



Angewandte Stochastik II

Universität Ulm
Institut für Stochastik

Vorlesungsskript
Dr. Ole Stenzel
Stand: Wintersemester 2014/15

ULM, IM OKTOBER 2014

Contents

1	Grundideen der statistischen Datenanalyse	3
1.1	Zufallsstichprobe	3
1.2	Stichprobenmittel	5
1.3	Stichprobenvarianz	7
1.4	Empirische Verteilungsfunktion	9
1.5	Dichteschätzer	12
2	Schätzung von Parametern	14
2.1	Eigenschaften von Parameterschätzern	16
2.2	Momentenmethode	17
2.3	Maximum-Likelihood-Methode	20

1 Grundideen der statistischen Datenanalyse

Zu den Zielen der mathematischen Statistik gehört die Beschreibung und Untersuchung der Eigenschaften bzw. Gesetzmäßigkeiten von (großen) Datensätzen. Dabei kann es sich einerseits um sogenannte

- *reale Daten*

handeln, die sich z.B. bei der Beobachtung (Messung) von Vorgängen bzw. Strukturen in Natur, Technik oder Wirtschaft ergeben oder es können

- *synthetische Daten*

sein, die bei der Simulation solcher Vorgänge bzw. Strukturen durch Computeralgorithmen erzeugt werden.

Die grundlegende Idee der Statistik, um diese Zielstellung zu erreichen, ist die *stochastische Modellierung* der vorliegenden Daten.

Die Vorlesung nutzt die Begriffe und Ergebnisse der Wahrscheinlichkeitsrechnung, die in der vorhergehenden Vorlesung ‘Angewandte Stochastik I’ diskutiert wurden, wie:

- Ereignis und Wahrscheinlichkeit
- Zufallsvariable und Verteilung
- Erwartungswert und Varianz
- stochastische Unabhängigkeit
- Gesetz der großen Zahlen
- zentraler Grenzwertsatz

1.1 Zufallsstichprobe

Der Vektor der vorliegenden Daten $(x_1, \dots, x_n)$ kann natürlich im Allgemeinen eine komplizierte Struktur aufweisen.

- Dabei muss der ‘Wert’ x_i nicht unbedingt eine Zahl sein, sondern x_i kann für jedes $i = 1, \dots, n$ selbst ein Vektor sein, der beispielsweise die Lage, Größe, Form und Orientierung eines geometrischen Objektes beschreiben kann.
- Im Rahmen dieser einführenden Vorlesung setzen wir jedoch stets voraus, dass - falls nicht anders angegeben - $x_i \in \mathbb{R}$ für jedes $i = 1, \dots, n$.

Wir nehmen an, dass die Daten $x_1, \dots, x_n$ die *Realisierung* eines stochastischen Modells sind.

- Und zwar sei $x_1, \dots, x_n$ die Realisierung einer Folge von unabhängigen und identisch verteilten Zufallsvariablen $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$, die über einem (im Allgemeinen nicht näher spezifizierten) Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$ definiert sind. Die Eigenschaft, dass die Zufallsvariablen $X_1, \dots, X_n$ unabhängig und identisch verteilt sind, wird oft mit ‘*iid*’ (independent and identically distributed) abgekürzt.
- Da $X_1, \dots, X_n$ unabhängig und identisch verteilt sind, folgt insbesondere, dass für ein $\omega \in \Omega$

$$X_i(\omega) = x_i \quad \forall i \in \{1, \dots, n\}.$$

- Außerdem setzen wir stets voraus, dass $\mathbb{E}(X_i^2) < \infty$ für $i = 1, \dots, n$.

Definition 1.1

1. Der Vektor $(x_1, \dots, x_n)$ heißt (konkrete) *Stichprobe*.
2. Die Menge $\mathbb{R}^n$ aller (potentiell möglichen) Stichproben $(x_1, \dots, x_n)$ heißt *Stichprobenraum*.
3. Der Zufallsvektor $(X_1, \dots, X_n)$ heißt *Zufallsstichprobe*.
4. Für jedes $i = 1, \dots, n$ heißt x_i *Stichprobenwert* von $(x_1, \dots, x_n)$. Analog hierzu nennt man X_i *Stichprobenvariable* von $(X_1, \dots, X_n)$.
5. Die Dimension n von $(x_1, \dots, x_n)$ bzw. $(X_1, \dots, X_n)$ heißt *Stichprobenumfang*.

Beachte

- Die allgemeine Zielstellung der statistischen Datenanalyse, die in der Einleitung diskutiert wurde, kann nun wie folgt präzisiert werden: Aus den vorliegenden Daten $x_1, \dots, x_n$ sollen Schlussfolgerungen über Eigenschaften der (unbekannten) Verteilung der Zufallsstichprobe $(X_1, \dots, X_n)$ gezogen werden.
- Weil wir voraussetzen, dass die Stichprobenvariablen $X_1, \dots, X_n$ unabhängig und identisch verteilt sind, wird die Verteilung von $(X_1, \dots, X_n)$ eindeutig durch die (Rand-) Verteilungsfunktion $F = F_{X_i}$ einer (einzelnen) Stichprobenvariablen bestimmt. Dieses ist leicht zu sehen, da für alle $t_1, \dots, t_n \in \mathbb{R}$ gilt:

$$F_{(X_1, \dots, X_n)}(t_1, \dots, t_n) \stackrel{\text{unabh.}}{=} F_{X_1}(t_1) \cdot \dots \cdot F_{X_n}(t_n) \stackrel{\text{ident. verteilt}}{=} F(t_1) \cdot \dots \cdot F(t_n).$$

- Aus den Daten $x_1, \dots, x_n$ sollen also Schlussfolgerungen über Eigenschaften der unbekannten Verteilungsfunktion F gezogen werden.

Beispiel

- Punktschätzer. Es soll anhand einer Umfrage von n Personen eine Wahlprognose für eine Stichwahl zwischen zwei Kandidaten A und B erstellt werden. Sei X die Anzahl derjenigen Personen, die Kandidat A wählen würden. Verteilungsannahme: $X \sim \text{Bin}(n, p)$. Gesucht ist eine Approximation für die unbekannte ‘Erfolgswahrscheinlichkeit’ p anhand einer konkreten Stichprobe $(x_1, \dots, x_n)$.

- Konfidenzintervall. Das Füllvolumen von 0,7l Wasserflaschen sei als normalverteilt angenommen. Sei $X = (X_1, \dots, X_n)$ eine entsprechende Zufallsstichprobe, d.h. $X_i \sim N(\mu, \sigma^2)$. Es soll ein anhand einer konkreten Stichprobe $(x_1, \dots, x_n)$ ein Intervall bestimmt werden, das die zufällige Füllmenge mit 95% Wahrscheinlichkeit enthält.
- Statistischer Test. Ein Hersteller von Lampen gibt eine durchschnittliche Lebensdauer von mindestens 1000 Stunden an. Diese Angabe soll anhand einer Realisierung $(x_1, \dots, x_n)$ einer Zufallsstichprobe $X = (X_1, \dots, X_n)$ untersucht werden. Es wird angenommen, dass die Lebensdauern exponentiell verteilt sind mit Mittelwert μ , d.h. $X_i \sim \text{Exp}(1/\mu)$. Anhand der Stichprobe $(x_1, \dots, x_n)$ soll überprüft werden, ob die erwartete Lebensdauer $E(X_i)$ größer gleich 1000 ist ($E(X_i) \geq 1000$).

