Stichprobenverfahren

Geschichtete Zufallsstichproben

Willi Mutschler (willi@mutschler.eu)

Sommersemester 2017

Motivation

- Verwendung der einfachen Zufallsstichprobe suboptimal, da üblicherweise ex ante Informationen zur Verfügung stehen
- Grundgesamtheit zerfällt auf natürliche Weise in Teilmengen: Staaten in Bundesländer, Städte in Stadtbezirke, Mitarbeiter eines Betriebes in verschiedene Abteilungen
- Zerlegung der Population in Untergruppen wird als **Schichtung** oder **Stratifizierung** bezeichnet und die entsprechenden Gruppen als **Schichten** oder **Strata**
- Beispiel: Stichprobe aus der Stadtbevölkerung.
  - Informationen über Stadtteile (reiche vs. arme), Gebiet der BRD hat um die 450 Kreise und kreisfreie Städte
  - Teile Grundgesamtheit in nichtüberlappende Gruppen, sogenannte Schichten (z.B. Stadtteile, Kreise) ein und ziehe aus den Schichten
- Geschichtete Stichprobenverfahren sorgen dafür, dass
  - wirklich aus allen Schichten gezogen wird
  - die Schätzung effizienter wird
  - mit Nonresponse oder Messfehlern besser umgegangen werden kann

Beispiel 1

Durchschnittsmiete

- Untersuche den durchschnittlichen Quadratmeterpreis von Mietwohnungen in einer Stadt
- Einfache Zufallsstichprobe bietet sich nicht an, da Mietpreise stark von dem Stadtviertel abhängig sind
- Besser: Ziehung auf der Ebene von Stadtvierteln bzw. Regionen, z.B.:
  - Region 1: „reiche Villengegend“
  - Region 2: „mittlere Lage“
  - Region 3: „Plattenbausiedlung“
- Mit einzelnen Stichproben aus allen drei Regionen lassen sich dann Aussagen fällen über:
  - Mietpreise in einzelnen Vierteln
  - Gesamtmittel des Quadratmeter-Mietpreises
- Erheblicher Effizienzgewinn (unter bestimmten Bedingungen)

Beispiel 2

Bibliotheksnutzung

- Untersuche wie oft und in welchem Umfang Studierende die Arbeitsräume der Bibliothek nutzen
- Sekundärinformation: Studierende in niedrigeren Semestern weniger in Bibliothek (Unterschätzung der Nutzung) als solche kurz vorm Abschluss (Überschätzung der Nutzung)
- Variabilität kann durch das Design der geschichteten Stichprobe verkleinert werden

Konstruktion einer Schicht

- Auswahl der Schichtvariable, an der wir die Grundgesamtheit nach Schichten unterteilen (Alter, Geschlecht, Berufsgruppen), bezeichnen wir mit X, sogennante Schichtungsmerkmal
  - Im Mietpreisbeispiel: Stadtteile, Wohnungsgröße
  - Im Bibliotheksbeispiel: Semesterzahl
- Wichtige Vorraussetzung: Kenntnis der relativen Schichtgrößen in der Grundgesamtheit, also Schichtzugehörigkeit und Schichtumfänge
- Herausforderungen:
  - Wie grenze ich das genau ab? Welche Intervalle, wie viele Schichten?
  - Unterschiedliche Stichprobenumfänge:
    - Stichprobenumfänge proportional zu Schichtgröße, also Ziehung entsprechend Anteilen an der Gesamtbevölkerung: Repräsentativität der Grundgesamtheit
    - Wahlverhalten in alten Bundesländern relativ stabil (kleiner Stichprobenumfang nötig), in neuen Bundesländern weniger Wahlkontinuität (größerer Stichprobenumfang)
  - Entscheidung für Stichprobenverfahren und Schätzmethodik innerhalb der Schichten (alle gleich oder unterschiedlich?)

Schichtungsprinzip

Die Schichten sollen so gewählt werden, dass die Variablen (oder Merkmalsträger) innerhalb einer Schicht so ähnlich wie möglich sind. Die einzelnen Schichten sollten sich untereinander so weit wie möglich unterscheiden.

