Einführung in die (induktive) Statistik

Typische Fragestellung der Statistik:

Auf Grund einer Problemmodellierung sind wir interessiert an:

Zufallsexperiment beschrieben durch ZV X.

Problem: Verteilung von X ist nicht vollständig bekannt.

Z.B.: $X \sim \mathcal{N}(\mu, 1)$ für unbekanntes μ.

Gegeben: Wiederholungen $X_1, X_2, \ldots, X_n$ des Experiments

D.h.: $X, X_1, \ldots, X_n$ sind identisch verteilt.

Häufige Annahme: $X_1, \ldots, X_n$ unabhängig.

ergeben Beobachtungen/Stichprobe $x_1, \ldots, x_n$.

- ▷ Versuche mit Hilfe der Stichprobe Rückschlüsse auf die Verteilung von X zu ziehen.

Beispiel: Anzahl Taxis

Situation:

Beobachten von vorbeifahrenden Taxis.

Jedes Taxi trägt eine eindeutige Nummer.

Nummer des i-ten beobachteten Taxis: $x_i \in \mathbb{N}$.

Fragestellung:

Wie viele Taxis gibt es?

Beispiel: Anzahl Taxis

Modellierung:

Wir treffen folgende (sinnvolle?) Annahmen:

- Beobachtung eines Taxis entspricht Zufallsexperiment X:

  Ziehung eines zufälligen Wertes aus $[M]$,

  wobei M die gesuchte Anzahl der Taxis (der unbekannte Parameter) ist.

- Beobachtung von n Taxis entspricht n unabhängigen Wiederholungen von X:

  Unabhängige ZVen $X_1, \ldots, X_n$ mit X_i ist gleichverteilt über $[M]$.

- Beobachtete Nummer x_i ist das Ergebnis des Experiments X_i.

Formalisierung des Problems:

Unter diesen Annahmen und gegeben die Stichprobe $x_1, \ldots, x_n$ bestimme einen Schätzwert für M.

Beispiel: Anzahl Taxis

Schätzer für M (1. Ansatz):

Für den Schätzwert

$$t_1 := \max\{x_1, \ldots, x_n\}$$

gilt sicherlich stets $t_1 \leq N$.

D.h.: Wir können nie von zuvielen Taxis ausgehen.

Beachte: Schätzwert t_1 ist das Ergebnis der ZV

$$T_1 = \max\{X_1, \ldots, X_n\}.$$

Konvention: T_1 wird als die zugehörige Schätzvariable (kurz: Schätzer) bezeichnet.

Offensichtliche Frage: Ist T_1 eine gute Schätzvariable?

Problem: Wie quantifiziert man die Güte von T_1?

Güte von Schätzvariablen

Übliche Gütekriterien für Schätzer T zu Parameter θ:

- Erwartungstreue (auch: Unverfälschtheit): $\mathbb{E}[T] = \theta$.

  „Im Mittel sollte T den richtigen Parameter θ liefern."

Güte von Schätzvariablen

Übliche Gütekriterien für Schätzer T zu Parameter θ:

- Erwartungstreue (auch: Unverfälschtheit): $\mathbb{E}[T] = \theta$.

  „Im Mittel sollte T den richtigen Parameter θ liefern."

- Mittlerer quadratischer Fehler: $\mathrm{mse}(T) := \mathbb{E}[(T-\theta)^2]$

  „Sollte im Mittel T kaum um θ schwanken."

  Idealerweise: $\mathrm{mse}(T) \to 0$ für wachsenden Stichprobenumfang n.

  Dann heißt T konsistent im quadratischen Mittel.

Güte von Schätzvariablen

Übliche Gütekriterien für Schätzer T zu Parameter θ:

- Erwartungstreue (auch: Unverfälschtheit): $\mathbb{E}[T] = \theta$.

  „Im Mittel sollte T den richtigen Parameter θ liefern."

- Mittlerer quadratischer Fehler: $\mathrm{mse}(T) := \mathbb{E}[(T-\theta)^2]$

  „Sollte im Mittel T kaum um θ schwanken."

  Idealerweise: $\mathrm{mse}(T) \to 0$ für wachsenden Stichprobenumfang n.

  Dann heißt T konsistent im quadratischen Mittel.

- Für gegebene Abweichung $\delta \geq 0$ die W'keit $\Pr[|T-\theta| \geq \delta]$.

