Stichprobenverfahren

Systematische Auswahl

Willi Mutschler (willi@mutschler.eu)

Sommersemester 2017

Motivation

Beispiel für systematische Auswahl

Ein Dozent muss 600 Klausuren korrigieren und möchte eine grobe Idee über die Durchfallquote bekommen. Hierzu wirft er einen Würfel einmal und wirft z.B. eine „2“. Dann korrigiert er die 2te, 8te, 14te, ... , 596te Klausur.

- Systematische Auswahl wird in der Praxis sehr häufig verwendet, da sie kostengünstig und einfach durchzuführen ist
  - Vgl. obiges Beispiel mit Bernoulli Sampling, d.h. 600 Mal würfeln und nur bei einer 6 die Klausur korrigieren
- Vorgehen: Das erste Element wird (mit gleicher Wahrscheinlichkeit für alle Elemente) zufällig gewählt, danach wird systematisch jedes $a$te Element in die Stichprobe aufgenommen
- Die ganzzahlige Zahl a wird Stichprobenintervall genannt

Definition

- Sei a das fixe Stichprobenintervall, N die Größe der Grundgesamtheit und n der ganzzahlige Teil von N/a, dann ist $N = na + c$, wobei $0 \leq c < a$.
- Algorithmus:
  1. Wähle mit gleicher Wahrscheinlichkeit $1/a$ ein zufällige ganzzahlige Zahl r mit $1 \leq r \leq a$; r bezeichnet man als zufälligen Start
  2. Die Stichprobe besteht dann aus

     $$s = \{k : k = r + (j-1)a \leq N; j = 1, 2, ..., n_s\} = s_r$$

     wobei die Stichprobengröße n_s entweder $n + 1$ (bei $r \leq c$) oder n (bei $c < r \leq a$) beträgt
- Im Beispiel: $r = 2$, $a = 6$, $c = 0$ und $n_s = n = 100$

Einschlusswahrscheinlichkeiten

- Die Menge $\mathcal{S}$ beinhaltet a mögliche nicht überlappende Mengen und ist im Vergleich zur einfachen Zufallsstichprobe ohne Zurücklegen sehr klein: $M = |\mathcal{S}| = a$.
- Es gilt: $U = \bigcup_{r=1}^{a} s_r$
- Sampling Design ist folglich:

$$p(s) = \begin{cases} 1/a & \text{falls } s \in \mathcal{S} \\ 0 & \text{sonst} \end{cases}$$

- Die Einschlusswahrscheinlichkeiten sind demnach

$$\pi_k = 1/a$$

$$\pi_{kl} = \begin{cases} 1/a & \text{falls } k \text{ und } l \text{ zur Stichprobe } s \text{ gehören} \\ 0 & \text{sonst} \end{cases}$$

- ACHTUNG: Die Voraussetzung $\pi_{kl} > 0$ ist nicht erfüllt!

Horvitz-Thompson Schätzer

- Der π-Schätzer für die Merkmalssumme $t_U = \sum_U y_k$ ist gegeben durch

$$\hat{t}_\pi = a t_s$$

  wobei $t_s = \sum_s y_k$ und s ist eine mögliche Stichprobe unter systematischer Auswahl.
- Die Varianz ist gegeben durch

$$V(\hat{t}_\pi) = a(a-1) \underbrace{\frac{1}{a-1} \sum_{r=1}^{a} (t_{s_r} - \bar{t})^2}_{S_t^2} = a \sum_{r=1}^{a} (t_{s_r} - \bar{t})^2$$

  mit $\bar{t} = \frac{1}{a} \sum_{r=1}^{a} t_{s_r}$ ist das Mittel der Stichprobensummen und $t_{s_r} = \sum_{s_r} y_k$
- Da $\pi_{kl} > 0$ nicht erfüllt ist, können wir die Formel für einen unverzerrten Varianzschätzer nicht verwenden
- Es gibt hier sogar keinen unverzerrten Schätzer! Mögliche Vorgehen
  - Verzerrten Schätzer verwenden, z.B. von einfacher Zufallsstichprobe
  - Modifizierung der systematischen Auswahl, z.B. mehrere zufällige Starts $m > 1$ und Stichprobenintervall ma

Effizienz

- $V(\hat{t}_\pi)$ ist klein, wenn die Stichprobensummen annähernd identisch sind
- Folglich hängt die Effizienz der systematischen Auswahl von der Annordnung der N Elemente der Grundgesamtheit ab
- Betrachte $N = an$:

$$\hat{t}_\pi = N \sum_{s_r} y_k / n = N \bar{y}_{s_r} \qquad \text{und} \qquad V(\hat{t}_\pi) = N^2 \frac{1}{a} \sum_{r=1}^{a} (\bar{y}_{s_r} - \bar{y}_U)^2$$

- Es gilt, dass die Variation der Grundgesamtheit zerlegt werden kann:

$$\underbrace{\sum_{U} (y_k - \bar{y}_U)^2}_{SST} = \underbrace{\sum_{r=1}^{a} \sum_{s_r} (y_k - \bar{y}_{s_r})^2}_{SSW} + \underbrace{\sum_{r=1}^{a} n(\bar{y}_{s_r} - \bar{y}_U)^2}_{SSB}$$

  SS Sum of Squares, T: total, W: within samples, B: between samples
- $SST = (N-1)S_{yU}^2$ ist fix, d.h. $SSW \uparrow \rightarrow SSB \downarrow$
- Für die Varianz folgt: $V(\hat{t_\pi}) = N \cdot SSB$
- Je homogener (Tendenz gleicher y Werte) die Elemente innerhalb der systematischen Stichproben sind, desto weniger effizient ist das systematische Auswahldesign

Homogenitätsmaß

- Homogenität lässt sich messen mit z.B.

$$\delta = 1 - \frac{N-1}{N-a}\frac{SSW}{SST}$$

mit intra-sample Varianz $SSW = (N-a)S^2_{yW}$ und Populationsvarianz $SST = (N-1)S^2_{yU}$ vereinfacht sich zu

$$\frac{S^2_{yW}}{S^2_{yU}} = (1-\delta)$$

- Die Extremwerte von δ sind

$$\delta_{min} = -\frac{a-1}{N-a} \qquad \text{und} \qquad \delta_{max} = 1$$

- Minimal falls $SSB = 0$, d.h. alle $\bar{y}_s$ konstant
- Maximal falls $SSW = 0$, d.h. komplette Homogenität