10. Stichproben

- 10.1 Allgemeines
  - Kriterien an eine Stichprobe
  - Schätzer
- 10.2 Zufallsauswahl
- 10.3 Quotenauswahl und geschichtete Stichprobe
- 10.4 Entzerrung

Personen

Haushalte

10.1 Allgemeines

Während die automatischen Messmethoden für Querschnittsdaten in der Regel eine **Vollerhebung** der Grundgesamtheit ermöglichen, sind die nur in Form von Befragungen erhebbaren Daten zum Mobilitätsverhalten (Verkehrsflusserhebung) lediglich als **Stichprobenerhebung** durchführbar.

Was ist eine Stichprobe?

Eine **Stichprobe** ist eine Teilmenge der **Grundgesamtheit**, also Teilmenge einer zeitlich, räumlich und sachlich eingegrenzten Menge von statistischen Einheiten

Kriterien an eine Stichprobe 1: Repräsentativität

Eine Stichprobe ist bezüglich des Merkmals Y **repräsentativ**, falls die Verteilungsfunktion dieses Merkmals innerhalb der Stichprobe im Mittel (bei vielen Ziehungen der Stichprobe) dieselbe ist wie in der Grundgesamtheit (GG).

Insbesondere gilt dann für Erwartungswert und Varianz (Y_i bezeichnet das i-te Element der Stichprobe bzw. der GG Zufallsvariable)

$$\begin{aligned} E(Y_i)_{\text{Stichpr}} = E(Y_i)_{\text{GG}} = E(Y) &:= \mu, \\ V(Y_i)_{\text{Stichpr}} = V(Y_i)_{\text{GG}} = V(Y) &:= \sigma^2. \end{aligned}$$

Daraus folgt insbesondere, dass das arithmetische Mittel ein erwartungstreuer *Schätzer* von μ ist:

$$E(\bar{X}) = \frac{1}{n} \sum_i E(X_i) = \frac{1}{n} \sum_i \mu = \mu.$$

Erwartungstreuer Schätzer der Varianz:

$$\hat{V}(X) = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

Kriterien an eine Stichprobe 2: Effizienz

Eine Stichprobe ist bezüglich des Merkmals Y **effizient**, wenn die Varianz des zugehörigen effizienten Schätzers $\hat{\mu}$ bei gegebenen Stichprobenumfang minimal ist.

- Der effiziente Schätzer ist dabei der erwartungstreue Schätzer mit der geringstmöglichen Varianz, z.B. das arithmetische Mittel für Schätzung von Erwartungswerten
- Für eine bestmögliche Ausnutzung der Stichprobe benötigt man eine effiziente Stichprobe *und* einen zugehörigen effizienten Schätzer
- Die Eigenschaften der Repräsentativität und Effizienz hängen vor allem ab von der Ziehungsgrundlage und Ziehungsmethode

Ziehungsgrundlage und Ziehungsmethode

Ziehungsgrundlage:

- Direkt die Grundgesamtheit
- eine Obermenge
  Beispiel: Ziehungsgrundlage Personenregister, aber sachliche Eingrenzung der GG auf Motorradfahrer
- eine abgeleitete Menge, aus der man die GG vollständig erschließen kann
  Motorradbeispiel: Zulassungsregister für Kfz oder Motorräder

Verwendet man jedoch nicht direkt die Grundgesamtheit als Ziehungsgrundlage, erhält man ohne zusätzliche Maßnahmen in der Regel nichtrepräsentative Stichproben, die man durch spezielle **Entzerrungs-Methoden** behandeln muss.

