Wahrscheinlichkeitsrechnung und Statistik

9. Vorlesung - 2017

Monte Carlo Methode für numerische Integration

Sei $g : [0,1] \to \mathbb{R}$ stetige Funktion; man möchte $\int_0^1 g(t)dt$ numerisch approximieren mit Hilfe von Zufallszahlen:

- Sei $(U_n)_n$ eine Folge von unabhängigen ZG mit $U_n \sim Unif[0,1]$. Sei $X_n = g(U_n)$.
- $(X_n)_n$ erfüllt das SGGZ (siehe Satz 18, Vorl. 8), d.h.

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - E(X_k)\Big) \xrightarrow{f.s.} 0.$$

- Es gilt $E(X_k) = \int_0^1 g(t)dt \ \forall k \in \mathbb{N}$

$$\Rightarrow \frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{f.s.} \int_0^1 g(t)dt$$

- in der Praxis:

$$\int_0^1 g(t)dt \approx \frac{1}{n}(g(U_1) + \cdots + g(U_n)) \text{ für } n \text{ hinreichend groß}$$

Monte Carlo Methode für numerische Integration

Sei $g : [a, b] \to \mathbb{R}$ stetige Funktion mit $0 \leq g(t) \leq M \ \forall \ t \in [a, b]$; man möchte $\int_a^b g(t)dt$ numerisch approximieren mit Hilfe von Zufallszahlen:

- Seien X_k, Y_k unabhängige ZG mit $X_k \sim Unif[a, b], \ Y_k \sim Unif[0, M]$ $\forall k = \overline{1, n}$.
- Sei $N_n = \#\{k \in \{1, ..., n\} : Y_k \leq g(X_k)\}$
- $\int_a^b g(t)dt =$ Flächeninhalt unter der Funktion g

$$P(\text{ zufälliger Punkt unter der Funktion } g) \approx \frac{N_n}{n}$$

$$\frac{\text{Fl. unter der Funktion } g}{\text{Fl. vom Rechteck } [a, b] \times [0, M]} = \frac{1}{(b-a)M} \int_a^b g(t)dt$$

$$\Rightarrow \int_a^b g(t)dt \approx \frac{\#\{k \in \{1, ..., n\} : Y_k \leq g(X_k)\}}{n} \cdot (b-a)M$$

für n hinreichend groß

Statistik

Statistik = wissenschaftliche Disziplin, deren Gegenstand die Entwicklung und Anwendung formaler Methoden zur Gewinnung, Beschreibung und Analyse und Beurteilung von Daten (Beobachtungen) ist.

Aufgaben und Ziele der Statistik

- Design von Experimenten:
  Wie sollen die Daten gewonnen werden?
- Beschreibende (deskriptive) Statistik:
  Wie sollen große Datensätze beschrieben werden, um die Gesetzmäßigkeiten und Strukturen in ihnen entdecken zu können?
- Schließende Statistik:
  Welche Schlußfolgerungen kann man aus den Daten ziehen?

Arbeitsweise in der Statistik

- Datenerhebung (Beobachtung, Befragung, Experiment)
- Visualisierung und beschreibende Datenanalyse: graphische Präsentation, Zusammenfassung (Darstellung der Daten in einer Tabelle oder Darstellung mit Hilfe von Grafiken)
- Explorative Datenanalyse (man sucht nach Gesetzmäßigkeiten in den Daten)
- Modellierung der Daten
- Modellanpassung (Schätzung von Modellparametern)
- Modellvalidierung (Wie gut war die Modellanpassung?)

Anwendung der Statistik:

- in politischen Umfragen: z.B. Befragung zur Beliebtheit von Politikern oder einer Partei
- in der Analyse von Finanzmarktdaten: z.B. Analyse von Aktien-, Zinskursen
- in klinischen und epidemiologischen Studien (Medizin und Pharmazie)
- in der Technik: z.B. die Lebensdaueranalyse oder die Zuverlässigkeit von elektronischen Systemen
- in der Wirtschaft: z.B. Data Mining und Data Warehousing sind zwei Bereiche, in denen man versucht mit Hilfe von statistischen Methoden aus einer Vielzahl von Kundendaten jene herauszufiltern und aufzubereiten, die für den Erfolg des Betriebs von Interesse sind

Grundbegriffe der Statistik

- statistische Einheiten = Objekte an denen interessierende Größen erfaßt werden
z.B. Bevölkerung einer Stadt; Schüler einer bestimmten Schule; Patienten einer Klinik
- Grundgesamtheit (Population) = Menge aller statistischen Einheiten über die man Aussagen erhalten will
z.B. wahlberechtigte Bevölkerung einer Stadt; Schüler der 10. Klasser einer bestimmten Schule; Patienten einer bestimmten Station einer Klinik
- statistisches Merkmal = Eigenschaft einer statistischen Einheit für die man sich bei einer statistischen Untersuchung interessiert
z.B. Alter; Geschlecht; Wert BMW Aktie 24.11.2016, 17 Uhr;
- Merkmalsausprägung = konkreter Wert des Merkmals
z.B. 25 Jahre; weiblich; 82.5 Euro
- Stichprobe = tatsächlich untersuchte Teilmenge der Grundgesamtheit

Ein Merkmal einer Grundgesamtheit ist eine Zufallsgröße X. Ziel der Statistik ist das Gesetz (die Verteilung) von X zu finden (zu schätzen), anhand der statistischen Daten.

