Einführung in die Induktive Statistik: Schätzen von Parametern und Verteilungen

Jan Gertheiss
LMU München

Sommersemester 2011

Vielen Dank an Christian Heumann für das Überlassen von TEX-Code!

Inhalt

- Stichproben
- Parameterchätzer und ihre Eigenschaften
- Konstruktion von Schätzfunktionen
- Konfidenzintervalle
- Nichtparametrische Dichteschätzung

Zufällige Stichproben

Wir unterscheiden:

- Zufällige Stichproben (Definition folgt.)
- Nichtzufällige Stichproben, wie zum Beispiel Auswahl aufs Geratewohl, Quotenverfahren, typische Fälle (Medizin)

Nur zufällige Stichproben ermöglichen Rückschlüsse im Sinne von Wahrscheinlichkeitsaussagen.

Zufällige Stichproben

Grundbegriffe

X Merkmal bzw. Zufallsvariable

Grundsätzlich zwei Typen von "Stichproben":

Fall A:

G konkrete (endliche) Grundgesamtheit mit N Elementen; daraus werden n Elemente zufällig gezogen.

- X_i die Zufallsvariable, die angibt, welchen Wert von X das i-te Element in der Auswahl (Stichprobe) haben wird, $i = 1, \ldots, n$.
  Situation **vor** der Ziehung.
- x_i beobachteter Wert von X beim i-ten Element, d.h. Realisierung von X_i.
  Situation **nach** der Ziehung.

Zufällige Stichproben

Grundbegriffe

Fall B:

Ein Zufallsvorgang wird n-mal wiederholt.

X_i ist die Zufallsvariable, die angibt, welchen Wert X beim i-ten Versuch annehmen wird.
Situation **vor** Durchführung des Zufallsvorgangs.

x_i der beim i-ten Versuch beobachtete Wert von X.
Situation **nach** Durchführung des Zufallsvorgangs.

In beiden Fällen:

$X_1, \ldots, X_n$ Stichprobenvariablen für X

$x_1, \ldots, x_n$ Stichprobenwerte oder (beobachtete) Stichprobe

n Stichprobenumfang

Zufällige Stichproben

Grundbegriffe

Situation in induktiver Statistik:
Verteilung von X nicht oder nicht vollständig bekannt.

Ziel: Schätzen der Verteilung oder von Parametern der Verteilung.

- Fall A: Verteilung gleich der Verteilung von X in Grundgesamtheit, kurz: Verteilung der Grundgesamtheit
- Fall B: Verteilung der Zufallsvariable

Zufällige Stichproben

Grundbegriffe

Im Fall A:

$G = \{1, \ldots, j, \ldots, N\}$,
ξ_j Ausprägung (Wert) des Merkmals X für $j \in G$.

Häufigkeitsverteilung/Verteilungsfunktion von X in G:

$$\begin{aligned} F_G(x) &= \frac{1}{N}\,[\text{ Anzahl der Elemente } j \text{ mit } \xi_j \leq x] \\ &= \text{ empirische Verteilung zu } \xi_1, \ldots, \xi_j, \ldots, \xi_N \\ &= \text{ diskrete Verteilung} \end{aligned}$$

Zufällige Stichproben

Grundbegriffe

Bei großem N betrachtet man oft eine Modellverteilung für X, mit Verteilungsfunktion $F(x)$. Im Sinne eines Modells stimmt F im Allgemeinen nicht exakt, sondern nur approximativ mit F_G überein:

$$F(x) \approx F_G(x)$$

Falls Abweichung vernachlässigbar, wird

$$F(x) \stackrel{!}{=} F_G(x)$$

gesetzt.

Zufällige Stichproben

Grundbegriffe

$\Rightarrow$ Für Fall A bzw. B: "Verteilung von X" und "Verteilung der Grundgesamtheit" identisch.

Für Fall A:

$$E(X) = \mu = \frac{1}{N} \sum_{j=1}^{N} \xi_j = \bar{\xi}$$

$$Var(X) = \sigma^2 = \frac{1}{N} \sum_{j=1}^{N} (\xi_j - \bar{\xi})^2$$

Für Fall B:

μ, σ^2 lassen sich nur als Verteilungsparameter von X auffassen

Zufällige Stichproben

Grundbegriffe

Spezialfall: X dichotom (binär)

Fall B:

$$X = \begin{cases} 1, & A \text{ tritt ein} \\ 0, & \bar{A} \text{ tritt ein} \end{cases}$$

$$\pi = P(X = 1) = P(A), \quad X \sim B(1, \pi)$$

$$\mu = \pi, \quad \sigma^2 = \pi(1 - \pi)$$

Fall A:

$$\xi_j = \begin{cases} 1, & \text{Element } j \in G \text{ hat Eigenschaft } A, \\ 0, & \text{Element } j \in G \text{ hat Eigenschaft } A \text{ nicht}, \end{cases}$$

$j = 1, \ldots, N$.

$$\mu = \pi = \frac{1}{N} \sum_{j=1}^{N} \xi_j$$

$$\Rightarrow \sum_{j=1}^{N}(\xi_j - \pi)^2 = \sum_{j=1}^{N}\xi_j^2 - 2\pi\sum_{j=1}^{N}\xi_j + N\pi^2 \overset{\xi_j^2=\xi_j}{=}$$

$$= \underbrace{\sum_{j=1}^{N}\xi_j}_{N\pi} - 2\pi N\pi + N\pi^2 =$$

$$= N\pi - N\pi^2 = N\pi(1-\pi)$$

$$\Rightarrow \sigma^2 = Var(X) = \pi(1-\pi) = \frac{1}{N}\sum_{j=1}^{N}(\xi_j - \pi)^2$$

Also: A und B passen für μ, σ^2 exakt zusammen.

