5.4 Der Kolmogorov–Smirnov Test

Grundlage für den **Kolmogorov–Smirnov Anpassungs–Test** *ist ein Satz von* KOLMOGOROV, *die* **asymptotische Verteilung einer Statistik Δ_n betreffend**. **Aus Δ_n** *ergibt sich* **durch Modifikation** *die* **Kolmogorov–Smirnov Prüfstatistik**, *die dem* **Kolmogorov–Smirnov Test** *zu Grunde liegt.* *Der Abschnitt schließt mit einem Sachverhalt zur zweckmäßigen Berechnung der KS–Statistik zu einer gegebenen Stichprobenrealisation.* **Der KS–Test** *ist im* **Gegensatz zum χ^2–Anpassungstest** *nur auf* **stetige W–Maße anwendbar**. *Überblick*

5.4.1 Das statistische Modell

Sei das statistische Modell aus 2.2.1 verabredet, wonach $(\mathbb{H}, \mathcal{H}, P) = (\times_1^\infty \mathbb{H}_j, \otimes_1^\infty \mathcal{H}_j, \otimes_1^\infty P_0) = (\mathbb{H}_0^\infty, \mathcal{H}_0^\infty, P_0^\infty)$ die abzählbare Potenz des W–Raumes $(\mathbb{H}_0, \mathcal{H}_0, P_0) = (\mathbb{R}, \mathcal{B}, P_0)$ ist. Die Verteilungsfunktion F_0 des W–Maßes P_0 wird dabei als stetig unterstellt.

Die Stichprobenvariablen der Stichprobe $X := (X_1, X_2, \ldots)$ unter P_0^∞ werden wiederum als die Projektionen $X_j : \mathbb{H} \to \mathbb{H}_j = \mathbb{H}_0$, $j \in \mathbb{N}$, definiert. Damit erweist sich die Stichprobe X als einfach, d.h. die Stichprobenvariablen X_j, $j \in \mathbb{N}$, sind unabhängig und identisch verteilt:

$$P_{X_j} = (P_0^\infty)_{X_j} = P_0 \ , \ j \in \mathbb{N}.$$

Im Folgenden bezeichnet $x^n = (x_1, \ldots, x_n) \in \mathbb{H}_0^n$ eine Realisation der einfachen Stichprobe $X^n = (X_1, \ldots, X_n) : \mathbb{H} \to \times_1^n \mathbb{H}_j = \mathbb{H}_0^n$.

5.4.2 Die empirische Verteilungsfunktion als Ausgangspunkt

Sei $Q_{0,n}$ die gemäß 2.2.2(1) definierte empirische Verteilung der Stichprobe X^n unter P_0^n und $F_{0,n}$ deren empirisch Verteilungsfunktion, vgl. 2.2.2(2).

$Q_{0,n}$ entspricht dem in 2.2.1(i) eingeführten W–Maß $Q_{0,n}^{x^n}$, während $F_{0,n}$ der in 2.2.1(ii) eingeführten Verteilungsfunktion $F_{0,n}^{x^n}$ entspricht.

$$Q_{0,n}^{x^n}(B) := \frac{1}{n} \left| \{ j \in \mathbb{N} |\, x_j \in B \} \right| = \frac{1}{n} \sum_{j=1}^{n} 1_B(x_j) \ , \ B \in \mathcal{B}, \qquad \text{(i)}$$

bzw.

$$F_{0,n}^{x^n}(t) = \frac{1}{n} \left| \{ j \in \mathbb{N} |\, x_j \leq t \} \right| = \frac{1}{n} \sum_{j=1}^{n} 1_{(\infty;t]}(x_j) \ , \ t \in \mathbb{R}. \qquad \text{(ii)}$$

Der Zusammenhang zwischen $Q_{0,n}$ und $Q_{0,n}^{x^n}$ bzw. $F_{0,n}$ und $F_{0,n}^{x^n}$ ergibt sich aufgrund von 2.2.2(5).

Grundlage für den Kolmogorov–Smirnov Anpassungtest ist der als Satz 5.4.4 wiedergegebene Sachverhalt. Dazu wird eine spezielle Statistik Δ betrachtet, aus der sich durch Modifikation die Kolmogorov–Smirnov Statistik ergibt.