Um Eigenschaften der Verteilungsfunktion F zu bestimmen, werden Funktionen $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ betrachtet, die der Stichprobe $(x_1, \dots, x_n)$ die ‘Bewertung’ $\varphi(x_1, \dots, x_n) \in \mathbb{R}$ zuordnen. Dies führt zu der folgenden Begriffsbildung.

Definition 1.2 Eine Borel-messbare Abbildung $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ heißt *Stichprobenfunktion*.

Beachte Es ist üblich, auch die zusammengesetzte Abbildung $\varphi(X_1, \dots, X_n) : \Omega \rightarrow \mathbb{R}$ mit $\varphi(X_1, \dots, X_n)(\omega) = \varphi(X_1(\omega), \dots, X_n(\omega))$ *Stichprobenfunktion* zu nennen, d.h., φ ist dann eine Funktion der Zufallsstichprobe $(X_1, \dots, X_n)$.

1.2 Stichprobenmittel

Wir diskutieren zunächst die Frage, wie der Erwartungswert $\mu = \mathbb{E} X_i$ der Stichprobenvariablen X_i aus den beobachteten Daten $x_1, \dots, x_n$ bestimmt werden kann.

- Hierfür betrachten wir die Stichprobenfunktion $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ mit $\varphi(x_1, \dots, x_n) = n^{-1} \sum_{i=1}^n x_i$.
- D.h., wir betrachten das arithmetische Mittel

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

der Stichprobenwerte $x_1, \dots, x_n$.

- Die Zahl $\bar{x}_n$ wird *Stichprobenmittel* der (konkreten) Stichprobe $(x_1, \dots, x_n)$ genannt.
- Außerdem betrachten wir das arithmetische Mittel der Stichprobenvariablen $X_1, \dots, X_n$, dessen Eigenschaften bereits im Zusammenhang mit dem Gesetz der großen Zahlen untersucht worden sind.

Definiton 1.3 Die Zufallsvariable

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

heißt *Stichprobenmittel* der Zufallsstichprobe $(X_1, \dots, X_n)$.

Theorem 1.4 Es gilt

$$\mathbb{E} \bar{X}_n = \mu \quad (3)$$

und

$$\text{Var} \bar{X}_n = \frac{\sigma^2}{n}, \quad (4)$$

wobei $\mu = \mathbb{E} X_i$ und $\sigma^2 = \text{Var} X_i$ den Erwartungswert bzw. die Varianz der Stichprobenvariablen X_i bezeichnen.

Beachte

- Wegen (3) sagt man, dass bei der Schätzung von μ durch $\bar{X}_n$ kein ‘systematischer Fehler’ begangen wird.
- Der Schätzer $\bar{X}_n$ kann dennoch sehr ungenau sein, wobei man den in (4) gegebenen Wert σ^2/n als Kennzahl für die Schätzgenauigkeit von $\bar{X}_n$ auffassen kann.
- Insbesondere bedeutet (4), dass die Schätzgenauigkeit mit wachsendem Stichprobenumfang n verbessert wird.
- Dabei ist σ^2 im allgemeinen ebenfalls unbekannt und muss aus den Daten geschätzt werden.

Neben den Formeln (3) und (4) für Erwartungswert und Varianz des Stichprobenmittels $\bar{X}_n$ sind noch weitere Aussagen über die Verteilung von $\bar{X}_n$ von Interesse bzw. über deren asymptotisches Verhalten für große n .

Aus dem starken Gesetz der großen Zahlen bzw. aus dem zentralen Grenzwertsatz ergibt sich unmittelbar

Theorem 1.5 Es gilt

$$P\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1 \quad (5)$$

und

$$\lim_{n \rightarrow \infty} P\left(\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq x\right) = \Phi(x) \quad (6)$$

für jedes $x \in \mathbb{R}$, wobei $\Phi(x)$ die Verteilungsfunktion der Standardnormalverteilung ist.

Beispiel

- Wir möchten die Wahrscheinlichkeit $P(|\bar{X}_n - \mu| > \varepsilon)$, dass das Stichprobenmittel $\bar{X}_n$ um mehr als einen vorgegebenen Schwellenwert $\varepsilon > 0$ von dem zu schätzenden Wert μ abweicht, approximieren. Es gilt
-

$$\begin{aligned} P(|\bar{X}_n - \mu| > \varepsilon) &= 1 - P(|\bar{X}_n - \mu| \leq \varepsilon) \\ &= 1 - P(-\varepsilon \leq \bar{X}_n - \mu \leq \varepsilon) \\ &= 1 - P\left(-\frac{\varepsilon\sqrt{n}}{\sigma} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq \frac{\varepsilon\sqrt{n}}{\sigma}\right) \\ &\stackrel{(6)}{\approx} 1 - \left(\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right) \\ &= 2\left(1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right). \end{aligned}$$

Es gilt also

$$P(|\bar{X}_n - \mu| > \varepsilon) \approx 2 \left(1 - \Phi \left(\frac{\varepsilon \sqrt{n}}{\sigma} \right) \right) \quad (7)$$

für große n .

- Falls σ^2 unbekannt ist, dann muß σ^2 (wie bereits oben erwähnt wurde) ebenfalls aus den vorliegenden Daten geschätzt werden, um mit Hilfe von (7) eine praktikable Näherungsformel für die Wahrscheinlichkeit $P(|\bar{X}_n - \mu| > \varepsilon)$ zu erhalten.

1.3 Stichprobenvarianz

Wir untersuchen nun die Frage, wie die Varianz $\sigma^2 = \text{Var } X_i$ der Stichprobenvariablen X_i aus den beobachteten Daten $x_1, \dots, x_n$ bestimmt werden kann. Dabei gehen wir ähnlich wie in Abschnitt 1.2 vor.

- Wir betrachten die Stichprobenfunktion $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ mit

$$\varphi(x_1, \dots, x_n) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2, \quad (8)$$

wobei $\bar{x}_n$ in (1) gegeben ist.

- Die in (8) eingeführte Größe wird *Stichprobenvarianz* der (konkreten) Stichprobe $(x_1, \dots, x_n)$ genannt und mit s_n^2 bezeichnet.

Definition 1.6 Die Zufallsvariable

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad (9)$$

heißt *Stichprobenvarianz* der Zufallsstichprobe $(X_1, \dots, X_n)$.

Theorem 1.7 Es gilt

$$\mathbb{E} S_n^2 = \sigma^2. \quad (10)$$

Falls $\mathbb{E}(X_i^4) < \infty$ für $i = 1, \dots, n$, dann gilt außerdem

$$\text{Var } S_n^2 = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right), \quad (11)$$

wobei $\mu_4 = \mathbb{E}(X_i^4)$ und $\sigma^4 = (\text{Var } X_i)^2$ das 4-te Moment bzw. die quadrierte Varianz der Stichprobenvariablen X_i bezeichnen.

