Zentraler Grenzwertsatz

Die Zufallsvariablen $X_1, \ldots, X_n$ seien voneinander unabhängig und besitzen jeweils dieselbe Verteilung (mit Erwartungswert μ und Varianz σ^2).

Dann sind die Zufallsvariablen

$$Z_n := \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

asymptotisch standardnormalverteilt, also $Z_n \sim \mathcal{N}(0,1)$ für $n \to \infty$. Formaler ausgedrückt gilt: Die Folge der zu Z_n gehörenden Verteilungsfunktionen F_n hat die Eigenschaft

$$\lim_{n\to\infty} F_n(x) = \Phi(x)$$

Beweis: siehe Literatur

Korollar des Zentralen Grentzwertsatzes:

Grenzwertsatz von DeMoivre:

$X_1, \ldots, X_n$ seien unabhängige Bernoulli-verteilte Zufallsvariablen mit gleicher Erfolgswahrscheinlichkeit p. [Bspw. Indikatorvariablen für die n-fache unabhängige Beobachtung eines Ereignis E]. Die Zufallsvariable H_n mit

$$H_n := X_1 + \ldots + X_n$$

nennen wir absolute Häufigkeit von E. Für $n \to \infty$ gilt: Die Verteilung der Zufallsgröße

$$\frac{H_n - np}{\sqrt{np(1-p)}}$$

konvergiert gegen die Standardnormalverteilung. Äquivalent dazu sind die Aussagen

- H_n besitzt (im Limes) die Verteilung $\mathcal{N}(np, np(1-p))$ bzw.
- H_n/n (die relative Häufigkeit von E) besitzt (im Limes) die Verteilung $\mathcal{N}(p, p(1-p)/n)$

Bsp: Ein Netzwerkadministrator steht vor folgender Aufgabe: Bei zwei Routern wurde gemessen, wie viele Datenpakete sie pro Tag bearbeitet haben. Für Router X wurden die acht Werte

$$(53, 53, 37, 73, 58, 61, 38, 54)$$

ermittelt. Für Router Y sind nur sieben Werte verfügbar, da das Messprogramm an einem Tag abgestürzt ist. Die Werte sind

$$(33, 66, 26, 43, 46, 55, 54).$$

Der Administrator möchte nun aus diesen Werten ablesen, ob einer der beiden Router im Durchschnitt mehr Last abarbeiten muss als der andere. Was muss er dazu tun?

Bsp: Ein Systemadministrator steht vor folgender Aufgabe: Bei einer neu gekauften Festplatte prüfte er die Zugriffszeit. Er erhält folgende Messwerte:

$$(7.1, 7.5, 11.2, 7.0, 6.3, 8.0, 6.9).$$

Darf er zu Recht davon ausgehen, dass die Festplatte die vom Hersteller zugesagte mittlere Zugriffszeit von $7\mu s$ einhält?

Definition: (*(erwartungstreuer) Schätzer*)

Gegeben sei eine Zufallsvariable X mit der Dichte $f(x;\theta)$. Eine *Schätzvariable* oder kurz *Schätzer* für den Parameter θ der Dichte von X ist eine Zufallsvariable, die aus mehreren Stichprobenvariablen zusammengesetzt ist.

Ein Schätzer U heißt *erwartungstreu*, wenn gilt

$$\mathbb{E}[U] = \theta.$$

Definition: (*Güte eines Schätzers*)

Ein wichtiges Maß für die Güte eines Schätzers ist die *mittlere quadratische Abweichung*, kurz *MSE* für mean squared error. Diese berechnet sich durch

$$MSE := \mathbb{E}[(U - \theta)^2].$$

Wenn U erwartungstreu ist, so folgt

$$MSE = \mathbb{E}[(U - \mathbb{E}[U])^2] = \mathrm{Var}[U].$$

Wenn die Schätzvariable A eine kleinere mittlere quadratische Abweichung besitzt als die Schätzvariable B, so sagt man, dass A *effizienter* ist als B.

Eine Schätzvariable heißt *konsistent im quadratischen Mittel*, wenn MSE $\to 0$ für $n \to \infty$ gilt. Hierbei bezeichne n den Umfang der Stichprobe.

Maximum-Likelihood-Schätzer

GEGEBEN: Unabhängige Stichprobenvariablen $X_1, \ldots, X_n$ eine Zufallsvariablen X.
X habe die Dichte $f(x;\theta)$, wobei θ ein a priori unbekannter Parameter der Verteilung sei.

GESUCHT: Schätzer für θ.

Der Ausdruck

$$L(\vec{x};\theta) := \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \Pr_\theta[X_i = x_i] \overset{\text{Unabh.}}{=} \Pr_\theta[X_1 = x_1, \ldots, X_n = x_n]$$

entspricht der Wahrscheinlichkeit, dass wir die Stichprobe $\vec{x}$ erhalten, wenn wir den Parameter mit dem Wert θ belegen. (Hier haben wir angenommen, dass X eine diskrete ZV ist.)

Definition. (*Maximum-Likelihood-Schätzwert*)

Ein Schätzwert $\widehat{\theta}$ für den Parameter einer Verteilung $f(x;\theta)$ heißt *Maximum-Likelihood-Schätzwert* (ML-Schätzwert) für eine Stichprobe $\vec{x}$, wenn gilt

$$L(\vec{x};\theta) \leq L(\vec{x};\widehat{\theta}) \quad \text{für alle } \theta.$$