Zufällige Stichproben

Grundbegriffe

Identisch und/oder unabhängig verteilte Stichproben

- Stichprobe heißt identisch verteilt (oder einfach):
  $\Leftrightarrow$ Stichprobenvariablen $X_1, \ldots, X_n$ sind identisch wie X verteilt.
- Stichprobe heißt unabhängig: $\Leftrightarrow X_1, \ldots, X_n$ unabhängig.
- $X_1, \ldots, X_n$ u.i.v./i.i.d. wie X verteilt: $\Leftrightarrow$ Stichprobe identisch und unabhängig verteilt (independent and identically distributed).
- Fall B: Wird Zufallsvorgang für X n-mal unabhängig wiederholt.
  $\Rightarrow X_1, \ldots, X_n$ i.i.d. wie $X \sim F(x)$.
- Fall A: Ob eine einfache und/oder unabhängige Stichprobe vorliegt, hängt vom Auswahlverfahren ab, siehe später.
- In jedem Fall gilt: Falls ohne Zurücklegen gezogen wird, sind $X_1, \ldots, X_n$ voneinander abhängig. Bei Ziehen mit Zurücklegen: "$X_1, \ldots, X_n$ unabhängig" als sinnvolle Annahme (vgl. Urnenmodell).

Zufällige Stichproben

Rein zufällige Stichproben aus endlichen Grundgesamtheiten

Erinnerung: Nur Zufallsstichproben erlauben induktive Schlüsse mit Wahrscheinlichkeitstheorie.

Einige nichtzufällige Stichproben: Auswahl aufs Geratewohl, "Experten"-Auswahl, Quotenverfahren, ...

Reine (oder uneingeschränkte) Zufallsstichprobe mit Zurücklegen:
$\Leftrightarrow$

- Einzelne Ziehungen sind voneinander unabhängig.
- Jedes Element hat bei jeder Ziehung dieselbe Wahrscheinlichkeit $\frac{1}{N}$ gezogen zu werden.

Es gilt: Eine reine Zufallsstichprobe mit Zurücklegen ist eine identisch und unabhängig verteilte Stichprobe, d.h. $X_1, \ldots, X_n$ i.i.d. wie X.

Zufällige Stichproben

Rein zufällige Stichproben aus endlichen Grundgesamtheiten

Reine Zufallsstichprobe ohne Zurücklegen: $\Leftrightarrow$

- n-mal ohne Zurücklegen ziehen
- Falls nacheinander gezogen wird, hat nach jeder Ziehung eines Elements jedes noch in der Grundgesamtheit vorhandene Element die gleiche Wahrscheinlichkeit als nächstes Element gezogen zu werden. Beim i-ten Zug ist diese Wahrscheinlichkeit

$$\frac{1}{N-(i-1)}, \quad i = 1, \ldots, n.$$

Äquivalent dazu ist: Jede Teilmenge von n Elementen aus G hat die gleiche Wahrscheinlichkeit, als Stichprobe aufzutreten, also $\binom{N}{n}^{-1}$.

Zufällige Stichproben

Rein zufällige Stichproben aus endlichen Grundgesamtheiten

Folgerungen aus der Definition:

1. Für jedes Element aus G ist die Wahrscheinlichkeit, in die Stichprobe zu gelangen (also bei n Ziehungen ausgewählt zu werden) gleich

$$\frac{n}{N} \quad (\text{“Auswahlsatz”}).$$

2. Vor Beginn der Ziehungen ist für jedes Element aus G die Wahrscheinlichkeit, genau beim i-ten Zug gewählt zu werden gleich

$$\frac{1}{N}.$$

3. $X_1, \ldots, X_n$ sind identisch wie X verteilt, d.h. eine reine Zufallsstichprobe ohne Zurücklegen ist eine identisch verteilte Stichprobe.

Zufällige Stichproben

Rein zufällige Stichproben aus endlichen Grundgesamtheiten

Auswahltechniken:

- Zufallszahlen am Rechner
- Weitere Techniken in Vorlesung “Stichproben”

Zufällige Stichproben

Rein zufällige Stichproben aus endlichen Grundgesamtheiten

Frage: Wie schätzt man $\mu = \frac{1}{N} \sum_{j=1}^{N} \xi_j = E(X)$?

Zufällige Stichproben

Rein zufällige Stichproben aus endlichen Grundgesamtheiten

Frage: Wie schätzt man $\mu = \frac{1}{N}\sum_{j=1}^{N} \xi_j = E(X)$?

