Kapitel 12

Elemente der Mathematischen Statistik

Beim Umgang mit zufälligen Erscheinungen ist es oft von Interesse, die Verteilungsfunktion F_X gewisser Zufallsgrößen X zu kennen. Daraus lassen sich Erwartungswert, Streuung, aber auch Wahrscheinlichkeiten der Form $P(X > c)$ berechnen. Diese Verteilungsfunktion ist in vielen Fällen jedoch nicht bekannt. Beispielsweise sind für ein Versicherungsunternehmen, das die Haftpflicht für Autofahrer versichert, die Wahrscheinlichkeitsverteilung der Anzahl der Unfälle pro Jahr und Versicherungsnehmer oder die Verteilung der Schadensumme pro Jahr und Versicherungsbestand Grundlagen für die Berechnung der Versicherungsprämie, die jeder Versicherungsnehmer im Jahr zu bezahlen hat.

Bekannt sind in vielen Fällen jedoch Daten, die Auskunft über die unbekannte Verteilungsfunktionen geben können. So verfügen Versicherungsunternehmen über umfangreiche Datensammlungen zeitlich zurück liegender Schadensfälle. Sie betreffen sowohl Schadenshäufigkeiten in einem Versicherungsbestand als auch Schadenshöhen.
In der klassischen Statistik geht man meist davon aus, dass der zugrunde liegende Datensatz die mehrfache voneinander unabhängige Realisierung einer Zufallsgröße X mit einer Verteilungsfunktion F_X ist, er bildet eine sogenannte "Stichprobe". Die Mathematische Statistik konstruiert und bewertet Verfahren, um aus Stichproben Rückschlüsse auf F_X oder Kenngrößen von F_X zu ziehen.
Zentrale Fragestellungen sind dabei das Schätzen von Parametern der zugrun-

de liegenden Verteilung und das Testen von Hypothesen über diese Parameter.

Eine prinzipielle Möglichkeit, wie man zu der Verteilungsfunktion F_X kommt, eröffnet der folgende Hauptsatz der Mathematischen Statistik. Er besagt, dass man F_X auf der Grundlage von Stichproben prinzipiell beliebig genau bestimmen kann.

12.1 Der Hauptsatz der mathematischen Statistik

Es seien F eine Verteilungsfunktion auf R_1 und $X^{(n)} := (X_1, \ldots, X_n)$ eine Folge unabhängiger identisch verteilter Zufallsgrößen über einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$ mit der Verteilungsfunktion F:

$$F(x) = P(X_k \leq x), \quad x \in R_1, \quad k = 1, \ldots, n.$$

Definition 12.1 *Man bezeichnet $X^{(n)}$ mit diesen Eigenschaften auch als mathematische Stichprobe vom Umfang n aus einer nach F verteilten Grundgesamtheit. Realisiert man die Zufallsgrößen $X_k, k = 1, \ldots, n$, so erhält man eine konkrete Stichprobe $x^{(n)} := (x_1, \ldots, x_n)$ vom Umfang n aus einer nach F verteilten Grundgesamtheit.*

Beispiel 12.2 Es sei $X^{(n)} = (X_1, X_2, \ldots, X_n)$ ein Bernoullischema $BS_n(p)$ mit $p \in (0,1)$. Der konkrete Wert von p sei unbekannt. Dann ist X im obigen Sinne eine mathematische Stichprobe aus einer zweipunktverteilten Grundgesamtheit mit den möglichen Werten 1 und 0 und den entsprechenden Wahrscheinlichkeiten p bzw. $1-p$. Jede Realisierung $x^{(n)}$ von $X^{(n)}$, zum Beispiel für $n = 5$

$$x^{(5)} = (0, 1, 1, 0, 1),$$

ist eine konkrete Stichprobe aus der erwähnten Grundgesamtheit.

Wir verbinden nun mit jeder Stichprobe eine neue Verteilungsfunktion.

Wir definieren

$$\hat{F}_n(x; X^{(n)}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i), \quad x \in R_1. \tag{12.1}$$

Die Funktion $\hat{F}_n(\cdot)$ (das Argument $X^{(n)}$ wird meist weggelassen) ist eine vom Zufall abhängige Verteilungsfunktion, die sogenannte *"empirische Verteilungsfunktion der mathematischen Stichprobe $X^{(n)} = (X_1, \ldots, X_n)$"*.

Da zu jeder Verteilungsfunktion F auf R_1 ein Wahrscheinlichkeitsmaß Q_F auf $\mathfrak{B}_1$ gehört, das F als seine Verteilungsfunktion besitzt, ist das auch für $\hat{F}_n$ der Fall. $Q_{\hat{F}_n}$ ist ein vom Zufall abhängiges diskretes Wahrscheinlichkeitsmaß und ordnet jedem Punkt $\{X_i(\omega)\}, i = 1, \ldots, n$, die Wahrscheinlichkeit

$$Q_{\hat{F}_n}(\{X_i(\omega)\}) = \frac{1}{n} \times \# \{j \in \{1, 2, \ldots, n\} \text{ mit } X_j(\omega) = X_i(\omega)\}$$

zu.

Setzt man in (12.1) anstelle $X^{(n)}$ eine Realisierung $x^{(n)}$, also eine konkrete Stichprobe, ein, so erhält man eine nichtzufällige Verteilungsfunktion

$$\hat{F}_n(x; x^{(n)}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(x_i), \quad x \in R_1. \tag{12.2}$$

Die dazu gehörende Wahrscheinlichkeitsverteilung ist die diskrete gleichmäßige Verteilung $Q_{\hat{F}_n}$ auf den Zahlen $\{x_1, \ldots, x_n\}$ mit

$$Q_{\hat{F}_n}(\{x_k\}) = \frac{1}{n} \times \#\{j \in \{1, \ldots, n\} : x_j = x_k\}.$$

Für festes $x \in R_1$ ist $\hat{F}_n(x; X^{(n)})$ die (zufällige) relative Häufigkeit, mit der die $\{X_k \leq x\}, k = 1, \ldots, n$ eintreten. Es gilt

$$E\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} P(X_i \leq x) = F(x)$$

$$D^2(\hat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n}.$$

Aus dem starken Gesetz der großen Zahlen folgt für jedes $x \in R_1$

$$\lim_{n\to\infty} \hat{F}_n(x) = F(x) \qquad P - f.s.$$

Darüber hinaus gilt der

Satz 12.3 (Hauptsatz der mathematischen Statistik)
Es seien F eine Verteilungsfunktion auf R_1 und $X^{(n)} = (X_1, X_2, \ldots, X_n)$ eine mathematische Stichprobe aus einer nach F verteilten Grundgesamtheit. $X^{(n)}$ sei definiert auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$.

Für die Zufallsgrößen $D_n, n \geq 1$, definiert durch

$$D_n := \sup_{x\in R_1} |\hat{F}_n(x) - F(x)|, \tag{12.3}$$

gilt

$$\lim_{n\to\infty} D_n = 0 \qquad\qquad P - f.s. \tag{12.4}$$

Beweis: Es seien N und j natürliche Zahlen mit $0 \leq j \leq N$,

$$x_{j,N} := \inf\{x : F(x) \geq \frac{j}{N}\}, x_{0,N} := -\infty,\ \inf\emptyset := \infty.$$

Ist $y < x_{j,N}$, so folgt $F(y) < \frac{j}{N}$,

und es gilt (wegen der Rechtsstetigkeit von F)

$$F(x_{j,N} - 0) \leq \frac{j}{N} \leq F(x_{j,N}).$$

Daraus ergibt sich für $0 \leq j < N$.

$$F(x_{j+1,N} - 0) \leq \frac{j+1}{N} \leq F(x_{j,N}) + \frac{1}{N} \tag{12.5}$$

Ist nun $x \in [x_{j,N}, x_{j+1,N})$, so erhalten wir wegen (12.5) und

$$F(x) \leq F(x_{j+1,N} - 0) \text{ die Ungleichung}$$

$$\hat{F}_n(x_{j,N}) - F(x_{j,N}) - \frac{1}{N} \leq \hat{F}_n(x) - F(x) \leq \hat{F}_n(x_{j+1,N} - 0) - F(x_{j,N}) \leq$$

$$\hat{F}_n(x_{j+1,N} - 0) - F(x_{j+1,N} - 0) + \frac{1}{N} \quad P - f.s.$$

Daraus ergibt sich für alle $x \in \bigcup_{i=0}^{N-1} [x_{j,N}, x_{j+1,N}) = [-\infty, x_{N,N})$ und alle x mit $x \geq x_{N,N}$

$$|\hat{F}_n(x) - F(x)| \leq \max_{0 \leq j < N} \{|\hat{F}_n(x_{j,N}) - F(x_{j,N})|, |\hat{F}_n(x_{j+1,N} - 0) - F(x_{j+1,N} - 0)|\}$$

$$+\frac{1}{N} \quad P - f.s.$$

Aus dem starken Gesetz der großen Zahlen folgen für jedes j mit $0 \leq j < N$ die Gleichungen $\lim_{n\to\infty} \hat{F}_n(x_{j,N}) = F(x_{j,N}) \quad$ und $\lim_{n\to\infty} \hat{F}_n(x_{j+1,N} - 0)$ $= F(x_{j+1,N} - 0) \quad P$-fast sicher.

Deshalb gilt:

$$D_n = \sup_{x \in R_1} |\hat{F}_n(x) - F(x)| \underset{n\to\infty}{\longrightarrow} 0 \quad P - \text{fast sicher.}$$

□

Der Hauptsatz der mathematischen Statistik ist von grundlegender Bedeutung für die praktische Anwendung der Wahrscheinlichkeitstheorie. Er besagt, dass man eine unbekannte Verteilungsfunktion F grundsätzlich beliebig genau bestimmen kann, wenn man sich eine hinreichend große konkrete Stichprobe $x^{(n)} = (x_1, \ldots, x_n)$ aus einer nach F verteilten Grundgesamtheit verschafft und

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{(-\infty,x]}(x_k), \quad x \in R_1$$

als Näherung für $F(\cdot)$ verwendet.

Als eine Verfeinerung des Hauptsatzes im Falle, dass F stetig ist, geben wir noch folgende Aussage an.

Aussage 12.4 *Ist F stetig, so gilt*

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq x) = K(x), \quad x \in R_1$$

mit

$$K(x) := \begin{cases} 0 & x \leq 0 \\ \sum\limits_{k=-\infty}^{\infty} (-1)^k \ \mathrm{e}^{-2k^2x^2}, & x > 0. \end{cases}$$

($K(\cdot)$ ist die Verteilungsfunktion der sogenannten Kolmogorov-Smirnov-Verteilung.)

Zum Beweis sei auf Winkler (1983) verwiesen. Für große n und für alle $y > 0$ kann man also $P(D_n \leq y)$ annähernd durch $K(\sqrt{n}y)$ ersetzen:

$$P(D_n \leq y) \approx K(\sqrt{n}y).$$

Wir haben gesehen, dass man prinzipiell auf der Grundlage von Stichproben die Verteilungsfunktion F_X einer Zufallsgröße X beliebig genau bestimmen kann. In praktischen Fällen wird dieses Verfahren jedoch selten angewandt. Vielfach hat man nämlich Vorabinformationen über F_X in dem Sinne, dass man weiß, dass F_X zu einer gewissen Klasse von Verteilungsfunktionen gehört. Zum Beispiel könnte aus inhaltlichen Gründen unter Verwendung eines zentralen Grenzwertsatzes geschlossen werden, dass F_X die Verteilungsfunktion einer Normalverteilung ist. Dann wären nur noch die Parameter μ und σ^2 zu bestimmen. Oder bei der Anzahl der Schäden, die ein Versicherungsnehmer pro Jahr verursacht, scheint in erster Näherung eine Poissonverteilung geeignet zu sein (Begründung?). Dann wäre nur noch ihr Parameter λ unbekannt. In vielen Fällen interessiert man sich auch nur für gewisse Kenngrößen der Verteilung, zum Beispiel für den Erwartungswert und/oder für die Streuung.

