

# Parameterschätzung

Karl Christoph Klauer

Die Neyman-Pearson-Theorie statistischer Tests (vgl. Willmes, in diesem Band) darf insofern als klassisch gelten, als weitreichende Übereinstimmung hinsichtlich ihrer praktischen und theoretischen Bedeutung besteht. Eine in ähnlicher Weise klassisch zu nennende Theorie zur Schätzproblematik gibt es leider nicht. Das liegt nicht an einem Mangel interessanter theoretischer Resultate, sondern vermutlich daran, daß über den praktischen Wert der verschiedenen Ansätze unterschiedliche Meinungen bestehen. Neyman-Pearson-Tests liefern eine Methode der Konstruktion von Tests, die ausgesprochen weitreichend anwendbar ist. In der Meinung vieler Statistiker spielt die Methode der *Maximum-Likelihood*-Schätzung eine vergleichbar universelle Rolle für die Konstruktion von Schätzern. Diese Statistiker werden jedoch von den Anhängern der Bayes-Methoden (vgl. Molenaar & Lewis, in diesem Band) stark attackiert, und als zusätzliche Komplikation mischen sich Kontroversen über das „wahre Wesen“ von Wahrscheinlichkeiten ein. Zudem läßt sich die *Maximum-Likelihood*-Methode nicht unmittelbar im Rahmen der heute vorherrschenden entscheidungstheoretischen Rekonstruktion der Statistik ableiten. Das gelingt, wie wir sehen werden, allerdings im Bereich der sogenannten asymptotischen Statistik.

## 1 Grundlegende Probleme und Begriffe

Das Schätzproblem beginnt bei einer Familie von Wahrscheinlichkeitsverteilungen  $\{P_\theta, \theta \in \Theta\}$  auf einem Grundraum  $\Omega$ , die durch einen möglicherweise vektorwertigen Parameter  $\theta$  indiziert ist. Das kann zum Beispiel die Familie der Normalverteilungen mit Varianz eins und Mittelwert  $\mu$  sein, der dann die Rolle des Parameters übernimmt. Im allgemeinen wird angenommen, daß die Beobachtungen  $\omega$  aus dem Grundraum  $\Omega$  auf der Grundlage einer bestimmten Wahrscheinlichkeitsverteilung aus der Familie generiert wurden. Diese entspricht einem bestimmten, leider aber unbekanntem  $\theta$ . Das Schätzproblem besteht darin, aus den Daten  $\omega$  einen Schätzwert  $\hat{\theta}$  für den unbekanntem, zugrundeliegenden Parameter  $\theta$  zu berechnen. Tatsächlich sollte man etwas allgemeiner sagen, daß es darum geht, eine Funktion  $f(\theta)$  des Parameters zu schätzen. Zum Beispiel kann es sein, daß  $\theta$  aus unbekanntem Mittelwert und unbekannter Standardabweichung einer Normalverteilung besteht, aber nur der Mittelwert, also eine Funktion des Parameters  $\theta$ , zu schätzen ist. Der Einfachheit halber wollen wir aber diese Komplikation zunächst außer acht lassen.

Eine Funktion  $\kappa$ , die jeder Beobachtung  $\omega$  einen Wert aus dem Parameterraum  $\Theta$  zuweist, heißt Schätzer. Ein bestimmter Wert  $\hat{\theta} = \kappa(\omega)$  dieser Funktion heißt Schätzwert. Nicht jeder Schätzer ist gleich gut. Die Güte eines Schätzers bewertet man anhand der Größe der Abweichungen des Schätzwerts vom zugrundeliegenden

Parameterwert. Die Größe der Abweichung wird durch eine Verlustfunktion  $W_\theta$  beziffert. Sie ist nicht nur eine Funktion des Schätzwerts  $\hat{\theta}$ , sondern hängt auch vom Parameter  $\theta$  ab, da dieser in ihre Berechnung eingeht. Sehr bekannte Beispiele sind bei eindimensionalem Parameter die quadratische oder Gaußsche Verlustfunktion:  $W_\theta(\hat{\theta}) = (\hat{\theta} - \theta)^2$ , und die absolute Abweichung:  $W_\theta(\hat{\theta}) = |\hat{\theta} - \theta|$ . Den mittleren Verlust,  $R_\theta(\kappa) = E_\theta(W_\theta(\kappa(\omega)))$ , der zu erwarten ist, wenn ein Schätzer  $\kappa$  eingesetzt wird, bezeichnet man als Risiko des Schätzers. Wird die Gaußsche Verlustfunktion zugrunde gelegt, so spricht man auch vom mittleren quadratischen Fehler des Schätzers. Für eine gegebene Verlustfunktion ist ein Schätzer natürlich umso besser, je geringer sein Risiko ist.

