University of Applied Sciences Northwestern Switzerland
School of Business

Einführung in die Methoden der Stichprobenerhebungen

Weiterbildungslehrgang in angewandter Statistik

ETH Zürich

Beat Hulliger

Fachhochschule Nordwestschweiz FHNW

21.11.2016

Ziele

- Kennen der Schritte einer Stichprobenerhebung
- Kennen der wichtigen Begriffe
- Verstehen des Paradigmas der Zufallsstichprobe
- Anwenden von Stichprobenlänen und Schätzverfahren für einfache und geschichtete Zufallsstichproben mit `R` und `survey`.
- Kennen der Probleme bei Datenaufbereitung und Auswertung
- Fähigkeit zur Beurteilung einer Erhebung

Inhalt

Einführung

Use of electronic information in the business

- Student project by Lea Bluntschli, Evelyne Lohrer, David Meyer, Roman Nussbaumer[1]
- December 2013 to June 2014
- 23 iterations to develop questionnaire (3 Sprachen)
- Online Survey
- Access to employees of two software companies in Spain an Switzerland through HR-departments.
- Reminder allowed in Switzerland but not in Spain.
- Data preparation and Analysis with SPSS.

---

[1] Bluntschli, L., Lohrer, E., Meyer, D., Nussbaumer, R. (2014) An Analysis of the Difference in Information Gathering of Generation X and Generation Y in the Business Environment, Master of Science International Management, School of Business FHNW, Olten

Tabelle: Response rates

	Switzerland	Spain	Total
Net sample size	109	154	263
Gross sample size	173	435	608
Response rate	63%	35%	43%

Exceptionally high response rate!

Data Preparation and Analysis

- Recoding (string/numeric)
- Scaling (Frequency of use)
- Weighting per country (reponse rate)
- Graphs and chisquare tests.

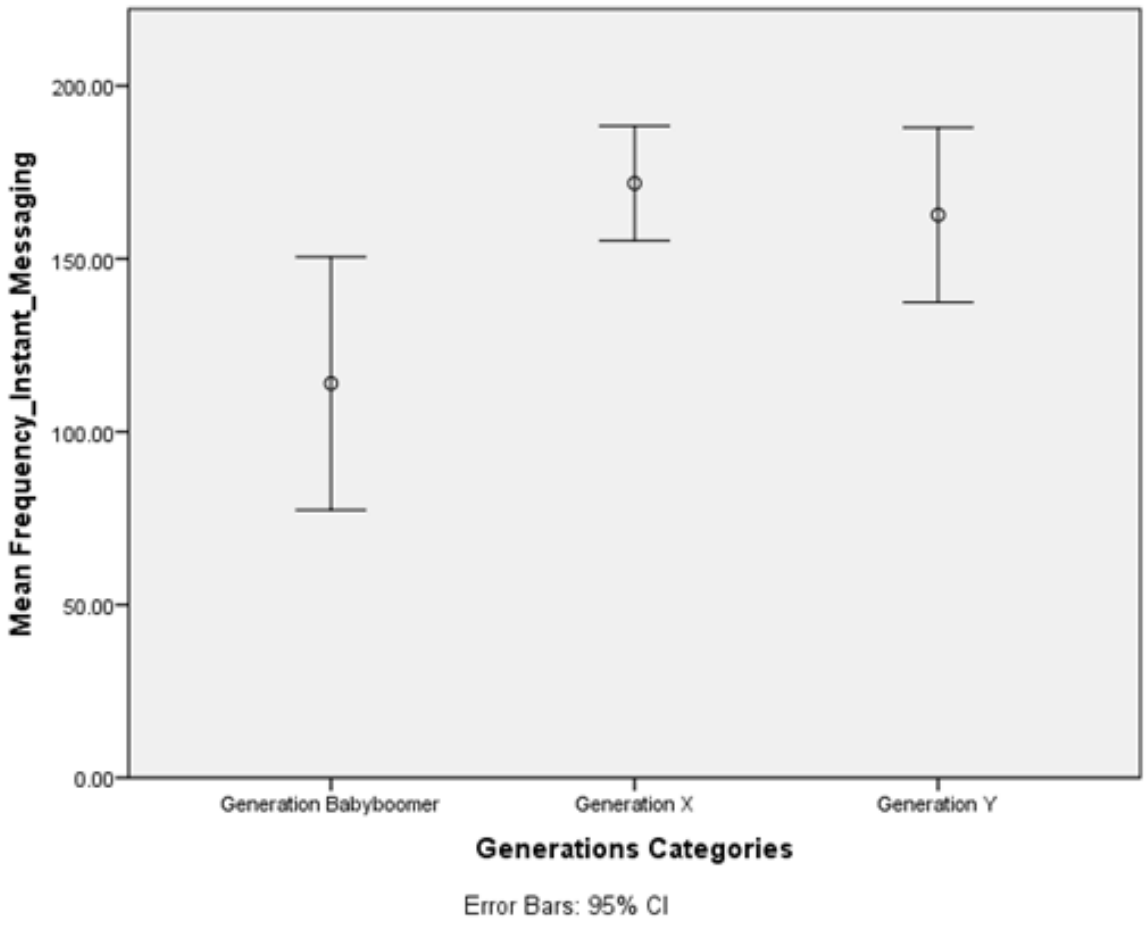


Abbildung: Use of Instant Messaging by Generations (Boomer 1940-1959, X 1960-1979, Y 1980-2000)

Erhebungen

"Statistics: Numerical data relating to an aggregate of individuals; the science of collecting, analysing and interpreting such data."
(Kendall and Buckland, Dictionary of Statistical Terms)

Mit der Erhebung ("collection") von Daten beschäftigen sich speziell zwei Teilgebiete der Statistik: Versuchsplanung und Stichprobentheorie.

Stichprobentheorie

- Randomisierung als wesentliches Element.
- Idee: Kiaer (International Statistical Institute, Bern 1895).
- Grundlegung der Stichprobentheorie durch Bowley, Tschuprow und Neyman 1920-1934.
- Anwendungsbereiche?

Stichprobenmethoden: Stichprobentheorie plus Methoden zur Lösung von praktischen Problemen, insbesondere Antwortausfälle. Kosten drücken sich in Stichprobengrössen aus und sind allgegenwärtig.

Eine Stichprobenerhebung wählt einen Teil, genannt Stichprobe, einer endlichen Population zufällig aus, untersucht die Elemente der Stichprobe, und schliesst dann auf Eigenschaften der gesamten Population.

POPULATION (N, θ)		$\hat{\theta} \approx \theta$
$\Downarrow$ Stichprobenplan	$\Longleftrightarrow$	Schätzer $\Uparrow$
STICHPROBE (n)	$\Rightarrow$ Datengewinnung $\Rightarrow$	DATEN

Andere Auswahlmethoden

Quotenstichprobe: Stichprobe als Abbild der Population gemäss Quoten. Kann zu verfälschten Schätzugen führen. Keine sinnvolle Varianzschätzung.

Selbstauswahl : Zeitungsumfragen (statistisch wertlos). Web-Panels.

Gelegenheitsstichprobe: Nach Belieben, oft nach Kostenkriterien (facebook..., Big Data).

Gezielte Auswahl: Suchen von Elementen mit bestimmten Eigenschaften, z.B. extreme.

Teilerhebung: Auswahl eines genau definierten Teils der Population, z.B. nur die grössten Unternehmen (Konzentrationsstichprobe).

Fall-Studie: Kleine Anzahl von Detailstudien. Exploratorisch sinnvoll, aber analytisch nicht auswertbar.

- Zufallsstichprobe liefert Basis für Schluss auf die Population und Varianzschätzung.
- Eine Vollerhebung (Zensus) untersucht die ganze Population.
- Bei Erhebungen mit Zufallsstichproben und Vollerhebungen gibt es Fehler und Antwortausfälle, generell Abweichungen vom Ideal.
- Ein Register wird nicht zum Zweck der Untersuchung der Population, sondern meistens für administrative Zwecke erstellt.
- Gelegenheits-Stichproben mit Modellierung: Bias?

Qualitätsunterschied nach Typ der Stichprobe

Kriterium	Stichprobe	Vollerhebung	Register	Gelegenheit
Umfang der Information	+	.	−	.
Qualität der Daten	+	.	−	−
Kosten	.	−	+	+
Schnelligkeit	+	−	.	+
Untergrupppen-Auswertung	−	+	.	.
Vollständigkeit	−	+	.	−
Bias	+	+	.	−
Varianz	−	+	+	.
Repräsentativität	+	+	+	−

Weitere wichtige Begriffe:

Charakteristik: (Parameter) Eigenschaft einer Population, Populations-Parameter, zu schätzende Grösse. Z.B. ein Populationsmittel oder ein Total, aber auch Varianzen und Quantile, Korrelationen

Stichprobenrahmen: Liste welche einen operationellen Zugang zu den einzelnden Einheiten der Population erlaubt.

Gewichtung: In der Praxis wird gerne mit linearen Schätzern gearbeitet, die als gewichtete Summen der einzelnen Beobachtungen geschrieben werden können.

Untersuchungsbereich: Teilpopulation für welche Auswertungen gemacht werden.

Schichtung: Aufteilung der Population in möglichst homogene Teilpopulationen.

Klumpung: Die Einheiten können nicht direkt erschlossen werden, sondern treten in Gruppen (Klumpen) auf.

Panel: Die Elemente einer Stichprobe werden in gewissen Zeitabständen wiederholt befragt.