Die Konstruktion und Beurteilung von Verfahren zur näherungsweisen Bestimmung von unbekannten Parametern auf der Grundlage von Stichproben ist

Aufgabe der sogenannten statistischen Schätztheorie, aus der wir im folgenden Abschnitt einige grundlegende Begriffe und Aussagen kennen lernen.

12.2 Statistische Schätzungen

12.2.1 Definitionen

Definition 12.5 *Es sei* $\mathfrak{P} = (P_\vartheta, \vartheta \in \Theta), \quad \Theta \subseteq R_k, \quad k \geq 1$*, eine Familie von Wahrscheinlichkeitsmaßen auf* $(\Omega, \mathfrak{A})$*. Dann heißt* $(\Omega, \mathfrak{A}, \mathfrak{P})$ *ein statistisches Modell.*

Für Θ wählt man irgendeine nichtleere Menge, meist eine offene oder abgeschlossene Menge. Angenommen, X ist eine reellwertige Zufallsgröße über $(\Omega, \mathfrak{A})$ und P_ϑ^X die zu X gehörende Verteilung unter P_ϑ:

$$P_\vartheta^X(B) := P_\vartheta(X \in B), \quad B \in \mathfrak{B}_1, \quad \vartheta \in \Theta.$$

Offenbar ist dann $(R_1, \mathfrak{B}_1, \mathfrak{P}^X)$ mit $\mathfrak{P}^X = (P_\vartheta^X, \vartheta \in \Theta)$ ebenfalls ein statistisches Modell. Den Erwartungswert von X oder irgendeiner anderen Zufallsgröße Y bezüglich der Verteilung P_ϑ bezeichnen wir mit $E_\vartheta X$ bzw. $E_\vartheta Y$.

Anschaulicher Hintergrund: Wir nehmen an, dass die Verteilung von X zu $\mathfrak{P}^X$ gehört, kennen aber den wahren Wert ϑ_0 des Parameters ϑ nicht.

Es sei $X^{(n)} = (X_1, X_2, \ldots, X_n)$ eine mathematische Stichprobe aus einer nach $P_\vartheta, \vartheta \in \Theta$, verteilten Grundgesamtheit.

Aufgabe: Man konstruiere auf der Grundlage einer Stichprobe eine Schätzung für den wahren Wert ϑ_0.

Häufig ist man gar nicht an ϑ selbst, sondern an einer gewissen Funktion von ϑ interessiert, zum Beispiel am Erwartungswert $\frac{1}{\lambda}$ einer $Exp(\lambda)$-verteilten Zufallsgröße.
Wir formulieren den Begriff der Schätzung deshalb zunächst einmal sehr allgemein. Auf Gütekriterien für Schätzungen gehen wir anschließend ein.

Definition 12.6 *Es seien* g *und* $\hat{G}_n$ *Borelmessbare Funktionen von* Θ *bzw. von* R_n *in* R_m*. Überdies sei* $\vartheta \in \Theta$*. Dann heißt* $\hat{G}_n(X_1, \ldots, X_n)$ *eine* Schätzung für

$g(\vartheta)$.
Durch Einsatz einer konkreten Stichprobe $x^{(n)} = (x_1, \ldots, x_n)$ in $\hat{G}_n$ erhält man einen Schätzwert $\hat{G}_n(x_1, x_2, \ldots, x_n)$ für *$g(\vartheta)$.*

Beispiel 12.7 Es sei X eine Zufallsgröße mit den möglichen Werten $1, 2, \ldots, N$ mit

$$P_N(X = k) = \frac{1}{N}, k = 1, 2, \ldots, N.$$

Der Parameter N sei unbekannt. Als Schätzung für N auf der Grundlage von $X^{(n)} = (X_1, \ldots, X_n)$ hat man zum Beispiel

$$\hat{N}_n = \max_{k=1,2,\ldots,n} X_k \quad \text{und} \quad \tilde{N}_n = \left[\frac{2}{n} \sum_{k=1}^{n} X_k \right].$$

12.2.2 Güteeigenschaften von Schätzungen

Wir verwenden die Terminologie des vorangegangenen Abschnittes.

Im Allgemeinen gibt es viele Schätzungen $\hat{G}_n(X_1, \ldots, X_n)$ für $g(\vartheta)$. Bei der Frage, welche Kriterien man bei der Auswahl anlegen sollte, bietet sich zuallererst die Eigenschaft der *Erwartungstreue* an.

Definition 12.8 *Die Schätzung $\hat{G}_n(X_1, \ldots, X_n)$ für $g(\vartheta)$ heißt* erwartungstreu*, falls gilt*

$$E_\vartheta \hat{G}_n(X_1, \ldots, X_n) = g(\vartheta) \text{ für alle } \vartheta \in \Theta.$$

Ist $\hat{G}_n(X^{(n)})$ irgendeine Schätzung für $g(\vartheta), \vartheta \in \Theta$, so nennt man die Funktion

$$E_\vartheta \hat{G}_n(X^{(n)}) - g(\vartheta), \quad \vartheta \in \Theta$$

die Verzerrung der Schätzung, ihren systematischen Fehler oder ihren Bias.
Eine erwartungstreue Schätzung heißt auch unverzerrt oder unbiased.

Erwartungstreue Schätzungen haben die Eigenschaft, dass sich ihre Werte bei häufiger (unabhängiger) Wiederholung der Schätzung um den Erwartungswert, also $g(\vartheta)$, gruppieren (Gesetz der großen Zahlen). Man kann also ein gewisses Vertrauen haben, dass die entsprechenden Schätzwerte in der Nähe des zu schätzenden Wertes $g(\vartheta)$ liegen.

Beispiele 12.9

1. Der Erwartungswert $\mu(\vartheta) := E_\vartheta X_1$ sei unbekannt. Dann ist für jeden Vektor $a = (a_1, \ldots, a_n)$ mit $a_k \geq 0, k = 1, \ldots, n,$ und $\displaystyle\sum_{k=1}^{n} a_k = 1$

   die Schätzung
   $$\hat{\mu}_{(a)}(X^{(n)}) := \sum_{k=1}^{n} a_k X_k$$
   eine erwartungstreue Schätzung für $\mu(\vartheta)$.

   Spezialfälle sind $\hat{\mu}_n := \frac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ und $\hat{\mu}_1 := X_1$.

2. $$\hat{\sigma}_n^2(X^{(n)}) = \frac{1}{n}\sum_{k=1}^{n}(X_k - \hat{\mu}_n)^2$$
   ist keine erwartungstreue Schätzung für $\sigma^2(\vartheta) = D_\vartheta^2 X_1$.
   Es gilt nämlich $E_\vartheta\, \hat{\sigma}_n^2(X^{(n)}) = \frac{n-1}{n}\sigma^2(\vartheta)$.

   Ihr Bias ist
   $$E_\vartheta\, \hat{\sigma}_n^2(X^{(n)}) - \sigma^2(\vartheta) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$
   Die Streuung σ^2 wird also bei häufiger Schätzung durch $\hat{\sigma}_n^2$ systematisch unterschätzt. Dagegen ist

$$\tilde{\sigma}_n^2(X^{(n)}) = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \hat{\mu}_n)^2$$

eine erwartungstreue Schätzung für σ^2.

Wie wir am Beispiel 12.9(1) gesehen haben, gibt es mitunter mehrere erwartungstreue Schätzungen für $g(\vartheta)$. Um unter ihnen eine Auswahl zu treffen, führen wir ein weiteres Gütekriterium ein.

Definition 12.10 *Sind $\hat{G}(X^{(n)})$ und $G^*(X^{(n)})$ zwei erwartungstreue Schätzungen für $g(\vartheta), \vartheta \in \Theta$, so heißt $\hat{G}(X^{(n)})$* besser *als $G^*(X^{(n)})$, falls*

$$D_\vartheta^2 \hat{G}(X^{(n)}) \leq D_\vartheta^2 G^*(X^{(n)}) \ \text{für alle} \ \vartheta \in \Theta \tag{12.6}$$

gilt. $\hat{G}(X_n)$ heißt beste erwartungstreue Schätzung *für $g(\vartheta), \vartheta \in \Theta$, oder* erwartungstreue Schätzung mit minimaler Streuung*, falls (6) für jede erwartungstreue Schätzung $G^*(X^{(n)})$ für $g(\vartheta), \vartheta \in \Theta$, gilt.*

Beispiel 12.11 *(Fortsetzung des Beispiels 12.9(1)):*

Es gilt $D_\vartheta^2(\hat{\mu}_{(a)}(X^{(n)}) = \sigma^2(\vartheta)\sum_{k=1}^{n} a_k^2$, und dieser Ausdruck wird minimal (unter der Nebenbedingung $a_k \geq 0, \sum a_k = 1$) für $a_k \equiv \frac{1}{n}$. Das arithmetische Mittel $\hat{\mu}_n(X^{(n)})$ ist also unter allen gewichteten Mitteln $\hat{\mu}_{(a)}(X^{(n)})$ die beste erwartungstreue Schätzung für $\mu(\vartheta)$.

Die Definition bester erwartungstreuer Schätzungen wirft die Frage auf nach der Existenz solcher Schätzungen und gegebenenfalls nach der Größe ihrer Streuung.
Ein Ergebnis in dieser Richtung ist die *sogenannte Ungleichung von Cramer-Rao.* Bevor wir auf sie eingehen, stellen wir noch einige Begriffe bereit.

Die Likelihoodfunktion

Es sei $X^{(n)} = (X_1, X_2, \ldots, X_n)$ eine mathematische Stichprobe aus einer nach $P_\vartheta^X, \vartheta \in \Theta$, verteilten Grundgesamtheit, wobei X eine reellwertige Zufallsgröße ist. Die Verteilung $P_\vartheta^{X(n)}$ ist also für jedes $\vartheta \in \Theta$ eine Verteilung auf $(R_n, \mathfrak{B}_n)$ mit

$$P_\vartheta^{X(n)}(B_1 \times \ldots \times B_n) = \prod_{k=1}^{n} P_\vartheta^X(B_k),\ B_1, \ldots, B_n \in \mathfrak{B}_1. \tag{12.7}$$

Um die sogenannte Likelihoodfunktion definieren zu können, unterscheiden wir zwei Fälle.