Da der zugrundeliegende Parameter  $\theta$  nicht bekannt ist, betrachtet man in der klassischen Statistik vorsorglich alle möglichen Parameterwerte und befaßt sich demzufolge mit ganzen Risikofunktionen  $R_\theta$  als Funktion des Parameters  $\theta$ . Liegen also zwei Schätzer  $\kappa$  und  $\kappa'$  vor, so müssen die Risiken beider Schätzer für jeden möglicherweise zugrundeliegenden Parameterwert  $\theta$  verglichen werden. Das ist ein ernstzunehmendes Problem, da es schwierig ist, zwei Funktionen verbindlich zu vergleichen.

Ein klarer Fall liegt vor, wenn der eine Schätzer  $\kappa$  für jeden Parameterwert ein höchstens ebenso großes Risiko wie der andere hat,  $R_\theta(\kappa) \leq R_\theta(\kappa')$ , und wenn er für wenigstens einen Wert  $\theta$  ein geringeres Risiko hat. Dann wird man  $\kappa$  immer  $\kappa'$  vorziehen, und  $\kappa$  wird auch formell „besser“ genannt als  $\kappa'$ . Auf der Suche nach guten Schätzern wird man also sicherlich solche ausscheiden können, für die ein in diesem Sinne besserer existiert. Die verbleibenden Schätzer heißen zulässige und bilden leider im allgemeinen immer noch eine sehr große Klasse von Schätzfunktionen. Einen besten Schätzer, der besser ist als alle anderen, muß es gar nicht geben, oder er mag sehr schwer zu finden sein.

Das allgemeine entscheidungstheoretische Schätzproblem erwies sich in vielen Fällen als unlösbar oder zu schwierig, und verschiedene heuristische Strategien wurden eingesetzt, um aus der Klasse möglicher Schätzfunktionen, wenn schon nicht insgesamt beste, so doch gute auszuzeichnen. Eine Klasse von Ansätzen reduziert die Komplexität des Problems, indem die Risikofunktionen durch einen Zahlenwert summarisch charakterisiert werden. Dann gilt es nur noch diese Zahlenwerte, aber nicht mehr ganze Funktionen zu vergleichen. Dieser Idee bedienen sich die sogenannten Minimax-Schätzer und die Bayes-Schätzer.

Das Minimax-Prinzip besteht darin, das maximale Risiko zu minimieren. Für einen Schätzer  $\kappa$  kann man das größte Risiko als den größten Wert angeben, den die Risikofunktion über die Parameterwerte  $\theta$  hinweg annimmt, beziehungsweise dem sie sich beliebig eng annähert. Ein Schätzer, der dieses maximale Risiko in der Klasse aller denkbaren Schätzer minimiert, heißt Minimax-Schätzer. Minimax-Schätzer begrenzen also das Risiko im ungünstigsten Fall. Manchmal gibt es eine Annahme darüber, wie wahrscheinlich das Auftreten einzelner Parameterwerte  $\theta$  als zugrundeliegender wahrer Wert ist. Ist diese Annahme so konkret, daß sie in eine Wahrscheinlichkeitsverteilung  $\nu$  für den Parameter  $\theta$  gegossen werden kann, so kommen Bayes-Methoden ins Spiel (vgl. Molenaar & Lewis, in diesem Band). Dann kann die Risikofunktion summarisch durch das mittlere Risiko bewertet werden, das der Erwartungswert des Risikos unter der Verteilung  $\nu$  ist. In den meisten Fällen gibt es einen Schätzer, der dieses mittlere Risiko minimiert. Er heißt Bayes-Schätzer. Bayes-

Schätzer minimieren also das an den Vorannahmen über die Wahrscheinlichkeit der Parameterwerte gewichtete Risiko. Eine Verallgemeinerung des Bayes-Ansatzes führt zu verwandten, in der einschlägigen Literatur diskutierten Schätzern wie den Pitman-Schätzern oder den *Maximum-Probability*-Schätzern (Strasser, 1985; Kapitel 7).