Naheliegend durch

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) \quad \text{arithmetisches Mittel der Stichprobenwerte.}$$

Dazu gehört die Zufallsvariable

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n).$$

Beim Ziehen mit und ohne Zurücklegen gilt

$$E(\bar{X}) = \frac{1}{n}(\underbrace{E(X_1)}_{=\mu} + \ldots + \underbrace{E(X_n)}_{=\mu}) = \frac{n\mu}{n} = \mu.$$

Zufällige Stichproben

Rein zufällige Stichproben aus endlichen Grundgesamtheiten

Beim Ziehen mit Zurücklegen

$$Var(\bar{X}) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}, \text{ da } X_1, \ldots, X_n \text{ unabhängig.}$$

Aber beim Ziehen ohne Zurücklegen

$$Var(\bar{X}) = ?$$

Zufällige Stichproben

Geschichtete Stichproben

Geschichtete Stichproben:

G in k Schichten $G_1, \ldots, G_j, \ldots, G_k$ zerlegt.

Dann: Reine Zufallsstichprobe ohne bzw. mit Zurücklegen separat in jeder Schicht.

Fragestellungen:

- Wie wählt man Schichten?
  "Schichten in sich möglichst homogen, untereinander möglichst heterogen bzgl. der x-Werte" $\rightarrow$ Vorlesung "Stichproben"
- Wie wählt man Stichprobenumfänge n_j in den Schichten G_j, $j = 1, \ldots, k$?
- Wie schätzt man μ?

Zufällige Stichproben

Geschichtete Stichproben

Notationen:

N_j = Anzahl der Elemente der j-ten Schicht in der Grundgesamtheit (Umfang der Schicht j)

ξ_{ji} = Wert von X, den das i-te Element in der j-ten Schicht besitzt

$\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \xi_{ji} =$ Mittelwert der j-ten Schicht

$\sigma_j^2 = \frac{1}{N_j} \sum_{i=1}^{N_j} (\xi_{ji} - \mu_j)^2 =$ Varianz der j-ten Schicht

n_j = Umfang der aus der j-ten Schicht gezogenen reinen Zufallsstichprobe

$\Rightarrow N = N_1 + \ldots + N_k,\ n = n_1 + \ldots + n_k$

$$\mu = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{N_j} \xi_{ji} = \frac{1}{N} \sum_{j=1}^{k} N_j \mu_j$$

Zufällige Stichproben

Geschichtete Stichproben

Sei $X_{j1}, \ldots, X_{jn_j}$ Teilstichprobe aus j-ter Schicht G_j.

Schätzung für μ ist gewichtetes Stichprobenmittel

$$\bar{\bar{X}} = \frac{1}{N} \sum_{j=1}^{k} N_j \bar{X}_j,$$

mit $\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji}$.

Zufällige Stichproben

Geschichtete Stichproben

- Frage: Wie legt man $n_1, \ldots, n_k$ fest?
  Im Wesentlichen zwei Varianten: Proportionale oder optimale Aufteilung.
- Proportional geschichtete Stichprobe: Auswahlsatz $\frac{n_j}{N_j}$ in jeder Schicht gleich groß, d.h.

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_k}{N_k}$$

$$\Rightarrow n_j = \frac{n}{N} N_j \quad \text{bzw.} \quad \frac{n_j}{n} = \frac{N_j}{N}$$

- Schätzung $\bar{\bar{X}}_{prop}$

$$\bar{\bar{X}}_{prop} := \frac{1}{N} \sum_{j=1}^{k} N_j \bar{X}_j = \frac{1}{N} \sum_{j=1}^{k} N_j \cdot \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ji}$$

Zufällige Stichproben

Geschichtete Stichproben

$\Rightarrow \bar{\bar{X}}_{prop}$ ungewichtetes Stichprobenmittel

- In einer geschichteten Stichprobe kann in den Schichten mit oder ohne Zurücklegen gezogen werden.
- In beiden Fällen gilt: Eine proportional geschichtete Stichprobe stellt eine gleichgewichtete, aber keine reine Zufallsstichprobe dar.
- Gleichgewichtet heißt: Vor Beginn der Ziehungen hat jedes Element die gleiche Wahrscheinlichkeit, in die Stichprobe zu gelangen.
- Optimal geschichtete Stichprobe: "Stichproben-Theorie"

Zufällige Stichproben

Klumpen-(Cluster-)Stichproben

Klumpen-(Cluster-)Stichproben:

- G in Klumpen (Cluster) zerlegt
- Klumpen in sich möglichst heterogen, untereinander möglichst homogen, d.h.: Jeder Klumpen möglichst repräsentativ für G.
- Aus M Klumpen werden m Klumpen durch reine Zufallsauswahl gewählt. Dann Totalerhebungen in den ausgewählten Klumpen.

Zufällige Stichproben

Klumpen-(Cluster-)Stichproben

Schätzen von μ:

Y_i Summe der x-Werte aller Elemente aus Klumpen i

$$\Rightarrow \hat{Y} = \frac{M}{m} \sum_{i=1}^{m} Y_i \quad \text{Schätzung für Gesamtsumme der } x\text{-Werte in } G$$

$$\Rightarrow \bar{X}_{km} = \frac{1}{N} \hat{Y} = \frac{1}{N} \frac{M}{m} \sum_{i=1}^{m} Y_i \quad \text{Schätzung für } \mu$$

Parameterschätzer und ihre Eigenschaften

Problemstellung/Definitionen

X Merkmal bzw. Zufallsvariable

Parameter:

- Kennwerte einer (unbekannten) Verteilung, z.B.