1. Fall: X besitzt für alle $\vartheta \in \Theta$ eine Dichte $f_\vartheta(\cdot)$ bezüglich des Lebesguemaßes.

In diesem Fall setzen wir

$$L^X(\vartheta, x) := f_\vartheta(x), \quad \vartheta \in \Theta, x \in R_1.$$

Es gilt nach Definition der Dichte

$$P_\vartheta^X(B) = \int_B L^X(\vartheta, x)dx, \quad \vartheta \in \Theta, B \in \mathfrak{B}_1.$$

2. Fall: X sei diskret verteilt unter P_ϑ mit den möglichen Werten $a_k, k \in N_0$, die nicht von ϑ abhängen. In diesem Fall sei

$$L^X(\vartheta, a_k) := P_\vartheta(X = a_k),\ k \in N_0.$$

Es gilt dann

$$P_\vartheta^X(B) = \sum_{a_k \in B} L^X(\vartheta, a_k).$$

Offenbar gilt in beiden Fällen

$$L^X(\vartheta, \cdot) \geq 0 \text{ und}$$

$$P_\vartheta^X(\{x : L^X(\vartheta, x) = 0\}) = 0. \tag{12.8}$$

Ist H im ersten Fall eine messbare nichtnegative Funktion auf R_n, so gilt

$$E_\vartheta H(X) = \int\limits_{R_1} H(x) P_\vartheta^X(dx) = \int\limits_{R_1} H(x) L^X(\vartheta, x) dx, \tag{12.9}$$

und ist H im zweiten Fall eine nichtnegative Funktion auf A, so haben wir

$$E_\vartheta H(X) = \sum_{a_k \in A} H(a_k) p_\vartheta(a_k) = \sum_{a_k \in A} H(a_k) L^X(\vartheta, a_k). \tag{12.10}$$

Definition 12.12 *Wir setzen voraus, es liegt der 1. oder 2. der eben eingeführten Fälle vor.*

Für jedes $x^{(n)} = (x_1, x_2, \ldots, x_n) \in R_n$ heisst die Funktion

$$\vartheta \to L_n(\vartheta; x^{(n)}) = \prod_{k=1}^{n} L^X(\vartheta, x_k),\ \vartheta \in \Theta,$$

Likelihoodfunktion des statistischen Modells

$\mathfrak{P}^X = (P_\vartheta^X, \vartheta \in \Theta)$ *(bei gegebener konkreter Stichprobe $x^{(n)}$).*

Bemerkung 12.13 Mit Hilfe der Likelihoodfunktion kann man die gemeinsame Verteilung von $X^{(n)} = (X_1, X_2, \ldots, X_n)$ ausdrücken (beachte die Schreibweise $x^{(n)} = (x_1, x_2, \ldots, x_n)$):

$$P_\vartheta^{X^{(n)}}(B_1, \times \ldots \times B_n) = \int\limits_{B_1} \cdots \int\limits_{B_n} L_n(\vartheta, x^{(n)})\, dx_1 \ldots dx_n$$

im ersten Fall und

$$P_\vartheta^{X^{(n)}}(B_1, \times \ldots \times B_n) = \sum_{x_1 \in A} \cdots \sum_{x_n \in A} L_n(\vartheta, x^{(n)}).$$

im zweiten Fall.

Offenbar gilt im ersten Fall für alle nichtnegativen messbaren H

$$E_\vartheta H(X^{(n)}) = \int \cdots \int_{R_n} H(x^{(n)}) L_n(\vartheta, x^{(n)}) dx_1 \ldots, dx_n$$

und im zweiten Fall für alle nichtnegativen Funktionen H

$$E_\vartheta H(X^{(n)}) = \sum_{x^{(n)} \in A^n} H(x^{(n)}) L_n(\vartheta, x^{(n)}).$$

Beispiele 12.14

a) Es sei $X \sim N(\mu, \sigma^2)$. Dann gilt mit $\vartheta = (\mu, \sigma^2)^T \in R_1 \times R_+ =: \Theta$

$$L_n(\vartheta; x^{(n)}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{k=1}^n (x_k - \mu)^2\Big] =$$

$$(\sigma^2)^{-\frac{n}{2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{k=1}^n x_k^2 + \frac{\mu}{\sigma^2}\sum_{k=1}^n x_k - \frac{n\mu^2}{2\sigma^2}\Big](2\pi)^{-\frac{n}{2}}, x^{(n)} \in R_n.$$

b) Es sei $X \sim$ Bin$\,(m, p)$. Dann ist mit $\vartheta = p \in (0,1) = \Theta$

$$L_n(\vartheta; x^{(n)}) = \prod_{k=1}^n \binom{m}{i_k} p^{i_k}(1-p)^{m-i_k}$$

$$x^{(n)} = (i_1, i_2, \ldots, i_n),\ 0 \le i_k \le m, k = 1, \ldots, n.$$

Aussage 12.15 (Cramer-Rao-Ungleichung) *Es sei vorausgesetzt:*

a) Die Likelihoodfunktion $\vartheta \to L_n(\vartheta, x^{(n)})$ ist für jedes $x^{(n)}$ differenzierbar bezüglich ϑ, grad $lnL_n(\vartheta, X^{(n)})$ ist ein zentrierter zufälliger Vektor und alle seine zweiten Momente bez. P_ϑ sind endlich ($\vartheta \in \Theta \subseteq R_k$).

$$\big(grad\ =\ grad_\vartheta = (\frac{\partial}{\partial\vartheta_1}, \frac{\partial}{\partial\vartheta_2}, \ldots, \frac{\partial}{\partial\vartheta_k})^T\big)$$

b) *Für jede reellwertige Borelmessbare Funktion h mit $E_\vartheta|h(X^{(n)})|^2 < \infty$ gilt im ersten Fall*

$$grad \int_{R_n} L_n(\vartheta; x^{(n)}) h(x^{(n)}) dx^{(n)} = \int_{R_n} grad\, L_n(\vartheta, x^{(n)}) h(x^{(n)}) dx^{(n)}$$

und im zweiten Fall

$$grad \sum_{x^{(n)} \in A^n} L_n(\vartheta, x^{(n)}) h(x^{(n)}) p_\vartheta(x^{(n)}) = \sum_{x^{(n)} \in A^n} grad L_n(\vartheta, x^{(n)}) h(x^{(n)}).$$

c) *Die Matrix $I_n(\vartheta)$, definiert durch*

$$(I_n(\vartheta))_{1 \le i,j \le k} = E_\vartheta \Big(\frac{\partial}{\partial \vartheta_i} ln L_n(\vartheta, X^{(n)}) \cdot \frac{\partial}{\partial \vartheta_j} ln L_n(\vartheta, X^{(n)}) \Big)_{1 \le i,j \le k}$$

ist invertierbar für jedes $\vartheta \in \Theta$.

(Es gilt $I_n(\vartheta) = E_\vartheta(grad\, ln\, L_n(\vartheta, X^{(n)}) grad^T ln\, L_n(\vartheta; X^{(n)}))$.)

Dann gilt für jede reellwertige Zufallsgröße Y der Form $Y = h(X_1, X_2, \ldots, X_n)$

$$E_\vartheta(Y - E_\vartheta Y)^2 \ge (grad\, E_\vartheta Y)^T [I_n(\vartheta)]^{-1} (grad\, E_\vartheta Y).$$

Ist insbesondere Y eine erwartungstreue Schätzung für $g(\vartheta)$, so gilt

$$E_\vartheta(Y - E_\vartheta Y)^2 \ge (grad\, g(\vartheta))^T [I_n(\vartheta)]^{-1} (grad\, g(\vartheta))$$

und für $k = 1$ erhalten wir:

$$D_\vartheta^2 Y \ge \frac{[g'(\vartheta)]^2}{I_n(\vartheta)}.$$

Definition 12.16 *Die Matrix $I_n(\vartheta)$ heißt Fisher'sche Informationsmatrix.*

Sie ist nichtnegativ definit, da sie die Kovarianzmatrix des Vektors $grad\ lnL_n(\vartheta, X^{(n)})$ ist.

Die Matrix $I_n(\vartheta)$ lässt sich durch $I_1(\vartheta)$ ausdrücken.
Es gilt nämlich wegen

$$lnL_n(\vartheta, X^{(n)}) = \sum_{k,l=1}^{n} lnL^X(\vartheta, X_k)$$

die Beziehung

$$(I_n(\vartheta))_{ij} = E_\vartheta\Big(\sum_{k,l=1}^{n} \frac{\partial}{\partial\vartheta_i} lnL^X(\vartheta, X_k) \cdot \frac{\partial}{\partial\vartheta_j} lnL^X(\vartheta, X_l)\Big) =$$

$$= E_\vartheta\Big(\sum_{k=1}^{n} \frac{\partial}{\partial\vartheta_j} lnL^X(\vartheta, X_k) \frac{\partial}{\partial\vartheta_j} lnL^X(\vartheta, X_k)\Big)$$

$$= n(I_1(\vartheta))_{ij}.$$

Beweis der Aussage 12.15: (Anstelle L_n schreiben wir hier auch kurz L.) Wir beschränken uns auf den ersten Fall. Der zweite wird völlig analog bewiesen. Aus der Voraussetzung b) folgt für $h \equiv 1$, dass

$$\int_{R_n} grad\ L(\vartheta, x^{(n)}) dx^{(n)} = 0$$

und damit haben wir

$$E_\vartheta\Big[grad\ lnL(\vartheta, X^{(n)})\Big] = E_\vartheta\Big[\frac{grad\ L(\vartheta, X^{(n)})}{L(\vartheta, X^{(n)})}\Big] = 0.$$

Weiterhin folgt damit aus b), falls $\int_{R_n} h^2(x^{(n)}) dx^{(n)} < \infty$ gilt,

$$grad\ E_\vartheta h(X^{(n)}) = grad \int_{R_n} L(\vartheta, x^{(n)}) h(x^{(n)}) dx^{(n)} =$$

$$E_\vartheta\Big[grad\ lnL(\vartheta, X^{(n)}) \cdot h(X^{(n)})\Big] =$$

$$E_\vartheta\Big[grad\ lnL(\vartheta, X^{(n)})(h(X^{(n)}) - E_\vartheta h(X^{(n)}))\Big].$$

Es sei nun $u \in R^k \setminus \{0\}$. Dann gilt ($< \cdot, \cdot >$ bezeichnet das Skalarprodukt):

$$< u, grad\ E_\vartheta h(X^{(n)}) >=$$

$$E_\vartheta[< u, grad\ lnL(\vartheta, X^{(n)}) > (h(X^{(n)}) - E_\vartheta h(X^{(n)}))].$$

Mittels der Schwarz'schen Ungleichung ergibt sich

$$E_\vartheta(h(X^{(n)}) - E_\vartheta h(X^{(n)}))^2 \geq$$

$$\frac{< u, grad\ E_\vartheta h(X^{(n)}) >^2}{E_\vartheta[< u, grad\ lnL(\vartheta) >^2]}$$

für alle $u \in R_k \setminus \{0\}$. Wir bestimmen den maximalen Wert der rechten Seite dieser Ungleichung für $u \in R_k \setminus \{0\}$.

Es sei $\vartheta \in \Theta$ fest gewählt und u so normiert, dass gilt

$$< u, grad\ E_\vartheta h(X^{(n)}) >= 1.$$

Man beachte in der Schreibweise des Skalarproduktes $< u, v >= u^T v$):

$$< u,\ grad\ lnL(\vartheta) >^2= (u^T\ grad\ lnL(\vartheta))^2 =$$

$$(u^T\ grad\ lnL(\vartheta)((grad\, L(\vartheta))^T u).$$

Somit gilt

$$E_\vartheta < u,\ grad\ lnL(\vartheta) >^2= u^T I_n(\vartheta) u.$$

Wir definieren:

$$\nu =\ grad\ E_\vartheta h(X^{(n)})$$

und haben folglich die quadratische Form

$$u^T I_n(\vartheta) u$$

unter der Nebenbedingung $< u, \nu >= 1$ zu minimieren.

