Stichprobenverfahren

Proportionale Auswahlwahrscheinlichkeiten

Willi Mutschler (willi@mutschler.eu)

Sommersemester 2017

Beispiel

In einer Stichprobe sollen die Ausgaben für Marketing und Werbemaßnahmen von Kreisen und kreisfreien Städten eines Landes erhoben werden. Ziel ist es, den mittleren Marketing-Etat pro Kreis (bzw. Stadt) zu schätzen. Daraus lässt sich dann der Gesamt-Etat des Landes für Marketing hochrechnen.

- Ein mögliches Vorgehen wäre n Kreise zufällig auszuwählen und nach ihrem Marketing-Etat zu befragen. Bei einer einfachen Zufallsstichprobe kann es dabei rein zufällig passieren, dass hauptsächlich kleine, bevölkerungsschwache Kreise gezogen werden, deren Etat generell kleiner ist als der von bevölkerungsreichen Städten.
- Große Städte sind bezüglich des Marketing-Etats bedeutender, d.h. informativer als kleine Kreise. Daher scheint es sinnvoll, die größeren Städte mit größerer Wahrscheinlichkeit zu ziehen.

$\hookrightarrow$ Einschlusswahrscheinlichkeiten sind proportional zu einer Hilfsvariablen

Effekt: Proportionale Auswahlwahrscheinlichkeiten können die Varianz verringern.

Proportionale Auswahlwahrscheinlichkeiten

Man unterscheidet

- πps Sampling:
  - feste Stichprobengröße
  - ohne Zurücklegen
  - verbunden mit dem π Schätzer
- *pps* Sampling
  - ggf. zufällige Stichprobengröße
  - mit Zurücklegen
  - verbunden mit dem *pwr* Schätzer
- Kombination aus πps und *pps*
  - Horvitz-Thompson Schätzer
  - Hansen-Hurwitz Varianzschätzer

πps Sampling (1)

- Der π Schätzer für die Merkmalssumme ist gegeben durch $\hat{t}_\pi = \sum_s y_k/\pi_k$
- Extremfall: ein Design bei dem $y_k/\pi_k = c$ mit c konstant und n fixe Stichprobengröße, dann gilt für eine beliebige Stichprobe: $\hat{t}_\pi = nc$
- $\hat{t}_\pi$ hat keine Variation:

$$V(\hat{t}_\pi) = -\frac{1}{2} \sum\sum_U \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

- Für so ein Design benötigen wir allerdings Informationen über alle y_k, die wir nicht haben
- Deswegen suchen wir eine Variable x, von der wir annehmen können, dass sie approximativ proportional zu y ist und bekommen so eine geringere Varianz des π Schätzers
- Im Beispiel haben pro-Kopf Ausgaben eine geringere Streuung als die absoluten Gesamtausgaben
- Ziel: $\pi_k \propto x_k$

πps Sampling (2)

- Einschlusswahrscheinlichkeiten: $\pi_k = n \frac{x_k}{\sum_{j=1}^{N} x_j}$ um Proportionalität und $\sum_U \pi_k = n$ zu gewährleisten; Annahme: $nx_k < \sum_{j=1}^{N} x_j$, damit $\pi_k \leq 1$
- Anforderung an Design und Algorithmus:
  1. Auswahlverfahren der Stichprobe ist relativ simpel
  2. Die Einschlusswahrscheinlichkeiten erster Ordnung π_k sind strikt proportional zu x_k
  3. Die Einschlusswahrscheinlichkeiten zweiter Ordnung sind positiv $\pi_{kl} > 0$ für alle $k \neq l$
  4. Die Berechnung der π_{kl} ist exakt und nicht allzu computerintensiv
  5. $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l < 0$ für alle $k \neq l$, so dass der Yates-Grundy-Sen Varianzschätzer stets nichtnegativ ist

πps Sampling (3)

