Stichprobenverfahren

Cluster-Stichproben

Willi Mutschler (willi@mutschler.eu)

Sommersemester 2017

Motivation

Bisher: Zugriff auf einzelne Untersuchungseinheiten ohne Probleme möglich und gleichzeitig kosteneffizient; in der Praxis häufig jedoch nicht möglich!

Zigarettenkonsum

- Zur Bestimmung des Zigarettenkonsums von Hauptschülern in der 8. Klasse soll eine Erhebung mit Hilfe von Fragebögen durchgeführt werden
- Ziehung von einzelnen Schülern ist sehr aufwendig, da eine Liste aller Schüler der 8. Klasse vorliegen müsste
- Eine derartige Liste ist jedoch selten vorhanden oder wird aus Datenschutzgründen nicht zur Verfügung gestellt
- Mögliches Vorgehen: Zufallsauswahl von Schulklassen und nicht von Schülern, da eine Liste der Schulklassen oder auch Schulen viel einfacher zu erhalten ist

Idee

Wir sprechen von einer sogenannten *Cluster-Ziehung* bzw. *Klumpen-Ziehung*, wenn

- Elemente der Grundgesamtheit (die Schüler) in natürlicher Weise sich in nicht überlappende Gruppen (die Klassen) zusammenfassen lassen, die wir als Cluster oder Klumpen bezeichnen
- Die Idee der Clusterstichprobe besteht nun darin, eine Zufallsstichprobe aus den Clustern zu ziehen und innerhalb der gezogenen Cluster eine Vollerhebung durchzuführen
- Ziehung somit nicht auf den Elementen der Population, sondern auf den Clustern
- Wichtigstes Argument für Cluster-Stichprobe: Kosteneffizienz!
- Wichtigstes Argument gegen Cluster-Stichprobe: Clusterbildung führt nicht notwendigerweise zu einer genaueren Stichprobe im Sinne einer reduzierten Varianz

Clusterbildung (I)

Clusterprinzip
Cluster sollten so gewählt werden, dass Beobachtungen innerhalb eines Clusters so heterogen wie möglich sind, sich einzelne Cluster aber so wenig wie möglich voneinander unterscheiden.

Clusterbildung (II)

Bemerkungen:

- Clusterprinzip bildet das Gegenteil zum Schichtungsprinzip
- Cluster werden häufig als lokale Gruppen gewählt: Straßenzüge, Gemeinden oder Schulen
- Aber Bewohner einer Straße sind homogen, wohingegen die Straßen einer Stadt von Seiten der Bevölkerungsstruktur her heterogen sind
- Ebenso sind Gemeinden (oder Schulen) in sich homogen und unterscheiden sich von anderen Gemeinden (oder Schulen)
- Die praktischen Vorteile einer Cluster-Stichprobe können im Widerspruch zum Clusterprinzip stehen $\Rightarrow$ Effizienzverlust
- Design der Cluster-Stichprobe wird folglich vor allem aufgrund der einfachen Umsetzbarkeit in der Praxis gewählt

Mehrere Stufen

- Einfache Cluster-Ziehung wird auch *single-stage cluster sampling* genannt
- *Two-stage cluster sampling*
  - Population wird gruppiert in nicht-überlappende Untergruppen, diese werden primary samling units (PSUs) genannt. Wir ziehen zufällig PSUs (first-stage sampling)
  - Für jedes PSU des first-stage samples werden nun wiederrum Elemente oder Cluster gezogen, man bekommt so die sogenannten *second-stage sampling units (SSUs)*
  - Falls jedes SSU ein Element ist, nennen wir dies *two-stage element sampling*, falls jedes SSU ein Cluster von Elementen ist, nennen wir es *two-stage cluster sampling*
- Erweiterung um *multi-stage sampling* möglich, z.B. bei drei Stufen sprechen wir dann von *third-stage sampling units* (TSU)

Notation

- Die Grundgesamtheit $U = \{1, ..., k, ..., N\}$ wird in N_I Cluster eingeteilt, diese werden mit $U_1, ..., U_i, ...U_{N_I}$ bezeichnet
- Die Menge der Cluster ist somit: $U_I = \{1, ..., i, ...N_I\}$
- N_i bezeichnet die Anzahl an Elementen im $i$ten Cluster U_i
- Es gilt: $U = \bigcup_{i \in U_I} U_i$ und $N = \sum_{i \in U_i} N_i$

Der Index I wird hier verwendet für die first-stage cluster sampling (II für second-stage usw.)