$$E(X), Var(X), \text{ Median}, \rho(X, Y), \ldots$$

- (unbekannte) Parameter eines Verteilungstyps, z.B.

$$\lambda \text{ bei } Po(\lambda);\ \mu, \sigma^2 \text{ bei } N(\mu, \sigma^2);\ \pi \text{ bei } B(n, \pi), \ldots$$

$X_1, \ldots, X_n$ Stichprobenvariablen, hier: i.i.d. wie X
$x_1, \ldots, x_n$ Stichprobenwerte

Parameterschätzer und ihre Eigenschaften

Problemstellung/Definitionen

Generelle Notation	Beispiel
$X \sim F(x\|\theta)$	$\theta = \mu = E(X)$
θ unbekannter Parameter(-vektor)	$X \sim N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$
$\theta \in \Theta$ Parameterraum	$\Theta = \mathbb{R}$ bzw. $\Theta = \mathbb{R} \times \mathbb{R}_+$

Gesucht: Schätzer bzw. Schätzwert für

θ : $\hat{\theta} \equiv t = g(x_1, \ldots, x_n)$

μ : $\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$

σ^2 : $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Parameterschätzer und ihre Eigenschaften

Problemstellung/Definitionen

Definition: Schätzer (Schätzfunktion, Schätzstatistik):

Zufallsvariable $T = g(X_1, \ldots, X_n)$ als (deterministische) Funktion der Stichprobenvariablen $X_1, \ldots, X_n$ heißt *Schätzer*.

Schätzwert $t = g(x_1, \ldots, x_n)$ ist Realisierung von T in der Stichprobe.

Beispiele für Schätzer/Schätzwerte

- Arithmetisches Mittel

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) \quad \text{Schätzer für } \mu = E(X)$$

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) \quad \text{Schätzwert für } \mu = E(X)$$

Parameterschätzer und ihre Eigenschaften

Problemstellung/Definitionen

- Spezialfall: X binär

$$P(X = 1) = \pi = E(X), \quad P(X = 0) = 1 - \pi$$

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) = \frac{H}{n} \quad \text{für } \pi = E(X),$$

wobei H die absolute Häufigkeit von Einsen in der Stichprobe ist.

$$\bar{X} = \frac{H}{n} \quad \text{relative Häufigkeit}$$

Parameterschätzer und ihre Eigenschaften

Problemstellung/Definitionen

- Stichprobenvarianz

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{für } \sigma^2 = Var(X)$$

- Oder: Empirische Varianz

$$\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Frage: Wie "gut" sind solche Schätzer?

Parameterschätzer und ihre Eigenschaften

Erwartungstreue

Beispiel: $\bar{X}$ Schätzer für $\mu = E(X)$

X Zufallsvariable mit $\mu = E(X)$; $X_1, \ldots, X_n$ i.i.d. wie X.

μ unbekannter, aber fester Wert

$$\Rightarrow E(\bar{X}) = E(\tfrac{1}{n}(X_1 + \ldots + X_n)) = \tfrac{1}{n}(\underbrace{E(X_1)}_{\mu} + \ldots + \underbrace{E(X_n)}_{\mu}) = \mu$$

Also: Unabhängig davon, welchen wahren (aber unbekannten) Wert μ tatsächlich besitzt, gilt

$$E(\bar{X}) = \mu\,.$$

D.h.: Der erwartete Wert von $\bar{X}$, in objektiver oder subjektiver Interpretation, ist μ. Damit:

Keine systematische "Verzerrung" beim Schätzen.

Parameterschätzer und ihre Eigenschaften

Erwartungstreue

Definition: Erwartungstreue und Verzerrung

- $T = g(X_1, \ldots, X_n)$ heißt *erwartungstreu (unverzerrt)* für θ: $\Leftrightarrow$

$$E(T) = \theta \quad \text{für alle } \theta \in \Theta$$

- T heißt verzerrt: $\Leftrightarrow$ $E(T) \neq \theta$

  $E(T) - \theta$ heißt *Verzerrung (Bias)*.

- $T_n = g(X_1, \ldots, X_n)$ heißt *asymptotisch erwartungstreu*: $\Leftrightarrow$

$$\lim_{n \to \infty} E(T_n) = \theta$$

Parameterschätzer und ihre Eigenschaften

Erwartungstreue

Beispiele:

- $E(\bar{X}) = \mu$, d.h. $\bar{X}$ für μ unverzerrt.
- $\frac{H}{n}$ für π unverzerrt.
- $E(\widetilde{S}^2) = E(\frac{1}{n}\sum_i (X_i - \bar{X})^2) = \frac{n-1}{n}\sigma^2$

  $\Rightarrow \widetilde{S}^2$ verzerrt

  $Bias(\widetilde{S}^2) = E(\widetilde{S}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$

  $\widetilde{S}^2$ asymptotisch ewartungstreu, da Verzerrung $-\frac{\sigma^2}{n} \to 0$ für $n \to \infty$
- $E(S^2) = \sigma^2, \quad S^2$ unverzerrt

Parameterschätzer und ihre Eigenschaften

Varianz, MSE und Konsistenz

Neben $E(T) - \theta$ ist auch $Var(T)$, d.h. die Varianz bzw. "Ungenauigkeit" des Schätzers ein Maß für die Güte von T.