Design für $n = 1$: Kumulative Totalmethode

- Kumuliere x_k wie folgt
  1. Setze $T_0 = 0$, berechne $T_k = T_{k-1} + x_k, k = 1, ..., N$
  2. Ziehe aus der Gleichverteilung eine zufällige Zahl ε zwischen 0 und 1.
  3. Falls $T_{k-1} < \varepsilon T_N \leq T_k$ wird das Element k ausgewählt.
- Da $\pi_k = Pr(T_{k-1} < \varepsilon T_N \leq T_k) = \frac{T_k - T_{k-1}}{T_N} = \frac{x_k}{\sum_U x_k}$.
- Aber, da $n = 1$, gilt $\pi_{kl} = 0$ für alle $k \neq l$

πps Sampling (4)

Design für $n = 2$ nach Brewer (1963, 1975):

- Sei $c_k = \frac{x_k(T_N - x_k)}{T_N(T_N - 2x_k)}$ mit $T_N = \sum_U x_k$
  1. Das erste Element k wird mit Wahrscheinlichkeit $p_k = c_k / \sum_U c_k$ ohne Zurücklegen gezogen
  2. Das zweite Element l wird mit Wahrscheinlichkeit $p_{l|k} = x_l/(T_N - x_k)$ ohne Zurücklegen gezogen
- Es lässt sich zeigen, dass $\pi_k = 2x_k/T_N$ und

$$\pi_{kl} = \frac{2x_k x_l}{T_N(\sum_U c_k)} \frac{T_N - x_k - x_l}{(T_N - 2x_k)(T_N - 2x_l)}$$

  und sogar $\Delta_{kl} < 0$
- Vereinfachende Annahme: $x_k < \sum_U x_k/2$, damit $\pi_k \leq 1$

πps Sampling (5)

Designs für $n > 2$:

- Algorithmisch schwierig genau nach π_k auszuwählen
- Einschlusswahrscheinlichkeiten zweiter Ordnung schwierig zu bekommen
- $\Delta_{kl} < 0$ schwierig zu gewährleisten
- Überblick, siehe z.B. Brewer und Hanif (1983), Tillé (2006) oder Rosén (1997)

Exkurs: Systematisches πps Sampling

- Sei $T_0 = 0$ und $T_k = T_{k-1} + x_k$, a Stichprobenintervall, n ganzahlige Teil von T_N/a, mit $T_N = \sum_U x_k = na + c$, $0 \leq c < a$
    - Falls $c = 0$ bekommen wir die Stichprobengröße n, falls $c > 0$ bekommen wir die Stichprobengröße n oder $n + 1$
- Annahme: $nx_k \leq \sum_U x_k$ und x_k (gerundete) ganze Zahl
- Systematisches πps Sampling:
    1. Wähle mit gleicher Wahrscheinlichkeit, $1/a$, eine Zahl r zwischen 1 und a (einschließlich).
    2. Die Stichprobe besteht dann aus

    $$s = \{k : k = T_{k-1} < r + (j-1)a \leq T_k \text{ für ein } j = 1, 2, ..., n_s\} = s_r$$

    wobei $n_s = n$ für $r \leq c$ oder $n_s = n + 1$ für $c < r \leq a$
- Intuition:
    - Distanzen x_k werden beginnend am Ursprung und Endend bei T_N eine nach der anderen auf einer horizontalen Achse ausgelegt
    - $c = 0$: totale Distanz T_N wird in n Intervalle mit gleicher Länge a unterteilt
    - Zufälliger Start für das erste Intervall, danach systematisch
    - im Prinzip fixe Stichprobengröße
- $\pi_k = \frac{nx_k}{T_N - c}$

Monetary Unit Sampling

Relevant bei Wirtschaftsprüfern, um eine Stichprobe von Konten für eine Prüfung auszuwählen. x_k ist z.B. die Größe des Kontos/Buchungen in Euro