Mittels der Methode des Lagrange'schen Multiplikators folgt als notwendige Bedingung

$$2I_n(\vartheta)u = \lambda\nu.$$

Nach Voraussetzung c) ergeben sich $< u, \nu >= 1$ und $u = \frac{\lambda}{2} I_n^{-1}(\vartheta)\nu$ als notwendige Bedingungen. Daraus folgt

$$1 =< u, \nu >= \frac{\lambda}{2}\nu^T I_n^{-1}(\vartheta)\nu \text{ und}$$

$$u^T I_n(\vartheta)u = \frac{\lambda^2}{4}\nu^T I_n^{-1}(\vartheta)I(\vartheta)I_n^{-1}(\vartheta)\nu =$$

$$\frac{\lambda^2}{4}\nu^T I_n^{-1}(\vartheta)\nu = \frac{1}{\nu^T I_n^{-1}(\vartheta)\nu}.$$

Somit ergibt sich für diese Wahl von u

$$E_\vartheta(h(X^{(n)}) - E_\vartheta h(X^{(n)})^2 \geq (grad\ E_\vartheta h(X^{(n)}))^T I_n^{-1}(\vartheta)(grad\ E_\vartheta h(X^{(n)}))).$$

□

Definition 12.17 *Jede erwartungstreue Schätzung $\hat{G}_n(X_n^{(n)})$ für $g(\vartheta)$, für die $D^2_{(\vartheta)}\hat{G}_n(X_n^{(n)})$ gleich der unteren Schranke in der Cramer-Rao-Ungleichung ist, heißt eine effiziente Schätzung für $g(\vartheta), \vartheta \in \Theta$.*

Effiziente Schätzungen sind offenbar beste erwartungstreue Schätzungen. Die Umkehrung gilt im Allgemeinen nicht.

Beispiel 12.18 (Effiziente Schätzung) Ist X eine Zufallsgröße mit $P_p(X = 1) = p, P_p(X = 0) = 1 - p = q, p \in (0, 1)$ unbekannt, und ist $X^{(n)}$ eine mathematische Stichprobe aus einer wie X verteilten Grundgesamtheit, so gilt

$$L_n(\vartheta; x^{(n)}) = p^{\sum x_l} q^{n-\sum x_l} \quad , x^{(n)} = (x_1, \ldots, x_n) \in \{0, 1\}^n,$$

und folglich ist

$$lnL_n(\vartheta; x^{(n)}) = \sum x_l lnp + (n - \sum x_l) ln(1-p) =$$

$$s_n lnp + (n - s_n) ln(1-p), \text{ mit } s_n = \sum_{l=1}^{n} x_l.$$

Daraus folgt

$$E_p\Big(\frac{d}{dp} lnL_n(X^{(n)})\Big)^2 = \frac{n}{p(1-p)} = I_n(\vartheta),\ I_1(\vartheta) = [p(1-p)]^{-1}.$$

Setzen wir $g(p) = p$, so erhalten wir mit $S_n = \sum\limits_{l=1}^{n} X_l$ für die erwartungstreue Schätzung $\hat{p}_n(X^{(n)}) := \frac{S_n}{n}$ für den Parameter p die Streuung:

$$D_p^2(\frac{Sn}{n}) = \frac{p(1-p)}{n} = I_n^{-1}(p).$$

Also ist $\frac{S_n}{n}$ eine effiziente Schätzung für p.
Die gleichfalls erwartungstreue Schätzung $\hat{G}_n(X^{(n)}) = X_1$ für p zum Beispiel hat dagegen eine wesentlich größere Streuung, nämlich $p(1-p)$.

12.2.3 Konstruktion von Schätzungen

Wir haben bisher Eigenschaften von Schätzungen angegeben und einige plausible Schätzungen kennen gelernt. Im Folgenden gehen wir auf zwei Methoden ein, Schätzungen zu konstruieren, die *Momentenmethode* und die *Maximum-Likelihood-Methode*. Keine dieser Methoden liefert universell beste Lösungen. Die mit ihrer Hilfe konstruierten Schätzungen müssen individuell auf ihre Eigenschaften untersucht werden. Einige allgemeine Aussagen lassen sich jedoch treffen.

1. Momentenmethode

Es sei $(\Omega, \mathfrak{A}, (P_\vartheta, \vartheta \in \Theta))$ ein statistisches Modell und X eine reellwertige Zufallsgröße über $(\Omega, \mathfrak{A})$. Für ein $k \geq 1$ gelte $E_\vartheta |X|^k < \infty, \vartheta \in \Theta$.
Wir setzen $\mu_l(\vartheta) := E_\vartheta X^l, \quad 1 \leq l \leq k, \vartheta \in \Theta$.

Dann ist, falls $X^{(n)} = (X_1, \dots, X_n)$ eine mathematische Stichprobe aus einer nach P_ϑ^X verteilten Grundgesamtheit bildet, $\hat{\mu}_l(X^{(n)}) := \frac{1}{n} \sum_{k=1}^{n} X_k^l$ eine Schätzung für $\mu_l(\vartheta)$. Das Prinzip besteht also darin, zur Schätzung des l-ten Momentes $\mu_l(\vartheta)$ der Zufallsgröße X bez. der Verteilung P_ϑ das l-te Moment der empirischen Verteilungsfunktion der mathematischen Stichprobe $X^{(n)}$ zu verwenden.

Diese Methode lässt sich auch zur Konstruktion von Schätzungen für Größen der Form

$$g(\mu_1(\vartheta), \dots, \mu_m(\vartheta))$$

ausnutzen, wobei g irgendeine stetige Funktion auf R_k ist. Man wählt in diesem Fall

$$\hat{G}_n(X^{(n)}) := g(\hat{\mu}_1(X^{(n)}) \dots, \hat{\mu}_m(X^{(n)}))$$

als Schätzung für $g(\mu_1(\vartheta), \dots, \mu_m(\vartheta))$. Dieses Vorgehen zur Konstruktion von Schätzungen bezeichnet man als *Momentenmethode*.
Diese Methode der Gewinnung von Schätzungen bezieht ihre Rechtfertigung aus der Gültigkeit des starken Gesetzes der großen Zahlen. Es gilt nämlich $P_\vartheta - f.s.$

$$\lim_{n\to\infty} \hat{\mu}_l(X^{(n)}) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k^l = E_\vartheta X^l = \mu_l(\vartheta), \vartheta \in \Theta \tag{12.11}$$

und

$$\lim_{n\to\infty} g(\hat{\mu}_1(X^{(n)}), \dots, \hat{\mu}_m(X^{(n)})) = g(\mu_1(\vartheta), \dots, \mu_m(\vartheta)), \vartheta \in \Theta. \tag{12.11'}$$

Man geht also bei großem Stichprobenumfang davon aus, dass $\hat{\mu}_l(X^{(n)})$ in der Nähe von $\mu_l(\vartheta)$ liegt, wobei ϑ der wahre Parameter ist.
Die Eigenschaft (12.11) bzw. (12.11') wird auch (*starke*) *Konsistenz der Schätzungen* $\hat{\mu}_l(X^{(n)}), n \geq 1$, bzw. $\hat{G}_n(X^{(n)}), n \geq 1$, genannt.

2. Maximum-Likelihood-Methode

Es sei $(\Omega, \mathfrak{A}, (P_\vartheta, \vartheta \in \Theta \subseteq R_k))$ ein statistisches Modell und X eine reellwertige Zufallsgröße über $(\Omega, \mathfrak{A})$. Mit F_ϑ werde die Verteilungsfunktion von X bez. P_ϑ bezeichnet, $\vartheta \in \Theta$. Weiterhin sei $X^{(n)} = (X_1, X_2, \ldots, X_n)$ eine mathematische Stichprobe aus einer nach $F_\vartheta, \vartheta \in \Theta$, verteilten Grundgesamtheit und $x^{(n)} = (x_1, \ldots, x_n)$ eine Realisierung von $X^{(n)}$ (konkrete Stichprobe). Wir nehmen der Einfachheit halber an, dass F_ϑ für jedes $\vartheta \in \Theta$ eine Dichte f_ϑ besitzt (Fall 1) oder für jedes $\vartheta \in \Theta$ eine diskrete Verteilung mit den Einzelwahrscheinlichkeiten $p_\vartheta(a_j) := P_\vartheta(X_1 = a_j), j \in \mathfrak{J} \subseteq N$ (Fall 2) darstellt. Die Menge $A = \{a_j | j \in \mathfrak{J}\}$ bildet im zweiten Fall die Menge der möglichen Werte von X.

Bei festem $x^{(n)}$ ist durch die Funktionen

$$\vartheta \longrightarrow L_n(\vartheta; x^{(n)}) = \prod_{k=1}^{n} f_\vartheta(x_k) \quad , \quad \vartheta \in \Theta \quad \text{(1. Fall)} \qquad :$$

bzw.

$$\vartheta \longrightarrow L_n(\vartheta; x^{(n)}) = \prod_{k=1}^{n} p_\vartheta(x_k), \vartheta \in \Theta \quad \text{(2. Fall)}$$

die *Likelihoodfunktion* $L_n(\vartheta, x^{(n)})$ der Familie $(P_\vartheta^X, \vartheta \in \Theta)$ gegeben.

Definition 12.19 *Als Maximum-Likelihood-Schätzwert bezeichnet man jeden Wert* $\hat{\vartheta}_n(x^{(n)})$ *mit*

$$L_n(x^{(n)}; \hat{\vartheta}_n(x^{(n)})) = \max_{\vartheta \in \Theta} L_n(x^{(n)}; \vartheta).$$

Man wählt den Parameter $\vartheta \in \Theta$ also so, dass die beobachtete Stichprobe $x^{(n)}$ im Fall 1. Ort der maximalen Dichte von $X^{(n)}$ bzw. im Fall 2. der Parameter ist, für den $X^{(n)}$ die maximale Wahrscheinlichkeit besitzt. Setzt man die mathematische Stichprobe $X^{(n)}$ anstelle $x^{(n)}$ ein, so erhält man eine *Maximum-Likelihood-Schätzung* $\hat{\vartheta}_n(X^{(n)})$. Dabei handelt es sich um eine Zufallsgröße mit Werten in Θ, deren Wert von der Stichprobe $X^{(n)}$ abhängt.

Das Prinzip der Maximum-Likelihood-Methode ist ein sehr allgemeines. Man könnte es so formulieren:

Kann eine Erscheinung mehrere Ursachen haben, so nimmt man diejenigen als die wahre Ursache an, für die die Wahrscheinlichkeit dafür, dass sie die Erscheinung nach sich zieht, am größten ist.

R.A. Fisher: "Finde diejenigen Voraussetzungen, die das Beobachtete mit großer Wahrscheinlichkeit nach sich ziehen und fasse Zutrauen, dass diese Voraussetzungen die wirksamen sind."

Anstelle L_n kann man auch $ln\ L_n$ bez. ϑ maximieren. Das führt häufig zu rechnerischen Vorteilen, da $ln\ L_n$ eine Summe, L_n dagegen ein Produkt von Funktionen von ϑ ist.