Die 5 Schritte einer Erhebung

1. Planung: Ziele, Ressourcen, Organisation, Grundgesamtheit, Stichprobenrahmen, Stichprobenplan)

Die 5 Schritte einer Erhebung

1. Planung: Ziele, Ressourcen, Organisation, Grundgesamtheit, Stichprobenrahmen, Stichprobenplan)
2. Erhebungsinstrument: Inhalte, Form (Fragebogen) und Methode (CATI etc.), Referenz- und Erhebungsperiode,

Die 5 Schritte einer Erhebung

1. Planung: Ziele, Ressourcen, Organisation, Grundgesamtheit, Stichprobenrahmen, Stichprobenplan)
2. Erhebungsinstrument: Inhalte, Form (Fragebogen) und Methode (CATI etc.), Referenz- und Erhebungsperiode,
3. Datengewinnung: Tests und Pilot,Stichprobenziehung, Feldarbeit, Rücklauf- und Qualitätskontrolle, Mahnungen, Erfassung

Die 5 Schritte einer Erhebung

1. Planung: Ziele, Ressourcen, Organisation, Grundgesamtheit, Stichprobenrahmen, Stichprobenplan)
2. Erhebungsinstrument: Inhalte, Form (Fragebogen) und Methode (CATI etc.), Referenz- und Erhebungsperiode,
3. Datengewinnung: Tests und Pilot,Stichprobenziehung, Feldarbeit, Rücklauf- und Qualitätskontrolle, Mahnungen, Erfassung
4. Auswertung: Datenaufbereitung, Rückfragen, Schätzverfahren, Deskriptive und analytische Statistiken

Die 5 Schritte einer Erhebung

1. Planung: Ziele, Ressourcen, Organisation, Grundgesamtheit, Stichprobenrahmen, Stichprobenplan)
2. Erhebungsinstrument: Inhalte, Form (Fragebogen) und Methode (CATI etc.), Referenz- und Erhebungsperiode,
3. Datengewinnung: Tests und Pilot,Stichprobenziehung, Feldarbeit, Rücklauf- und Qualitätskontrolle, Mahnungen, Erfassung
4. Auswertung: Datenaufbereitung, Rückfragen, Schätzverfahren, Deskriptive und analytische Statistiken
5. Kommunikation: Präsentation, Bericht, Datenschutz, Dokumentation, Archivierung, Sekundäranalysen

Kritische Punkte

- Management und Organisation
- Informatik und Datenlieferungen
- Statistisches Wissen und Können
- Aufwand
  - Schritte 1 bis 3: 60%
  - Schritte 4 und 5: 60%
- Vermittlung der Ergebnisse

Erhebungsinstrumente

- Befragung
- Messung
- Beobachtung

Befragungsmethoden

- Persönliches Interview, allenfalls unterstützt durch Computer (CAPI)
- Persönliches Interview über Telefon (CATI)
- Schriftlicher Fragebogen zum selbst Ausfüllen, Abgabe per Post oder direkt.
- Elektronischer Fragebogen zum selbst Ausfüllen: Fester Computer, Zusendung Fragebogen bzw. Programm, Internet (online-Erhebung) (CASI, CAWI)
- Touchtone Data Entry, SMS-Survey...

Instrument

Offene Antwort

Wie sind Sie heute morgen aufgestanden?

.....

.....

- Mehr Freiheit
- Neue Aspekte
- Codierung u.U. heikel
- Aufwändig (Beantwortung und Auswertung)

Geschlossene Antwort

Wie sind Sie heute morgen aufgestanden?
☐ Mit dem linken Fuss
☐ Mit dem rechten Fuss
☐ Mit beiden Füssen gleichzeitig
☐ Weiss nicht

- Einfache Auswertung
- Gute Vergleichbarkeit der Antworten
- Aufwändige Entwicklung
- Eingeschränkte Antwortmöglichkeit

Formulierung

- Eindeutige Fragen (keine Auswahlsendung)
- Eindeutige Antworten
- Kurze Fragen
- Spezifische Fragen
- Vorsicht bei Einleitungen!
- Vorsicht bei Verneinungen!
- Fachausdrücke vermeiden
- Sprache und Übersetzung

Antwortskalen

- Binär (Geschlecht)
- Numerisch (Anzahl Dienstjahre)
- Kategorien (ledig, verheiratet, geschieden, verwitwet)
- Likert Skalen (Stimme klar zu, stimme eher zu, weder-noch, lehne eher ab, lehne klar ab)
- Häufigkeiten/Klassen (Unter 18 Jahre, 18-30 Jahre, 31-45 Jahre, 45-65 Jahre, über 65 Jahre)
- Matrix

Struktur

- Einleitung
- Kapitel
- Sprungfragen
- Schluss

Entwicklung des Fragebogens

- Zuerst Planung und Konzept, dann Fragebogen entwickeln.
- Entwurf
- Kritische Hinterfragung jeder einzelnen Frage und der Struktur
- Argumentation für Fragen und Struktur
- Graphische Darstellung
- Kürzen!
- zwei bis zwanzig Versionen!

Testen und Pilot

- Kleine Testrunde mit Kollegen
- Diskussion (Test) mit Auftraggeber
- Kleine Testrunde bei möglichen Befragten
- Grosse Testrunde inklusive Erhebungsorganisation im Pilottest

Zufallsstichproben, einfache Zufallsstichprobe

Population

(Skript 1.5)

- Population $U = \{1, \ldots, N\}$
- Variable y mit Werten y_i, $i \in U$
  Beispiel: $y_i = 1$, falls AHV-Bezüger, sonst $y_i = 0$.
- Charakteristik $\theta(y_U)$ zu schätzen, z.B.
  Populations-Mittel $\bar{y}_U := \sum_{i \in U} y_i / N$
  Total $y_{U+} = \sum_{i \in U} y_i$.

Stichproben

- Stichprobenraum $\mathscr{S} = \{S \subset U\}$ (oder $\{0,1\}^N$).
- Stichprobenplan $p(S) : \mathscr{S} \to [0,1]$ mit $0 \leq p(S) \leq 1$ und $\sum_{\mathscr{S}} p(S) = 1$.
- Einschränkung auf $\{S \subset U : p(S) > 0\}$ oder $\{S \subset U : |S| = n\}$ (n feste Stichprobengrösse).
- Einschlusswahrscheinlichkeit für die Einheit i :

$$\pi_i = P[i \in S] = \sum_{S \ni i} p(S)$$

- Vektor von Einschlussindikatoren $I_i = 1$ oder $I_i = 0$: $P[I_i = 1] = \pi_i$ beschreibt Randverteilung.

Hier Stichproben ohne Zurücklegen: $S = \{i_1, ..., i_n\}$ mit $i_j \neq i_k (j \neq k)$.

Schätzer

- Gegeben Stichprobe $S = \{i_1, \ldots, i_n\}$ (oder $I \in \{0,1\}^N$).
- Schätzer $T(y_{i_1}, \ldots, y_{i_n}) = T(y_S)$.
- Erwartungswert des Schätzers

$$\mathrm{E}_S[T(y_S)] = \sum_{S \in \mathscr{S}} p(S) T(y_S).$$

- **Bias**: $\mathrm{E}_S[T(y_S)] - \theta(y_U)$.
  Falls der Bias 0 ist, heisst der Schätzer erwartungstreu.
- **Varianz**: $V_S[T(y_S)] = E_S\left[(T(y_S) - E_S[T(y_S)])^2\right]$

Stichproben-Paradigma und Strategie

Paradigma:

- Werte $y_{i_1}, \ldots, y_{i_n}$ sind fest!! (nicht wie bei klassischer Statistik!)
- Zufälliges Element ist S.

Strategie:

- Paar Stichprobenplan und Schätzer (p, T)
- Gute Strategie: kleiner Bias, kleine Varianz.
- Robuste Strategie: Einfacher Stichprobenplan
- Flexibilität: Modell-unterstützte Schätzverfahren

Einfache Zufallsstichprobe

Stichprobenplan ES

(Skript 2)

- Jede Teilmenge von U der Grösse n hat dieselbe Wahrscheinlichkeit, gezogen zu werden, nämlich

$$p(S) = 1/\left(\begin{array}{c} N \\ n \end{array} \right) = \frac{n!(N-n)!}{N!}.$$

- Einschlusswahrscheinlichkeit: $P[i \in S] = \pi_i = n/N, \forall i \in U$
- Stichprobenrate $f = n/N$.
- Urnenmodell: blindes Ziehen von i aus einer Urne.
  - Ohne Zurücklegen
  - Mit Zurücklegen

Gleiche Einschlusswahrscheinlichkeiten garantiert nicht einfache Zufallsstichprobe!
keine ES: Population in Reihe, zufälliger Start, S ist n nächste Elemente

Schätzer für das Populatonsmittel

Das Stichprobenmittel

$$T(y_S) = \bar{y}_S = \sum_{i \in S} y_i / n$$

schätzt $\bar{y}_U$ erwartungstreu:

$$\begin{aligned} \mathrm{E}\left[\sum_{i \in S} y_i / n\right] &= \mathrm{E}\left[\sum_{i \in U} y_i 1\{i \in S\} / n\right] \\ &= \frac{1}{n} \sum_{i \in U} y_i \mathrm{E}[1\{i \in S\}] = \frac{1}{n} \sum_{i \in U} y_i \pi_i = \bar{y}_U. \end{aligned}$$

Varianz des Stichprobenmittels

$$V[\bar{y}_S] = (1 - n/N)\frac{1}{n}D^2,$$

wobei

$$D^2 = \frac{1}{N-1}\sum_{i \in U}(y_i - \bar{y}_U)^2$$

- $f = n/N$ ist die Stichprobenrate.
- $1 - n/N = 1 - f$ heisst Endlichkeitskorrektur.
- Populationsgrösse N nur in Endlichkeitskorrektur!
- Standardabweichung des Stichprobenmittels $V[\bar{y}_S]^{1/2}$
- Halbieren (SA) braucht vier mal mehr Beobachtungen!