Eine andere Klasse von Ansätzen reduziert die Komplexität des Problems, indem die Menge von Schätzern, innerhalb derer nach einem besten gesucht wird, durch Nebenbedingungen von vornherein eingeschränkt wird. Hierbei bleibt es aber beim Vergleich ganzer Risikofunktionen, die Risikofunktionen werden also nicht summarisch durch einen Zahlenwert charakterisiert. Am bekanntesten sind die Nebenbedingungen, daß potentielle Schätzer „erwartungstreu“ sein sollen, beziehungsweise, daß sie erwartungstreu und linear sein sollen; auf die daraus resultierenden Theorien kommen wir weiter unten zu sprechen. Es gibt aber weitere, weniger weithin bekannte Nebenbedingungen, die untersucht werden, zum Beispiel Invarianzbedingungen oder die sogenannte Medianstreue (Strasser, 1985; Kapitel 7).

Zusammenfassend halten wir also erst einmal fest, daß es den klassischen oder besten Schätzer für einen gesuchten Parameterwert im allgemeinen nicht gibt. Die Wahl einer bestimmten Schätzfunktion hängt im allgemeinen von der Wahl einer bestimmten Verlustfunktion, einer oder mehrerer Nebenbedingungen und möglicherweise eines vereinfachenden Kriteriums wie des Minimax-Prinzips ab. Im folgenden sollen einige der klassischen Ergebnisse der statistischen Schätztheorie geschildert werden. Am bekanntesten sind wohl die Theorien des *best linear unbiased estimate* und die Theorien der erwartungstreuen Schätzer. Darüber hinaus soll das *Maximum-Likelihood*-Prinzip begründet werden. Schließlich werden sogenannte Konfidenzschätzungen erklärt.

## 2 Kleinste-Quadrat-Schätzung

In vielen Fällen ist aufgrund von Vorüberlegungen oder -untersuchungen bekannt, daß Zielgrößen  $x_1, x_2, \dots, x_n$  aus einem unbekanntem, möglicherweise vektorwertigen Parameter  $\theta$  nach bekannten Modellfunktionen  $f_1, f_2, \dots, f_n$  hervorgehen,  $x_i = f_i(\theta), i = 1, \dots, n$ , wenn da nicht noch der Meßfehler  $e_i$  wäre. Unter Berücksichtigung des Meßfehlers gilt tatsächlich nur,  $i = 1, 2, \dots, n$ :

$$x_i = f_i(\theta) + e_i. \quad (1)$$

Meßfehler und Zielgrößen sind Realisierungen von Zufallsvariablen  $E_i$  und  $X_i$ . Es wird angenommen, daß die Fehler im Mittel null sind:  $E_\theta(E_i) = 0$ . In dieser Situation erweist es sich manchmal als günstig, einen Schätzer für  $\theta$  als sogenannten Kleinste-Quadrat-Schätzer (*least squares estimator*) zu bestimmen. Der Kleinste-Quadrat-Schätzwert ist der Parameterwert  $\hat{\theta}$ , der die folgende Summe von Quadraten minimiert:

$$\sum_{i=1}^n (x_i - f_i(\theta))^2. \quad (2)$$

In vielen Fällen existiert ein solcher Parameterwert und ist eindeutig. Dann ist der Kleinste-Quadrat-Schätzer definiert. Wenn die Funktionen  $f_i$  differenzierbar sind, gelingt es oft, ihn mit Hilfe von Routineverfahren der Analysis zu berechnen.

Über die Eigenschaften solcher Schätzer ist im allgemeinen wenig bekannt. Eine Ausnahme liegt allerdings vor, wenn die Funktionen  $f_i$  linear sind. Dann läßt sich der Einsatz Kleinst-Quadrat-Schätzer entscheidungstheoretisch gut begründen. Dieser Fall ist in der Praxis auch besonders wichtig, er stellt das Allgemeine Lineare Modell (vgl. Andres, in diesem Band) dar. Varianz- und Regressionsanalyse arbeiten zum Beispiel mit Kleinste-Quadrat-Schätzern.

Nehmen wir an, der Parameter  $\theta = (\theta_1, \dots, \theta_p)$  habe  $p < n$  Komponenten und die  $n$  Linearkombinationen  $f_i$  spannen ein System vollen Rangs  $p$  auf. Ferner sei die Varianz-Kovarianzmatrix der Fehlervariablen  $E = (E_1, \dots, E_n)$  positiv definit (zur Definition dieser Begriffe vgl. das Kapitel über Grundlagen der multivariaten Datenanalyse von Andres, in diesem Band). Dann existiert der Kleinste-Quadrat-Schätzer  $\hat{\theta}$  von  $\theta$ . Er hat einige attraktive Eigenschaften. So ist er selber eine lineare Funktion der Daten  $x_i, i = 1, \dots, n$ . Außerdem ist er erwartungstreu (*unbiased*):