Beweis

- Wir zeigen nur die Gültigkeit von (10). Die Gültigkeit von (11) ergibt sich durch ähnliche Überlegungen.

- Aus der Definitionsgleichung (9) von S_n^2 ergibt sich, dass

$$\begin{aligned}
 S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\
 &= \frac{1}{n-1} \sum_{i=1}^n (X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2) \\
 &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - 2n\bar{X}_n^2 + n\bar{X}_n^2 \right) \\
 &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right).
 \end{aligned}$$

- Wegen der Linearität des Erwartungswertes folgt hieraus, dass

$$\begin{aligned}
 \mathbb{E} S_n^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n \mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}_n^2) \right) \\
 &= \frac{1}{n-1} \left(\sum_{i=1}^n (\text{Var } X_i + (\mathbb{E} X_i)^2) - n(\text{Var } \bar{X}_n + (\mathbb{E} \bar{X}_n)^2) \right) \\
 &= \frac{1}{n-1} \left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \right) \\
 &= \sigma^2,
 \end{aligned}$$

wobei sich das vorletzte Gleichheitszeichen aus (3) und (4) ergibt.

- Damit ist (10) bewiesen.

Beachte

- Wegen (10) wird bei der Schätzung von σ^2 durch S_n^2 kein systematischer Fehler begangen.

Neben den Formeln (10) und (11) für Erwartungswert und Varianz der Stichprobenvarianz S_n^2 sind erneut weitere Aussagen über die Verteilung von S_n^2 bzw. über deren asymptotisches Verhalten für große n von Interesse.

Aus dem starken Gesetz der großen Zahlen bzw. aus dem zentralen Grenzwertsatz ergibt sich

Theorem 1.8 Es gilt

$$P\left(\lim_{n \rightarrow \infty} S_n^2 = \sigma^2\right) = 1. \quad (12)$$

Falls $\mathbb{E}(X_i^4) < \infty$ für $i = 1, \dots, n$, dann gilt außerdem

$$\lim_{n \rightarrow \infty} P\left(\sqrt{n} \frac{S_n^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \leq x\right) = \Phi(x) \quad (13)$$

für jedes $x \in \mathbb{R}$, wobei $\Phi(x)$ die Verteilungsfunktion der Standardnormalverteilung ist.

Beweis

- Wir zeigen nur die Gültigkeit von (12). Die Herleitung von (13) erfordert Hilfsmittel, die über den Rahmen dieser einführenden Vorlesung hinausgehen.
- Im Beweis von Theorem 1.7 hatten wir gezeigt, dass

$$S_n^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right). \quad (14)$$

- Außerdem kann man zeigen, dass die Unabhängigkeit und identische Verteiltheit der Stichprobenvariablen $X_1, \dots, X_n$ impliziert, dass auch die Zufallsvariablen $X_1^2, \dots, X_n^2$ unabhängig und identisch verteilt sind.
- Deshalb ergibt sich aus dem starken Gesetz der großen Zahlen, dass mit Wahrscheinlichkeit 1

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i^2 = \mathbb{E}(X_i^2)$$

und

$$\begin{aligned} \lim_{n \rightarrow \infty} \bar{X}_n^2 &= \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \\ &= \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i \right)^2 \\ &= (\mathbb{E} X_i)^2. \end{aligned}$$

- Hieraus und aus (14) ergibt sich (12).

1.4 Empirische Verteilungsfunktion

Außer der Schätzung von Erwartungswert μ und Varianz σ^2 der Stichprobenvariablen X_i kann auch deren Verteilungsfunktion F aus den vorliegenden Daten $x_1, \dots, x_n$ geschätzt werden.

- Hierfür betrachten wir für jedes $x \in \mathbb{R}$ die Stichprobenfunktion $\varphi_x : \mathbb{R}^n \rightarrow \mathbb{R}$ mit

$$\varphi_x(x_1, \dots, x_n) = \frac{1}{n} |\{i : 1 \leq i \leq n, x_i \leq x\}|,$$

wobei $|A|$ so wie bisher die Anzahl der Elemente der Menge A bezeichnet.

- D.h. $\varphi_x(x_1, \dots, x_n)$ ist die relative Häufigkeit derjenigen Stichprobenwerte, die den Schwellenwert x nicht überschreiten.

Beachte

- Man kann sich leicht überlegen, dass für jeden Vektor $(x_1, \dots, x_n) \in \mathbb{R}^n$ die Abbildung

$$x \rightarrow \varphi_x(x_1, \dots, x_n) \quad (15)$$

die Eigenschaften einer Verteilungsfunktion hat.

- Die in (15) gegebene Abbildung wird deshalb *empirische Verteilungsfunktion* der (konkreten) Stichprobe $(x_1, \dots, x_n)$ genannt.

Dies führt zu der folgenden Begriffsbildung.

Definition 1.9 Die Abbildung $\hat{F}_n : \mathbb{R} \times \Omega \rightarrow [0, 1]$ mit

$$\hat{F}_n(x, \omega) = \frac{1}{n} |\{i : 1 \leq i \leq n, X_i(\omega) \leq x\}| \quad (16)$$

heißt *empirische Verteilungsfunktion* der Zufallsstichprobe $(X_1, \dots, X_n)$.

Beachte

- Die in (16) gegebene Abbildung kann man als eine Familie $\{\hat{F}_n(x), x \in \mathbb{R}\}$ von Zufallsvariablen $\hat{F}_n(x) : \Omega \rightarrow [0, 1]$ auffassen.
- Eine solche Familie von Zufallsvariablen wird *empirischer Prozeß* genannt.
- Empirische Prozesse sind eine spezielle Klasse *stochastischer Prozesse*.

Theorem 1.10 Für jedes $x \in \mathbb{R}$ gilt:

1. Die Zufallsvariable $n\hat{F}_n(x)$ ist binomialverteilt mit den Parametern n und $p = F(x)$.
D.h., für $k = 0, 1, \dots, n$ gilt

$$P(n\hat{F}_n(x) = k) = \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}. \quad (17)$$

2. Insbesondere gilt also

$$\mathbb{E} \hat{F}_n(x) = F(x) \quad \text{und} \quad \text{Var} \hat{F}_n(x) = \frac{1}{n} F(x)(1 - F(x)). \quad (18)$$

- 3.

$$P\left(\lim_{n \rightarrow \infty} \hat{F}_n(x) = F(x)\right) = 1. \quad (19)$$

4. Falls $0 < F(x) < 1$, dann gilt außerdem für jedes $y \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} P\left(\sqrt{n} \frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} \leq y\right) = \Phi(y), \quad (20)$$

wobei $\Phi(y)$ die Verteilungsfunktion der Standardnormalverteilung ist.

Beweis

- Da $n\hat{F}_n(x) = |\{i : 1 \leq i \leq n, x_i \leq x\}| = \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$, können wir die Zufallsvariable $n\hat{F}_n(x)$ als die Anzahl der Erfolge beim n -maligen Münzwurf mit den identischen Erfolgswahrscheinlichkeiten $a_1 = a_2 = \dots = a_n = F(x)$ auffassen. Hieraus ergibt sich (17).