Monetary Unit Sampling einfaches Beispiel

- Prüffeld aus drei Rechnungen (5, 10 und 15 Euro)
- Wahrscheinlichkeit bei einfacher Zufallsauswahl ist 1/3, unabhängig vom Rechnungswert
- Aber: in größeren Buchungen werden auch größere Fehler erwartet und in kleineren Buchungen die kleineren Fehler
- $p(R1) = 5/30$, $p(R2) = 10/30$ und $p(R3) = 15/30$
- Somit kann ein Prüfer geforderte Risikominimierung erreichen

pps Sampling (1)

Größenproportionale Designs mit Zurücklegen

- Verwendung des *pwr* Schätzers:

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{i=1}^{m} y_{k_i} / p_{k_i}$$

  mit m fixe Anzahl an Ziehungen mit Zurücklegen und p_{k_i} die Wahrscheinlichkeit, das Element k_i zu ziehen
- Wenn wir ein Design haben, bei dem $y_k / p_k = c$ mit c konstant, dann haben wir für jede geordnete Stichprobe, $os = \{k_1, ..., k_m\}$: $\hat{t}_{pwr} = c$
- $\hat{t}_{pwr}$ hat keine Variation:

$$V(\hat{t}_{pwr}) = \frac{1}{m} \sum_U p_k \left(\frac{y_k}{p_k} - t_U \right)^2 = \frac{1}{2m} \sum \sum_U p_k p_l \left(\frac{y_k}{p_k} - \frac{y_l}{p_l} \right)^2$$

- Für so ein Design benötigen wir Informationen über alle y_k
- Deswegen suchen wir eine Variable x, von der wir annehmen können, dass sie approximativ proportional zu y ist, und bekommen so eine geringere Varianz des *pwr* Schätzers
- Ziel: $p_k \propto x_k$

pps Sampling (2)

Für die Ein-Zug-Auswahlwahrscheinlichkeit gilt $p_k \propto x_k$, also

$$p_k = \pi_k / n = \frac{x_k}{\sum_U x_k}$$

- Für $n = 1$ ist kumulative Totalmethode äquivalent zu πps
- m-maliges Wiederholen der kumulativen Totalmethode ergibt pps geordnete Stichprobe $os = \{k_1, k_2, ..., k_m\}$
- pwr Schätzer:

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{p_{k_i}} = (\sum_U x_k) \frac{1}{m} \sum_{i=1}^{m} \frac{y_{k_i}}{x_{k_i}}$$

- Varianzschätzer lässt sich dann einfach berechnen gegeben obiges p_k:

$$\hat{V}(\hat{t}_{pwr}) = (\sum_U x_k)^2 \frac{1}{m(m-1)} \left[\sum_{i=1}^{m} \left(\frac{y_{k_i}}{x_{k_i}} \right)^2 - \frac{1}{m} \left(\sum_{i=1}^{m} \frac{y_{k_i}}{x_{k_i}} \right)^2 \right]$$

Kombination πps und pps Sampling

- Varianzformel des *pwr* Schätzers ist sehr simpel, Einschlusswahrscheinlichkeiten zweiter Ordnung werden nicht benötigt
- π Schätzer ist jedoch häufig effizienter als *pwr*
- Manchmal verbindet man πps und *pps*:
    1. Verwende πps Stichprobendesign mit fester Stichprobengröße m, so dass
    $$\pi_k = mp_k = m\frac{x_k}{\sum_U x_k}$$
    2. Verwende π Schätzer für Merkmalssumme t_U
    3. Die Varianz wird dann geschätzt mit der *pps* Formel:
    $$v = \frac{1}{m(m-1)} \sum_s \left(\frac{y_k}{p_k} - \frac{1}{m} \sum_s \frac{y_k}{p_k} \right)^2$$
    4. Wir haben hier jedoch eine Verzerrung: $BIAS(v) = E(v) - V(\hat{t}_\pi) = \frac{m}{m-1} \left[V(\hat{t}_{pwr}) - V(\hat{t}_\pi) \right]$
    5. Falls π Schätzer effizienter, dann überschätzen wir die Varianz mit v