Parameterschätzer und ihre Eigenschaften

Varianz, MSE und Konsistenz

Definition: Varianz und Standardabweichung eines Schätzers

$$T = g(X_1, \ldots, X_n) \text{ Schätzer}$$

$$Var(T) = Var\{g(X_1, \ldots, X_n)\} \text{ Varianz von } T$$

$$\sigma_T = +\sqrt{Var(T)} \text{ Standardabweichung von } T$$

Bemerkung:

Exakte analytische Formeln nur in einfachen Fällen angebbar; oft Approximation für großes n.

Beispiel: $\bar{X}$

$\sigma^2_{\bar{X}} = Var(\bar{X}) = \frac{\sigma^2}{n}, \quad \sigma^2 = Var(X), \quad \sigma^2$ aber unbekannt

Schätzer: $\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$

Parameterschätzer und ihre Eigenschaften

Varianz, MSE und Konsistenz

Definition: Erwartete quadratische Abweichung, Mean Squared Error

$$MSE(T) = E\{(T-\theta)^2\} = Var(T) + (Bias(T)^2)$$

Bemerkung:

Der Mean Squared Error $MSE(T)$ fasst als Erwartungswert der quadratischen Abweichung $(T-\theta)^2$ des Schätzers T vom zu schätzenden Parameter θ die Varianz und die quadrierte Verzerrung in einem gemeinsamen Gütekriterium für T zusammen.

Parameterschätzer und ihre Eigenschaften

Varianz, MSE und Konsistenz

Definition: Konsistenz

- T heißt (MSE-)konsistent für $\theta :\Leftrightarrow MSE(T) \to 0$ für $n \to \infty$
- T heißt (schwach) konsistent für $\theta :\Leftrightarrow P(|T - \theta| < \epsilon) \to 1 \ \ \forall \epsilon > 0$ und für $n \to \infty$

Bemerkungen:

- Damit $MSE(T) = Var(T) + (Bias(T))^2 \to 0$ geht, muss $Var(T) \to 0$ und $Bias(T) \to 0$ gelten.
- Aus MSE-Konsistenz folgt schwache Konsistenz mit Hilfe des Satzes von Tschebyscheff.

Parameterschätzer und ihre Eigenschaften

Varianz, MSE und Konsistenz

Beispiele

- $MSE(\bar{X}) = \frac{\sigma^2}{n} + \underbrace{(Bias(\bar{X}))^2}_{=0} = \frac{\sigma^2}{n} \to 0$ für $n \to \infty$

Annahme: $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$. Dann gilt:

- $MSE(S^2) = ?$
- $MSE(\widetilde{S}^2) = ?$

Parameterschätzer und ihre Eigenschaften

Effiziente (oder "wirksamste") Schätzstatistiken

$MSE(T)$ Maß für Güte von T

$Var(T)$ Maß für Varianz von T

$\Rightarrow$ Man kann zwei Schätzer T_1, T_2 bzgl. MSE (oder auch Var) vergleichen.

Definition: T_1 heißt (MSE-)effizienter als T_2:

$$\Leftrightarrow MSE(T_1) \leq MSE(T_2)$$

für alle zugelassenen Verteilungen.

Bei erwartungstreuen Schätzern T_1, T_2:

$$Bias(T_1) = Bias(T_2) = 0 \quad \Rightarrow \quad MSE(T_i) = Var(T_i),\ i = 1, 2$$

$$\Rightarrow T_1 \text{ effizienter als } T_2 \quad \Leftrightarrow \quad Var(T_1) \leq Var(T_2)$$

Parameterschätzer und ihre Eigenschaften

Effiziente (oder "wirksamste") Schätzstatistiken

Definition:

Ist ein Schätzer T besser als alle zur "Konkurrenz" zugelassenen anderen Schätzer $\widetilde{T}$, so heißt T (MSE-)effizient für θ.

Beispiele:

- $\bar{X}$ für μ, unter allen erwartungstreuen Schätzern für μ, wenn alle Normalverteilungen zugelassen sind.
- $\bar{X}$ für den Anteilswert π, unter allen erwartungstreuen Schätzern für π, wenn dichotome Grundgesamtheiten betrachtet werden und alle Bernoulli-Verteilungen zugelassen sind.

Konstruktion von Schätzfunktionen

Ziel:

Einführung in generelle Ansätze/Konzepte, wie man Schätzer insbesondere auch in komplexeren Nicht-Standardsituationen findet bzw. konstruiert und berechnet.

Konzepte/Methoden:

- Maximum-Likelihood-Schätzung
- Kleinste-Quadrate-Schätzung
- Bayes-Schätzung
- Momenten-Methode

Schwerpunkt: Maximum-Likelihood-Methode

Maximum-Likelihood-Schätzung

Voraussetzungen hier:

- Stichprobenvariablen $X_1, \ldots, X_n$ i.i.d. wie $X \sim f(x|\theta)$
- $f(x|\theta)$ diskrete Dichte (Wahrscheinlichkeitsfunktion) oder stetige Dichte

Konstruktion von Schätzfunktionen

Maximum-Likelihood-Schätzung

Grundidee für diskretes X:

Sei $X_1 = x_1, \ldots, X_n = x_n$ die konkrete Stichprobe.

