Einführung in Empirische Prozesse

Sebastian Nied

13. Mai 2013

Definition

Seien $X_1, X_2, \ldots$ unabhängige Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega, \Sigma, \mathbb{P})$ mit Verteilungsfunktion F. Dann wird die zufällige Verteilungsfunktion

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(X_i) \qquad \forall x \in (-\infty, \infty) \tag{1}$$

empirische Verteilungsfunktion der Zufallsstichprobe $(X_1, \ldots, X_n)$ genannt.
Hierbei ist $\mathbb{I}$ die Indikatorfunktion

$$\mathbb{I}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

Äquivalent definiert man

$$\hat{F}_n(x) = \frac{\sharp\{i : 1 \leq i \leq n, X_i \leq x\}}{n} \tag{2}$$

Verteilung von $n\,\hat{F}_n$

Sei $\hat{F}_n$ eine empirische Verteilungsfunktion von unabhängigen Zufallsvariablen $X_1, ..., X_n$, dann ist

$$n\,\hat{F}_n(x) = \sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(X_i)$$

binomialverteilt mit Parametern n und p = F(x).

Fast sichere Konvergenz

Da $n\,\hat{F}_n$ binomialverteilt ist, folgt aus dem starken Gesetz der großen Zahlen

$$\hat{F}_n \to_{f.s.} F\,. \tag{3}$$

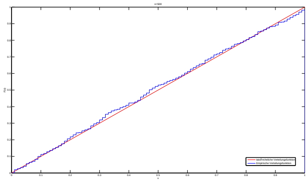
Gleichmäßige Konvergenz nach Glivenko-Cantelli

Im folgenden werden wir die Konvergenz der empirischen Verteilungsfunktion nach Glivenko-Cantelli zeigen:

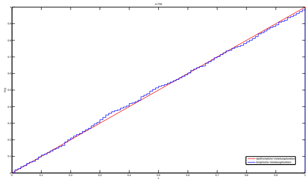
$$\hat{F}_n \rightarrow_{glm} F \qquad n \rightarrow \infty\,. \tag{4}$$

Dabei betrachten wir den gleichmäßigen Abstand D_n und dessen Konvergenz

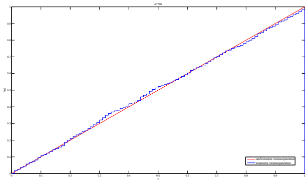
$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \rightarrow_{f.s.} 0\,. \tag{5}$$



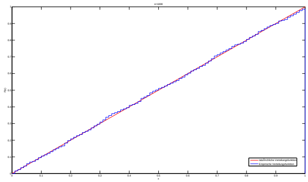
(a) n=250



(b) n=500



(c) n=750



(d) n=1000

Abbildung: Eine empirische Verteilungsfunktion für verschieden große n

Der Empirische Prozess

Definition

Aus der Standardisierung der empirischen Verteilungsfunktion folgt der empirische Prozess, definiert durch

$$\beta_n(x) := \sqrt{n}(\hat{F}_n(x) - F(x)) \qquad \forall x \in (-\infty, \infty)\,. \tag{6}$$

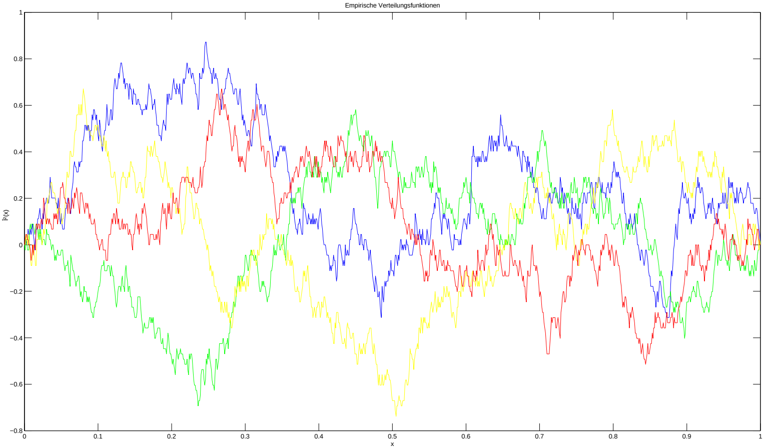


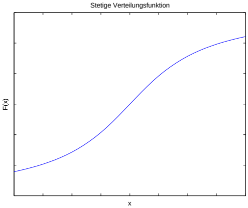
Abbildung: Verschiedene Realisierungen des empirischen Prozesses für gleichverteilte Zufallsvariablen und Stichprobengröße n = 1000

Die Inverse Transformation

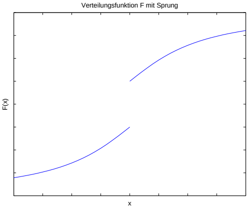
Definition (der linksstetigen Inversen)

Sei F eine Verteilungsfunktion. Dann ist ihre linksseitig stetige Inverse definiert durch

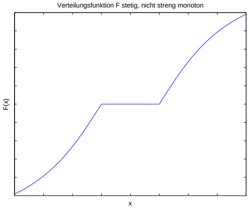
$$F^{-1}(t) := \inf\{x : F(x) \geq t\} \qquad \forall t \in (0,1)\,. \tag{7}$$