5.4.3 Die Statistik Δ_n

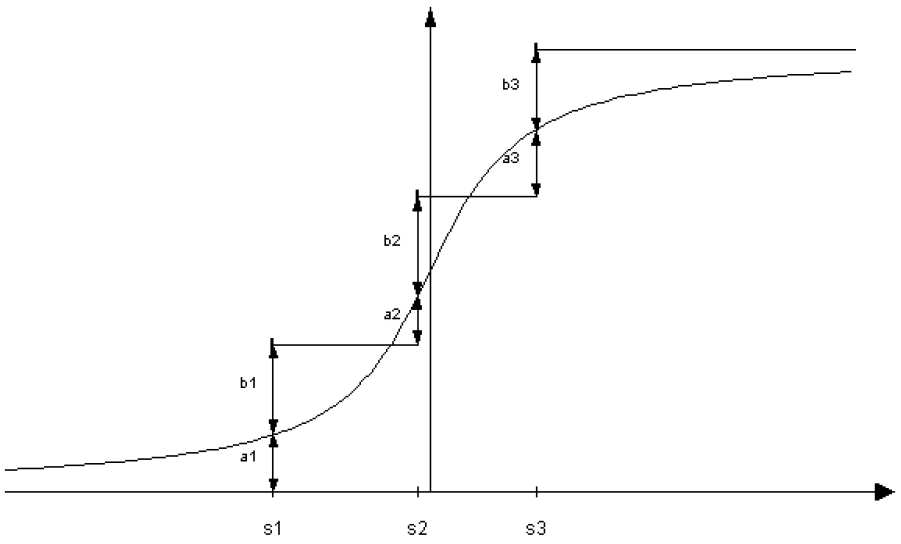
Seien 5.4.1 und 5.4.2 mit $P = P_0^\infty$ verabredet, weiter sei $n \in \mathbb{N}$.

Mit Blick auf Satz 5.4.4 wird die Statistik

$$\begin{gathered} \Delta_n : \mathbb{H} \to \mathbb{R}_+ \\ \Delta_n(x) := \sup |F_{0,n}(t) - F_0(t)| \end{gathered} \tag{iii}$$

eingeführt.

Die nachfolgende Abbildung liefert eine Veranschaulichung der Situation:



Als Ausgangspunkt: Die Graphen von F_0 bzw. $F_{0,3}$

Offenbar gilt

$$\Delta_3(x) = \max(a_1, a_2, a_3, b_1, b_2, b_3) \tag{iv}$$

bzw.

$$\Delta_n(x) = \sup(a_j, b_j | j \in \mathbb{N}_n), \tag{v}$$

wobei wir die (unmittelbar verständlichen) Zahlen $a_j, b_j,\ j \in \mathbb{N}_3$ als definiert betrachten.

Die folgende **asymptotische Aussage** über Δ_n geht auf A.N. KOLMOGOROV; (KOLMOGOROV, A.N. (1933): Sulla determinazione empirica di una legge di distribuzione; Giorn. Ist. Ital. Attuari 4, 83-91); (Attuari = Versicherungsmathematiker) zurück.

5.4.4 Satz

Seien 5.4.1 und 5.4.2 verabredet, mit $P = P_0^n$. Für $n \in \mathbb{N}$ sei Δ_n die in 5.4.3 eingeführte Statistik

$$\Delta_n : \mathbb{H} \to \mathbb{R}_+.$$

Dann konvergiert die Folge $(\sqrt{n}\Delta_n)_{n=1}^{\infty}$ in Verteilung gegen eine Verteilung H mit der Verteilungsfunktion $G : \mathbb{R} \to [0;1]$,

$$G(t) := \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2t^2) & \text{für} \quad t > 0 \\ 0 & \text{sonst} \end{cases};$$

d.h. es gilt

$$\lim_{n\to\infty} P\left(\{x \in \mathbb{H}_0^{\infty} | \sqrt{n}\Delta_n(x) \le t\}\right) = G(t)\ ,\ t \in \mathbb{R}.$$

Man beachtet, dass die **asymptotische Verteilung H von $\sqrt{n}\Delta_n$ von der Spezifikation des zugrunde gelegten W–Maßes P_0 unabhängig** ist.