$$E_{\theta}(\hat{\theta}) = \theta. \quad (3)$$

Schließlich ist er in der Klasse aller erwartungstreuen Schätzer, die lineare Funktionen der Daten sind, bezüglich der quadratischen Verlustfunktion

$$W_{\theta}(\hat{\theta}) = \sum_{i=1}^p (\hat{\theta}_i - \theta_i)^2 \quad (4)$$

ein bester Schätzer. Das heißt, er minimiert das Risiko für jeden denkbaren tatsächlich zugrundeliegenden Parameter  $\theta$ . Der Kleinste-Quadrat-Schätzer im Kontext linearer Modelle heißt deswegen auch oft *best linear unbiased estimator*.

Wichtige Kleinste-Quadrat-Schätzer sind zum Beispiel das Stichprobenmittel als Schätzwert für das Populationsmittel oder die Effektschätzungen der Varianzanalyse. Tatsächlich kann man aber bereits an einfachen Beispielen zeigen, daß der Kleinste-Quadrat-Schätzer in der Klasse aller denkbaren Schätzer im allgemeinen nicht optimal beziehungsweise nicht einmal zulässig ist (Stein, 1956).

### 3 Erwartungstreue Schätzungen

Historisch gesehen hat man sich in vielen Fällen von vornherein auf die Klasse erwartungstreuer Schätzer konzentriert und darin nach einem besten gesucht. Das hat vermutlich zwei Gründe. Zum einen ist eine seit Gauß vielfach verwendete Verlustfunktion die quadratische. Eine einfache Aufspaltung des Schätzrisikos für den Schätzer  $\kappa$  eines eindimensionalen Parameters  $\theta$  ist dabei möglich:

$$R_{\theta}(\kappa) = E_{\theta}(\kappa - \theta)^2 = E_{\theta}(\kappa - E_{\theta}(\kappa))^2 + (E_{\theta}(\kappa) - \theta)^2. \quad (5)$$

Der erste Term,  $E_{\theta}(\kappa - E_{\theta}(\kappa))^2$ , ist natürlich nichts anderes als die Varianz von  $\kappa$ ,  $Var_{\theta}(\kappa)$ , der zweite ist die Verzerrung oder der *bias* des Schätzers. Dieser zweite Term wird null, wenn der Schätzer erwartungstreu ist. Auf der Suche nach guten Schätzern scheint es daher vernünftig zu sein, sich zunächst auf Schätzer zu konzentrieren, die erwartungstreu sind. Für solche ist das Risiko durch die Varianz des Schätzers gegeben, und es gilt dann, einen Schätzer kleinster Varianz zu finden.

Zum anderen ergeben sich für erwartungstreue Schätzer besonders hilfreiche und allgemeine Resultate im Zusammenhang mit sogenannten suffizienten Statistiken.

### 3.1 Suffiziente Statistiken

Im allgemeinen werden Daten detaillierter aufgezeichnet als notwendig. Ein einfaches Experiment liefere zum Beispiel Realisierungen  $x_1, x_2, \dots, x_n$  von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, \dots, X_n$ , die die Werte Null und Eins annehmen können. Dabei sei die Wahrscheinlichkeit für eine Eins durch  $p$ ,  $0 < p < 1$ , gegeben. Es ist intuitiv klar, daß die Anzahl von Einsen beziehungsweise Nullen eine angemessene Zusammenfassung der Daten darstellt, während zum Beispiel die genaue Reihenfolge der Einsen und Nullen im Verlaufe der voneinander unabhängigen Beobachtungen wenig hergibt.

Kann man die Daten in einer Statistik so zusammenfassen, daß die wesentlichen Informationen über den Parameter darin kondensiert werden, so spricht man von einer suffizienten Statistik. Im Falle diskreter Verteilungen heißt eine Statistik  $T$  suffizient, falls die bedingte Verteilung

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid T = t) \quad (6)$$

nicht von  $\theta$  abhängt. Das heißt salopp gesprochen, daß die Daten nur über die suffiziente Statistik vom zugrundeliegenden Parameter beeinflusst werden. Im Beispiel oben kann gezeigt werden, daß die Anzahl der Einsen,  $T = \sum X_i$ , tatsächlich in diesem Sinne eine suffiziente Zusammenfassung der Daten darstellt. Bei kontinuierlichen Verteilungen gilt eine entsprechende Definition suffizienter Statistiken, deren Formulierung allerdings einige technische Vorarbeiten verlangt, die wir hier nicht leisten können.