In vielen Fällen ist die Likelihoodfunktion stetig differenzierbar bzw. ϑ und das Maximum bez. ϑ liegt nicht auf dem Rand von Θ. Dann sind die Gleichungen

$$\frac{\partial}{\partial \vartheta_m} L_n(x^{(n)}; \vartheta) = 0, \quad m = 1, 2, \ldots, k \tag{12.12}$$

notwendige Bedingung für $\vartheta = \hat{\vartheta}_n(x^{(n)})$ und liefern häufig bereits eine Lösung $\hat{\vartheta}_n(x^{(n)})$.
(Maximum-Likelihood-Gleichungen)

Äquivalent zu (12.12) sind folgende häufig besser zu behandelnde Gleichungen, die man ebenfalls als Maximum-Likelihood-Gleichungen bezeichnet.

$$\frac{\partial}{\partial \vartheta_m} ln\, L_n(x^{(n)}, \vartheta) = 0,\ m = 1, 2, \ldots, k. \tag{12.13}$$

Beispiele 12.20:

1) $\vartheta = (\mu, \sigma^2)^T \in R_1 \times (0, \infty), F_\vartheta = N(\mu, \sigma^2)$

$$lnL_n(x^{(n)}; \vartheta) = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{k=1}^{n}(x_i - \mu)^2$$

Aus den Maximum-Likelihood-Gleichungen (12.13) ergibt sich die eindeutige Lösung

$$\hat{\mu}_n = \tfrac{1}{n}\sum_{k=1}^{n} x_i, \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(x_i - \hat{\mu}_n)^2\ ,\ \hat{\vartheta}_n = (\hat{\mu}_n, \hat{\sigma}_n^2)^T.$$

2) Poissonverteilung:

$$Ł_n(x^{(n)}, \lambda) = C \cdot \exp\Big(\sum_{k=1}^{n} x_k \cdot ln\lambda - n\lambda)$$

mit einer nicht von λ abhängenden Konstanten C.

$$\tfrac{d}{d\lambda} lnL_n(x^{(n)}, \lambda) = 0$$

liefert $\hat{\lambda}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$.

3) Gleichmäßige Verteilung auf $[0, \vartheta]$:

$$L_n(\vartheta; x^{(n)}) = \tfrac{1}{\vartheta^n}\prod_{k=1}^{n} \mathbb{1}_{[0,\vartheta]}(x_k) = \tfrac{1}{\vartheta^n}\mathbb{1}_{[0,\vartheta]}\,(\max(x_1, \ldots, x_n)).$$

In diesem Fall ist L_n bez. ϑ nicht differenzierbar und wird maximal für $\vartheta = \max(x_1, \ldots, x_n)$.

Folglich lautet die Maximum-Likelihood-Schätzung hier

$$\hat{\vartheta}_n = \max(X_1, X_2, \ldots, X_n).$$

Maximum-Likelihood-Schätzungen sind i. Allg. nicht erwartungstreu, aber *(schwach) konsistent*, d. h., es gilt

$$\hat{\vartheta}_n(X_1, \ldots, X_n) \underset{P_\vartheta}{\longrightarrow} \vartheta, \vartheta \in \Theta.$$

Außerdem ist unter gewissen Regularitätsbedingungen an die zugrundeliegenden Verteilungen P_ϑ (der Einfachheit halber sei $\Theta \subseteq R_1$)

$$\sqrt{n}\Big(\hat{\vartheta}_n(X_1, \ldots, X_n) - \vartheta\Big) \xrightarrow{d} N\Big(0, \frac{1}{I_1(\vartheta)}\Big) \tag{12.14}$$

mit $I_1(\vartheta) = E_\vartheta\Big(\frac{d}{d\vartheta} ln f_\vartheta(X)\Big)^2 = \int \frac{(f'_\vartheta(x))^2}{f_\vartheta(x)} dx,$ falls F_ϑ die Dichte f_ϑ hat

bzw. $E_\vartheta\Big(\frac{d}{d\vartheta} ln p_\vartheta(X)\Big)^2 = \sum_x \Big(\frac{dp_\vartheta(x)}{d\vartheta}\Big)^2 \Big/ p_\vartheta(x),$ falls F_ϑ Verteilungsfunktion einer diskreten Verteilung mit den Einzelwahrscheinlichkeiten

$p_\vartheta(x), x \in A, \vartheta \in \Theta$ ist.

Das bedeutet insbesondere, Maximum-Likelihood-Schätzungen sind *asymptotisch effizient*. Für große n hat dann $\hat{\vartheta}_n$ nämlich annähernd die Varianz $(nI_1(\vartheta))^{-1}$.

Maximum-Likelihood-Schätzungen sind häufig einfach auszurechnen, existieren aber nicht immer bzw. sind eventuell nicht eindeutig bestimmt. Weitere Details und Beweise findet man z.B. in Winkler (1983) und Dacunha-Castelle, Band I, (1999).

Die Eigenschaft (12.14) kann man nutzen, sogenannte *Vertrauensintervalle* für die Schätzungen von ϑ zu konstruieren. Es gilt wegen (12.14) nämlich für $\alpha \in (0, 1)$

$$P_\vartheta\Big(\sqrt{n}\, I_1^{\frac{1}{2}}(\vartheta)(\hat{\vartheta}_n - \vartheta) \leq x\Big) \approx \Phi(x)$$

und somit

$P_\vartheta\Big(\hat\vartheta_n - \frac{x}{\sqrt{n}} I_1^{-\frac{1}{2}}(\vartheta) \le \vartheta \le \hat\vartheta_n + \frac{x}{\sqrt{n}}\, I_1^{-\frac{1}{2}}(\vartheta)\Big) \approx$

$1 - 2(1-\Phi(x)) = 2\Phi(x) - 1.$

Das bedeutet,

mit der Wahrscheinlichkeit $1-\alpha$ überdeckt das Intervall

$K_{\alpha,n} := \Big(\hat\vartheta_n - \frac{1}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}}\, I_1^{-\frac{1}{2}}(\vartheta), \hat\vartheta_n + \frac{1}{\sqrt{n}}\, q_{1-\frac{\alpha}{2}} I_1^{-\frac{1}{2}}\, (\vartheta)\Big)$

den unbekannten Parameter ϑ.

Hat man eine positive untere Schranke I_0 für $I_1^{\frac{1}{2}}(\vartheta), \vartheta \in \Theta$, so überdeckt auch

$\tilde K_{\alpha,n} := \Big(\hat\vartheta_n - \frac{1}{\sqrt{n}}\, q_{1-\frac{1}{2}}\, I_0^{-1}, \hat\vartheta_n + \frac{1}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\, I_0^{-1}\Big)$

den unbekannten wahren Parameter ϑ mit mindestens der P_ϑ -Wahrscheinlichkeit $1-\alpha$.

12.3 Elemente der Testtheorie

Wir gehen in diesem Punkt auf einige Grundbegriffe der statistischen Testtheorie ein und beschränken uns auf beispielhafte Ausführungen.
Gegeben sei ein zufälliges Experiment $(\Omega, \mathfrak{A}, P)$ mit einer Zufallsgröße X, die nur zwei mögliche Werte annehmen kann:

$$P(X=1) = p, P(X=0) = 1-p =: q, \quad p \in (0,1).$$

Die Wahrscheinlichkeit p sei unbekannt.

Beispiel 12.21: Zufälliges Werfen einer Münze.
Erscheint im k-ten Wurf das Wappen, so wird $X_k = 1$ gesetzt, anderenfalls $X_k = 0$.

Beim Münzenwurf liegt die Vermutung $p = \frac{1}{2}$ nahe. Man spricht von einer *Hypothese* $H_0 : p = \frac{1}{2}$, oder im Allgemeinen $H_0 : p = p_0$ für ein gegebenes p_0.

Zur Verfügung stehe eine konkrete Stichprobe $x^{(n)}$ vom Umfang n aus einer wie X verteilten Grundgesamtheit:

$x^{(n)} = (x_1, x_2, \ldots, x_n)$ mit $x_k \in \{0, 1\}, k = 1, \ldots, n.$
Anhand der Stichprobe soll geprüft werden, ob die Hypothese $H_0 : p = p_0$ zutrifft.

Grundidee: Wenn H_0 richtig ist, so sollte die relative Häufigkeit des Auftretens von Eins in $x^{(n)}$ auf Grund des Gesetzes der großen Zahlen etwa gleich p_0 sein.

Sollte diese relative Häufigkeit stark von p_0 abweichen, so sind Zweifel an der Richtigkeit der Hypothese angebracht, wir werden H_0 ablehnen.

12.3.1 Beispiel eines Alternativtests

"Tea tasting person" (siehe Krengel (2002))
Eine Person behauptet, anhand des Geschmackes bei jeder mit Zitrone und Zucker versehenen Tasse Tee in durchschnittlich 8 von 10 Fällen entscheiden zu können, ob zuerst die Zitrone oder zuerst der Zucker hinzu getan wurde. Wir bezweifeln diese Fähigkeit und vertreten die Hypothese, dass die Person ihre Aussage jedesmal rein zufällig trifft. Bezeichnet p die Wahrscheinlichkeit, mit der die Person die richtige Entscheidung trifft, so lautet unsere Hypothese $H_o : p = \frac{1}{2}$, die der Person $H_1 : p = 0,8$.
Um zu einer Entscheidung zu kommen welcher Hypothese Glauben zu schenken ist, werden $n = 20$ Tassen verkostet. Ist die Entscheidung der Person bei der k-ten Tasse richtig, so setzen wir $x_k = 1$, sonst $x_k = 0$. Im Ergebnis erhalten wir eine konkrete Stichprobe $x^{(n)} = (x_1, x_2, \ldots, x_n)$ aus Nullen und Einsen. Als Entscheidungsgröße berechnen wir die Anzahl $s_n = \sum_{k=1}^{n} x_k$ der Erfolge der Person beim n-maligen Prüfen. Ist $\frac{s_n}{n}$ wesentlich größer als $\frac{1}{2}$, etwa in der Nähe von $0,8$, würde man der Behauptung der Person Glauben schenken und unsere Hypothese $H_0 : p = \frac{1}{2}$ verwerfen. Ist dagegen $\frac{s_n}{n}$ in der Nähe von $\frac{1}{2}$ (oder sogar kleiner), so würde man H_0 annehmen und die Behauptung der Person

zurückweisen.

Um diese Vorgehensweise präzisieren zu können, gehen wir dazu über, die Situation vorab zu betrachten, bevor die Verkostung stattfindet. Dann wird das zukünftige Ergebnis der Verkostung durch einen zufälligen Vektor $X^{(n)} = (X_1, \dots, X_n)$ mit $X_k = 1$, falls die Person im k-ten Versuch recht hat, anderenfalls $X_k = 0$, modelliert. Wir nehmen an, $X^{(n)}$ bestehe aus unabhängigen Zufallsgrößen X_k mit $P^{X_k}(\{1\}) = p, P^{X_k}(\{0\}) = 1 - p, k = 1, \dots n$, und p sei unbekannt. Das heißt, $X^{(n)}$ bildet eine mathematische Stichprobe aus einer wie X_1 verteilten Grundgesamtheit. Unsere Hypothese ist $H_0 : p = \frac{1}{2}$, die der Person $H_1 : p = 0,8$. H_0 wird auch als *Nullhypothese*, H_1, als *Alternativhypothese* bezeichnet.

Es sei zunächst vermerkt, dass eine absolut sichere Entscheidung auf der Grundlage der Kenntnis von $X^{(n)}$ nicht möglich ist, da jede der 2^n Möglichkeiten für $X^{(n)}$ unter beiden Hypothesen mit positiver Wahrscheinlichkeit eintreten kann. Allerdings ist unter H_1 eine größere Anzahl richtiger Antworten wahrscheinlicher als unter H_0.

Entscheidungsvorschrift: Wenn die Anzahl $S_n = \sum_{k=1}^{n} X_k$ richtiger Antworten größer oder gleich einer noch festzulegenden Zahl n_0 ist, so wird H_0 abgelehnt und H_1 angenommen. Ist S_n kleiner als n_0, so wird H_1 abgelehnt und H_0 angenommen.