Varianzschätzer

- D mit d ersetzen.

$$v[\bar{y}_S] = (1 - \frac{n}{N})\frac{d^2}{n} = (1 - \frac{n}{N})\frac{1}{n}\left(\frac{1}{n-1}\sum_{i \in S}(y_i - \bar{y}_S)^2\right).$$

- $v[\bar{y}_S]$ schätzt $V[\bar{y}_S]$ erwartungstreu.

Zentraler Grenzwertsatz für ES

- ZGS gilt unter Bedingungen an Momente von y_U.
- $N \to \infty$, $n \to \infty$ und $n/N = O(1)$:

$$\mathscr{L}\left(\frac{\bar{y}_S - \bar{y}_U}{\sqrt{V[\bar{y}_S]}}\right) \to N(0,1)$$

- $\Rightarrow$ Vertrauensintervalle, Tests

Population 5, ES mit $n = 2$

(Skript Beispiel 2 und 3)

$U = \{1,2,3,4,5\}$, $y_U = (1,5,9,8,12)^\top$
$\mathscr{S} = \{S \subset U : |S| = 2\}$, alle mit gleicher Wahrscheinlichkeit

Aufgabe: $\mathscr{S}$ explizit aufzählen, je das Mittel berechnen:
$T_k, k = 1,\ldots,10$.

Berechne den Erwartungswert $E[T]$ der T_k und die Varianz $V[T] = \sum_{S\in\mathscr{S}}(T_k - E[T])^2/10$.

Keine ES

- Was passiert, wenn die Stichprobe $S = \{1, 5\}$ eine Wahrscheinlichkeit von $p(S) = 0.19$ erhält, während die anderen Stichproben je die gleiche Wahrscheinlichkeit haben?
- Was passiert, wenn zufällig ein Element gewählt wird und zusammen mit dem nächsten Element die Stichprobe bildet (i=5 nimmt i=1 in die Stichprobe)?

Schätzer für das Populationstotal

$$y_{U+} = \sum_U y_i$$

schätzen mit

$$N\bar{y}_S = \frac{N}{n}\sum_S y_i = \sum_S w_i y_i$$

“Hochrechnung" mit Gewichten

$$w_i = N/n.$$

Stichprobengrösse

Stichprobenmittel mit gewünschter Varianz V :

$$n_0 = \tilde{D}^2 / V$$

($\tilde{D}$ ist eine Abschätzung von D).
Stichprobenmittel mit gewünschtem Variationskoeffizient c:

$$n_0 = \frac{(\tilde{D}/\bar{y}_U)^2}{c^2}$$

(Variationskoeffizient: $c(T) = V[T]^{1/2}/\mathrm{E}[T]$, bzw. Schätzung)
Mit Endlichkeitskorrektur (falls n_0 nahe bei N):

$$n = \frac{n_0}{1 + n_0/N}$$

Anteile

Sei A eine bestimmte Teilmenge der Population.
$p_U = |A|/N = \sum_{i \in U} 1[i \in A]/N$. Anteil in der Stichprobe:

$$p_S = \sum_{i \in S} 1[i \in A]/n$$

schätzt p_U erwartungstreu (da Stichprobenmittel).

$$V(p_S) = \frac{p_U(1-p_U)}{n} \frac{N-n}{N-1}$$

$$v(p_S) = (1 - n/N) \frac{p_S(1-p_S)}{n-1}.$$

- Stichprobengrösse:

$$n_0 = \frac{\tilde{p}_U(1-\tilde{p}_U)}{V},$$

  wobei $\tilde{p}_U$ eine Abschätzung von p_U und V die gewünschte Varianz des Stichprobenanteils ist.
- Mit Endlichkeitskorrektur: $n = n_0/(1+(n_0-1)/N)$
- Da $p_U(1-p_U)$ bei 0.5 maximal ist, ist der schlimmste Fall $n_0 = 0.25/V$.
- Vertrauensintervall mit halber Länge 5% bei $p_U = 0.5$:

$$n_0 = 0.25/(0.05/2)^2 = 400$$

Untersuchungsbereiche

Auswertung nur für eine Teilpopulation: $U_B \subset U$:
Betrachte $S_B = S \cap U_B$ als einfache Zufallsstichprobe der Grösse n_B.
Schätze $\bar{y}_{U_B}$ mit

$$\bar{y}_{S_B} = \sum_{S_B} y_i / n_B.$$

(n_B ist zufällig!)
Total y_{U_B+} schätzen mit

$$N_B \cdot \bar{y}_{S_B}$$

oder, falls N_B unbekannt, mit $\frac{N}{n} \sum_{i \in S} y'_i = N \cdot \bar{y}'_S$, wobei $y'_i = y_i \cdot 1\{i \in S_B\}$.

Varianz von $\bar{y}_{S_B}$ wird erwartungstreu geschätzt durch

$$v(\bar{y}_{S_B}) = (1 - n_B/N_B)d_B^2/n_B$$

Varianz von $\bar{y}'_S$ mit normalem Var.-Schätzer für Stichprobenmittel.

Differenz des Populationsmittels zweier disjunkter Untersuchungsbereiche schätzen mit:

$$\bar{y}_{S_B} - \bar{y}_{S_C}$$

Varianz geschätzt durch

$$v(\bar{y}_{S_B}) + v(\bar{y}_{S_C}).$$

Auflösung einer Stichprobe

- **Grössen-Auflösung** R_s beschreibt Genauigkeit anhand des Anteils der kleinsten schätzbaren Gruppe (für $p < 0.1$) in einem Untersuchungsbereich U_B.
- Approximative Grössenauflösung $\tilde{R}_s = 4N_B/n_B$. Mit $p = \tilde{R}_s/N_B$ ergibt sich die Stichprobengrösse $n_B = 4/p$. Z.B. für $p = 0.01$ ist $n_B = 400$.
- Die **Differenz-Auflösung** R_d für die Unterscheidung zweier Anteile in zwei gleich grossen Untersuchungsbereichen beschreibt die Genauigkeit anhand der kleinst möglichen Differenz, die schätzbar ist.
- Grobe Approximation für Anteilsunterschied $\tilde{r}_d = \tilde{R}_d/N_B$ ergibt $n_B = 2/\tilde{r}_d^2$ in beiden Untersuchungs-bereichen. Z.B. für $\tilde{r}_d = 0.05$ ergibt sich $n_B = 800$!

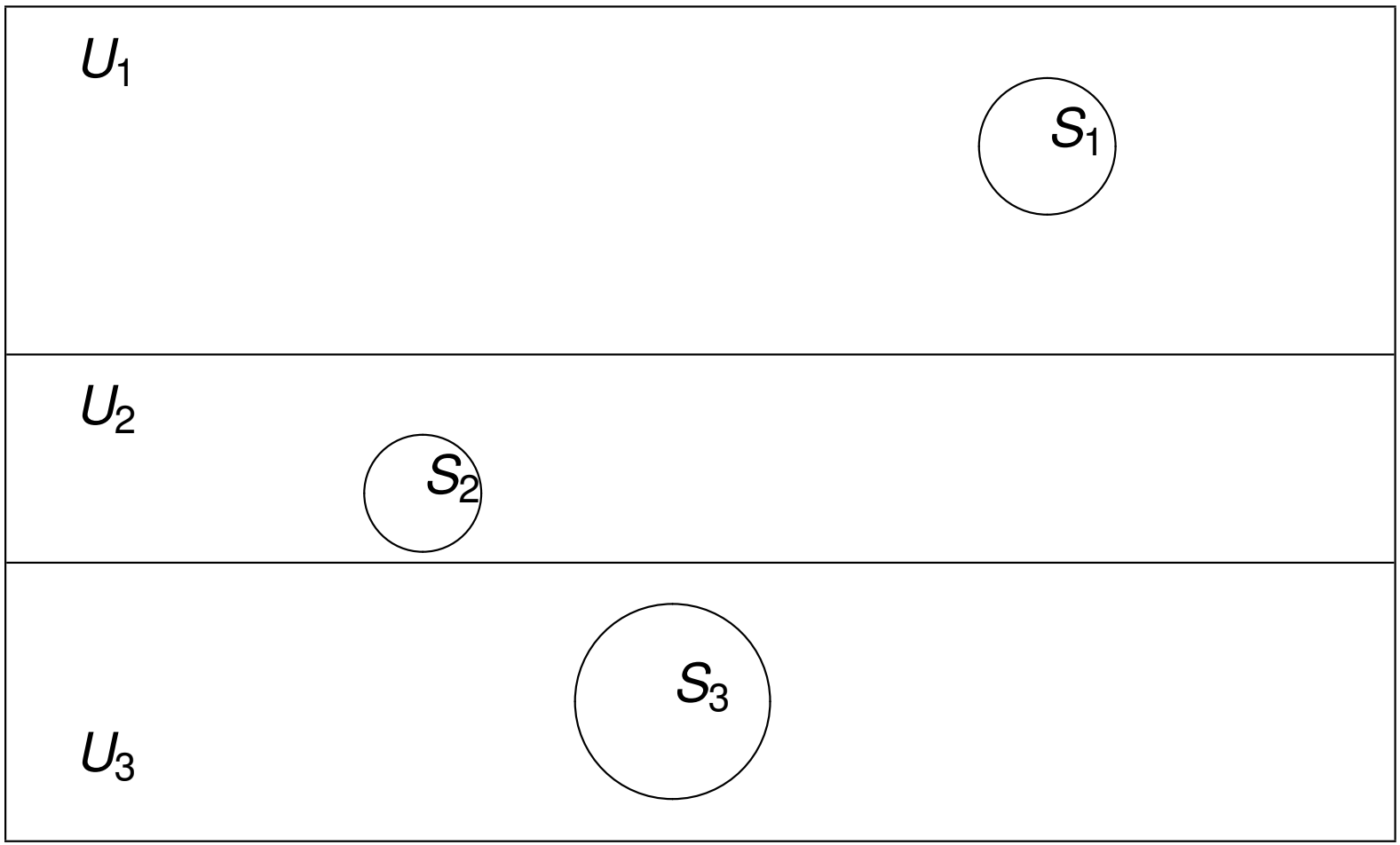
Geschichtete Zufallsstichproben

Die einfache geschichtete Zufallsstichprobe

(Skript Kapitel 3)