Ziehungsmethode:

- Zufallsauswahl, ggf mit Entzerrung
- geschichtete Auswahl bzw. Quotenauswahl
- effiziente Auswahl

Beispiel: Zwei große Mobilitätserhebungen

Erhebung	**Mobilität in Deutschland (MiD)**	**System relevanter Verkehrsbefragungen (SrV)**
Statistische Einheit bzw. Merkmalsträger	(i) Personen, (ii) Haushalte	wie MiD
Grundgesamtheit (GG)	zeitlich: 2002 räumlich: Deutschland sachlich: Inländer $\geq$ 14 J	zeitlich: 2003, je Di-Do räumlich: 34 Städte sachlich: alle Inländer
Stichprobe (räumlich/ sachlich wie GG)	7.12.2001 - 22.12.2002	April - Juni 2003
Auswahl	geschichtet	Zufallsauswahl mit Entzerrung
Stichprobenumfang	26 000 Haushalte 62 000 Personen	13 500 Haushalte (3 000 in DD) 33 000 Personen (6 000 in DD)
Erhebungsbasis (Ziehungsgrundlage)	Einwohnermelderegister	wie MiD

Schätzer

Ein **Schätzer** $\hat{Z} = f(\{Y_i\})$ einer Eigenschaft z der Grundgesamtheit ist eine Funktion der Merkmalswerte Y_i der Stichprobe. In der Regel beschreibt z eine globale Eigenschaft der Grundgesamtheit, die man gerne wissen möchte

Beispiele:

- das Stichprobenmittel $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ als Schätzer des Erwartungswertes $\mu = E(Y)$ eines Merkmals der Grundgesamtheit,
- die Stichprobenvarianz $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ als Schätzer der wahren Varianz $\sigma^2 = V(Y) = E(Y - \mu)^2$.

Merke:

- Schätzer werden üblicherweise mit demselben Symbol wie die zu schätzende Größe bezeichnet, aber mit "Dach"
- Als Funktion von Zufallsvariablen sind die Schätzer, im Gegensatz zu den zu schätzenden Größen der GG, *selbst* stochastische Größen

Kriterien an Schätzer

Ein Schätzer $\hat{Z}$ für die Größe z ist **erwartungstreu**, falls

$$E(\hat{Z}) = z$$

Ein Schätzer $\hat{Z}$ für die Größe z ist **effizient**, falls er erwartungstreu ist und falls, bei gegebener Stichprobe (insbesondere festen Stichprobenumfang) gilt,

$$V(\hat{Z}) = E((\hat{Z} - z)^2) \stackrel{!}{=} \min \tag{1}$$

bezüglich aller anderen möglichen erwartungstreuen Schätzer.

- Diese Kriterien stellen das Gegenstück der Repräsentativität und Effizienz von Stichproben dar
- Im Allgemeinen hängt der effiziente, also optimale Schätzer von der Art der Stichprobe ab. Ggf. auch von den Stichprobenwerten von Strukturmerkmalen wie Alter und Geschlecht, wenn deren Verteilung a priori bekannt ist

Beispiele für den Schätzer des wahren Erwartungswertes μ

- Wird die Stichprobe per Zufallsauswahl direkt aus der Grundgesamtheit gezogen und sind keine weiteren Informationen über die Struktur der Grundgesamtheit verfügbar (z.B. Altersstruktur), ist das arithmetische Mittel $\bar{Y}$ ein effizienter Schätzer für μ (Abschnitt 10.2)
- Dasselbe gilt, wenn die Stichprobe per Proportionalauswahl (Quotenverfahren) aus der Grundgesamtheit oder aus einer die Grundgesamtheit umfassenden Ziehungsgrundlage (Register) gezogen wird (Abschnitt 10.3)
- Wird die Stichprobe per Zufallsauswahl gezogen und sind weitere Strukturinformationen wie die Alters- und Geschlechtsstruktur der Grundgesamtheit bekannt, so ist der (nahezu) effiziente und ggf entzerrende Mittelwertschätzer ein gewichtetes arithmetisches Mittel $\hat{\mu} = \frac{1}{n} \sum_i E_i Y_i$. (Abschnitt 10.4)

10.2 Zufallsauswahl

Die N Elemente der Ziehungsgrundlage (ZG) werden willkürlich von 1 bis N nummeriert und dann die n Elemente der Stichprobe "blind" gezogen. Vorgehen, falls die GG (z.B. alle volljährigen Dresdner") vollständig in der ZG (Personenregister Dresdens, $N = 554\ 649$) enthalten ist:

- Indizierung der Ziehungsgrundlage, z.B. alphabetisch: Person 1 = Aae Lennert, ..., Person 554 649 = Zywitzski Olaf
- Ziehung einer (0,1)-gleichverteilten (Pseudo-) Zufallszahl $Z \sim G(0,1)$. Bei einer Ziehung erhält man z.B. die Realisierung z=0.9334678
- Ziehung der Person mit dem Index $i = \text{round}(zN + 0.5) = 517\,747$ aus der ZG
- Ist die Person in der GG und wurde nicht schon einmal gezogen, füge sie zur Stichprobe hinzu
- Fahre fort, bis der Stichprobenumfang n erreicht ist

Eine solcherart gezogene Stichprobe ist *repräsentativ*

Stichprobenfehler bei Zufallsauswahl

Da die Zufallsauswahl repräsentativ ist, ist das arithmetische Mittel der effiziente Schätzer des Erwartungswertes $E(Y_i) = \mu$ eines Merkmals Y_i der statistischen Einheit (Person, Stichprobenelement) i mit Varianz $V(Y_i) = \sigma^2$

- Schätzer:

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- Erwartungswert:

$$E(\hat{\mu}) = E(\bar{Y}) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu.$$

- Varianz:

$$V(\hat{\mu}) = V(\bar{Y}) = V\left(\frac{1}{n} \sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} V(Y_i) = \frac{\sigma^2}{n}$$

Das arithmetische Mittel ist ein effizienter Schätzer für den wahren Mittelwert μ, wenn die Stichprobe vom Umfang $n < N$ repräsentativ ist und keine a priori bekannten Strukturmerkmale wie Alter und Geschlecht verfügbar sind. Die Formel $V(\hat{Y}) = \sigma^2/n$ gilt jedoch nur, falls zusätzlich $n \ll N$. **(Warum?)**

Statistischer Hinweis zur Schreibweise

- Unterscheide zwischen **Zufallsvariablen** wie Y_i, $\bar{Y}$, $\hat{\mu}$ oder $\hat{\sigma}^2$ und den bei einer Ziehung dieser Zufallsvariablen sich ergebenden **Realisierungen** wie y_i und $\bar{y}$ oder den wahren Werten wie μ und σ^2, welche reine Zahlenwerte darstellen

- Nur auf Zufallsgrößen können Operatoren wie $E(.)$ oder $V(.)$ (sinnvoll) angewandt werden

Was ist der Erwartungswert und die Varianz eines festen Zahlenwertes?

der Zahlenwert bzw. Null

Anteilswerte

Wir definieren die **bernoulliverteilte** Dummyvariable

$$Y_i = \begin{cases} 1 & \text{“Ja”-Entscheidung} \\ 0 & \text{sonst} \end{cases}, \quad P(Y_i = 1) = \mu$$

Ist außerdem $n \ll N$, so sind die Y_i unabhängig und damit ist eine Summe aus n solcher Dummyvariablen (vgl. Statistik II-Vorlesung) binomialverteilt mit n Freiheitsgraden:

$$\sum_{i=1}^{n} Y_i \sim B(n, \mu), \quad E(Y_i) = \mu, \quad V(Y_i) = \sigma^2 = \mu(1-\mu) \tag{2}$$

mit

$$E\left(\sum_{i=1}^{n} Y_i\right) = n\mu, \quad V\left(\sum_{i=1}^{n} Y_i\right) = n\mu(1-\mu)$$

Zentraler Grenzwertsatz (ZGWS) für $n\mu(1-\mu) \geq 9$: Die Summe ist angenähert normalverteilt mit selben Erwartungswert und Varianz. Damit gilt für den Schätzer $\hat{\mu}$ eines Anteilswertes **(Warum?)**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N\left(\mu, \frac{\mu(1-\mu)}{n}\right)$$

Aufgabe

Wie groß muss man bei einer Mobilitätserhebung zur Frage "Benutzen Sie bei Arbeitswegen bevorzugt den ÖV" und eines erwarteten Anteils in der Größenordnung 50% die Stichprobe mindestens anlegen, damit der wahre Anteilswert mit einer Unschärfe von höchstens einem Prozentpunkt bestimmt werden kann? (Unterscheide Prozentpunkt von Prozent!) Definieren Sie dabei die Unschärfe

(i) als eine Standardabweichung,

(ii) als die halbe Breite des Konfidenzintervalls zur Fehlerwahrscheinlichkeit 5%.