### 3.2 Suffiziente Statistiken und erwartungstreue Schätzer

Ein wichtiges Resultat, das berühmte Theorem von Rao und Blackwell, besagt, daß uns die Kenntnis einer suffizienten Statistik eine Verbesserung unserer Schätzung erlaubt. Ist nämlich  $\kappa$  ein erwartungstreuer Schätzer eines eindimensionalen Parameters  $\theta$  und  $T$  eine suffiziente Statistik, so existiert eine von  $\theta$  unabhängige Festlegung des bedingten Erwartungswerts von  $\kappa$  bei gegebenem  $T = t$ ,

$$\kappa' = E(\kappa \mid T = t). \quad (7)$$

Diese ist ebenfalls erwartungstreu und hat für alle  $\theta$  eine nicht größere Varianz,  $\text{Var}_\theta(\kappa') \leq \text{Var}_\theta(\kappa)$ , und damit nach dem oben Gesagten ein nicht größeres Risiko bei quadratischer Verlustfunktion.

Im Beispiel aus Abschnitt 3.1: Die erste Beobachtung  $X_1$  ist für sich betrachtet bereits ein erwartungstreuer Schätzer für  $p$ , denn es gilt ja:  $E_p(X_1) = p$ . Eine suffiziente Statistik  $T$  ist die Anzahl der Einsen,  $T = \sum X_i$ . Die Anwendung des Satzes von Rao und Blackwell besteht darin, den bedingten Erwartungswert von  $X_1$  bei gegebener Summe  $T = t$  zu berechnen. Man kann zeigen, daß sich als bedingter Erwartungswert die relative Häufigkeit von Einsen ergibt:  $E(X_1 \mid T = t) = t/n$ . Die Varianz von  $X_1$  beträgt  $p(1-p)$ , die von  $\frac{1}{n}T$  dagegen  $\frac{1}{n}p(1-p)$ . Die Kenntnis einer suffizienten Statistik führt also zu einer deutlichen Verbesserung der Schätzung.

Tatsächlich gilt der Satz von Rao und Blackwell nicht nur bezüglich der quadratischen Verlustfunktion, sondern hinsichtlich beliebiger konvexer Verlustfunktionen.

Im allgemeinen ist es nicht schwierig, suffiziente Statistiken zu identifizieren. Wie gesagt, bestehen sie aus einer geeigneten Zusammenfassung der Daten. Es ist allerdings jedesmal denkbar, daß eine noch gröbere Kondensierung der Daten ebenfalls zu einer suffizienten Statistik führt. Es ist daher wünschenswert, Kriterien zu finden, die anzeigen, daß die größte Zusammenfassung, die ohne Informationsverlust möglich ist, gefunden wurde. Solche Kriterien sind die Minimalsuffizienz (Rohatgi, 1984; Kapitel 10) und die Vollständigkeit (Witting, 1978; Kapitel 3). Die Vollständigkeit stellt sicher, daß ein erwartungstreuer Schätzer kleinster Varianz gefunden wurde. Auf Vollständigkeit und Minimalsuffizienz soll hier nicht weiter eingegangen werden.

Ein wichtiges Resultat sei abschließend noch erwähnt. Es handelt sich um die sogenannte Cramer-Rao-Schranke für erwartungstreue Schätzer. Die Cramer-Rao-Schranke ist der inverse Wert der sogenannte Fisherschen Informationsfunktion  $F(\theta)$ . Die Cramer-Rao-Schranke ist eine untere Schranke der Varianz eines jeden erwartungstreuen Schätzers von  $\theta$ , sofern die zugrundeliegende Familie von Wahrscheinlichkeitsverteilungen  $\{P(\theta), \theta \in \Theta\}$  gewisse Regularitätsbedingungen erfüllt. Das heißt auch, daß ein Schätzer, dessen Varianz durch diese Schranke gegeben ist, einer mit geringster Varianz ist.

#### 4 Das *Maximum-Likelihood*-Prinzip

Eine sehr wichtige Schätzmethode ist die *Maximum-Likelihood*-Methode. Sie ist zum einen fast universell einsetzbar. Zum anderen führt sie, wenigstens im asymptotischen Bereich, das heißt bei zunehmender Stichprobengröße, zu Schätzern mit entscheidungstheoretisch gesehen hervorragenden Eigenschaften. Selbst in kleinen Stichproben erweisen sich *Maximum-Likelihood*-Schätzer vielfach als vernünftige Schätzer.

