Angewandte Stochastik

Prof. Dr. Evgeny Spodarev | Vorlesungskurs |

9. Thema

Heutiges Thema

- Eigenschaften der Ordnungsstatistiken und empirische Verteilungsfunktion

Eigenschaften der Ordnungsstatistiken

- Wir haben bereits die Ordnungsstatistiken $x_{(1)}, \ldots, x_{(n)}$ einer konkreten Stichprobe $(x_1, \ldots, x_n)$ betrachtet.
- Wenn wir nun auf der Modellebene arbeiten, also eine Zufallsstichprobe $(X_1, \ldots, X_n)$ von unabhängigen identisch verteilten Zufallsvariablen X_i mit Verteilungsfunktion $F(x)$ haben, welche Eigenschaften haben dann ihre Ordnungsstatistiken

$$X_{(1)}, \ldots, X_{(n)}\,?$$

Satz

1. Die Verteilungsfunktion der Ordnungsstatistik $X_{(i)}\,, \quad i = 1, \ldots, n$ ist gegeben durch

$$P\left(X_{(i)} \leq x\right) = \sum_{k=i}^{n} \binom{n}{k} F^k(x)(1 - F(x))^{n-k}\,, \qquad x \in \mathbb{R}\,.$$

2. Falls die X_i eine diskrete Verteilung besitzen, deren Wertebereich gegeben ist durch $E = \{\ldots, a_{j-1}, a_j, a_{j+1}, \ldots\}$, $i = 1, \ldots, n$, $a_i < a_j$ für $i < j$, dann gilt für die Zähldichte von $X_{(i)}$, $i = 1, \ldots, n$:

$$P(X_{(i)} = a_j) = \sum_{k=i}^{n} \binom{n}{k} \Big(F^k(a_j)\big(1 - F(a_j)\big)^{n-k} - F^k(a_{j-1})\big(1 - F(a_{j-1})\big)^{n-k}\Big)\,,$$

wobei

$$F(a_j) = \sum_{a_k \in E,\, k \leq j} P(X_i = a_k)\,.$$

3. Falls die X_i absolut stetig verteilt sind mit Dichte f, die stückweise stetig ist, dann ist auch $X_{(i)}$, $i = 1, \ldots, n$ absolut stetig verteilt mit der Dichte

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f(x) F^{i-1}(x) (1 - F(x))^{n-i}\,, \quad x \in \mathbb{R}\,.$$

Bemerkung

1. Für $i = 1$ und $i = n$ sieht die Formel 1 besonders einfach aus:

$$F_{X_{(1)}}(x) = 1 - (1 - F(x))^n, \qquad x \in \mathbb{R}$$

$$F_{X_{(n)}}(x) = F^n(x), \qquad x \in \mathbb{R}.$$

Diese Formeln lassen sich auch direkt herleiten:

$$\begin{aligned}
F_{X_{(1)}}(x) &= P(\min_{i=1,\dots,n} X_i \le x) = 1 - P(\min_{i=1,\dots,n} X_i > x) \\
&= 1 - P(X_i > x, \quad \forall i = 1, \dots, n) \\
&\underset{X_i \text{ uiv}}{=} 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - (1 - F(x))^n,
\end{aligned}$$

$$\begin{aligned}
F_{X_{(n)}}(x) &= P(\max_{i=1,\dots,n} X_i \le x) = P(X_i \le x, \quad \forall i = 1, \dots, n) \\
&\underset{X_i \text{ uiv}}{=} \prod_{i=1}^{n} P(X_i \le x) = F^n(x), \qquad x \in \mathbb{R}.
\end{aligned}$$

Bemerkung

2. Falls X_i absolut stetig verteilt sind mit einer stückweise stetigen Dichte f, so lassen sich Formeln für die gemeinsame Dichte der Verteilung von $(X_{(i_1)}, \ldots, X_{(i_k)})\,, \quad i \leq k \leq n$ herleiten.