Lösung

(i) Der zulässige Fehler $r = 1\% = 0.01$ wird gleich der Standardabweichung des Schätzers $\hat{\mu} = \bar{Y}$ gesetzt:

$$V(\hat{\mu}) = \frac{\mu(1-\mu)}{n} \stackrel{!}{=} r^2 \quad \Rightarrow \quad n = \frac{\mu(1-\mu)}{r^2} = \underline{\underline{2\,500}}.$$

(ii) ZGWS für $n\mu(1-\mu) \geq 9$:

$$Z = \frac{\hat{\mu} - \mu}{\sqrt{V(\hat{\mu})}} = \sqrt{n}\frac{\hat{\mu} - \mu}{\sqrt{\mu(1-\mu)}} \sim N(0,1)$$

Teilt man die Fehlerwahrscheinlichkeit symmetrisch auf (symmetrisches Konfidenzintervall), also je 2.5% links und rechts auf, ergibt sich für die obere Grenze in Z das tabellierte 97.5%-Quantil der Standardnormalverteilung, also $Z = z_{0.975} = 1.96$. In den Anteilswerten soll die obere Grenze der Abweichung $\hat{\mu} - \mu$ gleich $r = 0.01$ sein:

$$(\hat{\mu} - \mu)_{0.975} = z_{0.975}\sqrt{\frac{\mu(1-\mu)}{n}} \stackrel{!}{=} r = 0.01$$

bzw.

$$n = \frac{z_{0.975}^2\,\mu(1-\mu)}{r^2} = \underline{\underline{9\,600}}$$

10.3 Quotenauswahl und geschichtete Stichprobe

Bei einer durch **Quotenauswahl** bzw. **Proportionalauswahl** resultierenden **geschichteten Stichprobe** sind die Anteilswerte der Ausprägungen von bestimmten **Quotenmerkmalen** samt ihrer Korrelation dieselben wie in der Grundgesamtheit.

- Diese höhere Informationsgehalt schafft i.A. trennschärfere Stichproben.
- Die Quotenmerkmale sollten folgende Bedingungen erfüllen:
  - (i) Die Anteile der verschiedenen Ausprägungen der Quotenmerkmale sind in der Grundgesamtheit a priori (von vorneherein) bekannt.
  - (ii) Sie beeinflussen die zu untersuchenden Merkmale, indem sie z.B. deren Erwartungswerte und Varianzen ändern.
  - (iii) Die Anteile der verschiedenen Schichten (Klassen der Quotenmerkmale) in der Grundgesamtheit sollten nicht zu unterschiedlich sein.
- "Klassische" Quotenmerkmale sind insbesondere
  - das Geschlecht (binäres Merkmal),
  - das Alter (meist in mehrere Altersklassen eingeteilt),
  - die Größe des Haushalts (z.B.1, 2, 3, 4, 5 und mehr).

Quotenauswahl II

Eine Quotenauswahl ist vor allem bei einer Analyse mit einfachen Modellen vorteilhaft. In denr multiplen Regression und der diskreten Wahltheorie sind die wichtigsten Quotenmerkmale in der Regel als erklärende Variablen enthalten und eine Quotenauswahl meist entbehrlich.

Sind folgende Merkmale i.A. als Quotenmerkmale geeignet? Wenn nein, warum nicht?

- Anfangsbuchstabe des Nachnamens nein, da zwar in GG bekannt aber keinen Einfluss auf Y
- beruflicher Status (Student,erwerbstätig,arbeitslos etc) ja, da in GG bekannt und Einfluss auf Mobilitätsverhalten
- Geburtsmonat nein, da zwar in GG bekannt aber keinen Einfluss auf Y
- Wohnort ja, bekannt und Einfluss
- Kfz-Verfügbarkeit, Fahrrad-Verfügbarkeit nein, da zwar hochrelevant, aber in GG nicht bekannt
- Kfz-Besitz ja, da korreliert mit Verfügbarkeit und in GG bekannt
- Besitz einer ÖPNV-Dauerkarte nein, da in keinem Register