5.4.5 Ein Testproblem

Vorgelegt ist ein W–Maß $\overline{P}$ mit **stetiger** Verteilungsfunktion $\overline{F}$.
Anhand der Realisation $x^n = (x_1, \ldots, x_n)$ der einfachen Stichprobe $X^n = (X_1, \ldots, X_n)$ (unter dem unbekannten W–Maß P_0 mit ebenfalls **stetiger** Verteilungsfunktion, das hier die **Rolle der wahren Verteilung** übernimmt), soll geprüft werden, ob $\overline{P} = P_0$, also ob $\overline{P}$ die wahre Verteilung ist.

Zur systematischen Behandlung des formulierten Testproblems betrachtet

man das **Testexperiment**

$$\left.\begin{array}{lcl} & & (\mathbb{H}, \mathcal{H}, \mathcal{W}, \mathcal{W}_1, \mathcal{W}_2) \\ \text{mit} & & \\ \mathcal{W} & = & \{P^n | P \text{ ist ein stetiges W–Maß über } (\mathbb{R}, \mathcal{B})\} \\ \mathcal{W}_1 & := & \{\overline{P}^n\}, \\ \mathcal{W}_2 & := & \mathcal{W} \setminus \mathcal{W}_1. \end{array}\right\} \quad \text{(i)}$$

Die Stichprobe X^n ist gemäß dem unbekannten W–Maß P_0^n (wahre Verteilung) verteilt.

Bei dem formulierten Testproblem geht es um **dieselbe Fragestellung**, wie bei dem in 5.3 behandelten χ^2–Anpassungstest, wobei hier eine **Einschränkung von $\mathcal{W}$ auf Potenzen von stetigen W–Maßen** über $(\mathbb{R}, \mathcal{B})$ vorgenommen wird. Die statistische Fragestellung ist, wie man aufgrund des Testexperimentes (i) erkennt eine solche der **nicht–parametrischen** Statistik.

5.4.6 Entwicklung eines Testkonzeptes auf der Grundlage der KS–Statistik

Die Prüfung der Frage, ob die Hypothese zutrifft, ob die, die Hypothese definierende Verteilung $\overline{P}$ gleich der als wahren Verteilung auftretenden Verteilung P_0 ist, heißt letztendlich die beiden Verteilungen — wie auch immer — miteinander zu vergleichen.

Die **wahre Verteilung P_0 offenbart sich über die Stichprobenrealisationen, während die Verteilung $\overline{P}$ spezifiziert** ist.

Durch Modifikation ergibt sich auf der in 5.4.3 eingeführten Statistik Δ_n die sogenannte **Kolmogorov–Smirnov Prüfstatistik**

$$\begin{array}{l} D_n : \mathbb{H} \to \mathbb{R}_+ \\ D_n(x) := \sup_{t \in \mathbb{R}} |F_{0,n}(t) - \overline{F}(t)|, \end{array} \quad (1)$$

wobei $\overline{F}$ die Verteilungsfunktion des die Hypothese definierenden W–Maßes $\overline{P}$ und $F_{0,n}$ die empirische Verteilungsfunktion von P_0 bei n Stichprobenrealisationen ist.

Ersetzt man im Rechtsterm von D_n den Ausdruck $\overline{F}(t)$ durch $F_0(t)$, so ist man auf die Statistik Δ_n aus 5.4.3(i) zurückgeführt.

Die KS–Statistik D_n nimmt stets positive Werte an; die Aussage gilt ebenso für $\sqrt{n}D_n$.

Trifft die Hypothese zu, d.h. ist die in 5.4.6, die Hypothese definierende Verteilung $\overline{F}$ gleich der wahren Verteilung P_0, so gilt $D_n = \Delta_n$. Dabei strebt $D_n = \Delta_n$ nach dem Satz von GLIVENKO–CANTELLI gegen null.

Große positive Werte von $\sqrt{n}D_n(x)$ sind daher ein Indiz für das Abweichen von $\overline{P}$ von P_0, kleine positive Werte von $\sqrt{n}D_n(x)$ stützen die Hypothese, wonach $\overline{P}$ gleich der wahren Verteilung P_0 ist.

Für den hier zu testenden Fall, dass die Hypothese zutrifft, d.h. dass $P_0 = \overline{P}$ und somit $\sqrt{n}D_n = \sqrt{n}\Delta_n$ gilt, ist die asymptotische Verteilung H von $\sqrt{n}D_n = \sqrt{n}\Delta_n$ bekannt, vgl. Satz 5.4.5. Diese Verteilung wird als Näherung für die Verteilung von $\sqrt{n}D_n = \sqrt{n}\Delta_n$ benutzt.