Notation

- Partitionierung der Grundgesamtheit U in H Untergruppen/Schichten:

$$U_1, ..., U_h, ..., U_H \text{ mit } U_h = \{k : k \text{ gehört zur Schicht } h\}$$

- Ziehe für $h = 1, ..., H$ unabhängig voneinander nicht-überlappende Stichproben s_h aus U_h mithilfe Stichprobendesign $p_h(\cdot)$:

$$s = s_1 \cup s_2 \cup ... \cup s_H$$

- Aufgrund der Unabhängigkeit gilt $p(s) = p(s_1)p(s_2), \ldots, p(s_H)$
- Anzahl N_h an Elementen in Schicht h ist bekannt: $N = \sum_{h=1}^{H} N_h$
- Stichprobenumfang in Schicht h wird mit n_h bezeichnet: $n = \sum_{h=1}^{H} n_h$
- Die Merkmalssumme lässt sich zerlegen

$$t_U = \sum_U y_k = \sum_{h=1}^{H} t_{U_h} = \sum_{h=1}^{H} N_h \bar{y}_{U_h}$$

  mit $t_{U_h} = \sum_{U_h} y_k$ und $\bar{y}_{U_h}$ das Schichtmittel.
- Sei $W_h = N_h/N$ die (bekannte) relative Größe der Schicht U_h, dann

$$\bar{y}_U = \sum_{h=1}^{H} W_h \bar{y}_{U_h}$$

Einschlusswahrscheinlichkeiten

- Erster Ordnung:

$$\pi_k = Pr(k \in s) = Pr(k \in s_h) = \pi_{h,k}$$

- Zweiter Ordnung ($k \neq l$ und $h \neq g$):

$$\pi_{kl} = Pr(k, l \in s) = \begin{cases} Pr(k, l \in s_h) = \pi_{h,kl} \\ Pr(k \in s_h, l \in s_g) = \pi_{h,k}\pi_{g,l} \end{cases}$$

- Für k und l, die zu unterschiedlichen Schichten gehören, gilt also $\Delta_{kl} = 0$

Schätzung in Schichten (1)

- Der π-Schätzer für die Merkmalssumme der Grundgesamtheit ist

$$\hat{t}_\pi = \sum_{h=1}^{H} \hat{t}_{h\pi}$$

  wobei $\hat{t}_{h\pi}$ der π-Schätzer von t_{U_h} ist
- Die Varianz ist

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} V(\hat{t}_{h\pi})$$

  wobei $V(\hat{t}_{h\pi})$ die Varianz von $\hat{t}_{h\pi}$ ist
- Ein unverzerrter Schätzer für die Varianz ist

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{h\pi})$$

  wobei $\hat{V}(\hat{t}_{h\pi})$ ein unverzerrter Schätzer für $V(\hat{t}_{h\pi})$ ist
- Für Herleitung: Die Zufallsvariablen $\hat{t}_{h\pi}$ sind unabhängig

Schätzung in Schichten (2)

Falls in jeder Schicht eine einfache Zufallsstichprobe gezogen wird, gilt:

$$\hat{t}_\pi = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

$$V(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{yU_h}^2$$

$$\hat{V}(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{1 - f_h}{n_h} S_{ys_h}^2$$

wobei $f_h = n_h / N_h$ der Auswahlsatz in Schicht h und

$$S_{yU_h}^2 = \frac{1}{N_h - 1} \sum_{U_h} (y_k - \bar{y}_{U_h})^2$$

$$S_{ys_h}^2 = \frac{1}{n_h - 1} \sum_{s_h} (y_k - \bar{y}_{s_h})^2$$

Stichprobenumfang in den Schichten

- Proportionale Aufteilung
  - wenn keine weiteren Informationen vorhanden
- Optimale Aufteilung
  - klein in Schichten mit geringer Streuung
  - groß in Schichten mit hoher Streuung
- Kosten-optimale Streuung
  - Minimierung der Erhebungskosten zur Informationsgewinnung
- Abhängig von Rücklaufquoten
  - Hoher Umfang in Schichten mit wenig Bereitschaft
  - Kleiner Umfang in Schichten mit viel Bereitschaft

Stichprobenumfang in den Schichten: Proportionale Aufteilung

- Idee: größere Schichten erhalten einen größeren Anteil in der Stichprobe
- Also: Wähle Stichprobenumfang in den einzelnen Schichten proportional zur Schichtgröße N_h in der Population:

$$n_{h,prop} = \left[n \frac{N_h}{N} \right]$$

  wobei die eckige Klammer nächst gelegene ganze Zahl liefert
- Bezüglich Genauigkeit nicht notwendigerweise optimal