Ziehung einer geschichtete Stichprobe (Beispiel)

Alter \ Geschlecht	♂	♀
0-20 J	$\vartheta_{11} = 16\ \%$	$\vartheta_{12} = 15\ \%$
20-50 J	$\vartheta_{21} = 18\ \%$	$\vartheta_{22} = 15\ \%$
> 50 J	$\vartheta_{31} = 20\ \%$	$\vartheta_{32} = 16\ \%$

Wichtig ist, dass die ϑ_{kl} möglichst *exakt* bekannt sind. Es ist insbesondere zu beachten, dass selbst offizielle Register oft erschreckend ungenau sind. Ziehungsvorschrift:

1. Bestimme bei gegebenem Stichprobenumfang alle Teilumfänge $n_{lm} = n\vartheta_{lm}$
2. Ziehe aus der ZG per Zufallsauswahl und bestimme Zugehörigkeit zu Altersklasse l und Geschlecht m
3. Wurde die Person bisher nicht gezogen und sind in der entsprechenden "Schublade" (lm) noch weniger als n_{lm} Stichprobenteilnehmer, füge die Person hinzu, ansonsten verwerfe sie
4. Iteriere, bis alle "Schubladen" gefüllt sind

Statistischen Eigenschaften der Schichten

Festlegungen:

- Alle Schichten werden durch einen einzigen Index k gekennzeichnet (im Beispiel $k = \{l, m\}$). Der Anteilswert ϑ_k ist nach Konstruktion in der Stichprobe und der GG gleich
- Die Teilstichproben vom Umfang n_k repräsentativ für die entsprechende Teilmengen der GG (bei obiger Ziehungsmethode erfüllt)
- Die Schichtung ergibt nur Sinn, wenn Erwartungswerte und Varianzen der interessierenden Variable Y von der Schicht abhängen:

$$\mu_k = E(Y|k), \quad \sigma_k^2 = V(Y|k),$$

- Da der Erwartungswert-Operator linear ist, gilt

$$\sum_k \vartheta_k \mu_k = \mu,$$

während eine analoge Formel für die Varianzen *nicht* gilt

Schätzer bei geschichteten Stichproben

Auch bei geschichteten Stichproben ist der erwartungstreue und nahezu effiziente Schätzer das normale arithmetische Mittel $\bar{Y}$.
Wir betrachten zunächst für jede Schicht getrennt die Stichproben-Schätzer

$$\hat{\mu}_k = \bar{Y}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} Y_j^{(\mathsf{k})}$$

mit

$$E(\hat{\mu}_k) = \mu_k, \quad V(\hat{\mu}_k) = \frac{\sigma_k^2}{n_k}.$$

Schätzer als Summe über die Schichten

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_j^{(k)} = \sum_{k=1}^{K} \frac{n_k}{n} \frac{1}{n_k} \sum_{j=1}^{n_k} Y_j^{(k)} = \sum_{k} \vartheta_k \hat{\mu}_k$$

Erwartungswert

Arithmetisches Mittel $\hat{\mu} = 1/n \sum_i Y_i$ in der Form $\hat{\mu} = \sum_k \vartheta_k \hat{\mu}_k$:

$$\begin{aligned} E(\hat{\mu}) &= E\left(\sum_k \vartheta_k \hat{\mu}_k\right) \\ &= \sum_k \vartheta_k E(\hat{\mu}_k) \\ &= \sum_k \vartheta_k \mu_k \\ &= \underline{\underline{\mu}} \end{aligned}$$

Das einfache arithmetische Mittel ist bei geschichteten Stichproben erwartungstreu