- Aufteilung der Population in Unterpopulationen $U_h, (h = 1, ..., L)$ sogenannten Schichten.
- Unabhängige einfache Zufallsstichprobe innerhalb jeder Schicht.
- Nur noch Varianz innerhalb der Schichten zählt. Schichten möglichst homogen.

Schichtung



Bemerkungen:

- Reduktion der Varianz. Varianz zwischen den Schichten schlägt nicht auf Schätzer durch.
- Stichprobengrösse für interessante oder kleine Schichten erhöhen!
- Ungleiche Stichprobenraten f_h möglich: ungleiche Einschlusswahrscheinlichkeiten!
- Verschiedene und komplizierte Stichprobenverfahren in den verschiedenen Schichten.
- Benötigt Information: Strukturierte Population.
- Oft Untersuchungsbereiche als Schichten (z.B. Regionalisierung)

Schätzer

Sei N_h die Grösse der Schicht U_h und $W_h = N_h/N$ ihr Gewicht. Es gilt $\sum_{h=1}^{L} N_h = N$ und damit $\sum_{h=1}^{L} W_h = 1$.
Bei der geschichteten Stichprobe werden die Stichprobengrössen innerhalb der Schichten n_h durch den Stichprobenplan festgelegt.
Stichprobenmittel der Schicht h:

$$\bar{y}_{S_h} = \sum_{i \in S_h} y_i / n_h = \sum_{i=1}^{n_h} y_{hi} / n_h.$$

Das geschichtete Mittel:

$$T_{SS} = \left(\sum_{h=1}^{L} N_h \bar{y}_{S_h} \right) / N = \sum_{h=1}^{L} W_h \bar{y}_{S_h} = \frac{1}{N} \sum_{h=1}^{L} \sum_{i \in S_h} \frac{N_h}{n_h} y_{hi}$$

ist erwartungstreu für das Populationsmittel.

Varianz

Die Varianz des geschichteten Mittels ist

$$V(T_{SS}) = \sum_{h=1}^{L} W_h^2 (1 - n_h/N_h) D_h^2/n_h,$$

wobei

$$D_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_{U_h})^2.$$

Für die Schätzung von $V(T_{SS})$ wird D_h^2 durch

$$d_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_{S_h})^2$$

geschätzt.

Population 5

(Skript Beispiel 4)

- Schichtung: $U_1 = \{1,2,3\}, U_2 = \{4,5\}$
- $y_{U_1} = (1,5,9)^\top$ und $y_{U_2} = (8,12)^\top$.
- Populations-Mittel und Varianz in den Schichten?
- In beiden Schichten einfache Zufallsstichprobe der Grösse $n_1 = 1$ und $n_2 = 1$.
- Zähle die möglichen Stichproben auf und berechne je das geschichtete Mittel.

Aufteilung der Stichprobe auf die Schichten

Gesamt-Grösse der Stichprobe vorgegeben: n.

- **proportionale Aufteilung**

$$n_h = nN_h/N = nW_h.$$

- **optimale Aufteilung** (Neyman-Tschuprow)

$$n_h = n\frac{N_h\tilde{D}_h}{\sum_{j=1}^{L} N_j\tilde{D}_j}$$

($\tilde{D}_h$ eine Abschätzung von D_h). Die optimale Aufteilung liefert minimale Varianz für T_{SS}. Man kann auch die Kosten berücksichtigen!

- **uniforme Aufteilung**: $n_h = n/L$.

Stichprobengrösse

Sei V die gewünschte Varianz von T_{SS}.
Bei proportionaler Aufteilung

$$n_0 = \sum_{h=1}^{L} W_h D_h^2 / V,$$

Mit Endlichkeitskorrektur: $n = n_0/(1 + n_0/N)$.
Bei optimaler Aufteilung

$$n = \frac{(\sum_{h=1}^{L} W_h D_h)^2}{V + \sum_{h=1}^{L} W_h D_h^2/N}$$

Praxis: Mit proportionaler Aufteilung für gegebenes n starten, Varianz schätzen, n_h variieren, Kompromiss suchen.

Die Anzahl Schichten

$L = n$ optimal, aber:

- Varianzschätzung benötigt $n_h \geq 2$.
- Genauigkeitsgewinn flacht ab, wenn immer mehr Schichten gebildet werden. (Theoretisch auch Verlust möglich!)
- Ausfallrate bis zu 50%: Reserve notwendig.
- Hohe Variabilität für Untersuchungsbereiche, die quer zu Schichten liegen.

Bildung der Schichten

- möglichst homogen (Benötigt Proxy für D_h^2)
- oft durch Auswertungsbedürfnisse mitbestimmt (Regionalisierung)
- Verschiedene Variablen würden zu verschiedenen Schichtungen führen: Kompromisse notwendig.
- Antwortausfälle berücksichtigen!

Genauigkeitsvergleich

- Sei y_S die Variable, für die optimiert wurde.

$$V_{\text{SSopt}}(T(y_S)) \leq V_{\text{SSprop}}(T(y_S)) \leq V_{ES}(T(y_S)),$$

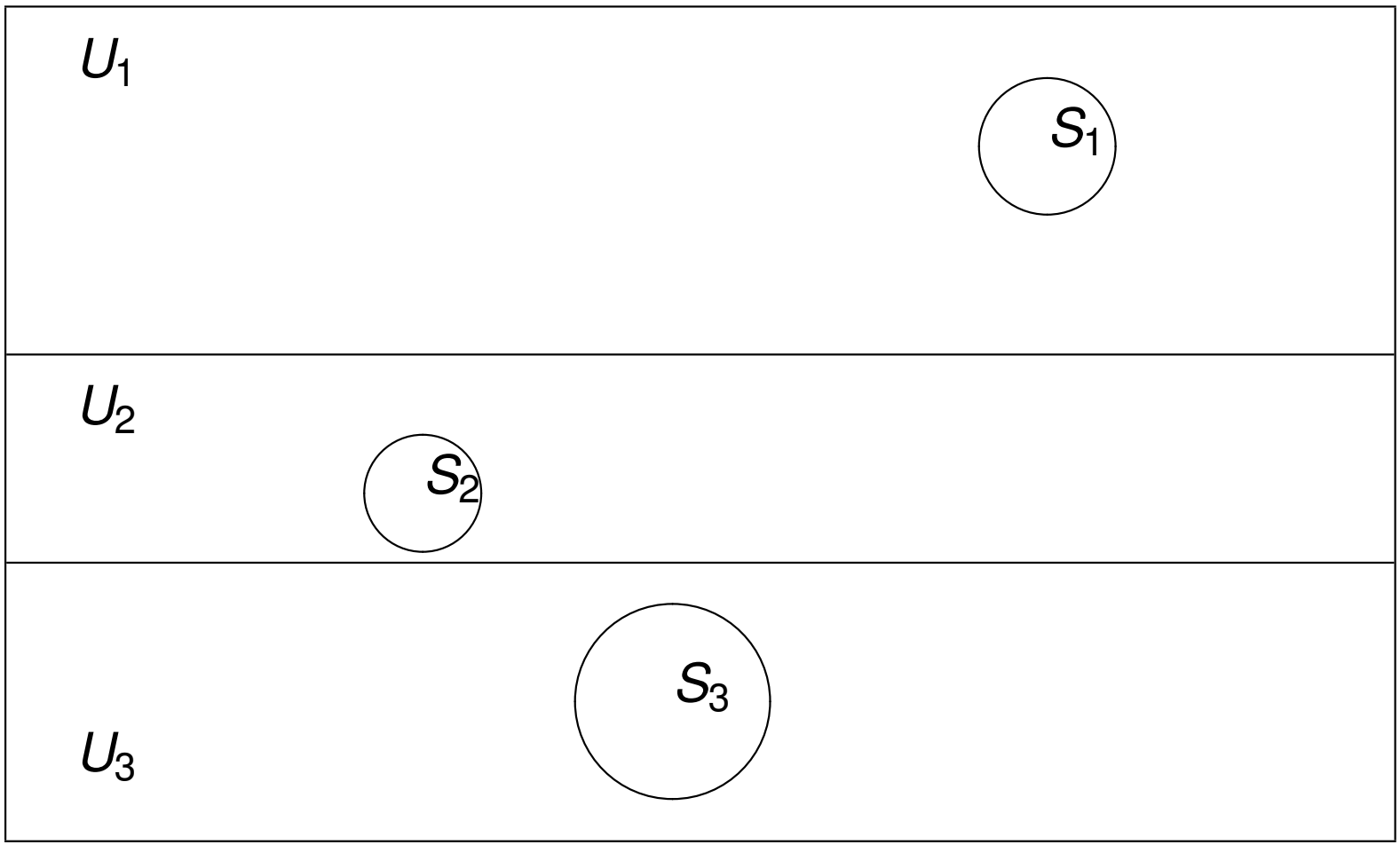
- $V_{\text{SSprop}}(T(x_S)) \leq V_{ES}(T(x_S))$ gilt (fast) immer.
- Bei nicht-proportionaler Aufteilung ist auch möglich $V_{\text{SS}}(T(x_S)) > V_{ES}(T(x_S))$.
- Bei Variablen, für die der Stichprobenplan nicht optimiert wurde, ist auch möglich $V_{\text{SSopt}}(T(x_S)) > V_{\text{SSprop}}(T(x_S))$.