Stichprobenumfang in den Schichten: Optimale Aufteilung

- Siehe Schichtungsprinzip!
- Unter Vernachlässigung des Auswahlsatzes ist die Varianz bei dem einfachen Stichprobenverfahren innerhalb der Schichten bestimmt durch:

$$Var(\hat{t}_\pi) = \sum_{h=1}^{H} N_h^2 \frac{S_{yU_h}^2}{n_h}$$

  je größer $N_h S_{yU_h}$ desto größer die Varianz
- Idee: Wähle n_h proportional zu $N_h S_{yU_h}$:

$$n_{h,opt} = \left[n \frac{N_h S_{yU_h}}{\sum_i^H N_i S_{yU_i}} \right]$$

- Optimale Aufteilung setzt Kenntnis von S_{yU_h} voraus (unrealistisch)
- Auswege:
  - Pilotstichprobe von kleinerem Umfang
  - vorherige Studien
  - bei Plausibilität, dass Varianzen in den einzelnen Schichten sehr ähnlich bzw. gleich sind, entspricht die proportionale Aufteilung annähernd der optimalen Aufteilung.
- Achtung: Optimalität bezieht sich lediglich auf ein Kriterium y, üblicherweise erheben wir sehr viele Kriterien

Stichprobenumfang in den Schichten: Kosten-optimale Aufteilung

- Optimale Stichprobe erstrebenswert, aber häufig mit Kosten für z.B. Pilotstudien verbunden
- Bezeichne c_0 die Fixkosten und c_h die jeweiligen Kosten, um Informationen über ein Individuum aus der $h$ten Schicht zu erhalten
- Gesamtkosten: $C = c_0 + \sum_{h=1}^{H} c_h n_h$
- Kosten-optimale Aufteilung ist dann

$$n_{h,kostopt} = \left[n \frac{N_h S_{yU_h} / \sqrt{c_h}}{\sum_{i=1}^{H} N_i S_{yU_i} / \sqrt{c_i}} \right]$$

- Die optimale relative Stichprobengröße für Schicht h ist größer, je (i) kleiner c_h (ii) größer N_h und (iii) größer S_{yU_h}
- Für $c_h = c$ bekommen wir die sogennante „Neyman (1934) Allokation“

A posteriori Schichtung (1)

- Generell bietet die geschichtete Stichprobe erhebliche Vorteile, wenn ein starker Design-Effekt vorliegt, d.h. wenn die Streuung innerhalb der Schichten deutlich geringer ist als die in der Grundgesamtheit
- Manchmal ist geschichteten Stichprobe aber nicht möglich, da die Schichtzugehörigkeit in der Grundgesamtheit nicht bekannt ist
- Ebenso kann es vorkommen, dass die Stichprobe bedingt durch unterschiedliche Rücklaufquoten bezüglich bekannter Sekundärmerkmale verzerrt ist, z.B. erhöhter Männeranteil
- Dies gilt es zu korrigieren im Rahmen einer „a posteriori Schichtung“: Höhergewichtung der Individuen, die in der Stichprobe bezüglich der Schichtungsmerkmale unterrepräsentiert sind.
- A posteriori Schichtung ist auch bekannt als Umgewichtung und ein häufig verwendetes Mittel in Befragungen

A posteriori Schichtung (2)

Rechnernutzung von Studenten

- Ziel einer Umfrage: wieviel Zeit pro Woche verbringen Studenten im Rahmen ihres Studiums vor dem Rechner
- Ziehung einer einfachen Zufallsstichprobe
- Bei der Auswertung zeigt sich, dass ein deutlicher Geschlechtsunterschied besteht: männliche Studenten verbringen weitaus mehr Zeit vor dem Rechner
- Nehmen wir an, dass der Anteil der männlichen Studierenden bei 50% liegt, in der Stichprobe hingegen befinden sich (bedingt durch die zufällige Auswahl) 60% Männer.
- Ignoriert man den geschlechtsspezifischen Effekt, so überschätzt man möglicherweise die Rechnernutzungszeit
- Schätzer kann jedoch unter Verwendung der Zusatzinformation zur Geschlechtsverteilung korrigiert werden.

A posteriori Schichtung (3)

- Ein unverzerrter Schätzer für den Mittelwert $\bar{y}_U$ der Population ist:

$$\hat{\bar{y}}_{U,post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{s_h}$$

  wobei $\bar{y}_{s_h}$ der Mittelwert in der h-ten Schicht ist

- Die Varianz lässt sich erwartungstreu schätzen durch

$$\hat{V}(\hat{\bar{y}}_{U,post}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h}$$

- Das Horvitz-Thompson-Theorem ist hier nicht anwenden, da die Gewichtung nicht durch die Auswahlwahrscheinlichkeiten, sondern durch die Schichtgrößen in Stichprobe und Grundgesamtheit erfolgt