Stichprobenverfahren

Einschlusswahrscheinlichkeiten

Willi Mutschler
willi@mutschler.eu

Sommersemester 2017

Notation (1)

- $U = \{u_1, \ldots, u_k, \ldots, u_N\}$ bezeichne die Grundgesamtheit mit individuellen Einheiten u_k, $k = 1, \ldots, N$, N ist üblicherweise bekannt
- Vereinfachend wird das $k$te Element über sein Label k repräsentiert, also $U = \{1, \ldots, k, \ldots, N\}$
- $s \subset U$ bezeichne die realisierte Stichprobe mit n Elementen aus U
- $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ bezeichne die Menge aller möglichen Stichproben
- Wir interessieren uns für das Merkmal Y: die spezifische Ausprägung y_k für Element u_k ist unbekannt
- Die Verteilung von Y in U kann mithilfe von Parametern beschrieben werden, z.B.

$$t_U = \sum_U y_k = \sum_{k=1}^{N} y_k$$

$$\bar{y}_U = \frac{1}{N} \sum_U y_k = \frac{1}{N} \sum_{k=1}^{N} y_k$$

$$s^2_{y,U} = \frac{1}{N-1} \sum_U (y_k - \bar{y}_U)^2$$

Notation (2)

- Das Stichprobendesign definiert die Wahrscheinlichkeitsverteilung $p(\cdot)$ für die zu ziehende Stichprobe S: $P(S = s) = p(s)$ für alle $s \in \mathcal{S}$
- Es gilt: $p(s) \geq 0$ und $\sum_{s \in \mathcal{S}} p(s) = 1$
- $p(s)$ wird auch „sampling design“ genannt

Einschlussindikator

- Die Indikatorvariable I is eine dichotome Zufallsvariable

$$I_k = \begin{cases} 1, & k \in \mathcal{S} \\ 0, & k \notin \mathcal{S} \end{cases}$$

- I_k ist folglich eine Funktion von S: $I_k = I_k(S)$

Einschlusswahrscheinlichkeit

- Die Einschlusswahrscheinlichkeit π_k ist die Wahrscheinlichkeit, dass eine zufällige Stichprobe S gezogen wird, die Element k enthält:

$$\pi_k = P(k \in S) = P(I_k = 1) = \sum_{s \ni k} p(s)$$

  $s \ni k$ bezieht sich auf alle Stichproben s, die k enthalten

- Summe über alle k:

$$\sum_{k \in U} \pi_k = \sum_{k \in U} \sum_{s \ni k} p(s) = n \sum_{s \in \mathcal{S}} p(s) = n \cdot 1 = n$$

- Einschlusswahrscheinlichkeit π_{kl} ist die Wahrscheinlichkeit, dass eine zufällige Stichprobe S gezogen wird, die Element k und l enthält:

$$\pi_{kl} = P(k \& l \in S) = P(I_k I_l = 1) = \sum_{s \ni k \& l} p(s)$$

- Es gilt: $\pi_{kl} = \pi_{lk}$ für alle k, l. Was ist mit $k = l$?

Eigenschaften des Einschlussindikators

- I_k ist Bernoulli verteilt
- Erwartungswert: $E(I_k) = 0 \cdot (1 - \pi_k) + 1 \cdot \pi_k = \pi_k = P(I_k = 1)$
- Varianz: $V(I_k) = E(I_k^2) - E(I_k)^2 = \pi_k(1 - \pi_k)$
- Kovarianz: $Cov(I_k, I_l) = E(I_k I_l) - E(I_k)E(I_l) = \pi_{kl} - \pi_k \pi_l$
- Bemerkung: Ein „sampling design“ wird „measurable“ genannt, wenn $\pi_k > 0$ und $\pi_{kl} > 0$ für alle $k \neq l \in U$