Die Zufallsgröße S_n heißt in diesem Zusammenhang die *Testgröße* und $K := \{n_0, n_0 + 1, \dots, n\}$ der *kritische Bereich*: Die Zahl n_0 nennt man *kritischen Wert*. Im Fall $S_n \in K$ wird H_0 abgelehnt.

Es gibt bei dem geschilderten Vorgehen zwei mögliche Fehlerarten:

Fehler erster Art: H_0 ist richtig und wird abgelehnt (d. h. in unserem Fall, H_1 wird angenommen).

Fehler zweiter Art: H_0 ist nicht richtig und wird nicht abgelehnt (in unserem Fall: H_1 ist richtig und H_0 wird angenommen).

Die durch die Wahl des kritischen Bereiches, hier also durch den "kritischen Wert" n_0, lassen sich die Wahrscheinlichkeiten der Fehler erster und zweiter Art beeinflussen. Je umfangreicher K (d.h. je kleiner n_0) ist, umso größer wird die Wahrscheinlichkeit des Fehlers erster Art und umso kleiner die Wahrscheinlichkeit des Fehlers zweiter Art.

In der Praxis legt man Wert darauf, dass die Wahrscheinlichkeit des Fehlers erster Art kleiner oder gleich einer vor dem Test festzulegenden *Irrtumswahrscheinlichkeit* α ist. Für α wählt man üblicherweise 0,05 oder, falls ein Fehler erster Art gravierende Schäden verursachen kann, 0,01, eventuell sogar noch kleiner. Häufig ist dadurch der kritische Bereich K und somit das Testverfahren schon festgelegt. Der Fehler zweiter Art ist dann bereits bestimmt und kann u. U. relativ groß sein. Es ist aber zunächst einmal von Interesse, die Wahrscheinlichkeit des Fehlers erster Art kleiner oder gleich α zu haben. Gemeinhin wählt man dabei als Nullhypothese diejenige Hypothese, deren Ablehnung, obwohl sie richtig ist, die schädlicheren Konsequenzen hat.

Angenommen H_0 in unserem Test ist richtig. Dann beträgt die Wahrscheinlichkeit eines Fehlers erster Art

$$p_1(n_0) := P_{\frac{1}{2}}(S_n \in K) = 2^{-n} \sum_{k=n_0}^{n} \binom{n}{k}.$$

Je kleiner n_0 ist, umso größer wird $p_1(n_0)$.
Der kritische Wert n_0 wird nun so groß gewählt, dass $p_1(n_0) \leq \alpha$ gilt. Allerdings vergrößert sich mit n_0 auch der Fehler zweiter Art:

$$p_2(n_0) := P_{0,8}(S_n \notin K) = \sum_{k=0}^{n_0-1} \binom{n}{k} 0,8^k 0,2^{n-k}.$$

Man wird also n_0 unter Einhaltung von $p_1(n_0) \leq \alpha$ möglichst klein wählen:

$$n_0 := \min\{m \in \{1, 2, \ldots, n\} : \sum_{k=m}^{n} \binom{n}{k} 2^{-n} \leq \alpha\}.$$

Als Wahrscheinlichkeit β des Fehlers zweiter Art ergibt sich dann $\beta = p_2(n_0)$.
Die Zahl $1 - \beta$ bezeichnet man auch als *Macht des Testes*.

Zahlenbeispiel:

$n = 20,\ p_0 = \frac{1}{2}$

m	14	15	16	17
$P_{\frac{1}{2}}(S_n \geq m)$	0,0577	0,0476	0,0207	0,0059

H_0 wird abgelehnt und H_1 angenommen (mit der Irrtumswahrscheinlichkeit $\alpha \leqq 0,05$), falls mindestens bei $n_0 = 15$ Tassen richtig entschieden wird.

Die Wahrscheinlichkeit des Fehlers zweiter Art beträgt in diesem Fall $P_{0,8}(S_n < 15) = p_2(n_0) = 0,196$. Sie ist also wesentlich größer als die Wahrscheinlichkeit des Fehlers erster Art.

Um die Wahrscheinlichkeiten der Fehler erster und zweiter Art in ihrer Abhängigkeit von α und n zu studieren, untersucht man die *Gütefunktion des Testes*:

$$g_{n_o}(p) := P_p(S_n \geq n_0) = \sum_{k=n_0}^{n} \binom{n}{k} p^k (1-p)^{n-k}, \quad p \in (0,1).$$

Für jedes $p \in (0,1)$ ist der Wert $g_{n_0}(p)$ die Wahrscheinlichkeit, bei dem oben konstruierten Test die Hypothese, $H_0 : p = \frac{1}{2}$ abzulehnen, falls die tatsächliche Wahrscheinlichkeit gleich p ist. Nach Konstruktion gilt in dem von uns betrachteten Fall

$$g_{n_0}(0) = 0, \quad g_{n_0}(1) = 1,$$

$$g_{n_0}(\tfrac{1}{2}) = p_1(n_0) \leq \alpha,$$

$$g_{n_0}(0,9) = 1 - p_2(n_0) = 1 - \beta.$$

Liegt p zwischen $\frac{1}{2}$ und $0,8$, so ist die Wahrscheinlichkeit des Fehlers zweiter Art noch größer als bei $p = 0,8$. Wenn in unserem Fall die Person gesagt hätte, sie rät durchschnittlich in sechs von zehn Fällen richtig, also $H_1 : p = 0,6$, so wäre die Wahrscheinlichkeit des Fehlers zweiter Art recht groß, nämlich $P_{0,6}(S_{20} < 15) = 0,874$. Wir würden also, falls H_1 richtig ist, trotzdem H_0 mit hoher Wahrscheinlichkeit annehmen. In einem solchen Fall sagt man, falls $S_n \notin K$ eintritt, nicht, dass H_1 falsch und H_0 richtig ist, sondern etwas zurückhaltender, dass auf Grund der Stichprobe gegen H_0 nichts einzuwenden ist. Gegebenenfalls zieht man weitere Entscheidungskriterien heran. Insbesondere wäre das der Fall, wenn die Person nur behauptet, dass sie mit einer Wahrscheinlichkeit p, die größer als $\frac{1}{2}$ ist, die richtige Entscheidung trifft.

12.3.2 Signifikanztests

Wir betrachten erneut die Situation, dass eine Zufallsgröße X gegeben ist, deren Verteilung P^X unbekannt ist, von der man aber weiß, dass sie zu einer Familie $\mathfrak{P}^X = (P_\vartheta^X, \vartheta \in \Theta)$ mit $\Theta \subseteq R_k$ gehört. Wir formulieren eine Hypothese $H_0 : \vartheta = \vartheta_0$, d. h., wir unterstellen, dass der wahre Parameter ϑ_0 ist, mit anderen Worten, dass $P^X = P_{\vartheta_0}^X$ gilt. Diese Hypothese soll an Hand einer Stichprobe $x^{(n)} = (x_1, x_2, \ldots, x_n)$ aus einer nach P^X verteilten Grundgesamtheit geprüft werden. Wie im vorigen Abschnitt bezeichnet man H_0 als Nullhypothese. Allerdings formulieren wir jetzt keine Alternativhypothese. Häufig ist nämlich die Alternative zur Hypothese H_0 nicht einmal genau festlegbar. Solche Tests nennt man Signifikanztests.

Mitunter setzt man Signifikanztests auch dazu ein, allgemeinere Hypothesen zu testen, zum Beispiel $H_0 : \vartheta \in \Theta_0$ bei vorgegebenem $\Theta_0 \subset \Theta$. Mittels der Stichprobe $x^{(n)}$ soll also entschieden werden, ob H_0 abzulehnen ist. Dabei soll die Wahrscheinlichkeit einer Fehlentscheidung, wenn H_0 richtig ist (Fehler erster Art) nicht größer als eine vorgegebene Zahl $\alpha \in (0, 1)$ sein. Die Zahl α heißt *Irrtumswahrscheinlichkeit*, die Zahl $1 - \alpha$ nennt man das *Signifikanzniveau*. (Ein Fehler zweiter Art ist hier mangels Alternativhypothese nicht vorhanden.)
Dazu konstruieren wir wie folgt einen statistischen Test.

1. Wir wählen eine Stichprobenfunktion $T_n = T_n(x_1, x_2, \ldots, x_n)$, wobei T_n eine Borelmessbare Funktion sein möge, die wir hier überdies als reellwertig annehmen.

2. Wir wählen einen *kritischen Bereich* K, d. h. eine Borelmessbare Teilmenge des Wertbereiches von T_n, so dass

   $$P_{\vartheta_0}^X(T_n \in K) \leq \alpha$$

erfüllt ist. (Hat T_n eine stetige Verteilung, wird man K so wählen, dass $P^X_{\vartheta_0}(T_n \in K) = \alpha$ gilt.)

3. Sodann vereinbaren wir die *Entscheidungsregel:*
   Die Hypothese $H_0 : \vartheta = \vartheta_0$ wird auf Grund der Stichprobe $x^{(n)}$ abgelehnt, falls $T_n(x_1, \ldots, x_n) \in K$. Anderenfalls, also wenn $T_n(x_1, \ldots, x_n) \notin K$ gilt, ist gegen H_0 auf Grund der Stichprobe nichts einzuwenden.

Man sagt im Fall der Ablehnung, dass sie zum *Signifikanzniveau* $1 - \alpha$ erfolge und bezeichnet den so konstruierten Test als *Signifikanztest der Hypothese* H_0 *zum Signifikanzniveau* $1 - \alpha$.

In der Wahl des kritischen Bereiches K steckt noch eine gewisse Willkür. Häufig ist er durch die konkreten Rahmenbedingungen nahegelegt. Allgemein sollte er so konstruiert werden, dass das Ereignis $\{T_n \in K\}$ unter H_0 eine derart kleine Wahrscheinlichkeit hat ($\leq \alpha$), dass man das Eintreten von $\{T_n \in K\}$ nicht als Zufall ansieht, sondern eher daran zweifelt, dass die Hypothese H_0 stimmt. Das wird umso mehr berechtigt sein, wenn das Ereignis $\{T_n \in K\}$ für den Fall, dass H_0 nicht stimmt, eine große Wahrscheinlichkeit besitzt.

Beispiele 12.22:

1. *Test des Mittelwertes einer* $N(\mu, \sigma^2)$*-verteilten Grundgesamtheit bei bekannter Streuung* σ^2

   Es sei $X \sim N(\mu, \sigma^2)$ und $X^{(n)} = (X_1, X_2, \ldots, X_n)$ eine mathematische Stichprobe aus einer wie X verteilten Grundgesamtheit. Die Varianz σ^2 sei bekannt, $\alpha \in (0, 1)$ sei vorgegeben. Wir konstruieren einen Signifikanztest der Hypothese $H_0 : \mu = \mu_0$ zum Niveau $1 - \alpha$.

   Als Testgröße wählen wir

   $T_n(X^{(n)}) = \frac{(\bar{X}_n - \mu_0)\sqrt{n}}{\sigma}$, wobei $\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k$ gesetzt wurde.

