Stichprobenverfahren

Einfache Zufallsstichprobe

Willi Mutschler (willi@mutschler.eu)

Sommersemester 2017

Einfache Zufallsstichprobe

- Man unterscheidet Modelle
  - ohne Zurücklegen: $y_1, \ldots, y_n$ sind identisch verteilt, aber stochastisch abhängig. Alle Stichproben haben den gleichen Umfang.
  - mit Zurücklegen: $y_1, \ldots, y_n$ sind unabhängig und identisch verteilt. Die Stichprobengröße ist zufällig.
- In der Theorie und Praxis betrachten wir meistens ohne Zurücklegen, aber bei mit Zurücklegen haben einige Schätzfunktionen extrem einfache statistische Eigenschaften, die wir approximativ ausnutzen können

Einfache Zufallsstichprobe: Mit Zurücklegen (1)

- Ziehe m Elemente unabhängig voneinander und derart, dass jede der N Grundgesamtheitselemente mit derselben Wahrscheinlichkeit $1/N$ gezogen wird. Alle N Elemente nehmen an jeder Ziehung teil.
- Bereits gezogene Elemente können erneut gezogen werden

$\hookrightarrow$ Stichprobengröße ist zufällig

- Die Wahrscheinlichkeit, dass ein Element genau r mal in den m Ziehungen auftritt ist

$$\binom{m}{r}\left(\frac{1}{N}\right)^r\left(1-\frac{1}{N}\right)^{m-r}$$

- Die Wahrscheinlichkeit, dass ein Element überhaupt nicht gezogen wird, ist $\left(1-\frac{1}{N}\right)^m$. Somit gilt, dass die Wahrscheinlichkeit, dass ein Element k mindestens einmal in der Stichprobe auftritt:

$$\pi_k = 1-\left(1-\frac{1}{N}\right)^m$$

- Die Einschlusswahrscheinlichkeiten zweiter Ordnung lauten

$$\pi_{kl} = 1-2\left(1-\frac{1}{N}\right)^m+\left(1-\frac{2}{N}\right)^m$$

Einfache Zufallsstichprobe: Mit Zurücklegen (2)

- Unterscheidung des Begriffs Stichprobe wichtig:
  1. Bezeichne k_i das Element, welches in der $i$ten Ziehung gezogen wird ($i = 1, \ldots, m$), dann nennen wir

     $$os = (k_1, \ldots, k_m)$$

     die „geordnete Stichprobe“ mit $p(os) = 1/N^m$. Informationen über Zeitpunkt der Ziehung und Multiplizität vorhanden.
  2. Die Menge rein verschiedener Elemente in os

     $$s = \{k : k = k_i \text{ für ein i}; i = 1, \ldots, m\}$$

     bezeichnen wir als Mengen-Stichprobe s mit Stichprobendesign $p(s)$. Die Kardinalität n_s von s ist eine Zufallsvariable, es gilt $Pr(n_s \leq m) = 1$. Informationen über Zeitpunkt der Ziehung und Multiplizität nicht vorhanden.

Einfache Zufallsstichprobe: Mit Zurücklegen (3)

Verallgemeinerung für Design mit ungleichen Wahrscheinlichkeiten

- Sei $Pr[\text{Ziehen von Element k}] = p_k$ mit $\sum_U p_k = 1$ und k wird bei jeder der m Ziehung ersetzt, dann gilt
  1. für das geordnete Stichprobendesign $Pr[(k_1, k_2, ..., k_m)] = p_{k_1} \cdot p_{k_2} \cdot ... \cdot p_{k_m}$
  2. für das Mengen-theoretische Stichprobendesign eine komplizierte Form
- Einschlusswahrscheinlichkeit: $\pi_k = 1 - (1 - p_k)^m$
- Mitteln über den „p-expanded“ Wert des $k$ten Elements $\frac{y_k}{p_k}$, ergibt

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

  den unverzerrten *pwr* Schätzer für die Merkmalssumme $t_U = \sum_U y_k$.
- Die Varianz ist

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

  und lässt sich unverzerrt schätzen mit

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

- Dies ist der „p-expanded with replacement“ Schätzer (Hansen und Hurwitz, 1943)

Einfache Zufallsstichprobe: Mit Zurücklegen (4)

- Man kann natürlich auch den üblichen π-Schätzer verwenden: $\hat{t}_\pi = \sum_s \check{y}_k$
- Beide Schätzer sind unverzerrt, welcher die kleinere Varianz hat, hängt von den y Werten ab