Sampford Sampling (1)

- Idee der Verwerfungsstichprobe:
  - Ziehe eine Stichprobe vom Umfang n mit Zurücklegen und den Ein-Zug-Auswahlwahrscheinlichkeiten p_i
  - Sind alle n Elemente verschieden, so wird die Stichprobe akzeptiert, ansonsten verworfen und eine neue gezogen
  - Approximativ gilt dann $p_i \approx \pi_i / n$
- Sampford (1967) Algorithmus sorgt für exakt $p_i = \pi_i / n$

Sampford Algorithmus

- Gegeben seien Auswahlwahrscheinlichkeiten π_k mit $\sum_U \pi_k = n$, eine Stichprobe der Größe n kann dann wir folgt gezogen werden:
  1. Ziehe das erste Element mit $p_k = \pi_k / n$
  2. In den weiteren $(n-1)$ Schritten werden aus allen Elementen mit Zurücklegen $(n-1)$ Elemente gezogen mit $\tilde{p}_k = \frac{\frac{\pi_k}{1-\pi_k}}{\sum_{j=1}^{N} \frac{\pi_j}{1-\pi_j}}$
  3. Falls die n gezogenen Elemente nicht paarweise verschieden, verwerfe die Stichprobe und beginne bei Schritt 1
- Auswahlwahrscheinlichkeiten zweiter Ordnung:

$$\pi_{kl} = K \frac{p_k}{1-\pi_k} \frac{p_l}{1-\pi_l} \sum_{t=2}^{n} (t - \pi_k - \pi_l) L_{n-t}(kl) \frac{1}{n^{t-2}}$$

mit $K := \left(\sum_{t=1}^{n} t L_{n-t}/n^t\right)^{-1}$, $L_m := \sum_{\text{s|s hat die Länge m}} \prod_{l \in s} \frac{\pi_l/n}{1-\pi_l}$,
$L_m(ij) := \sum_{\text{s|s hat die Länge m und enthält nicht i,j}} \prod_{l \in s} \frac{\pi_l/n}{1-\pi_l}$

- Es gilt: $\pi_{kl} < \pi_k \pi_l$, also Existenz positiver Varianzschätzung
- Bei großen Auswahlsätzen führt Sampford Sampling zu langen Rechenzeiten

Abschlussbeispiel

- Fischereistudie in England, siehe Cotter, Course, Buckland, und Garrod (2002)
- Ziel: Anzahl gefangener Fische schätzen
- Dabei wurden Kabeljau, Schellfisch und Weißfisch in der Nordsee in den Jahren von 1997 bis 1998 betrachtet.
- Gesamtzahlen nur sehr schwer zu erheben, daher Erhebung auf verschiedenen Fischerbooten
- y_k sind die in einem bestimmten Zeitraum auf dem Boot k gefangenen Fische
- Boote unterscheiden sich stark in Kapazität und Fangstrategie, hohe Streuung der y_k und damit wäre eine Schätzung basierend auf einer einfachen Zufallsstichprobe der Boote nur sehr ungenau.
- Verbesserung der Genauigkeit durch Hilfsmerkmal, das möglichst proportional zu den gefangenen Fischen y_k ist und vor der Stichprobenziehung bekannt ist:

$$X = \frac{VCU \cdot \text{Aufwand}}{\text{durchschnittliche Dauer der Ausfahrten in Tagen}}$$

- VCU: Kapazität der Schiffe („vessel capacity unit“)
- Aufwand: Stunden, die das Boot in den früheren Jahren unterwegs war
- Dauer der Ausfahrten ist indirekt proportional: kürzere Zeitspanne erlaubt mehr Ausfahrten
- Methodik: Ziehen der Boote mit Zurücklegen und Hansen-Hurwitz-Schätzer
- Da sich die Boote erheblich in ihren Fängen unterschieden, führte die PPS-Strategie hier zu einem erheblichen Effizienzgewinn