Offenbar besitzt $T_n = T_n(X^{(n)})$ eine $N(0,1)$-Verteilung, falls H_0 richtig ist. Stimmt H_0, so wird die Zufallsgröße $T_n(X^{(n)})$ bei ihrer Realisierung mit hoher Wahrscheinlichkeit einen Wert in der Nähe von Null annehmen. Stimmt H_0 nicht, ist also $\mu \neq \mu_0$, so hat $T_n = \frac{(\bar{X}_n - \mu)\sqrt{n}}{\sigma} + \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}$ eine $N\Big(\frac{(\mu-\mu_0)\sqrt{n}}{\sigma}, 1\Big)$-Verteilung. Ihre Realisierung würde stark von Null abweichen (falls μ sich stark von μ_0 unterscheidet). Deshalb wählen wir den kritischen Bereich K in der Form

$$K = \{t \mid |t| > z_{\alpha,n}\}$$

und bestimmen $z_{\alpha,n}$ so, dass unter H_0 gilt

$$P\big(|T_n(X^{(n)})| > z_{\alpha,n}\big) = \alpha.$$

Das ergibt wegen

$$P\big(|T_n(X^{(n)})| > z_{\alpha,n}\big) = 2(1 - \Phi(z_{\alpha,n}))$$

die Beziehung

$$z_{\alpha,n} = q_{1-\frac{\alpha}{2}}$$

(q_p bezeichnet das p-Quantil der Standard Normalverteilungsfunktion Φ).

Entscheidungsregel: $H_0 : \mu = \mu_0$ wird abgelehnt, falls für die konkrete Stichprobe $x^{(n)} = (x_1, x_2, \ldots, x_n)$ gilt

$|T_n(x^{(n)})| > q_{1-\frac{\alpha}{2}}$, d. h., falls gilt:

$$|\bar{X}_n - \mu_0| > \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}.$$

Anderenfalls ist gegen H_0 auf Grund der Stichprobe $x^{(n)}$ nichts einzuwenden.

Bemerkung 12.23: Ist der Stichprobenumfang n groß, so nimmt man für σ^2, falls nicht anders verfügbar, die Schätzung $\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x}_n)^2$. Das ist auf Grund des Gesetzes der großen Zahlen gerechtfertigt, da $E\hat{\sigma}_n^2 = \sigma^2$ gilt. Wie man im Fall kleiner Stichprobenumfänge verfährt, wird im folgenden Beispiel erläutert.

2. *Test des Mittelwertes einer $N(\mu, \sigma^2)$-verteilten Grundgesamtheit bei unbekannter Streuung*

Wir behandeln das gleiche Problem wie im vorangegangenen Beispiel, nehmen aber an, σ^2 ist nicht bekannt und der Stichprobenumfang n ist nicht allzu groß, so dass man bezweifeln kann, dass

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2$$

bereits eine gute Näherung für σ^2 ist. In diesem Fall verwendet man

$$T_n'(X^{(n)}) = \frac{(\bar{X}_n - \mu_0)}{\hat{\sigma}_n} \cdot \sqrt{n}$$

als Testgröße. Wir benötigen die Verteilung von $T_n'(X^{(n)})$ unter der Nullhypothese H_0, um den kritischen Bereich bestimmen zu können.

Lemma 12.24: *$\bar{X}_n$ und $\hat{\sigma}_n^2$ sind unter H_0 voneinander unabhängige Zufallsgrößen. $\bar{X}_n$ ist $N\left(\mu_0, \frac{\sigma^2}{n}\right)$-verteilt und $\frac{\hat{\sigma}_n^2}{\sigma^2} \cdot (n-1)$ besitzt eine χ^2-Verteilung mit $n-1$ Freiheitsgraden, d. h. eine Gammaverteilung $\Gamma(\alpha, \lambda)$ mit den Parametern $\alpha = \frac{n-1}{2}$ $\lambda = \frac{1}{2}$. (Siehe auch Abschnitt 12.3.3)*

Der Beweis soll in Übung 12.4 geführt werden (siehe auch Krengel (2002), Kap. II, § 14). Als Verteilung von T_n' ergibt sich damit die Verteilung mit der Dichte

$$f_{n-1}(x) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)\left(\frac{x^2}{n-1}+1\right)^{\frac{n}{2}}}\,, \quad x \in R_1.$$

Diese Verteilung trägt die Bezeichnung *t-Verteilung* (oder *Studentverteilung*) *mit $n-1$ Freiheitsgraden*, die Werte ihrer Verteilungsfunktion F_{n-1} bzw. ihre Quantile sind vertafelt und in vielen Büchern über Mathematische Statistik zu finden, siehe zum Beispiel Krengel (2002), Tabellen Seite 247.

Auch hier wählen wir den kritischen Bereich K in der Form

$$K = \{t|\ |t| > z_{\alpha,n}\}$$

und bestimmen $z_{\alpha,n}$ derart, dass unter H_0 gilt

$$P(|T'_n(X^{(n)})| > z_{\alpha,n}) = \alpha.$$

Das ergibt

$2(1 - F_{n-1}(z_{\alpha,n})) = \alpha$, also

$$z_{\alpha,n} = t_{n-1,1-\frac{\alpha}{2}}$$

wobei $t_{n-1,1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden ist.

12.3.3 Der χ^2-Test

Unter den Signifikanztests hat sich der sogenannte χ^2-Test als ein sehr flexibles statistisches Werkzeug seinen festen Platz erobert.

Es sei X eine diskret verteilte Zufallsgröße mit den endlich vielen möglichen Werten a_k aus der Menge $A = \{a_k : k \in \{1, 2, \ldots, r\}\}$ reeller Zahlen. Weiterhin sei $\{p_k : k \in \{1, 2, \ldots, r\}\}$ eine Wahrscheinlichkeitsverteilung auf A. Anhand einer Stichprobe $x^{(n)}$ aus einer nach P^X verteilten Grundgesamtheit soll die Hypothese H_0 geprüft werden, dass X die Verteilung $(a_k, p_k), k \in K$, besitzt, d. h.

$H_0 : P(X = a_k) = p_k, \quad 1 \le k \le r.$

Zu diesem Zweck bildet man die Testgröße

$$\chi^2 = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k},$$

wobei n_k die Anzahl derjenigen x_j aus der Stichprobe $x^{(n)} = (x_1, x_2, \ldots, x_n)$ mit $x_j = a_k$ bezeichne, $1 \le k \le r, 1 \le j \le n$.

Die Größe χ^2 ist eine gewichtete Summe der quadratischen Abweichungen zwischen den Anzahlen n_k und ihren bei richtiger Hypothese H_0 "zu erwartenden" Werte np_k.

Um die wahrscheinlichkeitstheoretischen Eigenschaften dieser Testgröße zu studieren, setzen wir in χ^2 an Stelle von n_k die Zufallsgrößen N_k ein, die sich aus der entsprechenden mathematischen Stichprobe $X^{(n)}$ genauso berechnen, wie die n_k aus der konkreten Stichprobe $x^{(n)}$.

Satz 12.25: *(R. A. Fisher) Die Wahrscheinlichkeiten p_k, $k = 1, 2, \ldots, r$, seien gegeben. Dann konvergieren die Verteilungsfunktionen F_n der Zufallsgrößen χ^2, falls die Hypothese H_0 richtig ist, mit wachsendem Stichprobenumfang n gegen eine Gammaverteilung $\Gamma(\alpha, \lambda)$ mit den Parametern*

$\alpha = \frac{r-1}{2}$ und $\lambda = \frac{1}{2}$:

$$\lim_{n\to\infty} F_n(x) = \frac{1}{2^{\frac{r-1}{2}}\Gamma\left(\frac{r-1}{2}\right)} \int_0^x y^{\frac{r-3}{2}} \mathrm{e}^{-\frac{y}{2}} dy, \quad x > 0$$
$$= 0, \qquad x \le 0.$$

Die Verteilung $\Gamma\left(\frac{r-1}{2}, \frac{1}{2}\right)$ trägt einen eigenen Namen und heißt χ^2-Verteilung mit $r-1$ Freiheitsgraden $(r \ge 1)$.

Den Beweis findet man z. B. in Krengel (2002), Kap. II, § 14.

Seine Grundidee besteht in der Beobachtung, dass der Vektor $(N_1, N_2, \ldots, N_r)$ eine Multinominalverteilung mit den Parametern $n, p_1, p_2, \ldots, p_r$ besitzt. Dann ist $(N_1 - np_1, \ldots, N_r - np_r)$ ein zentrierter zufälliger Vektor, mit $\sum\limits_{k=1}^{r} N_k = n$, also

$$\sum_{k=1}^{r} (N_k - np_k) = 0. \tag{12.15}$$

Die r-dimensionale Multinomialverteilung von $(N_1, N_2, \ldots, N_r)$ konvergiert für $n \to \infty$ ebenso wie die Binomialverteilung im globalen Grenzwertsatz von Moivre-Laplace gegen eine Normalverteilung, die wegen (12.15) auf einem $(r-1)$-dimensionalen Teilraum von R_r konzentriert ist.

Die Zufallsgröße

$$\chi^2 = \sum_{k=1}^{r} \frac{(N_k - np_k)^2}{np_k} \tag{12.16}$$

lässt sich damit durch Grenzübergang $n \to \infty$ zurückführen auf die Quadratsumme von $(r-1)$ Standard normalverteilten Zufallsgrößen. Dann erhält man mit folgendem Lemma die Aussage des Satzes.

Lemma 12.26: *Es seien $Y_1, Y_2, \ldots, Y_m$ $(m \geq 1)$ voneinander unabhängige $N(0,1)$-verteilte Zufallsgrößen. Dann besitzt*

$$S^2 = \sum_{k=1}^{m} Y_k^2$$

eine χ^2-Verteilung mit m Freiheitsgraden.

Beweis:

$P(Y_1^2 \leq y) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1$, woraus sich die Dichte $f_{Y_1^2}$ ergibt:

$f_{Y_1^2}(y) = \frac{2\varphi(\sqrt{y})}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi y}} \mathrm{e}^{-\frac{1}{2}y}, \; y > 0.$

Also besitzt jedes Y_k^2 eine $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$-Verteilung:

$Y_k^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right), \quad k = 1, 2, \ldots, m,$

und auf Grund der Unabhängigkeit der $Y_1, \ldots, Y_m$ folgt

$S^2 \sim \Gamma\left(\frac{m}{2}, \frac{1}{2}\right).$

Die Verteilungsfunktion der χ^2-Verteilungen ist nicht explizit berechenbar, sie bzw. ihre Quantile sind vertafelt, und man findet sie, wie oben bereits erwähnt zum Beispiel in Krengel (2002), Tabellen, Seite 249.

Für jede mit m Freiheitsgraden χ^2-verteilte Zufallsgröße Y gilt

$EY = m,\ D^2Y = 2m,\ \text{Modalwert}\,(Y) = \max(0, m-2).$

Die Testgröße $T_n(X^{(n)})$ wird also mit hoher Wahrscheinlichkeit (hier: $1-\alpha$) Werte annehmen, die in einem Intervall um den Modalwert liegen, z. B. in

$\left(\chi^2_{r-1,\frac{\alpha}{2}}, \chi^2_{r-1,1-\frac{\alpha}{2}}\right).$

Dabei bezeichnet $\chi^2_{r-1,p}$ das p-Quantil der χ^2-Verteilung mit $r-1$ Freiheitsgraden.

Eine erste Anwendung des χ^2-Test enthält das folgende Beispiel.

Beispiel 12.27 (χ^2-*Anpassungstest*): Es seien F eine Verteilungsfunktion auf R_1 und X eine reellwertige Zufallsgröße über einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{A}, P)$ mit

$P(X \leq x) = F_X(x), \quad x \in R_1.$

Wir wollen die Hypothese

$H_0 : F_X = F$

testen.