Komplexe Stichprobenpläne

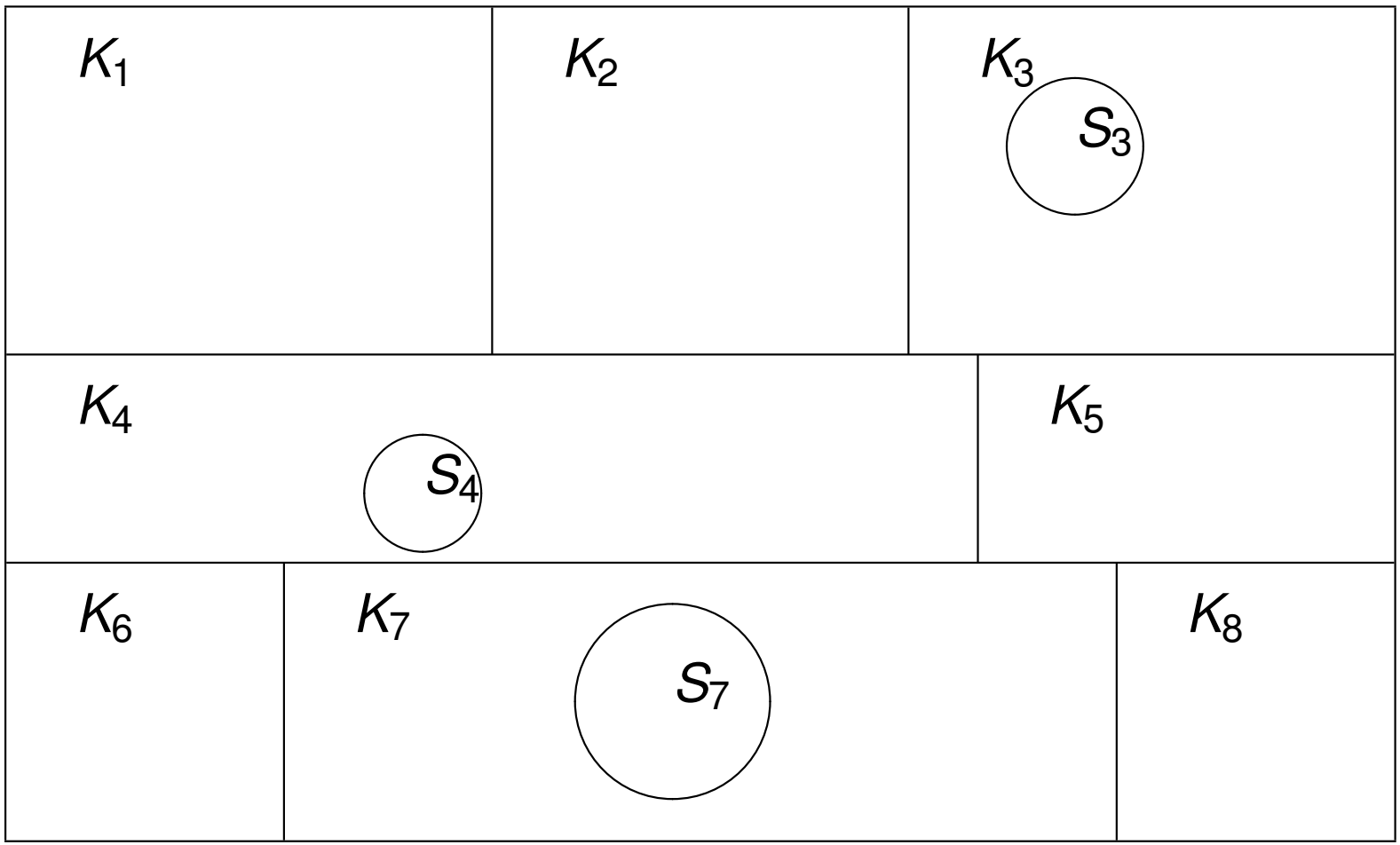
Komplexe Stichprobenpläne

- In der Praxis gibt es selten einen Stichprobenrahmen, der völlig unstrukturiert ist.
- Geschichtete Stichproben sind sehr häufig.
- Neben der Schichtung wird die sogenannte Klumpung als Strukturierungs-Element benützt.
- Horvitz-Thompson-Strategie mit ungleichen Einschlusswahrscheinlichkeiten proportional zu einem Grössenmass x_i sind eher selten.

Schichtung



Klumpung



Klumpung

(Skript Kapitel 5)

- Zufallsstichprobe von Klumpen.
- Innerhalb der Klumpen Vollerhebung (einstufige Klumpenstichprobe) oder Zufallsstichprobe (zweistufige Klumpenstichprobe).

Stichprobenpläne

- Klumpen gleicher Grösse: Oft einfache Zufallsstichprobe
- Klumpen verschiedener Grösse: oft IPPS (Horvitz-Thompson) oder Schichtung nach Grösse.
- Einheiten zweiter Stufe (innerhalb der Klumpen): Zuteilung der Stichprobe ist ein Optimalitätsproblem. Oft: fixe Stichprobe der Grösse *m* pro Klumpen.

Schätzer und Varianzschätzer

- Stichprobenplan erster und zweiter Stufe berücksichtigen!
- Varianz wird aufgebläht, wenn die Elemente eines Klumpens sich ähnlich sind.
- Schätzer für Pop.-mittel ssu bei ES psu

$$T_K = \frac{1}{M}\frac{N}{n}\sum_{i=1}^{n} M_i \bar{y}_i$$

- Varianzschätzer für T_K

$$v(T_K) = \frac{1}{M^2}\left[(1-\frac{n}{N})\frac{N^2}{n}\frac{1}{n-1}\sum_{i=1}^{n}(M_i\bar{y}_i - \frac{M}{N}T_K)^2 + \frac{N}{n}\sum_{i=1}^{n}(1-\frac{m_i}{M_i})\frac{M_i^2}{m_i}d_i^2\right],$$

Vor- und Nachteile der Klumpung

+ Braucht nur Stichprobenrahmen für Einheiten zweiter Stufen innerhalb der gezogenen Klumpen
+ Falls Klumpen=Regionen: Kostenreduktion bei persönlichen Interviews
– Varianz wird grösser als bei ES
– Schätzer werden komplizierter.

Komplexe Stichproben

- Verschachtelung der Methoden Schichtung und Klumpung
- Ungleiche Einschlusswahrscheinlichkeiten (Bei Schichtung und/oder Klumpung).
- Komplexe Varianzschätzungen.

Mehrphasige Stichproben

- Stichprobe $S_1 \subset U$ und Stichprobe $S_2 \subset S_1$.
- Erhebung der Elemente S_1: $x_{1i}, i \in S_1$.
- Stichprobenplan für S_2 unter Ausnützung von x_{1i}.
- Effizienzgewinn
- Screening

Panel

- Stehendes Panel: Erhebung an mehreren Zeitpunkten an der selben Stichprobe.
- Rotierendes Panel: Teilweise Erneuerung der Stichprobe.
- Einschluss als stochastischer Prozess
- Längs- und Querschnittsgewichtung notwendig
- Panel-Auszehrung: Abbau durch Antwortausfälle.

Mehrere Stichprobenrahmen und indirekte Stichproben

- Population nur über verschiedene Stichprobenrahmen zugänglich
- Kombination der Stichprobenpläne für Schätzer
- Indirekte Stichproben: Z.B. Haushalt-Stichprobe mit Hilfe von Personen-Stichprobe
- Netzwerk-Stichproben: Z.B. Kriminalitäts-Studien, HIV

Schätzer mit Modellunterstützung

(Skript Kapitel 4)

Modellunterstützung

- Idee: $x_i, i \in U$ oder wengistens $\bar{x}_U$ bekannt.
- Hilfsinformation x bei Auswertung benutzen.
- Regressionsmodelle: $Y_i = x_i^\top \beta + E_i$
- Kalibrierung: $\sum_{i \in S} w_i x_i = x_{U+}$
- (Balanced sampling: $\bar{x}_S = \bar{x}_U$.)

Nachschichtung

- Schichtgrössen N_h, bzw. $W_h = N_h/N$ bekannt.
- Schichtzugehörigkeit der Elemente erst dank Erhebung.
- Einfache Zufallsstichprobe.
- Schichten sind Untersuchungsbereiche.