- Damit ist auch (18) bewiesen, da

$$\mathbb{E}(\hat{F}_n(x)) = \frac{1}{n} \mathbb{E}(n\hat{F}_n(x)) = \frac{1}{n} nF(x) = F(x)$$

und

$$\text{Var}(\hat{F}_n(x)) = \frac{1}{n^2} \mathbb{E}(n\hat{F}_n(x)) = \frac{1}{n^2} nF(x)(1 - F(x)) = \frac{1}{n} F(x)(1 - F(x)).$$

- Da $n\hat{F}_n(x) = |\{i : 1 \leq i \leq n, x_i \leq x\}| = \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x\}}$, Wegen der Darstellungsmöglichkeit von $n\hat{F}_n(x)$ als die Anzahl der Erfolge beim n -maligen Münzwurf mit identischen Erfolgswahrscheinlichkeiten ist $n\hat{F}_n(x)$ die Summe von n unabhängigen und identisch (Bernoulli-) verteilten Zufallsvariablen. Deshalb ergibt sich (19) aus dem starken Gesetz der großen Zahlen (vgl. Theorem 4.22).
- Mit der gleichen Begründung ergibt sich (20) unmittelbar aus dem zentralen Grenzwertsatz.

Beachte

- Neben der fast sicheren Konvergenz (19), die für jedes (einzelne) $x \in \mathbb{R}$ gilt, kann man zeigen, dass sich auch die Verteilungsfunktion F insgesamt durch den empirischen Prozeß $\{\hat{F}_n(x), x \in \mathbb{R}\}$ im Sinne der fast sicheren Konvergenz approximieren läßt.
- Dabei ist die folgende Verschärfung des starken Gesetzes der großen Zahlen gemeint, die in der Literatur *Satz von Glivenko/Cantelli* genannt wird.

Theorem 1.11 Sei

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|. \quad (21)$$

Dann gilt

$$P\left(\lim_{n \rightarrow \infty} D_n = 0\right) = 1. \quad (22)$$

Beweis Der Beweis von Theorem 1.11 ist tiefgehend und geht über den Rahmen dieser einführenden Vorlesung hinaus.

Ähnlich wie beim zentralen Grenzwertsatz für Summen von unabhängigen und identisch verteilten Zufallsvariablen kann man zeigen, dass auch D_n bei entsprechend gewählter Normierung gegen einen nichtdeterministischen, d.h. zufälligen Grenzwert (im Sinne der Verteilungskonvergenz) strebt.

Dies ist die Aussage des folgenden Theorems, das *Satz von Kolmogorow/Smirnow* genannt wird.

Theorem 1.12 Falls die Verteilungsfunktion F der Stichprobenvariablen X_i eine stetige Funktion ist, dann gilt für $n \rightarrow \infty$

$$\sqrt{n}D_n \xrightarrow{d} D, \quad (23)$$

wobei D eine Zufallsvariable ist, deren Verteilungsfunktion gegeben ist durch

$$P(D \leq y) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 y^2) & \text{für } y \geq 0, \\ 0 & \text{für } y < 0. \end{cases} \quad (24)$$

Beweis Der Beweis von Theorem 1.12 ist tiefgehend und geht über den Rahmen dieser einführenden Vorlesung hinaus.

1.5 Dichteschätzer

Wir nehmen an, dass die (konkrete) Stichprobe $(x_1, \dots, x_n)$ die Realisierung einer Zufallsstichprobe $(X_1, \dots, X_n)$ sei, wobei $X_1, \dots, X_n$ unabhängig und identisch verteilte (iid) Zufallsvariablen seien mit Verteilungsfunktion F und Dichte f . Neben der Schätzung der Verteilungsfunktion (Abschnitt 1.4) ist es von Interesse, die Dichte f anhand der Stichprobe $(x_1, \dots, x_n)$ zu schätzen. Oft bieten Dichteschätzer (Histogramme oder Kerndichteschätzer) einen intuitiven Zugang, um einen Eindruck von der Form der Dichte zu gewinnen. Dieses ist hilfreich, um herauszufinden, wie die Stichprobenvariable X_i verteilt ist, z.B. ob es sich um eine Gamma- oder eine Normalverteilung handelt.

Histogramm

Ein erster sehr einfacher Schätzer für die Dichte f ist das *Histogramm*, welches wir in diesem Abschnitt einführen. Sei $(x_1, \dots, x_n)$ eine (konkrete) Stichprobe und $c_0 < c_1 < \dots < c_k \in \mathbb{R}$ eine Folge von Intervallgrenzen, so dass die Stichprobe in (c_0, c_k) enthalten ist, d. h. $x_i \in (c_0, c_k)$ für jedes $i = 1, \dots, n$. Falls die Abstände $c_i - c_{i-1}$ konstant sind, so heißt $h = c_i - c_{i-1}$ *Bandbreite*.

Die Idee ist es nun, die relative Häufigkeit von Stichprobenwerten x_i zu betrachten, die in die jeweiligen Intervalle $(c_{i-1}, c_i]$, $i = 1, \dots, k$ fallen. Sei also

$$n_i = \sum_{j=1}^n \mathbb{I}_{(c_{i-1}, c_i]}(x_j), \quad i = 1, \dots, k$$

die Anzahl an Stichprobenwerten im Intervall $(c_{i-1}, c_i]$. Normierung durch den Stichprobenumfang n ergibt die relativen Häufigkeiten

$$f_i = \frac{n_i}{n}, \quad i = 1, \dots, k.$$

Als *Histogramm* bezeichnet man die graphische Darstellung der relativen Häufigkeiten f_i , $i = 1, \dots, k$, wobei jedem Intervall $(c_{i-1}, c_i]$ ein Rechteck mit Flächeninhalt f_i zugeordnet wird, siehe Abbildung 1. Ein Histogramm kann als Dichteschätzer interpretiert werden, da die Dichte f durch $\hat{f}_n$ geschätzt wird, wobei

$$\hat{f}_n(x) = \sum_{i=1}^k \frac{1}{c_i - c_{i-1}} f_i \mathbb{I}_{(c_{i-1}, c_i]}(x), \quad x \in \mathbb{R}.$$

Man sieht, dass $\hat{f}_n$ die Eigenschaften einer Dichte besitzt, d.h. der Flächeninhalt der Dichte ergibt 1 und $\hat{f}_n(x) \geq 0$ für alle $x \in \mathbb{R}$.

Als Histogramm wird ebenso die graphische Darstellung von absoluten (und nicht relativen) Häufigkeitsverteilungen bezeichnet. Dann entspricht der Flächeninhalt dem Stichprobenumfang n .

Das Histogramm als Dichteschätzer hat zwei Nachteile. Zum einen ist die geschätzte Dichte $\hat{f}_n$ stückweise konstant - auch bei stetigen Dichten f . Zweitens hat die Wahl der Bandbreite h (bzw. die Anzahl an Intervallen) einen starken Einfluss auf das Histogramm, siehe Abbildung 2. *Kerndichteschätzer* umgehen den ersten Nachteil (bei geeigneter Wahl des Kerns) und sollen nun betrachtet werden.