Gesucht: Schätzwert $\hat{\theta}$ (bzw. T) für θ

Konzept: Bestimme/konstruiere $\hat{\theta}$ so, dass die Wahrscheinlichkeit für Auftreten der Stichprobe maximal wird, d.h. $\hat{\theta}$ so, dass

$$P(X_1 = x_1, \ldots, X_n = x_n | \theta) \to \max_{\theta}.$$

Es ist

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n | \theta) &= \underbrace{f(x_1, \ldots, x_n | \theta)}_{\text{gemeinsame W'fkt.}} \\ &= P(X_1 = x_1 | \theta) \cdot \ldots \cdot P(X_n = x_n | \theta) \\ &= f(x_1 | \theta) \cdot \ldots \cdot f(x_n | \theta). \end{aligned}$$

Konstruktion von Schätzfunktionen

Maximum-Likelihood-Schätzung

Definition: Likelihoodfunktion

Bei gegebenen $x_1, \ldots, x_n$ heißt

$$L(\theta) = f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta)$$

Likelihoodfunktion für θ.

Definition: Likelihood-Prinzip/Maximum-Likelihood-Schätzung

Bestimme $\hat{\theta}$ so, dass

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

Konstruktion von Schätzfunktionen

Maximum-Likelihood-Schätzung

Für stetige Zufallsvariablen X mit Dichte $f(x|\theta)$ überträgt man das Konzept in völliger Analogie:

Wähle θ so, dass die gemeinsame Dichte $L(\theta) = f(x_1, \ldots, x_n|\theta)$ $= f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta)$ maximal wird:

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

Im Allgemeinen ist $\hat{\theta}$ eine (komplizierte, nichtlineare) Funktion von $x_1, \ldots, x_n$:

$$\hat{\theta} = g(x_1, \ldots, x_n).$$

Setzt man statt der Realisierungen $x_1, \ldots, x_n$ die Stichprobenvariablen $X_1, \ldots, X_n$ ein, wird $T \equiv \hat{\theta}$ zum Maximum-Likelihood-Schätzer.

Konstruktion von Schätzfunktionen

Maximum-Likelihood-Schätzung

Konkrete Berechnung erfolgt meist durch Maximieren der log-Likelihood

$$\log L(\theta) = \log f(x_1|\theta) + \ldots + \log f(x_n|\theta) =$$
$$= \sum_{i=1}^{n} \log f(x_i|\theta)$$

Maxima $\hat{\theta}$ von $L(\theta)$ und $\log L(\theta)$ sind identisch, da log eine streng monotone Transformation ist.

Das Maximum wird i.d.R. durch Nullsetzen der ersten Ableitung berechnet.

Konstruktion von Schätzfunktionen

Maximum-Likelihood-Schätzung

Beispiele:

- Poisson-Verteilung: $X_1, \ldots, X_4$ i.i.d. $Po(\lambda)$ mit Realisierungen $x_1 = 2, x_2 = 4, x_3 = 6, x_4 = 3$.

  $\Rightarrow$ Likelihoodfunktion

$$L(\lambda) = f(x_1|\lambda) \cdots f(x_4|\lambda) = e^{-\lambda}\frac{\lambda^2}{2!}e^{-\lambda}\frac{\lambda^4}{4!}e^{-\lambda}\frac{\lambda^6}{6!}e^{-\lambda}\frac{\lambda^3}{3!}$$

$$= e^{-4\lambda}\lambda^{15}\frac{1}{2!\ 4!\ 6!\ 3!}$$

  $\Rightarrow$ Log-Likelihoodfunktion

$$\log L(\lambda) = -4\lambda + 15\log\lambda - \log(2!\ 4!\ 6!\ 3!)$$

  Ableiten und Nullsetzen

$$\Rightarrow \quad \frac{\partial \log L(\lambda)}{\partial \lambda} = -4 + \frac{15}{\hat{\lambda}} = 0 \Leftrightarrow \hat{\lambda} = \frac{15}{4}$$

Konstruktion von Schätzfunktionen

Maximum-Likelihood-Schätzung

- Normalverteilung: $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ mit Realisierungen $x_1, \ldots, x_n$.

$$\Rightarrow \quad L(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mathbf{x_1}-\mu)^2}{2\sigma^2}} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mathbf{x_n}-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \log L(\mu, \sigma) &= \sum_{i=1}^{n} \left[\log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= \sum_{i=1}^{n} \left[-\log \sqrt{2\pi} - \log \sigma - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \end{aligned}$$

Konstruktion von Schätzfunktionen

Maximum-Likelihood-Schätzung

$$\Rightarrow \quad \frac{\partial \log L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \hat{\mu}}{\hat{\sigma}^2} = 0$$

$$\frac{\partial \log L(\mu, \sigma)}{\partial \sigma} = \sum_{i=1}^{n} \left(-\frac{1}{\hat{\sigma}} + \frac{2(x_i - \hat{\mu})^2}{2\hat{\sigma}^3} \right) = 0$$

$$\Rightarrow \quad \hat{\mu} = \bar{x}, \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Bayes-Schätzung

Basiert auf subjektivem Wahrscheinlichkeitsbegriff; dennoch enge Verbindung zur Likelihood-Schätzung. Besonders für hochdimensionale, komplexe Modelle geeignet; "Revival" etwa seit 1990.