   Insbesondere gilt mit $k = n$ für die Dichte

$$
\begin{aligned}
&f_{(X_{(1)},\ldots,X_{(n)})}(x_1, \ldots, x_n) \\
&= \begin{cases} n! \cdot f(x_1) \cdot \ldots \cdot f(x_n)\,, & \text{falls } -\infty < x_1 < \ldots < x_n < \infty\,, \\ 0\,, & \text{sonst.} \end{cases}
\end{aligned}
$$

Beispiel

Seien $X_1, \ldots, X_n$ unabhängig identisch verteilt, $X_i \sim U[0, \theta]$, $\theta > 0$, $i = 1, \ldots, n$, dann gilt:

1. die Dichte von $X_{(i)}$ ist gleich

$$f_{X_{(i)}}(x) = \begin{cases} \frac{n!}{(i-1)!(n-i)!}\theta^{-n}x^{i-1}(\theta - x)^{n-i}\,, & x \in (0, \theta) \\ 0\,, & \text{sonst} \end{cases}$$

   und

2.

$$\mathrm{E}X_{(i)}^k = \frac{\theta^k n!(i+k-1)!}{(n+k)!(i-1)!}\,, \qquad k \in \mathbb{N}\,, \quad i = 1, \ldots, n.$$

Insbesondere gilt $\mathrm{E}X_{(i)} = \frac{i}{n+1}\theta$ und $\mathrm{Var}\ X_{(i)} = \frac{i(n-i+1)\theta^2}{(n+1)^2(n+2)}$.

Empirische Verteilungsfunktion

Im Folgenden betrachten wir die statistischen Eigenschaften der empirischen Verteilungsfunktion $\hat{F}_n(x)$ einer Zufallsstichprobe $(X_1, \ldots, X_n)$, wobei $X_i \stackrel{d}{=} X$ unabhängige identisch verteilte Zufallsvariablen mit Verteilungsfunktion $F(\cdot)$ sind.

Satz

Es gilt

1. $n\hat{F}_n(x) \sim Bin(n, F(x))\,, \quad x \in \mathbb{R}$.
2. $\hat{F}_n(x)$ ist ein erwartungstreuer Schätzer für $F(x)$, $x \in \mathbb{R}$ mit
$$\mathrm{Var}\ \hat{F}_n(x) = \frac{F(x)(1-F(x))}{n}\,.$$
3. $\hat{F}_n(x)$ ist stark konsistent.
4. $\hat{F}_n(x)$ ist asymptotisch normalverteilt:
$$\sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{d} Y \sim N(0,1)\,, \qquad \forall x:\ F(x) \in (0,1)\,.$$

- In Satz wird behauptet, dass

$$\hat{F}_n(x) \xrightarrow[n\to\infty]{\text{f.s.}} F(x), \qquad \forall x \in \mathbb{R}.$$

- Der nachfolgende Satz von Gliwenko-Cantelli behauptet, dass diese Konvergenz gleichmäßig in $x \in \mathbb{R}$ stattfindet.
- Um diesen Satz formulieren zu können, betrachten wir den *gleichmäßigen Abstand* zwischen $\hat{F}_n$ und F

$$D_n = \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right|.$$

- Dieser Abstand ist eine Zufallsvariable, die auch *Kolmogorow-Abstand* genannt wird.
- Er gibt den maximalen Fehler an, den man bei der Schätzung von $F(x)$ durch $\hat{F}_n(x)$ macht.

Lemma

Es gilt

$$D_n = \max_{i \in \{1,\ldots,n\}} \max \left\{ F\left(X_{(i)} - 0\right) - \frac{i-1}{n}, \frac{i}{n} - F\left(X_{(i)}\right) \right\},$$

weil $\hat{F}_n(x)$ eine Treppenfunktion mit Sprungstellen $X_{(i)}$, $i = 1, \ldots, n$ ist.

Satz

- (Gliwenko-Cantelli): Es gilt

$$D_n \xrightarrow[n\to\infty]{\text{f.s.}} 0.$$

- Für jede stetige Verteilungsfunktion F gilt

$$D_n \stackrel{d}{=} \sup_{y\in[0,1]} \left| \hat{G}_n(y) - y \right| ,$$

wobei

$$\hat{G}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y) , \quad y \in \mathbb{R}$$

die empirische Verteilungsfunktion der Zufallsstichprobe $(Y_1, \ldots, Y_n)$ mit unabhängigen identisch verteilten Zufallsvariablen $Y_i \sim U[0,1]$, $\quad i = 1, \ldots, n$ ist.

Bemerkung

Nach Lemma gilt, dass D_n für stetige F simuliert werden kann, z.B. für

$$D_n \stackrel{d}{=} \max_{i \in \{1,\dots,n\}} \max\left\{ \frac{i}{n} - U_{(i)}, -\frac{i-1}{n} + U_{(i)} \right\}$$

mit $U_{(1)} \leq \cdots \leq U_{(n)}$ Ordnungsstatistiken der Stichprobe $U_1, \dots U_n$, wobei $U_i \sim U[0,1]$ unabhängig.

