15. Vorlesung: Einführung in der Statistik

Nikolas Tapia

10. Juni 2024, Stochastik für Informatik(er)

15. Vorlesung: Einführung in der Statistik

Nikolas Tapia

10. Juni 2024, Stochastik für Informatik(er)

Statistik

Wahrscheinlichkeitstheorie

- Theoretische Grundlagen $\Rightarrow$ ZV
- Modellierung von Zufallsprozessen $(X_1, \dots X_n) \in \mathbb{R}^n$
- Axiomatische Definitionen

Statistik

- Analyse und Interpretation von Daten
- Aufstellung von Modellen $\downarrow$ $(x_1, \dots, x_n) \in \mathbb{R}^n$
- Fragen:
  - Bestimmung von Kenngrößen
  - Schätzung von Parametern
  - Hypothesentests

Hypothese

1. Die gemessenen Daten sind einzelne *Realisierungen* von unabhängigen, identisch verteilten Zufallsvariablen. $(X_1, \dots, X_n)$ u.i.v ZV $\Rightarrow (x_1, \dots, x_n) \in \mathbb{R}^n$
2. Die gemessenen Daten stellen eine *Stichprobe* aus einer *Population* dar.

Grundannahmen

Definition 15.1

Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ ein Vektor von Messwerten. Sei $x \in \mathbb{R}$. Die **absolute Häufigkeit** von x ist

$$H(x) := |\{i : x_i = x\}| .$$

Die **relative Häufigkeit** von x ist

$$h(x) := \frac{H(x)}{n}.$$

Klasseneinteilung

Definition 15.2

Sei $(x_1, \dots, x_n) \in \mathbb{R}^n$ ein Vektor von Messwerten. Eine **Klasseneinteilung** ist eine Zusammenfassung der Messwerte zu disjunkte Mengen $A_1, \dots, A_m \subseteq \mathbb{R}$ mit $m \leq n$, sodass jedes x_i zu genau einem A_j gehört.

Die **absolute Häufigkeit** einer Klasse A ist

$$H(A) := \left|\{i : x_i \in A\}\right|.$$

Die **relative Häufigkeit** einer Klasse A ist

$$h(A) := \frac{H(A)}{n}.$$

Stichproben

Beispiel

Dauer in Sekunden, gerundet auf eine Kommastelle.

$$(10{,}9;\ 6{,}8;\ 8{,}9;\ 5{,}6;\ 9{,}8;\ 2{,}3;\ 4{,}6;\ 2{,}8;\ 6{,}5;\ 3,\ 10{,}7;\ 8{,}0;\ 1{,}8;\ 8{,}9;\ 10{,}7)$$

$$H(10{,}9) = 1,\quad H(10{,}7) = 2,\quad H(4{,}0) = 0$$

$$h(10{,}9) = \frac{1}{14},\quad h(10{,}7) = \frac{1}{7},\quad h(4{,}0) = 0$$

Empirische Kenngrößen

Definition 15.3

Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ ein Vektor von Messwerten. Das **empirische Mittel** von $(x_1, \ldots, x_n)$ ist definiert als

$$\bar{\mu}_n(x_1, \ldots, x_n) := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Definition 15.4

Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ ein Vektor von Messwerten. Die **empirische Varianz** von $(x_1, \ldots, x_n)$ ist definiert als

statt $\frac{1}{n}$!

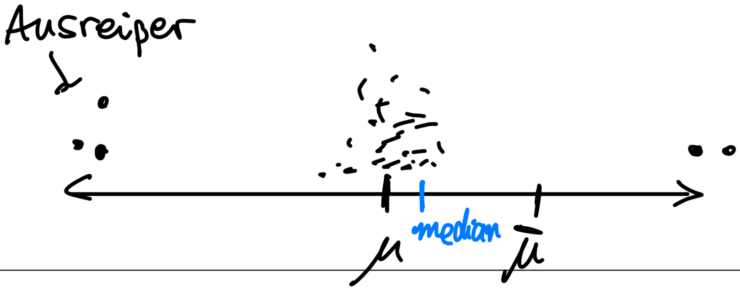
$$\bar{\sigma}_n^2(x_1, \ldots, x_n) := \boxed{\frac{1}{n-1}} \sum_{i=1}^{n} (x_i - \bar{\mu}_n(x_1, \ldots, x_n))^2.$$

Die **empirische Standardabweichung** von $(x_1, \ldots, x_n)$ ist $\bar{\sigma}_n := \sqrt{\bar{\sigma}_n^2}$.

Empirische Kenngrößen

Definition 15.5

Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ ein Vektor von Messwerten. Die **empirische Median** von $(x_1, \ldots, x_n)$ ist definiert als der Wert in der Mitte der geordneten Liste. Falls n gerade ist, ist der Median der Durchschnitt der beiden mittleren Werte.



Empirische Kenngrößen

Beispiel

$$x = (10{,}9;\ 6{,}8;\ 8{,}9;\ 5{,}6;\ 9{,}8;\ 2{,}3;\ 4{,}6;\ 2{,}8;\ 6{,}5;\ 3{,}1;\ 10{,}7;\ 8{,}0;\ 1{,}8;\ 8{,}9;\ 10{,}7)$$

$\bar{\mu}_n(x) = 8{,}014$ empirisches Mittel

$\bar{\sigma}_n^2(x) = 4{,}677$ empirische Varianz.