Single-stage Cluster

Eine single-stage Cluster-Stichprobe ist nun definiert durch:

1. Eine Stichprobe s_I an Clustern wird zufällig mit Design $p_I(\cdot)$ aus U_I gezogen. Die Größe von s_I bezeichnen wir mit n_I (bei fixierter Stichprobengröße) bzw. n_{s_I} (bei variabler Stichprobengröße).
2. Jedes Element in den ausgewählten Clustern wird beobachtet und voll erhoben.

Bemerkungen:

- p_I kann ein beliebiges Design sein: einfache Zufallsstichprobe ohne Zurücklegen, systematische Ziehung, Schichten,...
- Die Stichprobe ist $s = \bigcup_{i \in s_I} U_i$ mit $n_s = \sum_{s_I} N_i$
- Die Anzahl an beobachteten Elementen n_s ist im Allgemeinen nicht bekannt, da die Clustergrößen N_i unterschiedlich sein können

Einschlusswahrscheinlichkeiten

- Einschlusswahrscheinlichkeiten für Cluster:

$$\pi_{Ii} = \sum_{s_I \ni i} p_I(s_I)$$
$$\pi_{Iij} = \sum_{s_I \ni i\&j} p_I(s_I)$$

- Einschlusswahrscheinlichkeiten für Elemente:
    - $\pi_k = Pr(k \in s) = Pr(i \in s_I) = \pi_{Ii}$
    - Falls k und l im selben Cluster: $\pi_{kl} = Pr(k\&l \in s) = Pr(i \in s_I) = \pi_{Ii}$
    - Falls k und l in unterschiedlichen Clustern: $\pi_{kl} = Pr(k\&l \in s) = Pr(i\&j \in s_I) = \pi_{Iij}$

π-Schätzung

- $t_i = \sum_{U_i} y_k$ bezeichne die Merkmalssumme in Cluster i, dann ist die Populationssumme $t_U = \sum_U y_k = \sum_{U_I} t_i$
- Der π Schätzer für die Merkmalssumme t_U ist
$$\hat{t}_\pi = \sum_{s_I} \check{t}_i = \sum_{s_I} t_i/\pi_{Ii}$$
- Die Varianz ist gegeben durch
$$V(\hat{t}_\pi) = \sum\sum_{U_I} \Delta_{Iij} \check{t}_i \check{t}_j$$
- Die Varianz kann erwartungstreu geschätzt werden mit
$$\hat{V}(\hat{t}_\pi) \sum\sum_{s_I} \check{\Delta}_{Iij} \check{t}_i \check{t}_j$$
- Falls p_I ein Design mit fixierter Stichprobengröße ist, dann
$$V(\hat{t}_\pi) = -\frac{1}{2}\sum\sum_{U_I} \Delta_{Iij}(\check{t}_i - \check{t}_j)^2 \text{ und } \hat{V}(\hat{t}_\pi) = -\frac{1}{2}\sum\sum_{s_I} \check{\Delta}_{Iij}(\check{t}_i - \check{t}_j)^2$$

Achtung: Schätzung des Mittelwertes erfolgt hier nicht einfach durch Division mit N, da N üblicherweise unbekannt ist, somit ist t_U/N ein Quotient von zwei Zufallsvariablen.

Einfacher Cluster-Schätzer

- Betrachte einfache Zufallsstichprobe ohne Zurücklegen bei der Clusterauswahl
- Wir kriegen also eine Stichprobe s_I mit fixer Größe n_I, die aus den N_I Clustern U_I gezogen wird, wobei alle Elemente innerhalb der Cluster beobachtet werden
- Der π Schätzer für die Merkmalssumme t_U ist $\hat{t}_\pi = N_I \bar{t_{s_I}}$ mit $\bar{t_{s_I}} = \sum_{s_I} t_i / n_I$ ist die durchschnittliche Clustersumme in s_I
- Die Varianz ist gegeben durch

$$V(\hat{t}_\pi) = N_I^2 \frac{1 - f_I}{n_I} S^2_{tU_I}$$

mit $f_I = n_I / N_I$, $S^2_{tU_I} = \frac{1}{N_I - 1} \sum_{U_I} (t_i - \bar{t_{U_I}})^2$, wobei $\bar{t_{U_I}} = \sum_{U_I} t_i / N_I$

- Die Varianz kann erwartungstreu geschätzt werden mit

$$\hat{V}(\hat{t}_\pi) = N_I^2 \frac{1 - f_I}{n_I} S^2_{ts_I}$$

mit $S^2_{ts_I} = \frac{1}{n_I - 1} \sum_{s_I} (t_i - \bar{t_{s_I}})^2$

Design Effekt (I)