Im Beispiel aus Abschnitt 3 ist die Wahrscheinlichkeit eines beobachteten Musters von Nullen und Einsen,  $\mathbf{x} = (x_1, \dots, x_n)$ , gegeben durch

$$P_p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n p^{x_i} (1-p)^{(1-x_i)}.$$

Die Wahrscheinlichkeit hängt von dem Wert des unbekanntem Parameters  $p$  ab. Eine vernünftige Schätzung des Parameters scheint dann der Wert zu sein, der diese Wahrscheinlichkeit maximiert. Das ist der *Maximum-Likelihood*-Schätzwert. Im Beispiel ergibt sich so derselbe Schätzwert wie in Abschnitt 3, nämlich die mittlere Zahl von Einsen,  $t/n$ . Verschiedene Zugangsweisen zum Schätzproblem können also zu denselben Lösungen führen.

Die Wahrscheinlichkeit der Daten hängt vom Parameter  $p$  ab. Um diese Abhängigkeit zu betonen, definiert man die Likelihoodfunktion  $L(p) = P_p(\mathbf{X} = \mathbf{x})$ , bei der das beobachtete Datenmuster festgehalten wird und nur der Parameter variiert. Im kontinuierlichen Fall ist die Wahrscheinlichkeit jedes einzelnen Datums gleich Null. Dort ist es sinnvoll, die Likelihoodfunktion nicht anhand der Wahrscheinlichkeiten, sondern mit Hilfe der Dichtefunktionen analog zu definieren. Der *Maximum-Likelihood*-Schätzwert ist, wie gesagt, der Parameterwert, für den die Likelihoodfunktion ihr Maximum annimmt. Die Bestimmung des Schätzwerts ist also ein Maximierungsproblem, das mit Verfahren der Analysis oder der numerischen

Mathematik in den meisten Fällen gelöst werden kann. Einige einfache Regularitätsbedingungen, die für die meisten in der Psychologie relevanten Modelle zutreffen, garantieren, daß bei zunehmender Stichprobengröße ein Maximum der Likelihoodfunktion fast sicher existiert und daß dieses Maximum genau einem Parameterwert zugeordnet ist (Rao, 1973, Kapitel 5). Dann ist der *Maximum-Likelihood*-Schätzwert eindeutig definiert.

Bei komplizierten Modellen, die viele Parameter berücksichtigen, ist allenfalls die Eindeutigkeit des Maximums manchmal fraglich, wenn es nämlich mehrere Kombinationen von Parameterwerten gibt, für die die Likelihoodfunktion maximal wird. Das ist das sogenannte Identifizierbarkeitsproblem, das sich in vielen Fällen durch Reparametrisierung lösen läßt.

Über die Eigenschaften des *Maximum-Likelihood*-Schätzers  $\kappa$ , wenn er existiert, ist bei kleinen Stichproben wenig bekannt. Insbesondere ist er im allgemeinen nicht erwartungstreu. Immerhin kann er aber immer als Funktion einer suffizienten Statistik gewählt werden, wenn es eine suffiziente Statistik gibt. Wenn aber die Stichprobengröße  $n$  der unabhängigen und identisch verteilten Beobachtungen zunimmt, so zeigt er viele interessante Eigenschaften: Er ist konsistent, das heißt, die Wahrscheinlichkeit  $P_{\theta,n}(|\kappa_n - \theta| > \epsilon)$  einer beliebig kleinen Abweichung der Größe  $\epsilon > 0$  zwischen Schätzwert  $\kappa_n$  und Parameterwert  $\theta$  strebt gegen Null, wenn die Stichprobengröße zunimmt,  $n \rightarrow \infty$ . Der *Maximum-Likelihood*-Schätzwert nähert sich dem Parameterwert also mit zunehmender Stichprobengröße immer genauer an.

Es ist aber noch Genaueres über die Abweichungen  $\kappa_n - \theta$  vom Parameterwert bekannt. Sie werden im Mittel in der Größenordnung von  $\frac{1}{\sqrt{n}}$  kleiner, und die Verteilung der normierten Abweichungen

$$\sqrt{n}(\kappa_n - \theta)$$

kann mit zunehmender Stichprobengröße immer besser durch eine Normalverteilung mit Mittelwert Null und einer Varianz bestimmbarer Größe beschrieben werden: Der *Maximum-Likelihood*-Schätzer ist also asymptotisch normalverteilt. Das erlaubt es zum Beispiel bei großen Stichproben, Konfidenzintervalle für den gesuchten Parameterwert anzugeben.