Zu diesem Zweck unterteilen wir R_1 in r Intervalle $(r \geq 2)$

$I_1 = (-\infty, a_1],\ I_2 = (a_1, a_2], \ldots, I_{r-1} = (a_{r-2}, a_{r-1}],\ I_r = (a_{r-1}, \infty)$

und setzen

$p_k = F(a_k) - F(a_{k-1}), \quad k = 1, \ldots, r$

mit $a_0 = -\infty,\ F(a_0) = 0,\ a_r = +\infty,\ F(a_r) = 1.$

Ist H_0 richtig, so gilt

$P(X \in I_k) = p_k, \quad k = 1, 2, \ldots, r.$

Wir verwenden die Testgröße

$$\chi^2 = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k}$$

und den kritischen Bereich

$K = R_1 \backslash (\chi^2_{r-1,\frac{\alpha}{2}}, \chi^2_{r-1,1-\frac{\alpha}{2}})$
(zweiseitiger Test) bzw.

$K = (\chi^2_{r-1,1-\alpha}, \infty)$
(einseitiger Test).

Der so konstruierte Signifikanztest heißt χ^2-*Anpassungstest* zum Signifikanzniveau $1 - \alpha$.

Wir illustrieren diesen Test durch zwei Beispiele:

Zufallszahlen aus $[0, 1)$

Angenommen, $x^{(n)} = (x_1, x_2, \ldots, x_n)$ ist eine n-elementige Folge reeller Zahlen aus $[0, 1)$. Wir wollen die Hypothese prüfen, dass sie aus einer gleichmäßig auf $[0, 1)$ verteilten Grundgesamtheit stammen, und zwar zum Signifikanzniveau $0,95$. Dazu nehmen wir an, die konkrete Stichprobe $x^{(n)}$ ist Realisierung einer mathematischen Stichprobe $X^{(n)} = (X_1, X_2, \ldots, X_n)$, jedes X_k habe die Verteilungsfunktion F und formulieren die Hypothese

$H_0 : F = F_0$

mit $F_0(x) = x$ für $x \in [0,1], = 0$ für $x < 0, = 1$ für $x > 1$.

Es sei $n = 100$ und $\alpha = 0,05$.

Wir teilen $[0,1)$ in 10 Klassen

$I_k = \left[\frac{k-1}{10}, \frac{k}{10}\right), \; k = 1, 2, \ldots, 10$

ein. Dann gilt

$p_k = F_0\left(\frac{k}{10}\right) - F_0\left(\frac{k-1}{10}\right) = 0,1$ für $k = 1, 2, \ldots, 10$

und die Testgröße χ^2 ergibt sich zu

$$\chi^2 = \sum_{k=1}^{10} \frac{(N_k - 10)^2}{10} = 0,1 \cdot \sum_{k=1}^{10} (N_k - 10)^2.$$

Der kritische Bereich K wird für einen zweiseitigen Test wie folgt festgelegt:

$$K = (0, \chi^2_{0,025,9}) \cup (\chi^2_{0,975,9}, \infty)$$

$$= (0; 2,70) \cup (19,02, \infty).$$

Bei dieser Konstruktion wird die Hypothese H_0 abgelehnt, wenn die empirische Verteilung $\hat{F}_{100}$ zu weit von F_0 entfernt ist (d. h., wenn die Testgröße χ^2 groß ist), oder wenn $\hat{F}_{100}$ zu nahe an F_0 liegt (wenn χ^2 zu klein ist). Empfindet man sehr kleine χ^2 nicht als Mangel, so kann man K auch in der Form

$$K = (\chi^2_{1-\alpha,9}, \infty) = (16,92; \infty)$$

wählen (einseitiger Test).

Geburtenzahlen

Im Landkreis Teltow-Fläming wurden 1996 insgesamt 360 Kinder geboren, davon 175 Mädchen und 185 Jungen. Widerspricht diese Zahl der Hypothese, dass das Geschlecht von Neugeborenen mit gleicher Wahrscheinlichkeit weiblich bzw. männlich ist, zum Signifikanzniveau von $0,95$?

Bezeichnen wir mit p die Wahrscheinlichkeit, dass ein Neugeborenes ein Junge wird. Dann lautet die Hypothese

$H_0 : p = \frac{1}{2}$.

Die Testgröße berechnet sich zu

$\chi^2 = \frac{(185-180)^2}{180} + \frac{(175-180)^2}{180} = \frac{50}{180} = 0,2778.$

Da der kritische Bereich

$K = (\chi^2_{1-\alpha,1}, \infty) = (3,84; \infty)$

lautet, ist gegen H_0 auf Grund der Stichprobe nichts einzuwenden.

In Deutschland wurde 1991 insgesamt 911 600 Kinder geboren, davon 442 400 Mädchen und 468 000 Jungen.
Wendet man den gleichen Test wie eben auf

$H_0 : p = \frac{1}{2}$

an, so ergibt sich

$\chi^2 = 520,68 \gg 3,84.$

Die Überschreitung des kritischen Wertes 3,84 durch die Testgröße ist hochsignifikant, H_0 wird auf Grund dieser Stichprobe abgelehnt.

Wir kehren noch einmal zurück zum eingangs behandelten χ^2-Test einer diskreten Verteilung $P = (p_k, k = 1, \ldots, r)$ auf $A = \{a_1, a_2, \ldots, a_r\}$.

In manchen Fällen ist die Verteilung P nicht völlig festgelegt, sondern hängt noch von einem Parameter ϑ ab:

$P \in \mathfrak{P} = (p_k(\vartheta), k = 1, \ldots, r; \vartheta \in \Theta \subseteq R^l)$ für ein $l \geq 1$. Dadurch ist die Verteilung noch nicht eindeutig bestimmt und wir können den oben angegebenen Satz von Fisher nicht anwenden. Es gibt vielmehr die folgende allgemeinere Fassung.

Wir setzen voraus:

Satz 12.28: *Die Ableitungen*

$$\frac{\partial p_k}{\partial \vartheta_i}, \frac{\partial^2 p_k}{\partial \vartheta_i \partial \vartheta_j}, \; k = 1, 2, \ldots, r; \; i, j \in \{1, 2, \ldots, l\}$$

existieren und sind stetig bzgl. ϑ.

Die Matrix $\left(\frac{\partial p_k}{\partial \vartheta_i}\right)_{k,i}$ *habe den Rang* l.

Werden die unbekannten Parameter $\vartheta_1, \vartheta_2, \ldots, \vartheta_l$ *mit Hilfe der Stichprobe* $x^{(n)}$ *nach der Maximum-Likelihood-Methode geschätzt, so konvergieren die Verteilungsfunktionen* F_n *der Stichprobenfunktion* χ^2 *aus Formel (12.16) gegen eine* χ^2*-Verteilung mit* $r - l - 1$ *Freiheitsgraden.*

Für einen Beweis siehe z. B. Dacunha-Castelle, Duflo, Vol. II (1986).

Beispiel 12.29:

Test auf Unabhängigkeit in Kontingenztafeln

Gegeben seien zwei Zufallsgrößen X und Y, beide diskret verteilt mit r bzw. s möglichen Werten und den (unbekannten) Wahrscheinlichkeiten der gemeinsamen Verteilung

$$p_{ij} = P(X = x_i, Y = y_j), \quad i = 1, 2, \ldots, r; \; j = 1, 2, \ldots, s.$$

Es werde eine konkrete Stichprobe vom Umfang n realisiert:

n_{ij} = Häufigkeit des Auftretens des Paares (x_i, y_j).

$i\backslash^j$	1	2	...	s	
1	n_{11}	n_{12}		n_{1_s}	$n_{1\cdot}$
$\vdots$					
r	n_{r1}	n_{r2}		n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot s}$	n

Kontingenztafel

H_0 : "Die Merkmale X und Y sind voneinander unabhängig."

Das bedeutet mit den Bezeichnungen $p_{i\cdot} = \sum\limits_i p_{ij},\ p_{\cdot j} = \sum\limits_i p_{ij}$

$H_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$.

Durch diese Hypothese ist die Wahrscheinlichkeitsverteilung (p_{ij}) noch nicht festgelegt. Die Größen $p_{i\cdot}$ und $p_{\cdot j}$ $(1 \leq i \leq r, 1 \leq j \leq s)$ müssen geschätzt werden.
Die Maximum-Likelihood-Methode liefert $\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}$ (siehe unten).

Wegen $\sum\limits_i p_{i\cdot} = \sum\limits_j p_{\cdot j} = 1$ sind dies $(r-1)+(s-1)$ geschätzte Parameter.

Testgröße:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{ij} - p_{ij}\cdot n)^2}{n \cdot p_{ij}} = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(N_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n})^2}{N_{i\cdot}\cdot N_{\cdot k}}.$$

Diese Testgröße besitzt für $n \to \infty$ eine χ^2-Verteilung mit $r \cdot s - r - s + 2 - 1 = (r-1)(s-1)$ Freiheitsgraden.

Maximum-Likelihood-Schätzung der $p_{i\cdot}, p_{\cdot j}$:

Die Likelihoodfunktion ist unter der Hypothese H_0 gegeben durch

$L(\vartheta, X^{(n)}) = \prod_{i=1}^{r}\prod_{j=1}^{s} p_{ij}^{N_{ij}} = \prod_{i=r}^{r}\prod_{j=1}^{s} p_{i\cdot}^{N_{ij}} p_{\cdot j}^{N_{ij}} = \prod_{i=1}^{r} p_{i\cdot}^{N_{i\cdot}} \prod_{j=1}^{s} p_{\cdot j}^{N_{\cdot j}}$ mit

$N_{ij} = \#\{k \leq n : X_k = x_i, Y_k = y_j\}$,

$$N_{i\cdot} = \sum_{j=1}^{s} N_{ij}, \quad N_{\cdot j} = \sum_{i=1}^{r} N_{ij} \quad 1 \leq i \leq r, 1 \leq j \leq s$$

und $\vartheta = (p_{1\cdot}, \ldots p_{r-1\cdot}, p_{\cdot 1}, \ldots, p_{\cdot s-1})$.

Dabei wird gesetzt

$p_{r\cdot} = 1 - p_{1\cdot} - \ldots - p_{r-1\cdot}$ und

$p_{\cdot s} = 1 - p_{\cdot 1} - \ldots - p_{\cdot s-1}$.

Für die Maximum-Likelihood-Gleichungen ergibt sich

$$\frac{\partial}{\partial p_{i\cdot}} lnL = \frac{N_{i\cdot}}{p_{i\cdot}} - \frac{N_{r\cdot}}{1-p_{1\cdot}-\ldots-p_{r-1\cdot}} = 0, \tag{12.17}$$

$i = 1, \ldots, r-1$, also

$\frac{N_{i\cdot}}{\hat{p}_{i\cdot}} = \frac{N_{r\cdot}}{\hat{p}_{r\cdot}}$, mithin

$N_{i\cdot} = \hat{p}_{i\cdot} \cdot \frac{N_{r\cdot}}{\hat{p}_{r\cdot}}, i = 1, 2, \ldots, r.$

Summation über i liefert

$n = 1 \cdot \frac{N_{r\cdot}}{\hat{p}_{r\cdot}}$, also $\hat{p}_{r\cdot} = \frac{N_{r\cdot}}{n}$.

Daraus ergibt sich $\hat{p}_{i\cdot} = \frac{N_{i\cdot}}{n}$.

Die Schätzungen $\hat{p}_{\cdot j} = \frac{N_{\cdot j}}{n}$ ergeben sich analog aus

$\frac{\partial}{\partial p_{\cdot j}} lnL = 0,\ j = 1, \ldots, s-1.$

Index