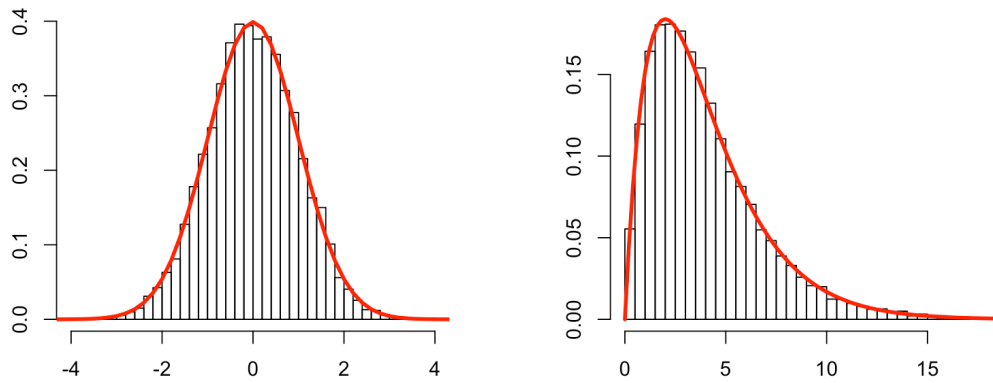


Figure 1: Histogramme von normalverteilten (links) und Gamma-verteilten (rechts) Stichproben mit einem Stichprobenumfang von $n = 10000$. Die Dichte ist als rote Kurve hinzugefügt.

Kerndichteschätzer

Definition 1.13 Ein *Kern* ist eine messbare Abbildung $K : \mathbb{R} \rightarrow \mathbb{R}$ mit folgenden Eigenschaften:

1. $K(x) \geq 0$ für alle $x \in \mathbb{R}$
2. $\int_{\mathbb{R}} K(x) dx = 1$

Beispiel Folgende Kerne werden häufig verwendet:

- Gausskern:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \quad x \in \mathbb{R}$$

- Epanechnikov-Kern

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2), & \text{falls } -1 \leq x \leq 1 \\ 0 & \text{sonst} \end{cases}$$

- Bisquare-Kern

$$K(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2, & \text{falls } -1 \leq x \leq 1 \\ 0 & \text{sonst} \end{cases}$$

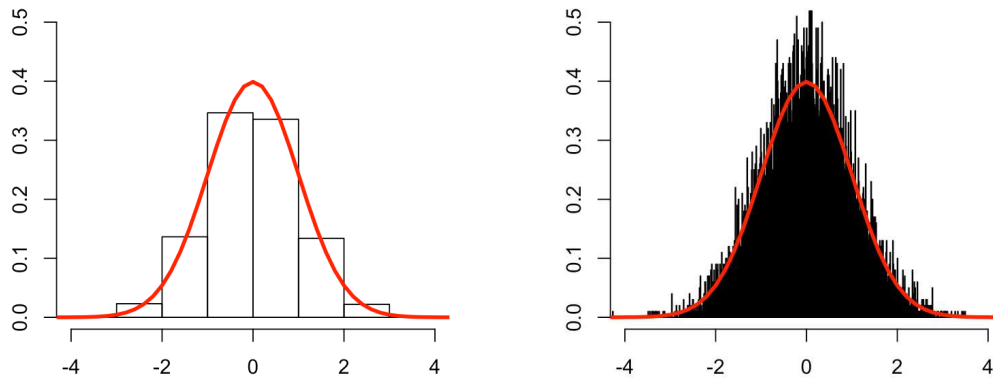


Figure 2: Histogramme von normalverteilten Stichproben (Stichprobenumfang $n = 10000$) mit einer zu gross gewählten Bandbreite (links) und einer zu klein gewählten Bandbreite (rechts). Die Dichte ist als rote Kurve hinzugefügt.

Definition 1.14 Sei K ein Kern, $h > 0$ ein Parameter (genannt Bandbreite) und $(x_1, \dots, x_n)$ eine Stichprobe. Ein *Kerndichteschätzer* ist definiert durch

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R}.$$

Beachte

- Die Idee des Kerndichteschätzers besteht darin, jedem Stichprobenwert x_i den Anteil

$$\frac{1}{nh} K\left(\frac{x - x_i}{h}\right)$$

zuzuordnen und die geschätzte Dichte als Summe der einzelnen Beiträge zu definieren.

- Wie bei den Histogrammen spielt auch bei Kerndichteschätzern die Wahl der Bandbreite h eine entscheidende Rolle. Grundsätzlich gilt: Je größer die Bandbreite h , desto mehr werden die Daten ‘geglättet’, siehe Abbildung 3.

2 Schätzung von Parametern

In diesem Abschnitt treffen wir folgende Annahmen:

- die Verteilungsfunktion F der Stichprobenvariablen X_i gehört zu einer vorgegebenen (d.h. bekannten) *parametrischen Familie* von Verteilungsfunktionen $\{F_\theta, \theta \in \Theta\}$,

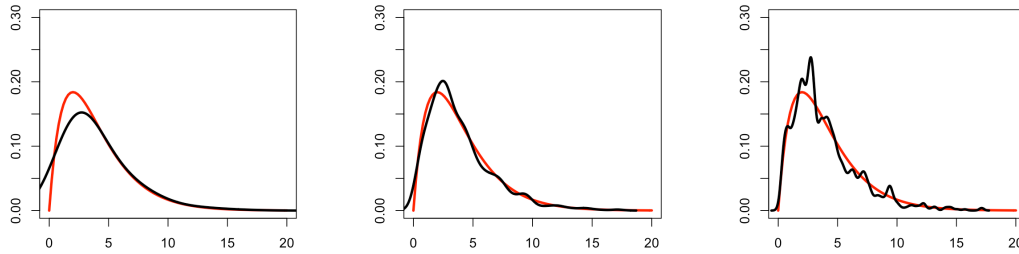


Figure 3: Kernerichteschätzungen von Gamma-verteilten Stichproben (Stichprobenumfang $n = 1000$) mit einer zu gross gewählten Bandbreite (links), einer ‘passenden’ Bandbreite (mittig) und einer zu klein gewählten Bandbreite (rechts). Die Dichte ist als rote Kurve hinzugefügt.

- wobei die Menge $\Theta \subset \mathbb{R}^m$ *Parameterraum* genannt wird und $m \in \mathbb{N}$ eine beliebige, jedoch vorgegebene natürliche Zahl ist.
- Mit anderen Worten: Es gelte $F = F_\theta$ für ein $\theta \in \Theta$, wobei jedoch der Parametervektor θ (bzw. ein Teil seiner Komponenten) unbekannt sei und aus den beobachteten Daten $x_1, \dots, x_n$ geschätzt werden soll.
- Weiterhin sei der Parameterraum Θ Borel-messbar.
- Die Parametrisierung $\theta \rightarrow F_\theta$ sei identifizierbar, d.h. es gelte $F_{\theta_1} \neq F_{\theta_2}$ falls $\theta_1 \neq \theta_2$.