- Homogenitätskoeffizient: $\delta = 1 - \frac{S^2_{yW}}{S^2_{yU}}$ mit

$$S^2_{yW} = \frac{1}{N - N_I} \sum_{U_I} \sum_{U_i} (y_k - \bar{y}_{U_i})^2 = \frac{\sum_{U_I} (N_I - 1) S^2_{yU_i}}{\sum_{U_I} (N_i - 1)}$$

  ist die *pooled within-cluster-variance* und $\bar{y}_{U_i} = \sum_{U_i} y_k / N_i$ ist der Mittelwert im Cluster i
- S^2_{yW} ist das gewichtete Mittel der N_I Cluster mit jeweiliger Varianz $S^2_{yU_i} = \frac{1}{N_i - 1} \sum_{U_i} (y_k - \bar{y}_{U_i})^2$
- Bemerkung: δ ist adjustiertes Bestimmtheitsmaß in der Regression von y auf N_I Dummy Variablen (Clusterzugehörigkeit)
- Für den Homogenitätsgrad δ gilt $-\frac{N_I - 1}{N - N_I} \leq \delta \leq 1$
- Ein hoher Wert für δ bedeutet, dass Elemente innerhalb eines Clusters sehr ähnlich sind, also eine hohe Homogenität aufweisen

Design Effekt (II)

- Sei $\bar{N} = N/N_I$ und $K_I = N_I^2(1-f_I)/n_I$ und $Cov = \frac{1}{N_I-1}\sum_{U_I}(N_i - \bar{N})N_i\bar{y}_{U_i}^2$ die Kovarianz zwischen N_i und $N_i\bar{y}_{U_i}^2$, dann

$$S_{tU_I}^2 = \bar{N}S_{yU}^2\left(1 + \frac{N - N_I}{N_I - 1}\delta\right) + Cov$$

- Die Varianz des einfachen Cluster-Schätzer, bezeichnen wir mit V_{SIC}, ist dann

$$V_{SIC} = \left(1 + \frac{N - N_I}{N_I - 1}\delta\right)\bar{N}K_I S_{yU}^2 + K_i Cov$$

- Die erwartete Anzahl an beobachtbaren Elementen mit n_I Clustern ist $E(n_s) = n_I\bar{N} = n$
- Betrachte nun einfache Zufallsstichprobe (SI) mit Stichprobengröße $n = n_I\bar{N}$, der π Schätzer ist ist dann $N\bar{y}_s$ und die Varianz

$$V_{SI} = \bar{N}K_I S_{yU}^2$$

- Der Design-Effekt ist dann also

$$deff(SIC, SI) = \frac{V_{SIC}}{V_{SI}} = 1 + \frac{N - N_I}{N_I - 1}\delta + \frac{Cov}{\bar{N}S_{yU}^2}$$

Design Effekt (III)

1. Annahme: Alle Clustergrößen identisch, $N_i = \bar{N}$, dann
   - $Cov = 0$ und
   $$deff = 1 + \frac{N - N_I}{N_I - 1}\delta$$
   - $V_{SIC} < V_{SI}$ nur wenn $\delta < 0$, also wenn es hinreichend große within-cluster Variation gibt
   - In Praxis $\delta > 0$ üblich, da Elemente innerhalb eines Clusters ähnliche Eigenschaften aufweisen
   - Effizienzverlust, insbesondere bei hohen Clustergrößen
2. Annahme: Unterschiedliche Clustergrößen und Korrelatoin zwischen N_i und $N_i \bar{y}^2_{U_i}$ ist positiv, dann
   - zweiter Term wird groß, Effizienzverlust groß
   - Extremfall: $\delta = \delta_{min}$, also alle $\bar{y}_{U_i}$ sind gleich $\bar{y}_U$ und V_{SIC} wird groß, wenn die Clustergrößenvarianz auch groß ist. Designeffekt ist hier:
   $$deff = \bar{N}\left(\frac{CV_N}{CV_y}\right)^2$$
   mit $CV_N = S_{NU_I}/\bar{N}$ und $CV_y = S_{yU}/\bar{y}_U$

Design Effekt (IV)

- Es zeigt sich, dass je kleiner die Varianz zwischen den Clustern ist, desto effizienter ist die Anwendung des Cluster-Schätzers
- Effizienz nimmt bei steigender Clustergröße ab
- Kosten für eine einfache Zufallsstichprobe in der Regel sehr viel höher als die einer Cluster-Stichprobe vom gleichen Umfang

Berücksichtigung der Clustergröße

- Clustergröße ist als Hilfsmerkmal geeignet
- Wähle also ein Design, bei dem die Auswahlwahrscheinlichkeiten proportional zur Clustergröße sind
- das Design ist in diesem Fall eine größenproportionale Ziehung