Nachgeschichtetes Mittel:

$$T_P = \sum_{h=1}^{L} W_h \bar{y}_{S_h} = \sum_{h=1}^{L} W_h \sum_{i=1}^{n_h} y_{hi}/n_h$$

Unterschied zu T_{SS}: Die n_h sind jetzt Zufallsvariablen.

- T_P setzt sich zusammen aus Schätzern für die Untersuchungsbereiche U_h.
- T_P ist erwartungstreu, wenn die N_h korrekt sind.
- Kalibrierung an bekannte demographische Grössen: Reduktion Nonresponse-Bias.

(Skript Beispiel 10)

Varianz des nachgeschichteten Mittels

$$V(T_P) \approx \sum_{h=1}^{L} W_h^2 (1 - nW_h/N_h) \frac{1}{nW_h} D_h^2 + \sum_{h=1}^{L} W_h^2 (1 - nW_h/N_h)(1 - W_h) \frac{1}{n^2 W_h^2} D_h^2.$$

- Der erste Term der Varianz ist gleich wie bei der geschichteten Stichprobe mit proportionaler Zuteilung!
- U.U. grosser Genauigkeitsgewinn verglichen mit einfacher Zufallsstichprobe
- Varianz-Schätzer:

$$v(T_P) = \sum_{h=1}^{L} W_h^2 (1 - n_h/N_h) \frac{1}{n_h} d_h^2$$

Quotientenschätzer

- Einfache Zufallsstichprobe.
- Populationsmittel von $\bar{x}_U$ bekannt.
- Superpopulations-Modell: $Y_i = \beta x_i + E_i$, $E_M[E_i] = 0$.
- Individuelle x_i nur bekannt dank Stichprobe.

Quotient in der Population:

$$R = \bar{y}_U / \bar{x}_U = y_{U+} / x_{U+}$$

Quotient in der Stichprobe:

$$\hat{R} = \bar{y}_S / \bar{x}_S$$

Falls $V(E_i) \propto x_i$, dann ist $\hat{R}$ der KQ-Schätzer von β.

Quotientenschätzer von $\bar{y}_U$

$$T_R = \bar{x}_U \frac{\bar{y}_S}{\bar{x}_S} = \bar{x}_U \hat{R} = \bar{y}_S \frac{\bar{x}_U}{\bar{x}_S}.$$

Quotientenschätzer ist Mittel der vorhergesagten Werte $\hat{y}_i = \hat{R} x_i$.

(Prediction Approach: Schätzer für Total $\sum_{i=S} y_i + \sum_{i \notin S} \hat{y}_i$ und Stichprobe als ancillary statistics.)

- Untersuchungsbereich: Schätzer $N_{U_B}\bar{y}_{S_B}$ für y_{U_B+} ist ein Quotientenschätzer.
- Verallgemeinerung mit mehreren Hilfsvariablen x: Regressionsschätzer.
- Spezialfall: T_P, das nachgeschichtete Mittel (Dummyvariablen).
- Bei geschichteter Stichprobe kann der Quotientenschätzer separat pro Schicht oder kombiniert über mehrere Schichten angewandt werden.

Linearisierung (Taylor-Approx.):

Bias des Quotientenschätzers:

$$E[T_R - \bar{y}_U] \approx \bar{y}_U \left[\frac{V(\bar{x}_S)}{\bar{x}_U^2} - \frac{\mathrm{Cov}(\bar{y}_S, \bar{x}_S)}{\bar{x}_U \bar{y}_U} \right] = O(1/n).$$

Varianz-Schätzung:
Anstatt $y_i - \bar{y}_S$ treten die Residuen $y_i - \hat{y}_i$ auf.

$$v(T_R) = (1 - n/N) \frac{1}{n} \left(\frac{1}{n-1} \sum_{i \in S} (y_i - \hat{R} x_i)^2 \right).$$

(Skript Beispiel 5)

Horvitz-Thompson Strategie

(siehe Skript Abschnitt 4.3)

- Hilfsvariablen x_i bekannt für ganze Population ($i \in U$).
- Vermutung: interessierende Variable y_i positiv korreliert mit x_i ($y_i = \beta x_i + e_i$ mit $\sum_U e_i = 0$ oder $E_M[E_i] = 0$).
- Einschlusswahrscheinlichkeiten π_i proportional zu x_i (IPPS): $\pi_i = n x_i / \sum_{i \in U} x_i$.

Horvitz-Thompson Schätzer für das Populationsmittel

$$T_{HT} = \frac{1}{N} \sum_{i \in S} \frac{y_i}{\pi_i}$$

Eigenschaften des HT-Schätzers

- Der HT-Schätzer ist erwartungstreu: $\mathrm{E}[T_{HT}] = \bar{y}_U$ (auch wenn y_i nicht positiv mit x_i korreliert!)
- Der HT-Schätzer hat kleine Varianz, wenn das Modell stimmt: Falls $y_i = \beta \cdot x_i$ dann gilt $T_{HT} = \beta \bar{x}_U = \bar{y}_U \quad \forall S$, also Varianz 0.
- Der HT-Schätzer ist ein universeller Schätzer mit "Hochrechnungs-Gewichten" $w_i = 1/\pi_i$ für das Populationstotal

$$T'_{HT} = NT_{HT} = \sum_{i \in S} w_i y_i$$

- Der HT-Schätzer ist der einzige erwartungstreue lineare Schätzer mit Gewichten, die nicht von der Stichprobe abhängen (unter IPPS).
- Grundlage für komplexe Stichprobenpläne mit Schichtungen und Klumpungen.
- Einfache Zufallsstichprobe mit Stichprobenmittel ist HT-Strategie: $\pi_i = n/N$.
- Einfache geschichtete Zufallsstichprobe mit geschichtetem Mittel ist HT-Strategie: $\pi_i = n_h/N_h$.
- Für vorgegebene π_i package `sampling` benützen. Es gibt viele Stichprobenpläne, die vorgegebene π realisieren. Probleme bereiten π_{ij}.

Varianz des Horvitz-Thompson Schätzers

- $\pi_{ij} = P[i \in S \wedge j \in S]$ gemeinsame Einschlusswahrscheinlichkeiten
- $$V(T_{HT}) = \sum_U \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i \neq j, \in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j$$
- Varianz-Schätzer, z.B. Sen-Yates-Grundy:
$$v(T_{HT}) = \frac{1}{2N^2} \sum_{i \neq j, \in S} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 .$$
- Doppelte Einschlusswahrscheinlichkeiten π_{ij} für $i, j \in U$ sind schwierig zu berechnen und meistens unbekannt.

Näherungen

- Hartley-Rao Näherung:

$$v_{HR}(T_{HT}) = \frac{1}{N^2}\frac{1}{2(n-1)}\sum_{i\neq j,\in S}\left(1-\pi_i-\pi_j+\frac{\sum_{i\in U}\pi_i^2}{n}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2.$$

  (Kott 2005)
- $\sum_{i\in U}\pi_i^2/n$ kann mit $\sum_{i\in S}\pi_i/n$ (arithmetisches Mittel der π_i) geschätzt werden.
- Oder Annahme "Mit Zurücklegen":

$$v_{WR}(T_{HT}) = \frac{1}{N^2}\frac{1}{n(n-1)}\sum_{i\in S}\left(\frac{y_i}{\pi_i/n}-T'_{HT}\right)^2$$

Hajek-Schätzer

- Hajek-Schätzer:

$$T_{\text{Hajek}} = \frac{\sum_{i \in S} y_i / \pi_i}{\sum_{i \in S} 1 / \pi_i} = \frac{\sum_S w_i y_i}{\sum_S w_i},$$

  für $w_i = 1/\pi_i$.
- Der Hajek-Schätzer ist ein Quotient von HT-Schätzern.
- Der Hajek-Schätzer ist ein gewichtetes Mittel, dessen Gewichte von S abhängen.
- Einfachere Schätzung für Untersuchungsbereiche:

$$\hat{\bar{y}}_{U_B} = \sum_{i \in S_B} w_i y_i / \sum_{i \in S_B} w_i$$

.

Generalized Regression Estimators (GREG)

- Modell: $Y_i = x_i^\top \beta + E_i$, $E_M[E_i] = 0$ und $V[E_i] = \lambda_i \sigma_E^2$.
- Schätzer für β (KQ mit Gew. λ_i und π_i):

$$\hat{\beta} = (X_S^\top V_S^{-1} \Pi_S^{-1} X_S)^{-1} X_S^\top V_S^{-1} \Pi_S^{-1} y_S,$$

  wobei $V_S = \text{diag}(\lambda_S)$ und $\Pi_S = \text{diag}(\pi_S)$
- GREG:

$$T_{GREG} = T_{HT}(y_S) + (\bar{x}_U - T_{HT}(x_S))^\top \hat{\beta}$$

- Residuen: $e_i = (y_i - x_i^\top \hat{\beta})$
- Alternative Form des GREG:

$$T_{GREG} = \bar{x}_U^\top \hat{\beta} + \sum_{i \in S} \frac{e_i}{N \pi_i}$$

- Wenn λ_i lineare Funktion der x_i, dann ist $\sum_{i \in S} e_i / \pi_i = 0$ (z.B. Quotienten-Schätzer)

GREG

- GREG ist asymptotisch erwartungstreu unabhängig vom Modell.
- GREG-Gewichte g_i in $T_{GREG} = \sum_S g_i y_i / \pi_i$:

$$g_i = 1 + (\bar{x}_U - T_{HT}(x_S))^\top (X_S^\top V_S^{-1} \Pi_S^{-1} X_S)^{-1} X_S^\top V_S^{-1} \Pi_S^{-1})$$

- g_i sind unabhängig von y_i: universelle Gewichtung.
- Varianz-Schätzer basiert auf Residuen e_i: Varianz Horvitz-Thompson Schätzer.