Bei zunehmender Stichprobengröße,  $n \rightarrow \infty$ , erweist sich der *Maximum-Likelihood*-Schätzer zudem anderen Schätzern gegenüber überlegen. Zum Beispiel kann gezeigt werden, daß er ein lokaler asymptotischer Minimax-Schätzer ist, das heißt, daß er um jeden Parameterwert  $\theta$  herum das maximale Risiko bezüglich beliebiger stetiger und konvexer Verlustfunktionen minimiert. Diese und weitere wichtige Optimalitätseigenschaften faßt Strasser (1985, Kapitel 13) zusammen.

Die *Maximum-Likelihood*-Methode findet weitreichende Anwendung. Insbesondere bei der Modellierung von Kontingenztabellen durch log-lineare und multinomiale Modelle ist sie die Standardmethode. *Maximum-Likelihood*-Schätzung spielt ferner für die Schätzung von Personen- und Itemparametern der probabilistischen Testtheorie eine herausragende Rolle. Viele mit anderen Ansätzen gewonnene Schätzer erweisen sich, wie eingangs exemplifiziert, zudem im Nachhinein auch als *Maximum-Likelihood*-Schätzer.

## 5 Konfidenzschätzung

Während die vorangegangenen Abschnitte mit sogenannten Punktschätzungen befaßt waren, soll es nun um Bereichsschätzungen gehen. Eine nützliche Darstellung von Beobachtungen bieten Konfidenzintervalle. Konfidenzintervalle sind Funktionen, die Beobachtungen  $\omega$  Intervalle  $I(\omega) \subset \Theta$  des Parameterraums zuweisen. Eine solche Funktion ist ein Konfidenzintervall vom Niveau  $1 - \alpha$ , wenn es den zugrundeliegenden Parameter  $\theta$  mit einer Wahrscheinlichkeit von wenigstens  $1 - \alpha$  überdeckt. Die Menge aller denkbaren Beobachtungen  $\omega$ , deren Intervall  $I(\omega)$  den Parameterwert  $\theta$  enthält, soll also eine Wahrscheinlichkeit von wenigstens  $1 - \alpha$  besitzen. Dies soll für jedes möglicherweise zugrundeliegende  $\theta$  gelten:

$$P_\theta(\theta \in I(\omega)) \geq 1 - \alpha. \quad (8)$$

Es macht im diskreten Fall keinen Sinn zu fordern, daß die Wahrscheinlichkeit von  $1 - \alpha$  genau eingehalten werden soll. Die Wahrscheinlichkeitsverteilung  $P_\theta$  nimmt ja im diskreten Fall nicht einmal jeden Wert zwischen Null und Eins an, vielleicht also auch nicht  $1 - \alpha$ . Im allgemeinen ist im diskreten Fall die linke Seite der Gleichung (8) auch eine Funktion von  $\theta$  und nicht konstant.

Offenbar ist es aber sehr einfach, ein Konfidenzintervall mit der in (8) angegebenen Eigenschaft zu konstruieren. Zum Beispiel ist  $I_1(\omega) = (-\infty, \infty)$ , das jeder Beobachtung das Intervall aller reellen Zahlen zuweist, trivialerweise ein Konfidenzintervall vom Niveau  $1 - \alpha$ , denn es gilt:  $P_\theta(\theta \in I_1) = 1$ . In Anwendungen ist man daran interessiert, Intervalle zu erhalten, die so klein wie möglich sind und immer noch eine Überdeckungswahrscheinlichkeit von wenigstens  $1 - \alpha$  garantieren. Wie schon bei den Punktschätzern der letzten Abschnitte gibt es verschiedene Konzeptualisierungen dafür, die Güte eines gegebenen Konfidenzintervalls zu bewerten und verschiedene Konfidenzintervalle zu vergleichen. Wir wollen uns auf einen Ansatz beschränken, der den meisten in der Praxis verwendeten Konfidenzintervallen zugrunde liegt und eine weithin anwendbare Methode der Konstruktion von Konfidenzintervallen erschließt. Es handelt sich um die Theorie der gleichmäßig genauesten Konfidenzintervalle (Mood, Graybill & Boes, 1974, p. 464).

Das Beispiel mit dem Intervall aller reellen Zahlen  $I_1$  zeigt, daß weitere Bedingungen notwendig sind, um derart schlechte Intervalle auszuschließen. Ein Konfidenzintervall heißt unverfälscht, wenn es falsche Parameter  $\theta'$ ,  $\theta' \neq \theta$ , mit einer Wahrscheinlichkeit von höchstens  $1 - \alpha$  überdeckt:

$$P_\theta(\theta' \in I) \leq 1 - \alpha. \quad (9)$$

Das Konfidenzintervall  $I_1$  ist offenbar nicht unverfälscht:  $P_\theta(\theta' \in I) = 1 > 1 - \alpha$ .