(a) F streng monoton und stetig



(b) F streng monoton mit Sprung



(c) F monoton und stetig

Abbildung: Verschiedene Verteilungsfunktionen

Satz (über die inverse Transformation)

Sei $\xi \sim U(0,1)$, F eine Verteilungsfunktion und F^{-1} ihre Inverse. Dann gilt

$$X := F^{-1}(\xi) \sim F\,. \tag{8}$$

Also ist

$$\mathbb{I}_{[X \leq x]} = \mathbb{I}_{[\xi \leq F(x)]} \quad \forall x \in (-\infty, \infty)\,. \tag{9}$$

Beweis.

Aus $\xi \leq F(x)$ folgt $X = F^{-1}(\xi) \leq$ x, durch die Definition der Inversen.

Ist $X = F^{-1}(\xi) \leq$ x, dann gilt $F(x+\varepsilon) \geq \xi \quad \forall \varepsilon > 0$.

Durch die rechtsseitige Stetigkeit der Verteilungsfunktion folgt $F(x) \geq \xi$.

$\square$

Satz

Seien $\xi_1, ..., \xi_n \sim$U(0,1) unabhängige Zufallsvariablen und seien $\hat{G}_n(x)$ und $\alpha_n(x)$ die empirische Verteilungsfunktion und der empirische Prozess dieser Zufallsvariablen mit

$$\hat{G}_n(z) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[\xi_i \le z]} \quad \textit{und} \quad \alpha_n(z) := \sqrt{n}(\hat{G}_n(z) - z) \ \ \forall z \in [0,1] \,. \tag{10}$$

Dann sind die Folgen von zufälligen Verteilungsfunktionen $\hat{F}_n$ und $\hat{G}_n(F)$ auf dem Intervall (-∞,∞) in Verteilung gleich.

$$\hat{F}_n(x) \cong \hat{G}_n(F(x))\,. \tag{11}$$

Und bzgl. des Empirischen Prozesses

$$\beta_n(x) \cong \alpha_n(F(x))\,. \tag{12}$$

Beweis.
$n\,\hat{F}_n \sim Bin(n, F(x))$.
Nun gilt für $\hat{G}_n(F)$

$$n\,\hat{G}_n(F(x)) = \sum_{i=1}^{n} \mathbb{I}_{[\xi_i \le F(x)]}\,. \tag{13}$$

Mit

$$p = \mathbb{P}\left[\mathbb{I}_{[\xi_i \le F(x)]} = 1\right] = \mathbb{P}\left[\xi_i \le F(x)\right] = F(x)$$

folgt $n\,\hat{G}_n(F) \sim Bin(n, F(x))$.
Daraus folgt, dass beide Folgen von Zufallsfunktionen die gleiche Verteilung haben. $\square$

Eigenschaften

Satz

Für eine Verteilungsfunktion F gilt

$$F \circ F^{-1}(t) \geq t \quad \forall t \in [0,1] \ . \tag{14}$$

Wobei die Gleichheit nicht erfüllt ist, wenn t nicht von F angenommen wird.

Satz (über die Dichte der Inversen)

Sei F eine Verteilungsfunktion mit positiver, stetiger Dichte definiert in der Umgebung von $F^{-1}(t_0)$ mit $0 < t_0 < 1$, dann ist F^{-1} in t_0 differenzierbar und es gilt

$$\frac{\partial F^{-1}}{\partial t_0}(t_0) = \frac{1}{f(F^{-1}(t_0))} \qquad \forall t_0 \in (0,1)\,. \tag{15}$$

Beweis.
Nach Definition gilt

$$\begin{aligned}
\frac{\partial F^{-1}}{\partial t_0}(t_0) &= \lim_{t \to t_0} \frac{F^{-1}(t_0) - F^{-1}(t)}{t_0 - t} \\
&= \lim_{x \to x_0} \frac{x_0 - x}{F(x_0) - F(x)} \\
&= \lim_{x \to x_0} \frac{1}{\frac{F(x_0) - F(x)}{x_0 - x}} \\
&= \frac{1}{f(x_0)}
\end{aligned}$$

Mit $x_0 = F^{-1}(t_0)$ folgt

$$\frac{1}{f(x_0)} = \frac{1}{f(F^{-1}(t_0))}\,.$$

$\square$

Theorem (von Skorokhod)

Seien $F_1, F_2, ...$ und F_0 Verteilungsfunktionen mit

$$F_n \to_d F_0 \qquad n \to \infty\,. \tag{16}$$

Seien

$$X_n := F_n^{-1}(\xi) \qquad n \geq 0\,, \tag{17}$$

wobei $\xi \sim U(0,1)$ bel. aber fest, s.d. $X_n \sim F_n$.
Dann gilt nach dem Satz von Skorokhod

$$X_n \to_{f.s.} X_0 \qquad n \to \infty\,. \tag{18}$$

Literatur

G.R. Shorak, J.A. Wellner.
Empirical Processes With Applications to Statistics.
Wiley, 1986.