Die Verteilung von X kann

1) vollständig spezifiziert sein
z.B. $X \sim Exp(3)$, $X \sim Bin(10, 0.3)$, $X \sim N(0, 1)$
2) spezifiziert sein, aber von einem oder mehreren unbekannten Parametern abhängen
z.B. $X \sim Exp(\lambda)$, $X \sim Bin(10, p)$, $X \sim N(0, \sigma^2)$
3) unbekannt sein, $X \sim ???$

- in den Fällen 2) und 3) werden der unbekannte Parameter oder die unbekannte Verteilung

$\hookrightarrow$ geschätzt $\rightarrow$ Schätztheorie
$\hookrightarrow$ getestet $\rightarrow$ Testen von statistischen Hypothesen

Schätztheorie

Sei X die zufällige Variable, welche das untersuchte statistische Merkmal darstellt. Seien $x_1, \ldots, x_n$ *statistische Daten* (Beobachtungen, Stichprobenwerte) für das Merkmal X, die anhand einer Stichprobe erhalten wurden.
Die Daten $x_1, \ldots, x_n$ können als Werte (Realisierungen) von n zufälligen Variablen $X_1, \ldots, X_n$ betrachtet werden; $X_1, \ldots, X_n$ heißen *Stichprobenvariablen* und sind unabhängige zufällige Variablen mit derselben Verteilung wie X.

Seien $X_1, \ldots, X_n$ Stichprobenvariablen und $g : \mathbb{R}^n \to \mathbb{R}$, so dass $g(X_1, \ldots, X_n)$ eine ZG ist.

- $g(X_1, \ldots, X_n)$ heißt Schätzfunktion (ist eine ZG)
- $g(x_1, \ldots, x_n)$ heißt Schätzwert (ist ein Wert)

Beispiele von Schätzfunktionen

- Stichprobenmittel (empirischer Mittelwert)

$$\bar{X}_n = g(X_1, \dots, X_n) = \frac{1}{n}(X_1 + \dots + X_n)$$

- Wert des Stichprobenmittels

$$\bar{x}_n = g(x_1, \dots, x_n) = \frac{1}{n}(x_1 + \dots + x_n)$$

- Stichprobenvarianz (empirische Varianz)

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$$

- Wert der Stichprobenvarianz

$$\tilde{s}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x}_n)^2$$

Beispiele von Schätzfunktionen

- empirische Standardabweichung

$$\tilde{S}_n = \left(\frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2 \right)^{\frac{1}{2}}$$

- Wert der empirischen Standardabweichung

$$\tilde{s}_n = \left(\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x}_n)^2 \right)^{\frac{1}{2}}$$

- empirische Verteilungsfunktion $\hat{F}_n : \mathbb{R} \times \Omega \to \mathbb{R}$

$$\hat{F}_n(x) = \frac{\#\{i \in \{1, \dots, n\} : X_i \leq x\}}{n}, x \in \mathbb{R}$$

- Wert der empirischen Verteilungsfunktion

$$\hat{F}_n(x) = \frac{\#\{i \in \{1, \dots, n\} : x_i \leq x\}}{n}, x \in \mathbb{R}$$

- X ist eine ZG $\hookrightarrow$ das untersuchte statistische Merkmal, welches von einem unbekannten Parameter θ abhängt
- $x_1, \ldots, x_n$ $\hookrightarrow$ *statistische Daten* (Beobachtungen, Stichprobenwerte) für das Merkmal X, die anhand einer Stichprobe erhalten wurden
- $X_1, \ldots, X_n$ sind *Stichprobenvariablen* $\hookrightarrow$ sind *unabhängige* ZG mit *derselben Verteilung* wie X.
- $g(X_1, \ldots, X_n)$ $\hookrightarrow$ Schätzfunktion (ist eine ZG; ist eine Funktion die von den Stichprobenvariablen abhängt)

Die Schätzfunktion $g(X_1, \ldots, X_n)$ ist **erwartungstreu** für den unbekannten Parameter θ, wenn

$$E(g(X_1, \ldots, X_n)) = \theta.$$

▷ Ein Schätzer ist dann erwartungstreu, wenn sein Erwartungswert gleich dem zu schätzenden Parameter ist.

Die Schätzfunktion $g(X_1, \ldots, X_n)$ ist **asymptotisch erwartungstreu** für den unbekannten Parameter θ, wenn

$$\lim_{n \to \infty} E(g(X_1, \ldots, X_n)) = \theta.$$

Die Schätzfunktion $g(X_1, \ldots, X_n)$ ist **konsistent** für den unbekannten Parameter θ, wenn

$$g(X_1, \ldots, X_n) \stackrel{f.s.}{\to} \theta.$$

Beispiele

- Stichprobenmittel (empirischer Mittelwert)

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

ist ein erwartungstreuer und konsistenter Schätzer für den Erwartungswert $E(X)$ des Merkmals X

- Stichprobenvarianz (empirische Varianz)

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$$

ist ein erwartungstreuer und konsistenter Schätzer für die Varianz $V(X)$ des Merkmals X

Übung: Ist

$$g(X_1, \dots, X_n) = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2$$

ein erwartungstreuer und konsistenter Schätzer für die Varianz $V(X)$ des Merkmals X?

- empirische Verteilungsfunktion $\hat{F}_n : \mathbb{R} \times \Omega \to \mathbb{R}$

$$\hat{F}_n(x) = \frac{\#\{i \in \{1, ..., n\} : X_i \leq x\}}{n}, x \in \mathbb{R}$$

ist eine ZG welche die Verteilungsfunktion F von X approximiert.

Eigenschaften der empirischen Verteilungsfunktion

Für alle $x \in \mathbb{R}$ gilt

$$E(\hat{F}_n(x)) = F(x),$$

$$V(\hat{F}_n(x)) = \frac{1}{n} F(x)(1 - F(x)),$$

$$\hat{F}_n(x) \overset{f.s.}{\to} F(x).$$

$\Rightarrow$ $\hat{F}_n(x)$ ist ein erwartungstreuer und konsistenter Schätzer für $F(x)$!