Kalibrierung

- Gegeben: d_i Stichproben-Gewicht (z.B. $d_i = 1/\pi_i$)
- Annahme $\hat{T}_{yd} = \sum_{i \in S} d_i y_i$ ein vorläufiger Schätzer des Totals
- Bekannt: Vektor $\mathbf{x}_{U+}$ von Populationstotalen der Hilfsvariablen.
- Gesucht: Gewichte w_i, nahe bei d_i, so dass

$$\hat{T}_{\mathbf{x}w} = \sum_{i \in S} w_i \mathbf{x}_i = \mathbf{x}_{U+}$$

Distanzfunktion

”nahe”: Distanz $G(w_i, d_i)$ wird unter Nebenbedingungen minimiert.

Kleinste Quadrate $G(w_i, d_i) = (w_i - d_i)^2/d_i$.
Lösung: GREG.

log-ratio: $G(w_i, d_i) = w_i \log(w_i/d_i) - w_i + d_i$
Lösung: iterative proportional fitting (raking)

Datenaufbereitung

(Skript Kapitel 6)

Datenaufbereitung

- Datenaufbereitung:
  - Kodierung
  - Kontrollen (Diagnostics)
  - Einsetzungen
- Kontrollen und Einsetzungen werden oft als “Plausibilisierung"bezeichnet.
- Englisch “Editing and Imputation"

Beispiel

Alter Jahre	Zivilstand 0:ledig,1:verh.	Gewicht kg	Grösse cm
241			165
10	1	30	120
43	0	89	105
	1	3	151

Kontrollen

Kontrollen dienen zum

- Fehlende Werte finden (und von strukturell fehlenden Werten zu unterscheiden)
- Fehlerhafte Beobachtungen finden
- Fehler lokalisieren, d.h. auf Variable einengen
- Beurteilen der Datenqualität ($\Rightarrow$ Gegenmassnahmen)

Untersuchung der Antwort-Ausfälle

Eingangskontrolle, Erfassung und Vollständigkeitskontrolle

- Gesamtausfall (unit-nonresponse)
  - Vergleich mit Stichprobenrahmen
  - Vergleich mit bekannten Populationsgrössen (sozio-demographisch etc.)
- Merkmals-Ausfall (item-nonresponse)
  - Response bzw. missingness patterns
  - Erklärung der Ausfälle mit Hilfe der anderen, beobachteten Variablen

Ausfallmechanismen

- Wenn die fehlenden Werte völlig zufällig sind (missing completely at random, MCAR): Problemlos, kann ignoriert werden.
- Wenn der Ausfallmechanismus nur von beobachteten Werten abhängt (missing at random. MAR): Mit Hilfe von Modellen kann der Effekt der Ausfälle kompensiert werden bzw. Einsetzungen können zumindest theoretisch die fehlenden Werte ersetzen.
- Wenn der Ausfallmechanismus von fehlenden Werten abhängt (Non-MAR): keine Möglichkeit Ausfälle zu kompensieren (Bias!).

Mikro-Kontrollen

- Eindimensionale Kontrolle einzelner Fragebogen: Erlaubte Art (alphanumerisch usw.), möglicher Wertebereich, Warngrenzen
- Mehrdimensionale Kontrolle einzelner Fragebogen: Widersprüche (verwittwete Jugendliche), Personalkosten höher als Gesamtkosten.

```
If Alter < 15 and Zivilstand = 1 then
RegelAlterZiv=0 else RegelAlterZiv=1
```

- Kontrollen mit Hilfe externer Information: Z.B. bei Panel Vergleich mit letztem gemeldeten Wert (Erhöhung der Miete um mehr als 20%).

Makro-Kontrollen

- Eindimensionale Kontrolle der Stichprobe: Vergleich Verteilung mit letzter bekannter Verteilung, eindimensionale Ausreisser
- Mehrdimensionale Kontrolle der Stichprobe: Cluster, mehrdimensionale Ausreisser.
- Vergleich Resultate mit bekannten Eckwerten und ähnlichen Statistiken, Diskussion mit Fachwissenschaftern.

Einsetzungen

- Rückfragen und "korrekten"Wert einsetzen (Korrektur)
- Prioritäten festlegen, deterministische "Korrekturen".
- Ausreisser behandeln
- Fehlerhafte Werte ersetzten.
- Einsetzungen: Schätzungen anstatt fehlender Werte oder Ausreisser einsetzen. (z.B. Imputation mit ”Hot Deck”)
- Einsetzung basiert auf Modellen und Annahmen, die z.T. nicht verifiziert werden können.

Datenaufbereitung ist schwierig!

- Kontrollen und Einsetzungen bilden ein logisches System: Es muss in sich konsistent sein und möglichst klein (Fellegi-Holt).
- Man kann durch Einsetzungen Bias erzeugen. Manchmal haben die Einsetzungen keinen Einfluss auf die Schätzer, waren also unnötig!
- Varianzschätzungen werden durch Einsetzungen verfälscht: Es sind spezielle Methoden nötig.
- EDIMBUS Manual.
- Der Aufwand ist u. U. sehr gross.

Auswertung

- Datenaufbereitung (Kontrollen und Einsetzungen)
- Stichprobenschätzer mit Gewichten entsprechend Stichprobenplan (d_i).
- Antwortausfälle (Unit-Nonresponse)
- Kalibrierung (Gewichte w_i)
- Ausreisser, robuste Schätzverfahren (Gewichte u_i).
- Varianzschätzung
- Analytische Statistik

Ausreisser, robuste Schätzverfahren

(Skript Kapitel 7)

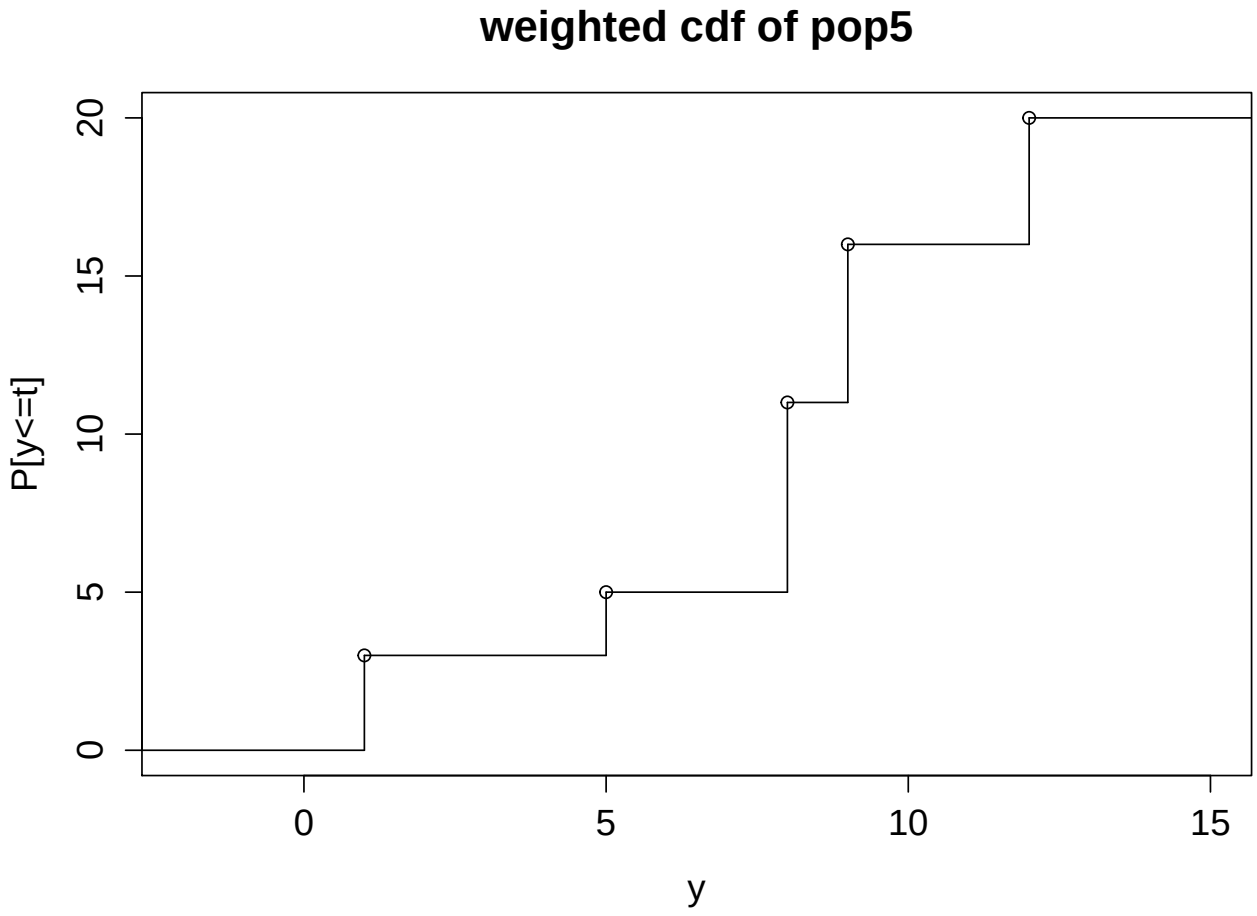
- Kein parametrisches Modell: Ausreisser ist weit weg vom Grossteil der Daten.
- Repräsentative und Nicht-repräsentative Ausreisser (Chambers 1986).
- Robuste Schätzung oder Ausreisser-Entdeckung und Einsetzung.