Obwohl die Bedingung (9) viele schlechte Konfidenzintervalle ausschließt, reicht sie nicht aus, denn im allgemeinen existiert immer noch eine große Klasse von unverfälschten Konfidenzintervallen eines gegebenen Niveaus. Ein gleichmäßig genauestes Intervall  $I^*$  ist aber ein Intervall aus dieser Klasse, das falsche Parameterwerte mit gleichmäßig kleinster Wahrscheinlichkeit überdeckt. Für alle  $\theta$  und  $\theta'$  soll also gelten:

$$P_\theta(\theta' \in I^*) = \inf, \quad (10)$$



wobei das Infimum über alle alternativen unverfälschten Intervalle  $I$  vom Niveau  $1 - \alpha$  berechnet wird (Witting, 1978; Kapitel 1.6).

Zwischen den gleichmäßig besten Konfidenzintervallen und den gleichmäßig besten Tests der Neyman-Pearson-Theorie (vgl. Willmes, in diesem Band) bestehen enge Beziehungen. Das Problem, ein gleichmäßig bestes Konfidenzintervall zu konstruieren, kann in ein gleichwertiges Testproblem umgeformt werden. Alles, was daher für die Existenz bester Tests gilt, ist direkt auf die Konfidenzintervalle übertragbar. Insbesondere können beste Konfidenzintervalle also immer dann gefunden werden, wenn die zugrundeliegende Familie von Wahrscheinlichkeitsverteilungen  $\{P_\theta, \theta \in \Theta\}$  eine Exponentialfamilie von Verteilungen darstellt (Barndorff-Nielsen, 1978; Witting, 1978; Kapitel 2).

Konfidenzintervalle haben gegenüber Punktschätzungen einige Vorteile. Sie erlauben eine anschauliche Bewertung der Genauigkeit der Schätzung durch die Größe des Intervalls. Wegen der engen Beziehungen zwischen Tests und Konfidenzintervallen werden zudem Inferenzen über den zugrundeliegenden Parameterwert ermöglicht. Wegen Gleichung (8) zum Beispiel ist der zugrundeliegende Parameterwert  $\theta$  von außerhalb des Intervalls liegenden Parameterwerten  $\theta'$  auf dem  $\alpha$ -Niveau signifikant verschieden.

## 6 Weiterführende Literatur

Einen Überblick über gebräuchliche Schätzstatistiken in verschiedenen, für die Sozialwissenschaften relevanten Bereichen gibt Rohatgi (1984). Das Buch richtet sich an Anwender und ist dementsprechend vergleichsweise einfach zu lesen, wenn es auch auf der konzeptuellen Ebene nicht zufriedenstellt. Die Theorie der Kleinst-Quadrat-Schätzer im Kontext linearer Modelle beschreibt Rao (1973). Er bietet auch eine Zusammenfassung wichtiger Eigenschaften und Voraussetzungen der *Maximum-Likelihood*-Methode. Das Buch von Rao setzt schon mehr Kenntnisse der Mathematik voraus, ist aber für Anwender immer noch lesbar. Standardwerke der mathematischen Statistik, die den modernen entscheidungstheoretischen Ansatz konsequent verfolgen, sind Witting (1978) und Strasser (1985). Witting stellt die wichtigsten Ergebnisse für erwartungstreue Schätzer dar und liefert eine hervorragende Erklärung und Ableitung von Konfidenzschätzern. Strasser bemüht sich um stringente, allgemeine Formulierungen und behandelt das Schätzproblem, inklusive jüngster Entwicklungen, wohl am umfassendsten. Leider ist er für Anwender kaum zu verstehen.

### Literaturverzeichnis

- Barndorff-Nielsen, O. (1978). *Information and exponential families*. New York: Wiley.
- Mood, A. M., Graybill, F. A. & Boes, D. C. (1974). *Introduction to the theory of statistics*. Tokio: McGraw Hill.
- Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.
- Rohatgi, V. K. (1984). *Statistical inference*. New York: Wiley.
- Stein, C. (1956). Inadmissability of the usual estimator for the mean of a multivariate normal distribution. *Proceedings 3rd Berkeley Symposium*, 1, 197–206.
- Strasser, H. (1985). *Mathematical theory of statistics*. Berlin: de Gruyter.
- Witting, H. (1978). *Mathematische Statistik*. Stuttgart: Teubner.