Varianz

Wieder nutzen wir das arithmetisches Mittel $\hat{\mu} = 1/n \sum_i Y_i$ in der Form $\hat{\mu} = \sum_k \vartheta_k \hat{\mu}_k$:

$$
\begin{aligned}
V(\hat{\mu}) &= V\left(\sum_k \vartheta_k \hat{\mu}_k\right) \\
&= \sum_k \vartheta_k^2 V(\hat{\mu}_k) + 2 \sum_k \sum_{l<k} \vartheta_k \vartheta_l \mathrm{Cov}(\hat{\mu}_k \hat{\mu}_l) \\
&\overset{\text{unkorreliert}}{=} \sum_k \frac{\vartheta_k^2 \sigma_k^2}{n_k} \\
&\overset{n_k = n\vartheta_k}{=} \frac{1}{n} \sum_k \vartheta_k \sigma_k^2
\end{aligned}
$$

Statt der Gesamtvarianz bei Zufallsstichproben erscheint bei der geschichteten Stichprobe das gewichtete Mittel der Einzelvarianzen im Zähler. Dieses ist deutlich geringer, da die Anteile $(\mu_k - \mu_l)^2$ fehlen, die in der Gesamtvarianz auftauchen

Fragen

? Warum kann man die obige Rechnung nicht auch bei der Zufallsauswahl durchführen, indem man dort einfach anstelle der wahren Anteile ϑ_k der Schichten die in der Stichprobe vorgefundenen relativen Häufigkeiten f_k nimmt?

! Bei einer reinen Zufallsauswahl sind die Quoten nicht fest. Vielmehr spielen die relativen Häufigkeiten f_k die Rolle von (erwartungstreuen) Schätzern für die wahren Anteile, $f_k = \hat{\vartheta}_k$. Damit gilt

$$\hat{\mu} = \sum_k f_k \hat{\mu}_k = \sum_k \hat{\vartheta}_k \hat{\mu}_k$$

Man hat also in jedem Summand *zwei* Zufallsgrößen statt einer, womit letztendlich die größere Streuung desselben Schätzers bei Zufallsauswahl resultiert.

Beispiel: Studentenstadt

In einer "Studentenstadt" wie Erlangen bietet sich die Schichtung bezüglich des Quotenmerkmals "Ist der Einwohner Student?" an. Es soll die Frage der MIV -Bevorzugung anhand einer Stichprobe ($n = 2\ 500$) untersucht werden. Bei den Studenten (Anteil 40 %) liege der tatsächliche (aber unbekannte und durch die Stichprobe abzuschätzende) Anteil bei 12.5% und bei den übrigen Einwohnern bei 75%.

? Warum ist das gewählte Quotenmerkmal für diese Untersuchung sinnvoll?

! Es ist

(i) aus der Grundgesamtheit (z.B. dem Einwohnermeldeamt) exakt bestimmbar,
(ii) das zu untersuchende Merkmal hängt deutlich vom Quotenmerkmal ab,
(iii) Die GG wird durch das Quotenmerkmal in nicht zu ungleich starke Schichten aufgeteilt. Gäbe es z.B. nur 10% Studierende, würde die Schichtung wenig Verbesserung der Trennschärfe bringen.

Studentenstadt II

? Wie groß ist die Varianz des geschätzten Gesamtanteils der MIV-Bevorzugung, wenn man das Quotenmerkmal ignoriert und die Stichprobe durch Zufallsauswahl gewinnt?

! Es sei die Dummyvariable $Y_i = 1$, falls die MIV-Bevorzugung zutrifft und $=0$ sonst. Der wahre Gesamtanteil beträgt

$$E(Y) = \mu = \vartheta_1\mu_1 + \vartheta_2\mu_2 = 0.4 * 0.125 + 0.6 * 0.75 = \underline{\underline{50\%}}.$$

Stichprobenvarianz bei Zufallsauswahl:

$$\sigma^2_{\text{Zufall}} = \frac{\mu(1-\mu)}{n} = \underline{\underline{0.0001}}, \quad \sigma = \underline{\underline{1\%}}.$$

Studentenstadt III

? Wie ist die geschichtete Stichprobe aufgebaut und wie groß ist ihre Stichprobenvarianz?