Das Gesagte führt zu folgendem Test.

5.4.7 Der Test

Sei $D_n(x)$ eine Realisation der KS–Prüfstatistik D_n und $\mathrm{fr}_{(1-\alpha)}(H)$ das $(1-\alpha)$–Fraktil der asymptotischen Verteilung H für $\alpha \in (0;1)$, z.B. für $\alpha = 0,05$, so wird die Hypothese **verworfen**, falls

$$\sqrt{n}D_n(x) > \mathrm{fr}_H(1-\alpha);$$

während sie für

$$\sqrt{n}D_n(x) \leq \mathrm{fr}_H(1-\alpha);$$

nicht verworfen wird.

Ein Test, der auf der Statistik $\sqrt{n}\Delta_n$ und ihrer asymptotischen Verteilung H basiert, heißt **Kolmogorov-Smirnov (Anpassungs–)Test**.

5.4.8 Bemerkung

Die Verteilungsfunktion G von H ist, vgl. Satz 5.4.4, über eine Reihe definiert, so dass man sich bei der numerischen Bestimmung von $\mathrm{fr}_H(1-\alpha)$ zweckmäßigerweise bereits berechneter, d.h. vertafelter Werte bedient.

5.4.9 Zur effizienten Bestimmung der KS–Prüfstatistik

Seien 5.4.1 und 5.4.2 verabredet mit P_0 als wahrer Verteilung. Sei $\overline{P}$ das W–Maß, welches in 5.4.5(i) die Hypothese definiert; $\overline{F}$ bezeichne die (stetige) Verteilungsfunktion von $\overline{P}$.

Sei $x^n = (x_1, \ldots, x_n) \in \mathbb{H}_0^n$ eine Realisation der einfachen Stichprobe $X^n : \mathbb{H} \to \mathbb{H}_0^n$ vom Umfang n, wobei X_j gemäß der Verteilung P_0 über $(\mathbb{H}_0, \mathcal{H}) = (\mathbb{R}, \mathcal{B})$ verteilt ist.

Eine Schwierigkeit, der man bei der Bestimmung des Wertes der KS–Prüfstatistik $D_n(x)$ begegnet, besteht darin, dass $D_n(x)$ formal als das Supremum einer unendlichen Menge definiert ist:

$$D_n(x) = \sup_{t \in \mathbb{R}} |F_{0,n}^{x^n}(t) - \overline{F}(t)|. \tag{i}$$

Im Folgenden wird aufgezeigt, wie man dieses Supremum in (i) durch ein Maximum einer endlichen Menge ersetzen kann.

Dazu benötigt man eine modifizierte Darstellung von $F_{0,n}^{x^n}$, die im folgenden Lemma angegeben wird:

5.4.10 Lemma

Sei $x^n = (x_1, \ldots, x_n) \in \mathbb{H}_0^n$ eine Realisation der einfachen Stichprobe $X^n : \mathbb{H} \to \mathbb{H}_0^n$. Da die Verteilung von X_j stetig ist, $j \in \mathbb{N}_n$, sind $x_1, \ldots, x_n$ mit Wahrscheinlichkeit 1 paarweise verschieden; sei

$$x_{(1)} < \ldots < x_{(n)}$$

die Anordnung von x^n gemäß "¡".
Dann gilt:

$$F_{0,n}^{x^n}(t) = \begin{cases} 0 & \text{falls} \quad t < x_{(1)} \\ \frac{k}{n} & \text{falls} \quad x_{(k)} \le t < x_{(k+1)} \ , \ k \in \mathbb{N}_{n-1} \\ 1 & \text{falls} \quad x_{(n)} \le t. \end{cases}$$

Beweis:

Ist $t < x_{(1)}$, dann ist die Anzahl der Realisationen x_j mit $x_j \le t$ gleich 0. Ist $x_{(k)} \le t < x_{(k+1)}$ für ein $k \in \mathbb{N}_{n-1}$, dann entspricht die Anzahl der Realisationen $x_j \le t$ der Anzahl der gemäß "¡" angeordneten $x_j \le t$; diese

sind aber $x_{(1)} < \ldots < x_{(k)}$, so dass die gesuchte Anzahl gleich k ist. Ist $x_{(n)} \leq t$, dann ist

$$|\{j \in \mathbb{N}_n | \, x_j \leq t\}| = n.$$

□

Das folgende Lemma liefert die angekündigte Darstellung der KS–Prüfstatistik D_n als das Maximum einer endlichen Menge von reellen Werten. Die Verwendung dieser Darstellung bietet sich stets bei Verwendung des KS–Tests an.