  Schwache Konsistenz: $\Pr[|T-\theta| \geq \delta] \xrightarrow{n\to\infty} 0$ für jedes $\delta \geq 0$.

  Markov: Konsistenz im quadr. Mittel impliziert schwache Konsistenz.

  Verwandte Frage (später): Konfidenzbereich

  Gegeben W'keit α bestimmte kleinstes δ, so dass $\Pr[|T-\theta| \geq \delta] \leq \alpha$.

Beispiel: Anzahl Taxis

Erinnerung: $X_1, \ldots, X_n$ unabhängig, jeweils gleichverteilt auf $[M]$.

Ist $T_1 = \max\{X_1, \ldots, X_n\}$ erwartungstreu? D.h.: Gilt $\mathbb{E}[T_1] = M$?

Dichte von T_1:

$$\Pr[T_1 = k] = \Pr[T_1 \leq k] - \Pr[T_1 \leq k-1] = \frac{k^n - (k-1)^n}{M^n}.$$

Somit:

$$\mathbb{E}[T_1] = \sum_{k=1}^{M} k \frac{k^n - (k-1)^n}{M^n} \in \frac{1}{M^n} \sum_{k=1}^{M} k\mathcal{O}(n \cdot k^{n-1}).$$

Man kann zeigen, dass für große M: $\mathbb{E}[T_1] \approx \frac{n}{n+1} M$.

- T_1 ist somit nicht erwartungstreu (aber fast).

Ist T_1 konsistent im quadr. Mittel? D.h.: Gilt $\mathbb{E}[(T_1 - M)^2] \xrightarrow{n\to\infty} 0$?

$$\mathbb{E}[(T_1 - M)^2] = \sum_{k=1}^{M} (k-M)^2 \frac{k^n - (k-1)^n}{M^n} \in \mathcal{O}\left(\frac{M^2}{n^2}\right).$$

Beispiel: Anzahl Taxis

Schätzer für M (2. Ansatz): Maximum-Likelihood-Prinzip

Wähle M so, dass die W'keit für die beobachtete Stichprobe $x_1, \ldots, x_n$ maximiert wird.

Wegen Unabhängkeit:

$$\Pr[X_1 = x_1, \ldots, X_n = x_n] = \Pi_{i=1}^{n} \Pr[X_i = x_i] = M^{-n}.$$

Wahrscheinlichkeit wird mit wachsendem M kleiner.

Allerdings muss auch $M \geq t_1 = \max\{x_1, \ldots, x_n\}$ gelten, damit alle $\Pr[X_i = x_i] > 0$.

Maximierender Schätzwert somit wieder $t_1 = \max\{x_1, \ldots, x_n\}$.

$T_1 = \max\{X_1, \ldots, X_n\}$ also auch ML-Schätzer.

- Maximum-Likelihood-Prinzip liefert i.A. keine erwartungstreue Schätzer.

Beispiel: Anzahl Taxis

Schätzer für M (3. Ansatz):

Wegen Gleichverteilung gilt: $\mu := \mathbb{E}[X] = \frac{M+1}{2}$.

Idee: Entwerfe Schätzer $\overline{X}$ für μ. Setze dann $T_2 = 2\overline{X} - 1$.

Natürliche Wahl für Schätzwert für μ: Stichprobenmittelwert

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Zugehörige Schätzvariable: Stichprobenmittel

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Stichprobenmittel

Ist Stichprobenmittel $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. erwartungstreu?

$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Also immer erwartungstreu, egal welche Verteilung X hat.

Ist Stichprobenmittel konsistent im quadr. Mittel?

$$\mathbb{E}[(\overline{X}-\mu)^2] = \mathbb{E}[(\overline{X}-\mathbb{E}[\overline{X}])^2] = \mathsf{Var}[\overline{X}] = \frac{1}{n}\mathsf{Var}[X].$$

Also immer konsistent im quadr. Mittel, wenn $\mathsf{Var}[X] < \infty$ gilt.

Zurück zu unserem Beispiel: $T_2 = 2\overline{X} - 1$.

Erwartungstreue? Konsistenz im quadr. Mittel?

Beispiel: Anzahl Taxis

Schätzer für M (letzter Ansatz):

Wegen Gleichverteilung gilt: $\sigma^2 := \mathsf{Var}[X] = \frac{M^2-1}{12}$.

Idee: Entwerfe Schätzer S^2 für σ^2. Setze dann $T_4 = \sqrt{12 \cdot S^2 + 1}$.

