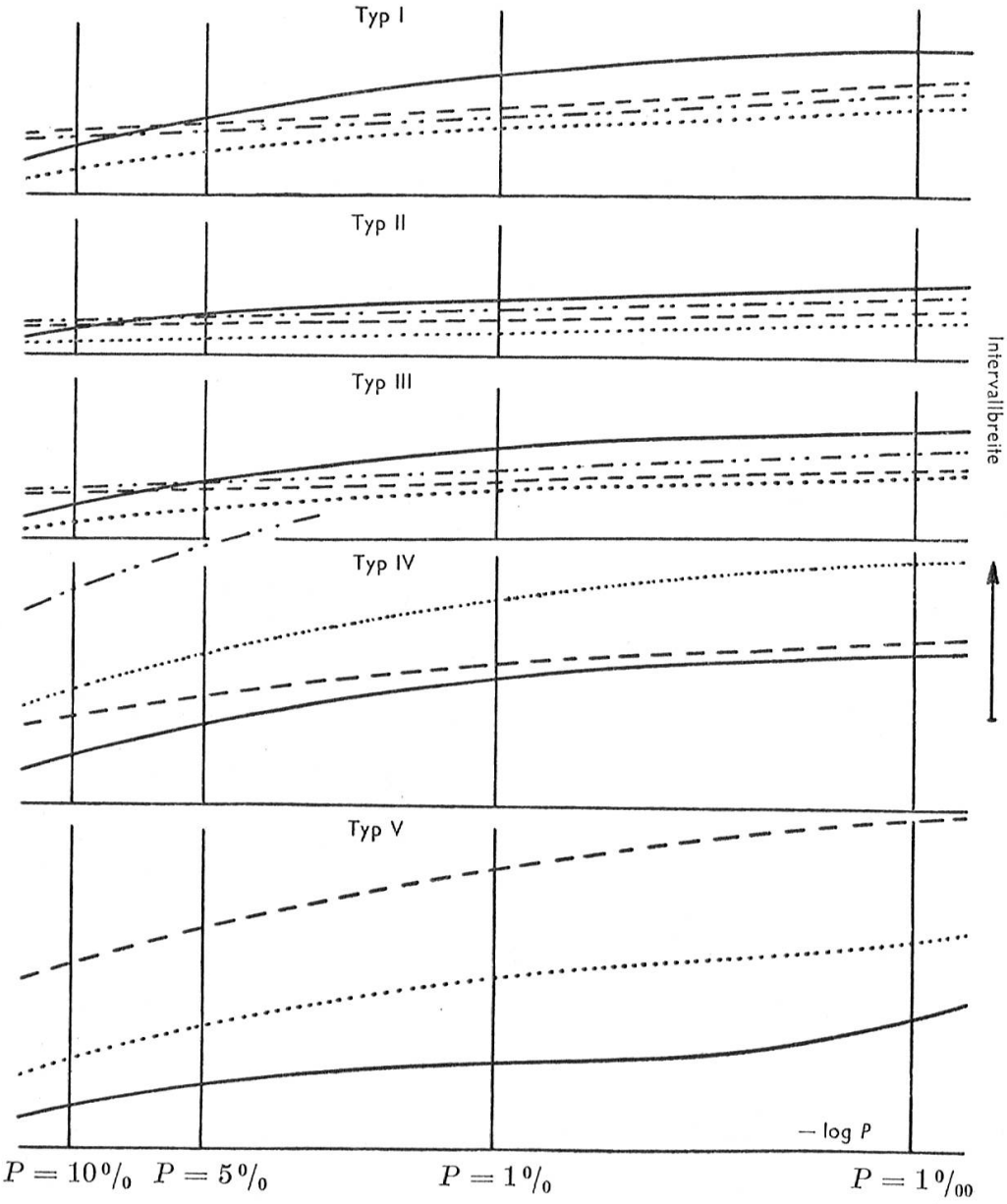
Verlauf der Masszahlen χ^2, $(I\chi)^2$, λ_I, λ_{II} und ψ_1^2 für Parametervariationen bei Makehamschen Sterbetafeln