Wir nehmen an, dass der Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, P)$, über den die Stichprobenvariablen $X_1, X_2, \dots$ definiert sind, der sogenannte *kanonische Wahrscheinlichkeitsraum* ist. Das heißt, wir setzen

- $\Omega = \mathbb{R} \times \mathbb{R} \times \dots = \{\omega : \omega = (\omega_1, \omega_2, \dots), \omega_i \in \mathbb{R} \ \forall i = 1, 2, \dots\}$
- und $\mathcal{F} = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) \otimes \dots$,
- wobei das Wahrscheinlichkeitsmaß P gegeben ist durch:

$$P(\{\omega : \omega \in \Omega, \omega_{i_1} \leq x_{i_1}, \dots, \omega_{i_k} \leq x_{i_k}\}) = F_\theta(x_{i_1}) \cdot \dots \cdot F_\theta(x_{i_k}). \quad (25)$$

Die Stichprobenvariablen $X_i : \Omega \rightarrow \mathbb{R}$ sind gegeben durch $X_i(\omega) = \omega_i$, d.h. durch die Projektion auf die i -te Komponente ω_i von ω .

Beachte

- Das in (25) definierte Wahrscheinlichkeitsmaß P bezeichnen wir im Folgenden gelegentlich mit P_θ , um zu betonen, dass P_θ von dem (unbekannten) Parametervektor $\theta \in \Theta$ abhängt.
- Entsprechend verwenden wir die Bezeichnungen $\mathbb{E}_\theta$ und Var_θ für Erwartungswert bzw. Varianz.

- Falls keine Verwechslung möglich ist, dann bezeichnen wir auch die Verteilung einer (einzelnen) Stichprobenvariablen X_i mit P_θ , d.h. $P_\theta(X_i \leq x_i) = P_\theta((-\infty, x_i]) = F_\theta(x_i)$.
- Für jede Borel-messbare Funktion

$$\hat{\theta} : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

wird der Zufallsvektor $\hat{\theta}(X_1, \dots, x_n)$ *Stichprobenfunktion* bzw. *Statistik* genannt.

- Bei Schätzung von Parametern heißt $\hat{\theta}(X_1, \dots, X_n)$ *Punktschätzer* von θ . Es wird in der Regel vorausgesetzt, dass $P_\theta(\hat{\theta}(X_1, \dots, X_n) \in \Theta) = 1$

Beispiel

- Die wichtigste parametrische Familie $\{P_\theta, \theta \in \Theta\}$ von Verteilungen, die in diesem Abschnitt betrachtet wird, ist die *Normalverteilungsfamilie*

$$\{P_\theta, \theta \in \Theta\} = \{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}. \quad (26)$$

In diesem Fall ist $m = 2$, $\Theta = \mathbb{R} \times (0, \infty)$ und $(\theta_1, \theta_2) = (\mu, \sigma^2)$.

- Eine andere parametrische Familie ist die *Binomialverteilungsfamilie*

$$\{P_\theta, \theta \in \Theta\} = \{Bin(n, p), n \in \mathbb{N}, p \in [0, 1]\}. \quad (27)$$

In diesem Fall ist $m = 2$, $\Theta = \mathbb{N} \times [0, 1]$ und $(\theta_1, \theta_2) = (n, p)$.

Beachte Weitere Beispiele von Familien parametrischer Verteilungen sind die

- Poisson-Verteilung,
- Exponentialverteilung,
- (diskrete bzw. stetige) Gleichverteilung.

2.1 Eigenschaften von Parameterschätzern

- Wir betrachten eine beliebige Stichprobenfunktion $\hat{\theta} : \mathbb{R}^n \rightarrow \Theta \subset \mathbb{R}^m$
- und den zugehörigen Schätzer $\hat{\theta}(X_1, \dots, X_n)$ des Parametervektors θ .

Definition 2.1

- Es gelte $\mathbb{E}_\theta |\hat{\theta}(X_1, \dots, X_n)| < \infty$ für jedes $\theta \in \Theta$, wobei $|\hat{\theta}(X_1, \dots, X_n)|$ die Norm des Zufallsvektors $\hat{\theta}(X_1, \dots, X_n)$ bezeichnet.
- Dann heißt der Schätzer $\hat{\theta}(X_1, \dots, X_n)$
 1. *erwartungstreu* oder *unbiased*, falls $\mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) = \theta$ für jedes $\theta \in \Theta$,
 2. *asymptotisch erwartungstreu*, falls $\lim_{n \rightarrow \infty} \mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) = \theta$ für jedes $\theta \in \Theta$.
 3. Die Differenz $\mathbb{E}_\theta \hat{\theta}(X_1, \dots, X_n) - \theta$ heißt *Bias* (Synonym: *Verzerrung*).

4. Falls $\mathbb{E}_\theta(|\hat{\theta}(X_1, \dots, X_n) - \theta|^2) < \infty$, dann heißt $\mathbb{E}_\theta(|\hat{\theta}(X_1, \dots, X_n) - \theta|^2)$ *MSE* (mean squared error) bzw. *erwartete quadratische Abweichung* des Schätzers $\hat{\theta}(X_1, \dots, X_n)$ von dem zu schätzenden Parametervektor $\theta \in \Theta \subset \mathbb{R}^m$.

Beispiel

- Die Stichprobenvariablen X_i seien normalverteilt, d.h. $X_i \sim N(\mu, \sigma^2)$.
- Aus (3) und (10) ergibt sich, dass der Schätzer $(\bar{X}_n, S_n^2)$ des Parametervektors (μ, σ^2) erwartungstreu ist.

Definition 2.2 Der Schätzer $\hat{\theta}(X_1, \dots, X_n)$ des Parametervektors θ heißt

1. *schwach konsistent*, falls für $n \rightarrow \infty$

$$\hat{\theta}(X_1, \dots, X_n) \xrightarrow{\text{st}} \theta,$$

d.h.

$$\lim_{n \rightarrow \infty} P_\theta(|\hat{\theta}(X_1, \dots, X_n) - \theta| > \varepsilon) = 0 \quad \forall \varepsilon > 0$$

für jedes $\theta \in \Theta$.

2. *stark konsistent*, falls für $n \rightarrow \infty$

$$\hat{\theta}(X_1, \dots, X_n) \xrightarrow{\text{f.s.}} \theta,$$

d.h.

$$P_\theta\left(\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) = \theta\right) = 1$$

für jedes $\theta \in \Theta$.

Beispiel

- Die Stichprobenvariablen X_i seien normalverteilt, d.h. $X_i \sim N(\mu, \sigma^2)$.
- Dann ergibt sich aus (5) und (12), dass der Schätzer

$$\hat{\theta}(X_1, \dots, X_n) = (\bar{X}_n, S_n^2)$$

des Parametervektors $\theta = (\mu, \sigma^2)$ sowohl schwach als auch stark konsistent ist.

2.2 Momentenmethode

- Die Momentenmethode ist eines der ältesten Verfahren zur Gewinnung von Schätzern für die unbekannten Komponenten des Parametervektors $\theta = (\theta_1, \dots, \theta_m)$.
- Die Methode wurde von Karl Pearson (1857-1936) gegen Ende des 19. Jahrhunderts eingeführt und
- beruht auf dem Vergleich von Momenten der Stichprobenvariablen $X_i, \dots, X_n$ mit den entsprechenden empirischen Momenten der konkreten Stichprobe $(x_1, \dots, x_n)$.

