# Stichprobenverfahren

## Schätzung

Willi Mutschler
willi@mutschler.eu

Sommersemester 2017

# Notation

- $\theta$ bezeichnet einen Populationsparameter
- $\hat{\theta} = \hat{\theta}(S)$ bezeichnet Schätzfunktion für $\theta$ basierend auf einer zufälligen (noch zu ziehenden) Stichprobe $S$
- $\hat{\theta} = \hat{\theta}(s)$ bezeichnet Schätzwert für $\theta$ basierend auf einer realisierten Stichprobe s
- In der Stichprobentheorie interessieren wir uns für die Eigenschaften von Schätzfunktionen
- Die Verteilung einer Schätzfunktion lässt sich durch Betrachtung aller $M = |\mathcal{S}|$ möglichen Stichproben $s$ bewerkstelligen

# Eigenschaften von Schätzfunktionen

- Erwartungstreue: $E(\hat{\theta}) = \sum_{s \in \mathcal{S}} \hat{\theta}(s) p(s) = \theta$
- Varianz: $V(\hat{\theta}) = \sum_{s \in \mathcal{S}} \left(\hat{\theta}(s) - E(\hat{\theta})\right)^2 p(s)$
- Bias: $B(\hat{\theta}) = E(\hat{\theta}) - \theta$
- Mean Square Error (mse):
  $E\left[\hat{\theta} - \theta\right]^2 = \sum_{s \in \mathcal{S}} \left(\hat{\theta}(s) - \theta\right)^2 p(s) = V(\hat{\theta}) + [B(\hat{\theta})]^2$
- Variationskoeffizient:

$$cve(\hat{\theta}) = \frac{[V(\hat{\theta})]^{1/2}}{E(\hat{\theta})} \approx \frac{[\hat{V}(\hat{\theta})]^{1/2}}{\hat{\theta}}$$

  Indikator für die Präzision, je kleiner desto besser

**Frequentistische Betrachtungsweise**

In einer langen Serie wiederholter Ziehungen von Stichproben aus der Grundgesamtheit mit Stichprobendesign $p(s)$, werden die Durchschnittswerte einer Statistik $\theta(s)$ und die Varianz der Werte von $\theta(s)$ annähernd ihrem theoretischen Gegenstück entsprechen.

## Der Horvitz-Thompson Schätzer

- Der Horvitz-Thompson-Schätzer für z.B. die Merkmalssumme $t_U$ der Grundgesamtheit $U$ lässt sich basierend auf Stichprobe $s$ der Größe $n$ mit Einschlusswahrscheinlichkeiten $\pi_k$, die bekannt sind, folgendermaßen berechnen:

$$\hat{t}_\pi = \sum_U I_k \frac{y_k}{\pi_k} = \sum_U I_k \check{y}_k = \sum_s \frac{y_k}{\pi_k} = \sum_s \check{y}_k$$

- $\check{y} = y_k/\pi_k$ ist der sogenannte *expanded value* von $y_k$ für $k$ in Stichprobe $s$
- Idee: Da die Stichprobe weniger Elemente enthält als die Grundgesamtheit, wird eine „Expansion“ benötigt. Das $k$te Element in der Stichprobe repräsentiert folglich $1/\pi_k$ Elemente der Grundgesamtheit.
- Voraussetzung: $\pi_k > 0$ für alle $k \in U$
- Der Horvitz und Thompson (1952) Schätzer ist definiert für alle Stichprobendesigns, häufig findet man auch die Bezeichnung $\pi$-estimator oder inverse probability estimator.

# Eigenschaften des Horvitz-Thompson Schätzers (I)

- Einschlussindikator ist die einzige Zufallsvariable, deren Wert (1 oder 0) annimmt, hängt nur ab von $S$
- Horvitz-Thompson Schätzer ist unverzerrt:

$$E\left(\sum_U \frac{I_k}{\pi_k} y_k\right) = \sum_U \frac{y_k}{\pi_k} E(I_k) = \sum_U \frac{y_k}{\pi_k} \pi_k = \sum_U y_k$$

- Für die Varianz betrachten wir Ziehen ohne Zurücklegen:

$$Cov(I_k, I_l) = \Delta_{kl} =: \begin{cases} \pi_k(1-\pi_k), & k = l \\ \pi_{kl} - \pi_k \pi_l, & k \neq l \end{cases}$$

$$\check{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}} = \begin{cases} 1 - \pi_k, & k = l \\ 1 - \frac{\pi_k \pi_l}{\pi_{kl}}, & k \neq l \end{cases}$$