Einfache Zufallsstichprobe: Ohne Zurücklegen

- Einschlusswahrscheinlichkeiten: $\pi_k = \frac{n}{N}$ und $\pi_{kl} = \frac{n}{N}\frac{n-1}{N-1}$
- Der π-Schätzer für die Merkmalssumme der Grundgesamtheit U vereinfacht sich zu:

$$\hat{t}_\pi = N\bar{y}_s = \frac{1}{f}\sum_s y_k$$

$$V(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{y,U}$$

$$\hat{V}(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{y,s}$$

mit

$$f = n/N, \qquad \text{(sampling fraction)}$$

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

$$S^2_{y,s} = \frac{1}{n-1}\sum_s (y_k - \bar{y}_s)^2 \qquad \text{(Stichprobenvarianz)}$$

- Für den π-Schätzer für den Mittelwert der Grundgesamtheit U wird durch N geteilt, bei der Varianz des Schätzers durch N^2

Designeffekt

Das Framework der einfachen Zufallsstichprobe ohne Zurücklegen wird häufig als Referenzwert für alternative Schätzmöglichkeiten verwendet

- Bezeichne p ein alternatives Design mit π Schätzer $\hat{t}_\pi$ und SI das Design der einfachen Zufallsstichprobe ohne Zurücklegen mit π-Schätzer $\hat{t}_{SI}$, dann bezeichnen wir das Varianzverhältnis

$$deff = \frac{V(\hat{t}_\pi)}{V(\hat{t}_{SI})} = \frac{\sum\sum_U \Delta_{kl}\check{y}_k\check{y}_l}{N^2\left(\frac{1}{n}-\frac{1}{N}\right)S^2_{y,U}}$$

  als „Designeffekt“
- $deff < 1$ bedeutet, dass das alternative Design präziser ist

Schätzung von Domains (1)

- In den meisten Umfragen werden Schätzwerte für Untergruppen der Grundgesamtheit, sogenannte „Domains“, erwünscht
- Beispiele:
  - Anteil von Personen über 65 Jahren
  - Durchschnittliche Einkommen von Haushalten mit drei oder mehr Kindern
- Notation:
  - $U_d \subset U$ bezeichne eine Unterpopulation der Größe N_d
  - $P_d = N_d/N$ bezeichne die relative Größe von U_d
- Annahme, dass N bekannt und N_d unbekannt ist
- Definiere Domain-Indikatorvariable

$$z_{dk} = \begin{cases} 1 & \text{falls } k \in U_d \\ 0 & \text{sonst} \end{cases} \qquad (k = 1, \ldots, N)$$

  dann

$$\sum_U z_{dk} = N_d \text{ und } \bar{z}_{dU} = \sum_U z_{dk}/N = N_d/N = P_d$$

- Also N_d ist Populationssumme und P_d der Populationsmittelwert von z_d

Schätzung von Domains (2)

- Im Rahmen der einfachen Zufallsstichprobe ohne Zurücklegen lassen sich die absolute und relative Größe einer Domain recht einfach schätzen
- Definiere $Q_d = 1 - P_d$, $n_d = \sum_s z_{dk}$, $p_d = n_d/n$ und $q_d = 1 - p_d$
- Es folgt, dass

$$S_{z_d U} = \frac{N}{N-1} P_d Q_d \qquad \text{und} \qquad S_{z_d s} = \frac{n}{n-1} p_d q_d$$

- Für den π-Schätzer dann

$$\hat{N}_d = N p_d, \qquad V(\hat{N}_d) = N^2 \frac{N-n}{N-1} \frac{P_d Q_d}{n}, \qquad \hat{V}(\hat{N}_d) = N^2 (1-f) \frac{p_d q_d}{n-1}$$

  wobei $\hat{N}_d$ und $\hat{V}(\hat{N}_d)$ unverzerrte Schätzer sind.
- Die relative Domaingröße, $P_d = N_d/N$, lässt sich mit $\hat{P}_d = p_d = n_d/n$ schätzen. Die Varianzen sind N^2 mal kleiner als die obigen Ausdrücke

Schätzung von Domains (3)

- Für die Schätzung der Summe $t_d = \sum_{U_d} y_k$ und Mittelwertes $\bar{y}_{U_d} = \sum_{U_d} y_k / N_d$ einer Untergruppe, definiere

$$y_{dk} = \begin{cases} y_k & \text{falls } k \in U_d \\ 0 & \text{sonst} \end{cases}$$

dann gilt $t_d = \sum_{U_d} y_k = \sum_U y_{dk}$ und lässt sich schätzen mit

$$\hat{t}_{d\pi} = \sum_s y_{dk}/\pi_k = \frac{N}{n} \sum_s y_{dk} = \frac{N}{n} \sum_{s_d} y_k$$

mit $s_d = U_d \cap s$, d.h. s_d ist die Untermenge an Elementen von s, die in die Domain U_d fallen