! Die Stichprobe besteht aus $n_1 = n\vartheta_1 = 1000$ Studierende und $n_2 = n - n_1 = 1\,500$ Nicht-Studierende. Stichprobenvarianz bei Quotenauswahl:

$$\begin{aligned}
\sigma^2_{\text{Quoten}} &= \frac{\vartheta_1\sigma_1^2 + \vartheta_2\sigma_2^2}{n} \\
&= \frac{\vartheta_1\mu_1(1-\mu_1) + \vartheta_2\mu_2(1-\mu_2)}{n} \\
&= \underline{\underline{6.2510^{-5}}}, \quad \sigma_{\text{Quoten}} = \underline{\underline{0.791\%}}.
\end{aligned}$$

Studentenstadt IV

? Wie groß muss der Stichprobenumfang mindestens sein, dass die Standardabweichung des geschätzten Anteils der MIV-Bevorzugung höchstens 1% beträgt?

! Zufallsauswahl: $n \geq 2\ 500$.
Quotenauswahl:

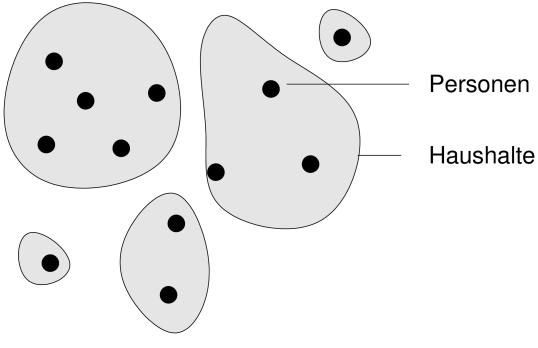
$$\frac{\vartheta_1\mu_1(1-\mu_1) + \vartheta_2\mu_2(1-\mu_2)}{n} \leq 10^{-4} \ \Rightarrow \ \underline{\underline{n \geq 1\ 562}}.$$

Die Stichprobe kann bei gleicher Genauigkeit auf 62.5% des ursprünglichen Umfangs reduziert werden.

10.4 Entzerrung

Gründe für eine Verzerrung:

1. **Zufällige Verzerrung**: Zieht man eine Stichprobe trotz bekannter Quoten mit dem reinen Zufallsverfahren, erhält man i.A. eine zufällige Verzerrung bezüglich der Quotenanteile. Arithmetische Mittel bleiben unverzerrt, aber sind nicht so effizient wie bei einer geschichteten Stichprobe
2. Durch *Entzerrung* **kontrollierbare systematische Verzerrung**: Beispielsweise, wenn die ZG nicht identisch zur GG ist wie bei Personen *vs* Haushalte. Diskussion der Abbildung? 2/12, 2/12, 3/12, 0, 5/12 vs 2/5, 1/5, 1/5, 0, 1/5
3. **Unkontrollierbare systematische Verzerrungen**: Beispielswiese durch unvollständigen Rücklauf. Diskussion am Beispiel des Mobilitätsverhaltens von Managern und Rentnern



Entzerrung II

- Im Ggs zur geschichteten Stichprobe (z.B. Mobilitätserhebung MiD) wird nicht *im Vorfeld* die Stichprobe angepasst, sondern *nachträglich* die Gewichtung der Stichprobenelemente durch **Entzerrungsfaktoren** (z.B. Mobilitätserhebung SrV)
- Die Methode setzt die Kenntnis von Quotenmerkmalen bei den Stichprobenelementen voraus
- Sie ist für alle Verzerrungsursachen gleich, funktioniert aber nur, wenn *innerhalb* jeder Schicht die Stichprobe repräsentativ ist

? Machen Sie sich klar, dass diese Voraussetzung bei den obigen Verzerrungsursachen 1 und 2 erfüllt sind, nicht aber bei der Ursache 3

! Problematik der Rückläufe von Managern und Rentnern. Kann teilweise durch den Berufsstatus als Quotenmerkmal entzerrt werden, aber nicht ganz

Durchführung der Entzerrung

- Tabelliere die bekannten wahren Anteile ϑ_k der (ggf mehrdimensionalen) Schichtung der GG
- Bestimme aus der vorliegenden Stichprobe die relativen Häufigkeiten $f_k = n_k/n$
- Rechne die relativen Häufigkeiten auf die wahren Anteile durch die Entzerrung $E_k = \vartheta_k/f_k$ hoch

