Stochastik für die Informatik, Vorlesung 16

Inhalt

- Einführung in die Statistik
- Parameterschätzung
- Maximum Likelihood Schätzung

Lernziele

- Grundprinzipien der Statistik kennen
- Wichtige Schätzer für Parameter kennen
- Maximum Likelihood Schätzer berechnen können

Vorkenntnisse Erwartungswert, Varianz, Gesetz der großen Zahlen; Differentialrechnung, Logarithmengesetze

Kapitel 8: Parameterschätzung

Wahrscheinlichkeitstheorie: Allgemeine Theorie angewandt auf konkrete Modelle.

Woher kommt das konkrete Modell?

Statistik: Durch Analyse und Interpretation von Daten aus Beobachtungen können Modelle aufgestellt, getestet und kalibriert werden

Statistik: Grundproblem

Gegeben: Große Anzahl von Messwerten (Daten) $x_1, ..., x_n$ mit $x_i \in \mathbb{R}$

Allgemeines Ziel der Statistik: Aufstellen eines mathematischen Modells, welches diese Daten beschreibt, und welches mit wahrscheinlichkeitstheoretischen Methoden untersucht werden kann.

Grundaufgaben:

- Schätzer, Bestimmung von Kenngrößen
- Konfidenzintervalle
- Hypothesentests

Grundannahmen der Statistik:

Betreffend der gemessenen Daten $x_1, ..., x_n$ gehen wir von einer der beiden (sich nicht ausschließenden) Grundannahmen aus:

- Die gemessenen Daten sind einzelne Realisierungen von (unabhängigen, identisch verteilten) Zufallsvariablen $X_1, ..., X_n$
- Die gemessenen Daten stellen eine Stichprobe aus einer (noch viel größeren) Population dar.

Unter dieser Prämisse will man mittels der Stichprobe Aussagen über die zugrundeliegende Zufallsvariablen bzw. über die gesamte Population machen

Beschreibende Statistik

Beispiel 8.1: Eine Messreihe. Messung der Zeit bis zur Betriebsbereitschaft eines elektronischen Geräts (in Sekunden)

Messung Nr.	1	2	3	4	5	6	7	8
Wert	10.9	6.8	9.5	6.9	8.2	3.4	6.2	8.6

Messung Nr.	9	10	11	12	13	14
Wert	5.3	10.7	8.1	8.0	8.9	10.7

- Wie können solche Daten geeignet dargestellt werden?
- Welche Informationen können aus diesen Daten abgelesen werden?
- Um welche "Art" von Daten handelt es sich hier, und inwiefern sind sie mit unseren Grundannahmen kompatibel?
- Welche weiterführenden Fragestellungen ergeben sich möglicherweise?

Häufigkeiten

(Def. 8.1) Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten). Sei $x \in \mathbb{R}$. Die absolute Häufigkeit von x ist

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

d.h. sie gibt an, wie oft der Wert x im Vektor $(x_1, ..., x_n)$ vorkommt. Die relative Häufigkeit von x ist

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

- $H(x) \in \mathbb{N}_0, h(x) \in [0, 1]$.
- (Beispiel 8.1)
- (Bem. stetige und diskrete Merkmale)

(Def. 8.2) Ein Histogramm der Daten ist ein Plot der Funktion $x \mapsto H(x)$ oder $x \mapsto h(x)$, oder, im Falle einer Einteilung in Klassen, der Funktion $A \mapsto H(A)$ bzw. $A \mapsto h(A)$.

Kenngrößen von Daten

(Def. 8.4) Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Das empirische Mittel von $(x_1, ..., x_n)$ ist definiert als

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

(Def. 8.5) Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Der Median von $(x_1, ..., x_n)$ ist definiert als der Wert in der Mitte der geordneten Liste. Falls n gerade ist, wird der Durchschnitt der beiden mittleren Werten gebildet.

(Def. 8.6) Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Die empirische Varianz von $(x_1, ..., x_n)$ ist definiert als

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

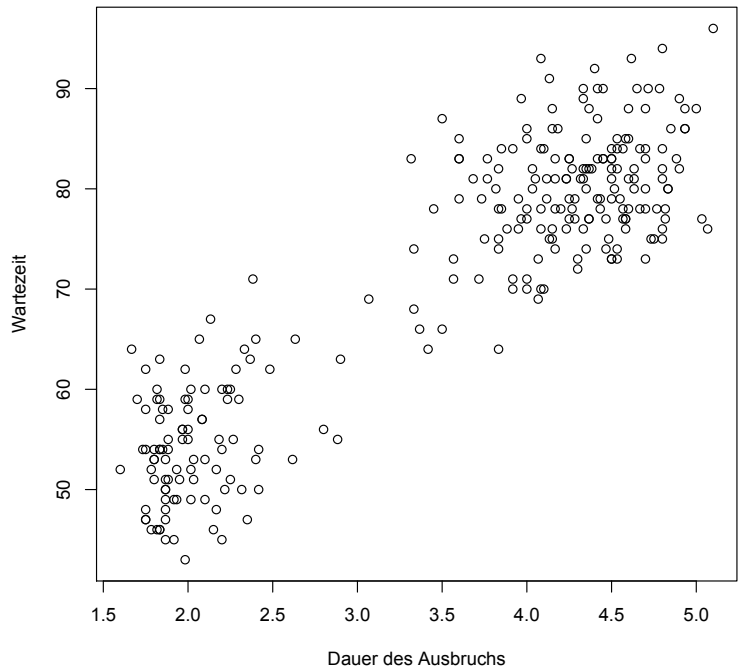
Kenngrößen von Daten

- Empirisches Mittel: Durchschnittswert
- Empirische Varianz: Maß für die Streuung