"Subjektives" Grundverständnis:

- θ wird als Realisierung einer Zufallsvariablen Θ aufgefasst
- Unsicherheit/Unkenntnis über θ wird durch eine sog. **priori**-Verteilung (stetige oder diskrete Dichte) $f(\theta)$ bewertet. Meist: Θ als stetige Zufallsvariable, $f(\theta)$ als stetige Dichte.

Die Bayes-Inferenz beruht nun auf der sog. **posteriori**-Verteilung von Θ, gegeben die Daten $x_1, \ldots, x_n$. Dazu benötigen wir den Satz von Bayes für Dichten.

Konstruktion von Schätzfunktionen

Bayes-Schätzung

Notation

$f(x\|\theta)$	bedingte Wahrscheinlichkeitsfunktion bzw. Dichte von X, gegeben $\Theta = \theta$.
$f(x)$	Randverteilung oder -dichte von X.
$f(\theta)$	a priori Wahrscheinlichkeitsfunktion oder a priori Dichte von Θ (d.h. die Randverteilung von Θ).
$f(\theta\|x)$	a posteriori (oder bedingte) Wahrscheinlichkeitsfunktion oder Dichte von Θ, gegeben die Beobachtung $X = x$.
$f(x, \theta)$	gemeinsame Wahrscheinlichkeitsfunktion oder Dichte.

Konstruktion von Schätzfunktionen

Bayes-Schätzung

Dann gilt folgende Form des **Satzes von Bayes**:

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{f(x)}$$

Θ und X diskret:

$$\Rightarrow \quad P(X = x) = f(x) = \sum_j f(x|\theta_j)f(\theta_j),$$

wobei über die möglichen Werte θ_j von Θ summiert wird.

Θ stetig:

$$\Rightarrow \quad f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)\,d\theta} = \frac{f(x|\theta)f(\theta)}{f(x)}$$

Dabei kann X stetig oder diskret sein.

Konstruktion von Schätzfunktionen

Bayes-Schätzung

Für Stichprobe $x = (x_1, \ldots, x_n)$ aus $f(x_1, \ldots, x_n|\theta)$:

$$f(x) \rightarrow f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdot \ldots \cdot f(x_n|\theta) = L(\theta)$$

$\Rightarrow$ **Bayes-Inferenz, Bayesianisches Lernen**:

Die Wahrscheinlichkeitsfunktion oder Dichte von X, gegeben θ, sei

$$f(x|\theta),$$

und

$$L(\theta) = f(x_1, \ldots, x_n|\theta)$$

die gemeinsame Dichte bzw. Likelihoodfunktion für n unabhängige Wiederholungen von X.

Konstruktion von Schätzfunktionen

Bayes-Schätzung

Für den unbekannten Parameter wird eine a priori Dichte $f(\theta)$ spezifiziert.

Dann ist die a posteriori Dichte über den Satz von Bayes bestimmt durch

$$f(\theta|x_1,\ldots,x_n) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)f(\theta)}{\int f(x_1|\theta)\cdots f(x_n|\theta)f(\theta)\,d\theta} = \\ = \frac{L(\theta)f(\theta)}{\int L(\theta)f(\theta)\,d\theta}\,.$$

Konstruktion von Schätzfunktionen

Bayes-Schätzung

Bayes-Schätzer

- *a posteriori Erwartungswert:*

$$\hat{\theta} = E(\theta|x_1, \ldots, x_n) = \int \theta f(\theta|x_1, \ldots, x_n)\, d\theta$$

- *a posteriori Modus* oder *maximum a posteriori (MAP)* Schätzer: Wähle denjenigen Parameterwert $\hat{\theta}$, für den die a posteriori Dichte maximal wird, d.h.

$$L(\hat{\theta}) f(\hat{\theta}) = \max_{\theta} L(\theta) f(\theta)$$

bzw.

$$\log L(\hat{\theta}) + \log f(\hat{\theta}) = \max_{\theta} \{\log L(\theta) + \log f(\theta)\}\,.$$

Konfidenzintervalle

Bisher:

(Punkt-)Schätzer T bzw. $\hat{\theta}$ für θ liefert einen Schätzwert t bzw. $\hat{\theta}$; i.A. $\hat{\theta} \neq \theta$.

Jetzt:

Angabe eines Intervalls, das θ mit hoher Wahrscheinlichkeit $1 - \alpha$ überdeckt. Irrtumswahrscheinlichkeit α z.B. $= 0.1, 0.05, 0.01$.

$1 - \alpha$: Sicherheits- oder Konfidenzwahrscheinlichkeit

Konfidenzintervalle

Allgemeine Definition

Definition: $(1-\alpha)$-Konfidenzintervall (KI)

Irrtumswahrscheinlichkeit α vorgegeben.

Untere und obere Intervallgrenzen

$$G_u = g_u(X_1, \ldots, X_n) \qquad \text{und} \qquad G_o = g_o(X_1, \ldots, X_n)$$

bilden $(1-\alpha)$-Konfidenzintervall (Vertrauensintervall): $\Leftrightarrow$

$$P(G_u \leq G_o) = 1,\ P(G_u \leq \theta \leq G_o) = 1 - \alpha$$

Die Intervallgrenzen G_u und G_o sind Zufallsvariablen! Somit ist auch das Intervall $[G_u, G_o]$ zufällig.