– – – – $P = 1\%$
——— $P = 5\%$
} Parameterintervalle der zulässigen Tafeln

Figur 3

Intervallbreite der zulässigen Makehamparameter als Funktion der Wesentlichkeitsschranke P



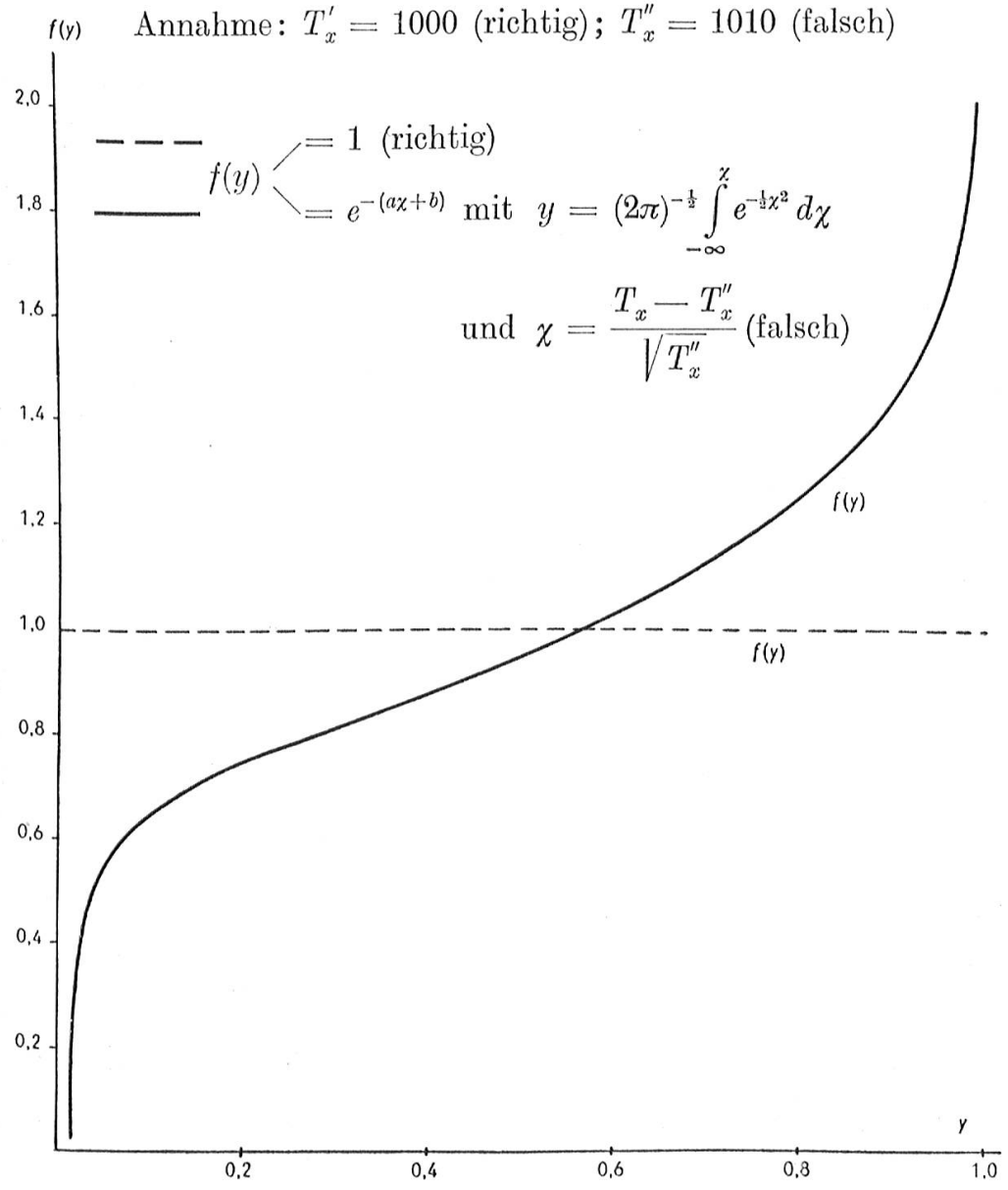
Figur 4

Verteilungsfunktionstransformation bei der Verteilung

$$f(T_x) = (2\pi\,\bar{T}_x)^{-\frac{1}{2}}\, e^{-\frac{1}{2}\frac{(T_x-\bar{T}_x)^2}{\bar{T}_x}}$$

bei richtiger und falscher Annahme über den Erwartungswert $\bar{T}_x$

$$y_x = \int_{-\infty}^{T_x} f(T_x)\, dT_x$$



Literaturverzeichnis

A. Lehrbücher

[1] *O. Anderson:* Einführung in die mathematische Statistik. Wien 1935.
[2] *H. Cramér:* Mathematical methods of statistics. Princeton 1946.
[3] *W. P. Elderton:* Frequency curves and correlation. Cambridge 1938.
[4] *R. A. Fisher:* Statistical methods for research workers. London 1922.
[5] *G. M. Kendall:* The advanced theory of statistics. London 1947.
[6] *A. Linder:* Statistische Methoden. Basel 1944.
[7] *R. v. Mises:* Wahrscheinlichkeitsrechnung. Leipzig 1931.
[8] *E. T. Whittaker* and *G. Robinson:* The calculus of observations. London 1940.

B. Einzelarbeiten

[9] *H. Ammeter:* Untersuchungen über die jährlichen Sterblichkeitsschwankungen in einem Versicherungsbestand. Mitteilungen 1945.
[10] — A generalized χ^2-distribution and its applications for testing mortality table graduations by moving averages. Proceedings of the XIIIth international actuarial congress Scheveningen 1951.
[11] — Ein neues Testverfahren für geordnete Beobachtungsreihen. Mitteilungen 1951.
[12] *H. Cramér:* On the composition of elementary errors. Skand. Aktuarietidskrift 1928.