R-Befehle:

- mean() empirisches Mittel
- var() empirische Varianz
- median() Median
- sort() Liste aufsteigend sortieren
- hist() Zeichnet Histogramm.

Beispiel 8.13: R-Datensatz



- Daten von zwei gleichzeitig gemessenen Größen: Paare von Messwerten $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Grundannahme: Realisierungen zweier Zufallsvariablen X und Y.

Beispiel 8.13: R-Datensatz

Beispieldatensatz in R: Old faithful Geysir, Yellowstone

- Aufrufen mit Befehl `faithful`
- Dauer des Ausbruchs und Wartezeit zwischen Ausbrüchen in Minuten
- 272 Datenpaare
- Beispiel: Berechnung von empirischem Mittel und empirischer Varianz von Dauer und Wartezeit:

```
> data<-faithful
> x<-faithful$eruptions
> y<-faithful$waiting
> mx<-mean(x)
> my<-mean(y)
> vx<-var(x)
> vy<-var(y)
```

- Ergebnis: $\bar{\mu}_x = 3.487783, \bar{\mu}_y = 70.89706$, $\bar{\sigma}_x^2 = 1.302728, \bar{\sigma}_y^2 = 184.8233$

Parameterschätzung

Grundprinzip der Parameterschätzung: Aus den gemessenen Daten die Kenngrößen der (unbekannten) zugrundeliegenden Verteilung schätzen. Dafür müssen gewisse Annahmen getroffen werden (z.B. Unabhängigkeit).

Es gibt viele verschiedene Methoden für die Parameterschätzung. Wir lernen klassische Schätzer für wichtige Kenngrößen kennen, sowie die Methode der "Maximum Likelihood"-Schätzung.

Schätzfunktion

(Def. 8.7) Seien $X_1, ..., X_n$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$. Eine Schätzfunktion zur Stichprobengröße n ist eine Funktion

$$\theta_n : \mathbb{R}^n \to \mathbb{R}, \ (X_1, ..., X_n) \mapsto \theta_n(X_1, ..., X_n).$$

- Für praktische Zwecke sollte eine Schätzfunktion einen Zusammenhang mit einem Parameter oder einer Kenngröße der Verteilung der $X_1, ..., X_n$ haben
- Ist dieser Parameter unbekannt, so erhält man, nach Messung der Daten $x_1, ..., x_n$ als Realisierungen von $X_1, ..., X_n$ einen Schätzer oder Schätzwert für den Parameter.
- Damit die Schätzfunktion und der Schätzer nützlich sind, sollten sie gewisse günstige Eigenschaften haben.
- (Bem. Statistisches Modell)

Klassische Schätzer

Beispiel 8.3: Empirisches Mittel. Seien $(x_1, ..., x_n) \in \mathbb{R}^n$ Messwerte. Das empirische Mittel (vgl. Def. 8.4) ist definiert als

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

- Sind die x_i Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen $X_1, ..., X_n$, so gilt mit dem Gesetz der großen Zahlen

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

- $\bar{\mu}_x$ ist ein Schätzwert für den Erwartungswert der zugrundeliegenden Zufallsvariablen.

Klassische Schätzer

Beispiel 8.4: Empirische Varianz. Seien $(x_1, ..., x_n) \in \mathbb{R}^n$ Messwerte. Die empirische Varianz (vgl. Def. 8.6) ist definiert als

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

- $\bar{\sigma}_x^2$ ist ein Schätzwert für die Varianz der zugrundeliegenden Zufallsvariablen.

Eigenschaften von Schätzern

(Def. 8.7) Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor von Daten, welche als Realisierungen von identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit Parameter θ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$. Sei θ_n eine Schätzfunktion zur Stichprobengröße n.

- θ_n ist ein erwartungstreuer Schätzer für θ, falls

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

  ist (engl: unbiased).
- θ_n ist ein konsistenter Schätzer für θ, falls

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

  gilt.
- θ_n ist ein effizienter Schätzer für θ, falls

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

  gilt.
- (Beispiele)

Maximum Likelihood Schätzung (MLE)

Grundidee: Gegeben Daten $x_1, ..., x_n$ einer Verteilung mit einem unbekannten Parameter. Unter allen (theoretisch möglichen) Werten des Parameters ist der Maximum Likelihood-Schätzer derjenige, für welchen die Wahrscheinlichkeit, die gemessenen Werte $x_1, ..., x_n$ zu finden maximal wird.