Gewichteter Median

- Populationsmedian $\text{med}(y_U)$: Population nach y sortieren und in untere und obere Hälfte teilen; der Median ist die Grenze.
- Geordnete Stichprobe $0 \leq y_1 \leq y_2 \cdots \leq y_n$ mit Gewichten $w_1, \ldots, w_n$. Linearer Schätzer für Populationsmittel: $T = \sum_{i=1}^{n} w_i y_i \,/ \sum_{i=1}^{n} w_i$.
- $i_1 = \min\{i : \sum_{j=1}^{i} w_j \Big/ \sum_{j=1}^{n} w_j \geq 0.5\}$ und

  $i_2 = \min\{i : \sum_{j=1}^{i} w_j \Big/ \sum_{j=1}^{n} w_j > 0.5\}$
- Gewichteter Median

$$\text{med}(y_S) = \frac{w_{i_1} y_{i_1} + w_{i_2} y_{i_2}}{w_{i_1} + w_{i_2}}$$


weighted cdf of pop5
P[y<=t]
0
5
10
15
20
0
5
10
15
y

- Vertrauensintervall mit Inversion der Verteilungsfunktion (Woodruff)
- Median ist sehr robust gegen Ausreisser und z.B. für Einkommensdaten interessant.
- Analog können gewichtete Quantile berechnet werden. Auch winsorisierte Mittel und M-Schätzer können adaptiert werden.

Winsorisiertes Mittel

Geordnete Stichprobe $0 \leq y_1 \leq y_2 \cdots \leq y_n$ mit Gewichten $w_1, \ldots, w_n$.

$$T = \sum_{i=1}^{n} w_i y_i.$$

Zu $\alpha \in [0, 0.5]$ suche Index

$$i_u = \min\{i : \sum_{j=1}^{i} w_j \Big/ \sum_{j=1}^{n} w_j \geq \alpha/2\} \text{ und analog } i_o.$$

Winsorisierung: $y_i = \begin{cases} y_i, & i_u \leq i \leq i_o; \\ y_{i_u}, & i < i_u; \\ y_{i_o}, & i > i_o \ . \end{cases}$

Robuste Schätzer für Stichprobenerhebungen

- Robustifizierter Horvitz-Thompson Schätzer (Hulliger 1995)
- Robuster Quotienten-Schätzer
- M-Schätzer und Einschritt-M-Schätzer
- Multivariate Ausreisser-Entdeckung und Imputation (`modi`)
- Wahl der Abstimmung-Konstanten: Bias ist viel wichtiger als bei klassischer Statistik.

Antwortausfälle (unit-nonresponse) und Kalibrierung

(Skript Kapitel 8)

- Haushalte: 20 bis 50% Nonresponse (totale Antwortausfälle, unit nonresponse).
- Unternehmen: 40 bis 80 % Antwortausfälle.
- Gründe: Mangelhafter Stichprobenrahmen, Erreichbarkeit, Invalidität und Verweigerung.
- Beispiel: Drogengebrauch.
- Antwortausfälle sind meistens mit dem Untersuchungsgegenstand korreliert und führen daher zu einem Bias (MAR oder non-MAR, aber nicht MCAR)

Bias durch Antwortausfälle

Bei einfacher Zufallsstichprobe und festem Antwortverhalten:
U_r Untersuchungsbereich der Antwortenden
U_n Untersuchungsbereich der Nicht-Antwortenden.

$$E[\bar{y}_{S_r} - \bar{y}_U] = \bar{y}_{U_r} - (N_r y_{U_r} + (N - N_r) y_{U_n})/N$$
$$= (1 - N_r/N)(\bar{y}_{U_r} - \bar{y}_{U_n})$$

Varianzerhöhung durch Antwortausfälle

- Reduktion der Stichprobengrösse: Antwortrate $r \in [0,1]$ führt zu Nettostichprobengrösse $n_r = n_b * r$.
- Wird meist in der Bruttostichprobengrösse vorgesorgt.
- Eventuell Reservestichproben vorsehen

Vermeidung von Antwortausfällen

- Gute und kurze Fragebogen
- Gute Information der Befragten
- Mahnaktionen (telefonisches Nachhaken).

Auswertung bei Antwortausfällen

- Modell der „Antwortschicht": Schicht von Leuten, die potentiell antworten. Erhebung liefert im strikten Sinn nur Aussagen über diese.
- Bei $y_i \in \{0, 1\}$ kann man wenigstens Grenzen angeben.
- Mehr Information über Ausfälle: Nonresponse-Studie
- Korrekturen bei der Schätzung: Modelle (Nachschichtung, Kalibrierung, Quotientenschätzer, Regressionsschätzer, Antwortneigung).

Nonresponse-Studie und Antwortneigung

- Kleine Stichprobe aus Antwortausfällen
- Logistische Regression für Antwortneigung $p_{r,i}$ (propensity scores) mit erklärenden Variablen, die die Nonresponse differenzieren.
- $T = \sum_S w_i y_i$ ersetzen durch

$$T' = \sum_s w_i y_i / p_{r,i} \frac{\sum_S w_i}{\sum_S w_i / p_{r,i}}.$$

  (Horvitz-Thompson)
- Bei item-nonresponse oft propensity scores matching: Einsetzen innerhalb Klassen, die durch $\hat{p}_{r,i}$ definiert werden.

Nachschichtung

- Idee: Homogene Nachschichten bezüglich den Antwortausfällen, zwischen den Schichten Unterschiede im Antwortverhalten und in interessierenden Variablen.
- Das nachgeschichtete Mittel T_P hat im Normalfall kleineren Bias als das Stichprobenmittel der Antwortenden, aber die Variabilität der (Nachschichtungs-) Gewichte geht u.U. in den Schätzer ein!
- Faustregel: Varianzerhöhung um bis zu $(1 + cv(w_i)^2)$.

(Beispiel 10 im Skript)

Kalibrierung

- Kalibrierung ist eine Verallgemeinerung der Nachschichtung mit mehreren Variablen.
- Beispiel: Nachschichtung nach Alter, Geschlecht, Zivilstand und Nationalität wünschbar
- Besetzung der gekreuzten Zellen in der Stichprobe zu klein ($n_h < 20$).
- Genaue Grösse der Zelle in der Population unbekannt oder nicht genau bekannt.
- Man kann nur auf Randsummen (Haupteffekte) oder nur auf gewisse Kreuzungen kalibrieren.

Inferenz

(Skript Kapitel 9)

- Vertrauensintervalle
- Hypothesentests (inbesondere χ^2-test)
- Regressionsmodelle
- Multivariate Analyse

Varianzschätzung

- Bei komplexen Stichproben und schwierigen Gewichtungen ist die Schätzung der Varianz u. U. sehr schwierig.
- Die üblichen Varianzschätzungen, welche Statistikprogrammpakete liefern, sind falsch.
- In SAS die Survey-Prozeduren, in SPSS die Complex Sampling Funktionen und in R das package `survey` berechnen für viele Fälle gültige Varianzen.

Methoden für Varianzschätzungen

- Herleitung einer direkten, geschlossene Formel
- Linearisierung: Varianzapproximation mit Hilfe von Taylor
- “resampling” - Verfahren , z.B. **Jackknife**

Vertrauensintervalle

- $T \quad \pm \quad 2 \quad v(T)^{1/2}$ schliesst mit 95% Wahrscheinlichkeit die von T geschätzte Populationscharakteristik ein. (Annahme: Normalverteilung ... , t_{60})
- Vergleich von Untersuchungsbereichen: Ueberlappen sich die Vertrauensintervalle eines Schätzers für zwei Untersuchungsbereiche nicht, dann ist die Differenz der geschätzten Charakteristiken signifikant.
- Achtung: Vertrauensintervalle sind schwierig zu erklären. Standardabweichung ist einfacher.

Tests und Modelle bei komplexen Stichproben

- Die üblichen χ^2-Tests angewandt auf komplexe Stichproben ergeben oft falsche Signifikanzen.
- Konsistente Schätzer (mit HT-Schätzer/Gewichtung)
- Korrektur für Design-Effekt (1. und 2. Ordnung).

Regression

- Die üblichen Tests in Regressionsmodellen geben falsche Signifikanzen bei Daten aus komplexen Stichproben.
- Wenn in Regressionsmodellen die Untersuchungsvariablen mit dem Stichprobenplan korrelieren, ergibt sich eine Auswahl-Effekt (Selection-Bias) ergeben. Test mit und ohne Gewichtung.
- Variablen des Stichprobenplans in Modell aufnehmen kann Bias reduzieren, ist aber oft nicht sinnvoll.

Multivariate Statistik

- Kovarianzen zweier Variablen X und Y, $C[X, Y]$, schätzen mit

$$c(X, Y) = \sum_{i \in S} w_i' \left(x_i - \sum_{i \in S} w_i' x_i \right) \cdot \left(y_i - \sum_{i \in S} w_i' y_i \right),$$

  wobei $w_i' = w_i / \sum_{i \in S} w_i$, so dass $\sum_{i \in S} w_i' = 1$.
- $c(X, Y)$ kann in Faktor-Analyse etc. verwendet werden.
- Problem: Ausreisser und fehlende Werte (siehe `modi`).