[13] *H. Cramér* und *H. Wold:* Mortality variations in Sweden. Skand. Aktuarietidskrift 1935.
[14] *F. N. David:* A 'χ^2-smooth-test' for goodness of fit. Biometrika 1947.
[15] *R. A. Fisher:* Contributions to mathematical statistics. New York 1950.
[16] *F. R. Helmert:* Über die Wahrscheinlichkeit der Potenzsummen und einige damit im Zusammenhang stehende Fragen. Zeitschrift für Mathematik und Physik 1876.
[17] *J. Neyman:* Smooth-test for goodness of fit. Skand. Aktuarietidskrift 1937.
[18] *J. Neyman* and *E. S. Pearson:* On the problem of the most efficient tests of statistical hypotheses. Phil. Transact. A 1933.
[19] *E. S. Pearson:* The probability integral transformation for testing goodness of fit and combining independent tests of significance. Biometrika 1938.
[20] *K. Pearson:* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag. 1900.
[21] — On a method of determining whether a sample of given size n supposed to have been drawn from a parent population having a known probability integral, has probably been drawn at random. Biometrika 1933.
[22] *H. L. Seal:* Tests of a mortality table graduation. Journal of the Institute of Actuaries 1943.
[23] — A probability distribution of deaths at age x, when policies are counted instead of lives. Skand. Aktuarietidskrift 1947.
[24] — A note on the χ^2-smooth-test. Biometrika 1948.
[25] *M. N. Smirnoff:* Sur la distribution de ω^2. Comptes rendus de l'académie des sciences 1936.
[26] *W. Wegmüller:* Grundlagen der Schweiz. Volkssterbetafeln 1931/41 und 1939/44. Schweiz. Zeitschrift für Volkswirtschaft und Statistik 1949.