Modellannahmen

- Es gelte $\mathbb{E}_\theta(|X_i|^r) < \infty$ für jedes $\theta \in \Theta$ und für eine natürliche Zahl $r \geq m$.
- Ausserdem nehmen wir an, dass für jedes $k = 1, \dots, r$ das k -te Moment $\mu_k = \mathbb{E}_\theta(X_i^k)$ der Stichprobenvariablen X_i eine (bekannte) Funktion des Parametervektors $\theta = (\theta_1, \dots, \theta_m)$ ist.
- D.h., für jedes $k = 1, \dots, r$ gelte

$$\mu_k = g_k(\theta_1, \dots, \theta_m), \quad (28)$$

wobei $g_k : \mathbb{R}^m \rightarrow \mathbb{R}$ eine Borel-messbare Funktion sei.

Lösungsansatz

- Für jedes $k = 1, \dots, r$, sei

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^n x_i^k$$

das k -te *empirische Moment*.

- Löse das Gleichungssystem

$$\hat{m}_k = g_k(\theta_1, \dots, \theta_m), \quad k = 1, \dots, r \quad (29)$$

nach den Unbekannten $\theta_1, \dots, \theta_m$ auf, wobei

$$\hat{\theta}(x_1, \dots, x_n) = \left(\hat{\theta}_1(x_1, \dots, x_n), \dots, \hat{\theta}_m(x_1, \dots, x_n) \right) \quad (30)$$

die Lösung sei.

- Dabei wird vorausgesetzt, dass $\hat{\theta}(x_1, \dots, x_n)$ eindeutig bestimmt ist und dass die Abbildung

$$(x_1, \dots, x_n) \rightarrow \hat{\theta}(x_1, \dots, x_n)$$

eine Stichprobenfunktion ist, die den Stichprobenraum $\mathbb{R}^n$ in den Parameterraum $\Theta \subset \mathbb{R}^m$ abbildet.

Definition 2.3 Der Zufallsvektor $\hat{\theta}(X_1, \dots, X_n)$ heißt *Momentenschätzer* (bzw. *M-Schätzer*) des Parametervektors θ , wobei in (30) die Zufallsstichprobe $(X_1, \dots, X_n)$ anstelle der konkreten Stichprobe $(x_1, \dots, x_n)$ eingesetzt wird.

Beachte

- Es gibt Beispiele parametrischer Verteilungsfamilien, so dass das Gleichungssystem (29) für $r = m$ keine eindeutig bestimmte Lösung besitzt. Dann ist $r > m$, d.h., die Anzahl der betrachteten Momente kann größer als die Anzahl der (unbekannten) Parameterkomponenten sein. In vielen Fällen ist jedoch das Gleichungssystem (29) für $r = m$ eindeutig lösbar.
- Mit Hilfe des starken Gesetzes der großen Zahlen kann die folgende Begründung für die Anwendung der Momentenmethode gegeben werden.

- Für das empirische Moment $\hat{m}_k(X_1, \dots, X_n)$ der Zufallsstichprobe $(X_1, \dots, X_n)$ gilt

$$\hat{m}_k(X_1, \dots, X_n) \xrightarrow{\text{f.s.}} \mu_k, \quad (31)$$

falls $n \rightarrow \infty$.

- Falls die (28) gegebene Abbildung $g = (g_1, \dots, g_m) : \Theta \rightarrow C$ mit $C = g(\Theta) = \{g(\theta) : \theta \in \Theta\} \subset \mathbb{R}^m$ eindeutig ist und falls die Umkehrabbildung $g^{-1} : C \rightarrow \Theta$ stetig ist, dann gilt wegen (31) für den M-Schätzer $\hat{\theta}(X_1, \dots, X_n)$, dass für jedes $\theta \in \Theta$ mit Wahrscheinlichkeit 1

$$\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) = \theta. \quad (32)$$

- D.h. der Schätzer $\hat{\theta}(X_1, \dots, X_n)$ ist stark konsistent.

Beispiel Die Stichprobenvariablen X_i seien normalverteilt, d.h., $X_i \sim N(\mu, \sigma^2)$. Dann ist

- $m = 2$ mit $(\theta_1, \theta_2) = (\mu, \sigma^2)$
- Außerdem ist

$$\mu_1 = \mu = g_1(\mu, \sigma^2) \quad \text{und} \quad \mu_2 = \sigma^2 + \mu^2 = g_2(\mu, \sigma^2).$$

- Das Gleichungssystem (29) hat also die Form

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n x_i &= \mu \\ \frac{1}{n} \sum_{i=1}^n x_i^2 &= \hat{\sigma}^2 + \hat{\mu}^2. \end{aligned}$$

- Durch Umstellen ergibt sich

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

und

$$\begin{aligned} \hat{\sigma}^2 &= \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 \\ &\vdots \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2. \end{aligned}$$

- D.h., bei normalverteilten Stichprobenvariablen X_i ergeben sich mit der Momentenmethode die Schätzer

$$\hat{\mu}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i \quad (33)$$

und

$$\hat{\sigma}^2(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (34)$$

für die Modellparameter μ bzw. σ^2 .

Beachte

- Aus den Theoremen 1.4 und 1.5 ergibt sich, dass der in (33) gewonnene Schätzer $\hat{\mu}(X_1, \dots, X_n)$ des Parameters μ erwartungstreu und konsistent ist.
- Aus den Theoremen 1.7 und 1.8 ergibt sich, dass der in (34) gewonnene Schätzer $\hat{\sigma}^2(X_1, \dots, X_n)$ des Parameters σ^2 asymptotisch erwartungstreu und konsistent ist.
- Wegen des im Vergleich zu S_n^2 modifizierten Normierungsfaktors $1/n$ ist $\hat{\sigma}^2(X_1, \dots, X_n)$ jedoch *nicht* erwartungstreu.
- Somit ist auch der Schätzer $(\hat{\mu}(X_1, \dots, X_n), \hat{\sigma}^2(X_1, \dots, X_n))$ des Parametervektors (μ, σ^2) lediglich asymptotisch erwartungstreu und konsistent.

Beispiel Die Stichprobenvariablen X_i seien Gamma-verteilt, d.h., $X_i \sim \Gamma(b, p)$ mit $b > 0, p > 0$. Die Dichte der Gammaverteilung ist gegeben durch

$$f_{b,p}(x) = \begin{cases} \frac{b^p}{\Gamma(p)} x^{p-1} e^{-bx} & \text{für } x > 0, \\ 0 & \text{sonst.} \end{cases}$$

- Dann ist $m = 2$ mit $(\theta_1, \theta_2) = (b, p)$.
- Man kann ausrechnen, dass

$$\mu_1 = \frac{p}{b}, \quad \mu_2 = \frac{p(p+1)}{b^2}.$$

- Hieraus ergibt sich das Gleichungssystem

$$\hat{m}_1 = \frac{p}{b}, \quad \hat{m}_2 = \frac{p(p+1)}{b^2}$$

mit der Lösung

$$\hat{b} = \frac{\hat{m}_1}{\hat{m}_2 - (\hat{m}_1)^2}, \quad \hat{p} = \frac{(\hat{m}_1)^2}{\hat{m}_2 - (\hat{m}_1)^2}$$

2.3 Maximum-Likelihood-Methode

- Eine andere Methode zur Gewinnung von Schätzern für die unbekannten Komponenten des Parametervektors $\theta = (\theta_1, \dots, \theta_m)$ ist die Maximum-Likelihood-Methode.
- Genauso wie bei der Momentenmethode wird auch bei der Maximum-Likelihood-Methode das Ziel verfolgt, θ so zu schätzen, dass eine möglichst gute Anpassung der Modellcharakteristiken P_θ bzw. F_θ an die beobachteten Daten erreicht wird.
- Dabei wird bei der Maximum-Likelihood-Methode die Wahrscheinlichkeit für das Eintreten bestimmter (infinitesimaler) Ereignisse maximiert.
- Als der ‘Erfinder’ der Maximum-Likelihood-Methode gilt Sir Ronald Aylmer Fisher (1890–1962).