Realisiertes Konfidenzintervall:
$[g_u, g_o]$; $g_u = g_u(x_1, \ldots, x_n),\ g_o = g_o(x_1, \ldots, x_n)$

$\Rightarrow$ Warnung: Eine Aussage wie "θ liegt mit Wahrscheinlichkeit $1-\alpha$ in $[g_u, g_o]$" ist Unsinn!

Konfidenzintervalle

Konfidenzintervalle für Erwartungswert, Varianz und Anteilswert

Beispiele:

- $X \sim N(\mu, \sigma^2)$, σ^2 *bekannt*; $X_1, \ldots, X_n$ i.i.d. wie X.
  $(1-\alpha)$-Konfidenzintervall für μ:

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- $X \sim N(\mu, \sigma^2)$, σ^2 *unbekannt*; $X_1, \ldots, X_n$ i.i.d. wie X.
  $(1-\alpha)$-Konfidenzintervall für μ:

$$\left[\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right],$$

d.h. ersetze σ durch Schätzer $S = \sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}$
und $z_{1-\alpha/2}$ durch $t_{1-\alpha/2}$.

Konfidenzintervalle

Konfidenzintervalle für Erwartungswert, Varianz und Anteilswert

- $X \sim N(\mu, \sigma^2)$, μ *unbekannt*; $X_1, \ldots, X_n$ i.i.d. wie X. Zweiseitiges $(1-\alpha)$-Konfidenzintervall für σ^2:

$$\left[\frac{(n-1)S^2}{q_{1-\alpha/2}}, \frac{(n-1)S^2}{q_{\alpha/2}}\right],$$

mit $q_{\alpha/2}$ als dem $(\alpha/2)$-Quantil der $\chi^2(n-1)$-Verteilung, und $q_{1-\alpha/2}$ als dem $(1-\alpha/2)$-Quantil der $\chi^2(n-1)$-Verteilung.

Konfidenzintervalle

Konfidenzintervalle für Erwartungswert, Varianz und Anteilswert

- Konfidenzintervall für μ ohne Normalverteilungsannahme; approximatives $(1-\alpha)$-Konfidenzintervall (für $n \geq 30$):

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}},\ \bar{X} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right],$$

  (Falls Varianz σ^2 bekannt, ersetze S durch σ.)

- $X \sim B(1, \pi)$, Konfidenzintervall für Anteilswert π; approximatives $(1-\alpha)$-Konfidenzintervall (für $n \geq 30$):

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Nichtparametrische Dichteschätzung

Bisher:	Funktionale Form der Dichte $f(x\|\theta)$ bis auf unbekannte Parameter θ bekannt; z.B. $X \sim N(\mu, \sigma^2), X \sim Po(\lambda)$, etc.
Jetzt:	Kein parametrischer Verteilungstyp vorausgesetzt; $X \sim f(x)$ stetig.
Ziel:	"Nichtparametrische" Schätzung von $f(x)$.
Bekannt:	$\hat{f}(x)$ über Histogramm.

Nichtparametrische Dichteschätzung

Bisher: Funktionale Form der Dichte $f(x|\theta)$ bis auf unbekannte. Parameter θ bekannt; z.B. $X \sim N(\mu, \sigma^2), X \sim Po(\lambda)$, etc.

Jetzt: Kein parametrischer Verteilungstyp vorausgesetzt; $X \sim f(x)$ stetig.

Ziel: "Nichtparametrische" Schätzung von $f(x)$.

Bekannt: $\hat{f}(x)$ über Histogramm.

Besser: Gleitendes Histogramm:

$$\hat{f}(x) = \frac{\frac{1}{n} \cdot \text{ Anzahl der Daten } x_i \text{ in } [x-h,\, x+h)}{2h}$$

Nichtparametrische Dichteschätzung

Darstellung des gleitenden Histogramms durch Rechteckfenster

- Einheitsrechteckfenster/Einheits-"Kern"

$$K(u) = \begin{cases} \frac{1}{2} & \text{für } -1 \leq u < 1 \\ 0 & \text{sonst} \end{cases}$$

- Rechteckfenster über x_i

$$\frac{1}{h} K\left(\frac{x - x_i}{h}\right) = \begin{cases} \frac{1}{2h} & x_i - h \leq x < x_i + h \\ 0 & \text{sonst} \end{cases}$$

$$\Rightarrow \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

Nichtparametrische Dichteschätzung

Weitere Kerne:

- Epanechnikov-Kern:

$$K(u) = \frac{3}{4}(1 - u^2) \text{ für } -1 \leq u < 1,\ 0 \text{ sonst}$$

- Bisquare-Kern:

$$K(u) = \frac{15}{16}(1 - u^2)^2 \text{ für } -1 \leq u < 1,\ 0 \text{ sonst}$$

- Gauß-Kern:

$$K(u) = \frac{1}{\sqrt{2\pi}}\ exp\left(-\frac{1}{2}u^2\right) \text{ für } u \in \mathbb{R}$$

Nichtparametrische Dichteschätzung

Definition: Kern-Dichteschätzer

Sei $K(u)$ eine Kernfunktion. Zu gegebenen Daten $x_1, \ldots, x_n$ ist dann

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

ein *(Kern-)Dichteschätzer* für $f(x)$.