Daraus können wir für jedes n empirische Quantile der Verteilung von D_n berechnen.

Monte-Carlo-Test für F

- Sei $(x_1, \ldots, x_n)$ eine konkrete Stichprobe.
- Teste, ob x_i als Realisierung von einer Zufallsvariable $X \sim F$ (F stetig) interpretiert werden kann.
- Gehe dabei wie folgt vor:

1. Berechne $d_n = \sup\limits_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$ basierend auf dem obigen Lemma.
2. Für gegebenes n berechne die empirischen Quantile des Kolmogorov-Abstandes D_n wie in obiger Bemerkung: $d_{n,\alpha}$, $\alpha \in (0,1)$.
3. Man testet

$$
\begin{aligned}
&H_0 : X_i \sim F \,,\; i = 1, \ldots, n \text{ (Nullhypothese)} \\
&\quad \text{vs.} \\
&H_1 : X_i \nsim F \,,\; i = 1, \ldots, n \text{ (Alternativhypothese)} \;.
\end{aligned}
$$

Monte-Carlo-Test für F

4. Hierbei gilt folgende Testregel: Falls $d_n \notin \left[d_{n,\frac{\alpha}{2}}, d_{n,1-\frac{\alpha}{2}}\right]$, für α klein, dann wird H_0 verworfen.
5. dabei gilt:

$$\begin{aligned} P\,(H_0 \text{ wird abgelehnt } | H_0) &= P_{H_0}\left(d_n \notin \left[d_{n,\frac{\alpha}{2}},\ d_{n,1-\frac{\alpha}{2}}\right]\right) \\ &= 1 - P_{H_0}\left(d_n \in \left[d_{n,\frac{\alpha}{2}},\ d_{n,1-\frac{\alpha}{2}}\right]\right) \\ &= 1 - F_{D_n}(d_{n,1-\frac{\alpha}{2}}) + F_{D_n}(d_{n,\frac{\alpha}{2}}) \\ &= 1 - (1 - \frac{\alpha}{2}) + \frac{\alpha}{2} = \alpha. \end{aligned}$$

Satz (Kolmogorow-Smirnow)

Falls die Verteilungsfunktion F der unabhängigen und identisch verteilten Stichprobenvariablen X_i, $i = 1, \ldots, n$ stetig ist, dann gilt

$$\sqrt{n} D_n \xrightarrow[n\to\infty]{d} Y \,,$$

wobei Y eine Zufallsvariable mit der Verteilungsfunktion

$$K(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2x^2} \,, & x > 0 \,, \\ 0 \,, & \text{sonst} \end{cases}$$

(Kolmogorow-Verteilung) ist.

Kolmogorow-Smirnow-Anpassungstest

- Mit Hilfe des letzten Satzes ist es möglich, folgenden *asymptotischen Anpassungstest von Komogorow-Smirnow* zu entwickeln:
- Es wird die Haupthypothese
  • $H_0 : F = F_0$ (die unbekannte Verteilungsfunktion der Stichprobenvariablen $X_1, \ldots, X_n$ ist gleich F_0, F_0–stetig)
  gegen die Alternative
  • $H_1 : F \neq F_0$ getestet.
- Dabei wird H_0 verworfen, falls

$$\sqrt{n} D_n \notin [k_{\frac{\alpha}{2}} , k_{1-\frac{\alpha}{2}}]$$

ist, wobei

$$D_n = \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right|$$

und k_α das α-Quantil der Kolmogorow-Verteilung ist.

Kolmogorow-Smirnow-Anpassungstest

- Somit ist die Wahrscheinlichkeit, die richtige Hypothese H_0 zu verwerfen (Wahrscheinlichkeit des *Fehlers 1. Art*) asymptotisch gleich

$$P\left(\sqrt{n}D_n \notin [k_{\frac{\alpha}{2}}\,,\, k_{1-\frac{\alpha}{2}}] \,|\, H_0\right) \underset{n\to\infty}{\longrightarrow} 1-K(k_{1-\frac{\alpha}{2}})+K(k_{\frac{\alpha}{2}}) = \alpha\,.$$

- In der Praxis wird α klein gewählt, z.B. $\alpha \approx 0,05$.
- Somit ist im Fall, dass H_0 stimmt, die Wahrscheinlichkeit einer Fehlentscheidung in Folge des Testens klein.
- Dieser Test ist nur ein Beispiel dessen, wie der Satz von Kolmogorow in der statistischen Testtheorie verwendet wird.