Wir betrachten nur die beiden (grundlegenden) Fälle, dass die Stichprobenvariablen $X_1, \dots, X_n$ entweder diskret oder absolutstetig sind. D.h., für jedes $\theta \in \Theta$ gelte entweder

- $P_\theta(X_i \in C) = 1$ für eine abzählbare Menge $C \subset \mathbb{R}$,
- wobei wir mit $\{p(x; \theta), x \in C\}$ die Wahrscheinlichkeitsfunktion von X_i bezeichnen; $p(x; \theta) = P_\theta(X_i = x)$

oder

- $F_\theta(x) = \int_{-\infty}^x f(y; \theta) dy$ für jedes $x \in \mathbb{R}$,
- wobei $f(\cdot; \theta)$ die Dichte von X_i ist.

Definition 2.4 Die Abbildung $L : \mathbb{R}^n \times \Theta \rightarrow [0, \infty)$ sei durch die folgende Vorschrift gegeben.

- Falls X_i diskret ist, dann gelte

$$L(x_1, \dots, x_n; \theta) = p(x_1; \theta) \dots p(x_n; \theta) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n, \theta \in \Theta. \quad (35)$$

- Falls X_i absolutstetig ist, dann gelte

$$L(x_1, \dots, x_n; \theta) = f(x_1; \theta) \dots f(x_n; \theta) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n, \theta \in \Theta. \quad (36)$$

Für jeden Vektor $(x_1, \dots, x_n) \in \mathbb{R}^n$ heißt die Abbildung $\theta \rightarrow L(x_1, \dots, x_n; \theta)$ die *Likelihood-Funktion* der Stichprobe $(x_1, \dots, x_n)$.

Die Idee der Maximum-Likelihood-Methode besteht nun darin, für jede (konkrete) Stichprobe $(x_1, \dots, x_n)$ einen Parametervektor $\theta \in \Theta$ zu bestimmen, so dass der Wert $L(x_1, \dots, x_n; \theta)$ der Likelihood-Funktion möglichst groß wird. Etwas vereinfacht bedeutet dies, dass der Parameter θ so gewählt wird, dass die (konkrete) Stichprobe $(x_1, \dots, x_n)$ möglichst 'plausibel' erscheint. Dies führt zu der folgenden

Definition 2.5 Sei $\hat{\theta} : \mathbb{R}^n \rightarrow \Theta \subset \mathbb{R}^m$ eine Stichprobenfunktion mit

$$L(x_1, \dots, x_n; \theta) \leq L(x_1, \dots, x_n; \hat{\theta}(x_1, \dots, x_n)) \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n, \theta \in \Theta. \quad (37)$$

Der Zufallsvektor $\hat{\theta}(X_1, \dots, X_n)$ wird dann *Maximum-Likelihood-Schätzer (ML-Schätzer)* von θ genannt.

Beachte

- Es gibt Beispiele parametrischer Verteilungsfamilien, so dass der mittels (37) definierte Maximum-Likelihood-Schätzer $\hat{\theta}(X_1, \dots, X_n)$ nicht eindeutig bestimmt ist.
- In vielen Fällen lässt sich das Optimierungsproblem (37) jedoch durch Differenzieren eindeutig lösen.
- Anstelle (37) wird dabei oft die (äquivalente) Bedingung

$$\log L(x_1, \dots, x_n; \theta) \leq \log L(x_1, \dots, x_n; \hat{\theta}(x_1, \dots, x_n)), \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n, \theta \in \Theta \quad (38)$$

betrachtet.

- Für jeden Vektor $(x_1, \dots, x_n) \in \mathbb{R}^n$ wird die Abbildung $\theta \rightarrow \log L(x_1, \dots, x_n; \theta)$ die *Loglikelihood-Funktion* der Stichprobe $(x_1, \dots, x_n)$ genannt.
- Sie bietet den Vorteil, dass beim Übergang zur Loglikelihood-Funktion die Produkte in (35) bzw. (36) in (einfacher handhabbare) Summen übergehen.

Beispiele

1. Wer war der Absender?

- Eine Warenlieferung eines unbekannten Herstellers bestehe aus 12 Exemplaren eines Artikels.
- Dabei sei festgestellt worden, dass eines der 12 Exemplare Ausschuss ist.
- Es sei bekannt, dass nur drei potentiell mögliche Hersteller in Frage kommen und dass deren Lieferungen erfahrungsgemäss jeweils einen Ausschussanteil von $\theta_1 = 0,05$, $\theta_2 = 0,10$ bzw. $\theta_3 = 0,15$ aufweisen.
- *Frage:* Welcher der drei Hersteller war vermutlich der Absender der Warenlieferung?
- *Modell:* Betrachten die Stichprobenvariablen $X_1, \dots, X_{12}$ mit

$$X_i(\omega) = \begin{cases} 1, & \text{falls das } i\text{-te Exemplar der Lieferung Ausschuss ist} \\ 0 & \text{sonst} \end{cases}$$

und die Familie der drei Bernoulli-Verteilungen $\{\text{Bin}(1, \theta_1), \text{Bin}(1, \theta_2), \text{Bin}(1, \theta_3)\}$, d.h. $m = 1$ und $\Theta = \{\theta_1, \theta_2, \theta_3\}$.

- *Lösung:* Die Stichprobenfunktion

$$\hat{\theta} : \{0, 1\}^{12} \rightarrow \{\theta_1, \theta_2, \theta_3\}$$

wird so gewählt, dass für jeden Vektor $(x_1, \dots, x_{12}) \in \{0, 1\}^{12}$ mit $|\{i : x_i = 1\}| = 1$ die Wahrscheinlichkeit

$$P_\theta((X_1, \dots, X_{12}) = (x_1, \dots, x_{12})) = \theta(1 - \theta)^{11}$$

maximal ist.

- Es gilt

θ	$P_\theta((X_1, \dots, X_{12}) = (x_1, \dots, x_{12}))$
0,05	0,028
0,10	0,031
0,15	0,025

- Das Maximum 0,031 steht in der zweiten Zeile dieser Tabelle.
- Also ist $\hat{\theta}(x_1, \dots, x_{12}) = \theta_2$, d.h., der Hersteller mit dem Ausschussanteil $\theta_2 = 0,10$ war vermutlich der Absender der Lieferung.