<u>median:</u>

$$(2{,}3;\ 2{,}8;\ \ldots) \Rightarrow \frac{8{,}1 + 8{,}2}{2} = 8{,}15 = \text{median} \simeq \bar{\mu}_n$$

Schätzfunktion

Definition 15.6

Seien $X_1, \dots, X_n$ Zufallsvariablen. Eine **Schätzfunktion** zur Stichprobengröße n ist eine Funktion $\theta_n \colon \mathbb{R}^n \to \mathbb{R}$, welche die Zufallsvariablen $X_1, \dots, X_n$ auf $\theta_n(X_1, \dots, X_n)$ abbildet.

<u>Bem:</u> $\theta_n(X_1, \dots, X_n)$ ist eine ZV.

$\overline{\theta}_n := \theta_n(x_1, \dots, x_n) \in \mathbb{R} \Rightarrow$ Schätzer für θ

Parameter der Verteilung der X_i.

Schätzfunktion

Beispiel

$$\bar{\mu}_n(x_1, \dots, x_n) := \frac{1}{n} \sum_{i=1}^{n} x_i$$

Realisierung

$$\left((x_1, \dots, x_n) \overset{\downarrow}{=} (X_1, \dots, X_n) \right)$$

$$\bar{\mu}_n(X_1, \dots, X_n) \approx \mu \quad \text{falls} \quad \mathbb{E}[X_1] = \dots = \mathbb{E}[X_n] = \mu.$$

Satz der Großen Zahlen:

$$\bar{\mu}_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mathbb{E}[X_1] = \mu$$

Schätzfunktion

$$\mathbb{E}\left[\bar{\mu}_n(X_1, \dots, X_n)\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu.$$

Beispiel

Schätzfunktion

$$\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$$

Beispiel

$$\mathbb{E}\left[\tilde{\sigma}_n(X_1, \ldots, X_n)\right] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{\mu}_n(X_1, \ldots, X_n)\right)^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \bar{\mu}_n)^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2 - 2\bar{\mu}_n X_i + \bar{\mu}_n^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\underbrace{\mathbb{E}\left[X_i^2\right]}_{\mu^2+\sigma^2} - \frac{2}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[\bar{\mu}_n X_i\right] + \frac{n}{n-1}\mathbb{E}\left[\bar{\mu}_n^2\right]$$

Schätzfunktion

$$\boxed{(a+b)^2 = a^2+b^2+2ab}$$

Beispiel

$$\mathbb{E}\left[\bar{\mu}_n^2\right] = \frac{1}{n^2}\,\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right]$$

$$= \frac{1}{n^2}\left(n(\mu^2+\sigma^2) + 2\sum_{i<j}\mathbb{E}[X_i X_j]\right)$$

$$= \frac{1}{n}\mu^2 + \frac{1}{n}\sigma^2$$

$$\Rightarrow \quad \mathbb{E}\left[\bar{\sigma}_n^2(X_1,\ldots,X_n)\right] = \mathbb{V}(X_1) = \sigma^2$$

Schätzfunktion

Beispiel

Eigenschaften von Schätzern

Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ ein Vektor von Messwerten, welche als Realisierungen von identisch verteilten Zufallsvariablen $(X_1, \ldots, X_n)$ mit Parameter θ gelten.
Sei $\theta_n \colon \mathbb{R}^n \to \mathbb{R}$ eine Schätzfunktion zur Stichprobengröße n.

Definition 15.7

Der Schätzer θ_n heißt **erwartungstreu für** θ, falls

Schätzfunktion

$$\mathbb{E}[\theta_n(X_1, \ldots, X_n)] = \theta.$$

$\Longrightarrow \theta_n(x_1, \ldots, x_n) \approx \theta$

Parameter.

Definition 15.8

Der Schätzer θ_n heißt **konsistent**, falls

$$\lim_{n\to\infty} \theta_n(x_1, \ldots, x_n) = \theta.$$

Definition 15.9

Der Schätzer θ_n heißt **effizient** falls

$$\lim_{n\to\infty} \mathbb{V}[\theta_n(X_1, \ldots, X_n)] = 0.$$

$\theta_n(X_1, \ldots, X_n) \approx$ kst.

für große n.

Schätzern

Erwartungswert $\mathbb{E}[X_1] = \mu$.

$\mathbb{E}[\bar{\mu}_n(X_1, \ldots, X_n)] = \mu \Rightarrow \bar{\mu}_n$ erwartungstreu

Satz der großen Zahlen $\rightarrow \Rightarrow \bar{\mu}_n$ konsistent.

$\bar{\mu}_n$ ist auch effizient.

$\mathbb{E}[\bar{\sigma}_n^2(X_1, .., X_n)] = \sigma^2 \Rightarrow \bar{\sigma}_n^2$ erwartungstreu

$= \mathbb{V}(X_1) \Rightarrow \bar{\sigma}_n^2$ konsistent.

Beispiel

Schätzern

$$\tilde{\sigma}_n^2(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{\mu}_n(x_1, \ldots, x_n)\right)^2$$

$\tilde{\sigma}_n^2$ ist konsistent, aber nicht erwartungstreu.

Beispiel