Üblicher Schätzwert für σ^2:

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

- Warum nicht Faktor $1/n$? Wird auch verwendet, dann aber nicht erwartungstreu (gleich).

Zugehörige Schätzvariable: Stichprobenvarianz

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Stichprobenvarianz

Erwartungstreue von S^2:

$$\mathbb{E}[S^2] = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[(X_i - \overline{X})^2] = \frac{n}{n-1}\mathbb{E}[(X_n - \overline{X})^2].$$

Mit $\mu = \mathbb{E}[X_i] = \mathbb{E}[\overline{X}]$:

$$\begin{aligned}\mathbb{E}[(X_n - \overline{X})^2] &= \mathbb{E}[(X_n - \mu + \mu - \overline{X})^2] \\ &= \mathsf{Var}[X_n] + \mathsf{Var}[\overline{X}] - 2\mathbb{E}[(X_n - \mu)(\overline{X} - \mu)].\end{aligned}$$

Bleibt:

$$\begin{aligned}\mathbb{E}[(X_n - \mu)(\overline{X} - \mu)] &= \mathbb{E}[(X_n - \mu)\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(X_i - \mu)] \\ &= \tfrac{1}{n}\mathbb{E}[(X_n - \mu)^2] + \tfrac{1}{n}\textstyle\sum_{i=1}^{n-1}\mathbb{E}[(X_n - \mu)(X_i - \\ &= \tfrac{1}{n}\mathsf{Var}[X_n],\end{aligned}$$

da $\mathbb{E}[(X_n - \mu)(X_i - \mu)] = 0$ für $i < n$, da X_i, X_n unabhängig.

Insgesamt somit:

$$\mathbb{E}[S^2] = \frac{n}{n-1}(\mathsf{Var}[X] + \frac{1}{n}\mathsf{Var}[X] - \frac{2}{n}\mathsf{Var}[X]) = \mathsf{Var}[X] = \sigma^2.$$

Stichprobenvarianz

Anmerkung: Mit Hilfe des (schwachen) Gesetzes der großen Zahlen kann man zeigen, dass S^2 schwach konsistent ist, falls $\mathrm{Var}[X]$ existiert.

Für unser Beispiel: $T_4 = \sqrt{12S^2 + 1}$.

Im Allgemeinen gilt nur:

$$\mathbb{E}[T_4] \leq \sqrt{\mathbb{E}[12S^2 + 1]} = M.$$

Z.B. für $n = 2, M = 10$: $\mathbb{E}[T_4] \approx 8.260 < 10$.

Zusammenfassung

Übliches Vorgehen: Fragestellung $\rightarrow$ Modellierung mit Hilfe der W'Theorie $\rightarrow$ Schätzvariable T für gesuchten Parameter θ.

- Unangemessene Modellierung $\rightarrow$ Schlechter Schätzer.

  Z.B.: Nummern der Taxis könnten nicht durchgehend sein.

  Annahme der Gleichverteilung von X nicht mehr sinnvoll.

Maximum-Likelihood-Prinzip zur Bestimmung von Schätzern:

Wähle Schätzwert so, dass W'keit $\Pr[X_1 = x_1, \ldots, X_n = x_n]$ maximiert wird.

- Falls X stetig verteilt: maximiere gemeinsame Dichte der $X_1, \ldots, X_n$.

Zusammenfassung

Verschiedene Kriterien, um die Güte von Schätzern T zu messen:

- Erwartungstreue: $\mathbb{E}[T] = \theta$.
- Mittlerer quadr. Fehler: $\mathsf{mse}(T) = \mathbb{E}[(T-\theta)^2]$.
  - Falls T erwartungstreu: $\mathsf{mse}(T) = \mathsf{Var}[T]$.

  Konsistenz im quadr. Mittel: $\mathsf{mse}(T) \xrightarrow{n\to\infty} 0$.
- Schwache Konsistenz: $\Pr[|T-\theta| \geq \delta] \xrightarrow{n\to\infty} 0$ für jedes $\delta \geq 0$.
- Vergleich von zwei verschiedenen Schätzer T_1, T_2:

  Vergleich Konvergenzgeschwindigkeiten von $\mathsf{mse}(T_i)$ bzw. $\Pr[|T_i-\theta| \geq \delta]$.

  Oder bestimmte für gegebenes α das kleinste δ_i mit $\Pr[|T_i-\theta| \geq \delta_i] \leq \alpha$.