# Eigenschaften des Horvitz-Thompson Schätzers (II)

- Für die Varianz von $\hat{t}_\pi$ gilt dann:

$$V(\hat{t}_\pi) = \sum\sum_U \Delta_{kl}\check{y}_k\check{y}_l$$

- Diese lässt sich basierend auf einer realisierten Stichprobe $s$ schätzen mit:

$$\hat{V}(\hat{t}_\pi) = \sum\sum_s \check{\Delta}_{kl}\check{y}_k\check{y}_l$$

- Es gilt, dass $E(\hat{V}(\hat{t}_\pi)) = V(\hat{t}_\pi)$
- Mithilfe von Einschlusswahrscheinlichkeiten lässt sich die Varianz und ihr Schätzer auch schreiben als:

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

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

- Voraussetzung: $\pi_{kl} > 0$ für alle $k \neq l \in U$

# Yates-Grundy Varianz

- Alternativ können wir die Varianz auch schreiben als

$$V(\hat{t}_\pi) = \sum\sum_U \Delta_{kl} \check{y}_k \check{y}_l = \frac{-1}{2} \sum\sum_U \Delta_{kl} \left(\check{y}_k - \check{y}_l\right)^2$$

  und schätzen mit

$$\hat{V}(\hat{t}_\pi) = \frac{-1}{2} \sum\sum_s \check{\Delta}_{kl} \left(\check{y}_k - \check{y}_l\right)^2$$

- Auch hier gilt, dass $E(\hat{V}(\hat{t}_\pi)) = V(\hat{t}_\pi)$
- Voraussetzung: $\pi_{kl} > 0$ für alle $k \neq l \in U$
- Dies ist die sogenannte Yates-Grundy(-Sen) Varianz, benannt nach Yates und Grundy (1953) und Sen (1953)
- Gilt jedoch nicht für zufällige Stichprobengrößen!

# Stichprobengröße

- Auch die Stichprobengröße $n_S$ ist eine Statistik, die wir betrachten können:

$$n_S = \sum_U I_k$$

$$E(n_S) = \sum_U \pi_k$$

$$V(n_S) = \sum_U \pi_k(1-\pi_k) + \sum\sum_{\substack{U\\k\neq l}}(\pi_{kl} - \pi_k\pi_l) = \sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum\sum_{\substack{U\\k\neq l}}\pi_{kl}$$

- Zufällige Stichprobengrößen treten z.B. bei Stichproben mit Zurücklegen, Bernoulli Sampling oder single-stage cluster sampling auf
- Üblicherweise versucht man dies zu vermeiden und $p(s)$ derart zu gestalten, dass jede Stichprobe genau $n$ Elemente enthält. Dann gilt:

$$\sum_U \pi_k = n, \qquad \sum\sum_{\substack{U\\k\neq l}}\pi_{kl} = n(n-1), \qquad \sum_{\substack{l\in U\\l\neq k}}\pi_{kl} = (n-1)\pi_k$$

## Zusammenfassung Horvitz-Thompson-Schätzer

- $\hat{t}_\pi$ ist der $\pi$-Schätzer für die Merkmalssumme der Grundgesamtheit U

$$\hat{t}_\pi = \sum_U I_k \frac{y_k}{\pi_k} = \sum_s \frac{y_k}{\pi_k} = \sum_s \check{y}_k$$

$$V(\hat{t}_\pi) = \sum\sum_U \Delta_{kl} \check{y}_k \check{y}_l$$

$$\hat{V}(\hat{t}_\pi) = \sum\sum_s \check{\Delta}_{kl} \check{y}_k \check{y}_l$$

- $\hat{\bar{y}}_\pi$ ist der $\pi$-Schätzer für den Mittelwert der Grundgesamtheit U

$$\hat{\bar{y}}_\pi = \frac{1}{N} \sum_U I_k \frac{y_k}{\pi_k} = \frac{1}{N} \sum_s \frac{y_k}{\pi_k} = \frac{1}{N} \sum_s \check{y}_k$$

$$V(\hat{\bar{y}}_\pi) = \frac{1}{N^2} \sum\sum_U \Delta_{kl} \check{y}_k \check{y}_l$$

$$\hat{V}(\hat{\bar{y}}_\pi) = \frac{1}{N^2} \sum\sum_s \check{\Delta}_{kl} \check{y}_k \check{y}_l$$

- Bemerkung: Alle Aussagen, die wir über die Merkmalssumme $t_U = \sum_U y_k$ treffen, lassen sich analog mit $1/N$ für den Mittelwert $\bar{y}_U$ treffen, wobei bei der Varianz mit $N^2$ geteilt wird.