Wir betrachten zwei verschiedene Situationen:

- Die zugrundeliegenden Zufallsvariablen $X_1, ..., X_n$ sind diskret, und haben Verteilung p_θ in Abhängigkeit von einem unbekannten Parameter θ
- Die zugrundeliegenden Zufallsvariablen $X_1, ..., X_n$ haben Dichte f_θ in Abhängigkeit von einem unbekannten Parameter θ.

Maximum Likelihood Schätzung (MLE)

Beispiel: Diskrete Verteilung

- Gegeben: $(x_1, ..., x_n)$
- Annahme: Dies sind Realisierungen von n unabhängigen identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit unbekanntem Parameter θ.
- Sei $p_\theta(x)$ die Verteilung der $X_1, ..., X_n$, mit unbekanntem θ.
- Betrachte

$$\mathbb{P}_\theta(X_1 = x_1, ..., X_n = x_n) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n).$$

Grundidee:

- Bestimme θ so, dass diese Wahrscheinlichkeit für das vorgegebene $x = (x_1, ..., x_n)$ maximal ist (z.B. duch Ableiten).
- (Beispiel 8.8)

Maximum Likelihood Schätzung (MLE)

Beispiel: Diskrete Verteilung

- Gegeben: $(x_1, ..., x_n)$
- Annahme: Dies sind Realisierungen von n unabhängigen identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit unbekanntem Parameter θ.
- Sei $p_\theta(x)$ die Verteilung der $X_1, ..., X_n$, mit unbekanntem θ.
- Betrachte

$$\mathbb{P}_\theta(X_1 = x_1, ..., X_n = x_n) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n).$$

Grundidee:

- Bestimme θ so, dass diese Wahrscheinlichkeit für das vorgegebene $x = (x_1, ..., x_n)$ maximal ist (z.B. duch Ableiten).
- (Beispiel 8.8)

Likelihood-Funktion

(Def. 8.9) Seien $x_1, ..., x_n \in \mathbb{R}$ Messwerte, welche als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen $X_1, ..., X_n$ aufgefasst werden können. Die Likelihood-Funktion ist die Funktion $L((x_1, ..., x_n); \theta)$ von $\mathbb{R}^n \times \mathbb{R}$ nach $[0, 1]$, für welche gilt:

- Falls die X_i eine diskrete Verteilung p_θ mit Parameter θ besitzt, so ist

$$L((x_1, ..., x_n); \theta) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n) = \prod_{i=1}^{n} p_\theta(x_i).$$

- Falls die X_i eine Dichte f_θ mit Parameter θ besitzt, so ist

$$L((x_1, ..., x_n); \theta) = f_\theta(x_1) \cdot ... \cdot f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Maximum Likelihood Schätzung (MLE)

(Def. 8.10) Seien $x_1, ..., x_n$ Messwerte, welche Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen sind, welche entweder der diskreten Verteilung p_θ folgen, oder die Dichte f_θ haben. Der Maximum Likelihood-Schätzer für θ ist $\theta_* = \theta_*(x_1, ..., x_n)$, welches die Likelihood-Funktion $L((x_1, ..., x_n); \theta)$ unter allen $\theta \in \mathbb{R}$ maximiert. Formal gesprochen ist

$$\theta_*(x_1, ..., x_n) = \operatorname{argmax}_{\theta \in \mathbb{R}} L((x_1, ..., x_n); \theta).$$

- Der Maximum Likelihood-Schätzer wird dabei manchmal auch kurz als MLE bezeichnet, für das englische "maximum likelihood estimator".

Log-likelihood

(Def. 8.11) Die Log-Likelihood-Funktion ist definiert als

$$l((x_1, ..., x_n); \theta) = \ln L((x_1, ..., x_n); \theta).$$

(Satz 8.24) Die Funktion $l((x_1, ..., x_n); \theta)$ ist genau dann maximal (in θ), wenn $L((x_1, ..., x_n); \theta)$ maximal ist.

- $l(x_1, ..., x_n; \theta)$ ist oft einfacher zu maximieren als $L(x_1, ..., x_n; \theta)$.

Maximum Likelihood Schätzung: Vorgehen

Das allgemeine Vorgehen zur Bestimmung des MLE wird somit:

1. Likelihood-Funktion $L(x_1, ..., x_n; \theta)$ aufstellen (nach Definition 10.8), abhängig davon ob die Verteilung diskret ist oder eine Dichte besitzt
2. Die log-Likelihood-Funktion $l(x_1, ..., x_n; \theta)$ durch logarithmieren von $L(x_1, ..., x_n; \theta)$ berechnen und so weit wie möglich vereinfachen (Logarithmengesetze!)
3. Den Parameter θ so bestimmen, dass $l((x_1, ..., x_n); \theta)$ für das vorgegebene $x = (x_1, ..., x_n)$ maximal ist. Meistens geschieht das durch Ableiten nach θ.
4. Überprüfen, dass es sich beim Ergebnis tatsächlich um ein Maximum handelt (z.B. durch Test mit der 2. Ableitung).

(Beispiel 8.9)