5.4.11 Lemma

Sei $x^n = (x_1, \ldots, x_n) \in \mathbb{H}_0^n$ eine Realisation der einfachen Stichprobe $X^n : \mathbb{H} \to \mathbb{H}_0^n$, die gemäß P_0^n verteilt ist. Sei

$$x_{(1)} < \ldots < x_{(n)},$$

die Anordnung von $x^n = (x_1, \ldots, x_n)$ gemäß "¡". Sei weiter

$$x_{(0)} := -\infty \qquad , \qquad x_{(n+1)} := \infty$$

und

$$F_{0,n}^{x^n}(\infty) := \overline{F}(-\infty) := 0 \quad F_{0,n}^{x^n}(\infty) := \overline{F}(\infty) := 1,$$

wobei $\overline{F}$ die stetige Verteilungsfunktion des die Hypothese definierenden W–Maßes $\overline{P}$ bezeichnet. Die KS–Prüfstatistik $D_n(x)$ sei gemäß 5.4.6(i) definiert. Es gilt:

$$D_n(x) = \max_{k=1}^{n} \max \left\{ \frac{k}{n} - \overline{F}(x_{(k)}), \overline{F}(x_{(k+1)}) - \frac{k}{n} \right\}.$$

Beweis:

Wegen der Stetigkeit und der Monotonie von $\overline{F}$ gilt nach Lemma 5.4.10

$$\begin{aligned} & \sup_{x_{(k)} \leq t < x_{(k+1)}} |F_{0,n} x^n - \overline{F}(t)| \\ = \; & \max \left\{ \left| \tfrac{k}{n} - \overline{F}(x_{(k)}), \tfrac{k}{n} - \overline{F}(x_{(k+1)}) \right| \right\}, \; n = 0, 1, \ldots, n; \end{aligned} \tag{i}$$

die getrennte Betrachtung der drei Fälle

(1) $\overline{F}(x_{(k+1)}) \leq \frac{k}{n}$

(2) $\overline{F}(x_{(k)}) \leq \frac{k}{n} < \overline{F}(x_{(k+1)})$

(3) $\frac{k}{n} < \overline{F}(x_{(k)})$

liefert

$$\begin{array}{rl} & \sup\limits_{x_{(k)} \leq t < x_{(k+1)}} |F_{0,n}^{x^n}(t) - \overline{F}(t)| \\ = & \max\left\{ \left| \frac{k}{n} - \overline{F}(x_{(k)}), \overline{F}(x_{(k+1)}) - \frac{k}{n} \right| \right\},\ n = 0, 1, \ldots, n. \end{array} \tag{ii}$$

Wegen

$$D_n(x) = \max_{k=0}^{n} \sup_{x_{(k)} \leq t < t(k+1)} |F_{0,n}^{x^n}(t) - \overline{F}(t)|$$

impliziert (i) das Gewünschte. □

5.4.12 Bemerkung

Man beachtet, dass die in der Abbildung in 5.4.3 illustrierte Situation durch den Eintritt von Fall 2 des Beweises zu 5.4.11 für die drei Sprungstellen s_1, s_2, s_3 gekennzeichnet werden kann. Insofern bezieht sich diese Abbildung im Kontext der in 5.4.6(i) eingeführten KS–Prüfstatistik D_n auf einen Spezialfall der Stichprobenrealisationen $x^n = (x_1, \ldots, x_n)$.

Selbstbeurteilung

Obwohl es sich beim KS–Test ebenfalls wie beim χ^2–Anpassungstest um einen Anpassungstest handelt, sind die Akzente in 5.3 und 5.4 — bedingt durch technische Fragen — etwas anders gesetzt.

Nehmen Sie sich die Zeit und vergleichen Sie die beiden Testkonzepte, wobei Sie die Entsprechungen exakt zu sehen versuchen.

Was ist der Unterschied von D_n und Δ_n? Unter welcher Maßgabe ist D_n gemäß H verteilt?