Ergebnis: Entzerrender Schätzer

$$\hat{\mu}^{(\mathsf{E})} = \frac{1}{n}\sum_{i=1}^{n} E_{k(i)} Y_i, \quad E_{k(i)} = \frac{\vartheta_k}{f_k} \text{ falls Element } i \in \text{Schicht } k$$

Alternative Darstellung zur Analyse:

$$\begin{aligned}
\hat{\mu}^{(\mathsf{E})} &= \frac{1}{n}\sum_{i=1}^{n} E_i Y_i = \frac{1}{n}\sum_{k}\sum_{j=1}^{n_k} E_k Y_j^{(\mathsf{k})} \\
&= \frac{1}{n}\sum_{k} E_k n_k \hat{\mu}_k = \frac{1}{n}\sum_{k}\frac{\vartheta_k}{f_k} n_k \hat{\mu}_k \overset{f_k = n_k/n}{=} \sum_{k} \vartheta_k \hat{\mu}_k
\end{aligned}$$

Erwartungswert und Varianz des entzerrenden Schätzers

Nach Voraussetzung gilt in jeder Schicht Repräsentativität, also $E(\hat{\mu}_k) = \mu_k$:

$$E(\hat{\mu}^{(\mathsf{E})}) = \sum_k \vartheta_k E(\hat{\mu}_k) = \sum_k \vartheta_k \mu_k = \mu$$

Das entzerrte Stichprobenmittel $\hat{\mu}^{(\mathsf{E})}$ ist also erwartungstreu

Varianz: Mit verschwindender Kovarianz zwischen den einzelnen Schichten ergibt sich

$$\begin{aligned} V(\hat{\mu}^{(\mathsf{E})}) &= \sum_k \vartheta_k^2 V(\hat{\mu}_k) = \sum_k \vartheta_k^2 \frac{\sigma_k^2}{n_k} = \sum_k \vartheta_k^2 \frac{\sigma_k^2}{n f_k} \\ &\overset{f_k = \vartheta_k / E_k}{=} \frac{1}{n} \sum_k \vartheta_k E_k \sigma_k^2 \end{aligned}$$

Die Varianz des entzerrenden Schätzers ist vergleichbar der des normalen arithmetischen Mittels bei Quotenauswahl, aber mit den Wichtungen E_k

Was gilt bei geschichteter Stichprobe? $\hat{\mu}^{(\mathsf{E})}$ wird zum normalen arithmetischen Mittelwert und die Varianz zu der der geschichteten Stichprobe

Beispiel: Studentenstadt

Die Stichprobe $n = 2\,500$ aus der Studentenstadt der Quotenauswahl wird nun mit Zufallsauswahl durchgeführt und dabei zufällig 1 050 Studenten (statt der durch die Quotenbedingung festgelegten $n\vartheta_1 = 1\,000$) sowie 1 450 Nicht-Studenten gezogen. Wie groß sind die Entzerrungsfaktoren und die Standardabweichung des entzerrenden Schätzers für die MIV-Bevorzugung? (in der GG angenommen 12.5 % bei den Studenten und 75 % bei den anderen)?

Wir haben $\vartheta_1 = 0.4$, $\vartheta_2 = 0.6$, $n_1 = 1050$ und $n_2 = 1450$:

$$E_1 = \frac{\vartheta_1}{f_1} = \frac{n\vartheta_1}{n_1} = \underline{\underline{\frac{20}{21}}}, \quad E_2 = \frac{\vartheta_2}{f_2} = \underline{\underline{\frac{30}{29}}}.$$

Varianz des entzerrenden Schätzers (mit $\mu_1 = 0.125$ und $\mu_2 = 0.75$):

$$V(\hat{\mu}^{(\mathsf{E})}) = \frac{1}{n}\Big[\vartheta_1 E_1 \mu_1(1-\mu_1) + \vartheta_2 E_2 \mu_2(1-\mu_2)\Big] = \underline{\underline{6.32 * 10^{-5}}}$$

Dies entspricht einer Standardabweichung von 0.795% und ist nahezu identisch zum Ergebnis des Quotenverfahrens, aber deutlich trennschärfer als das Zufallsverfahren ohne Entzerrung.