Stichprobenverfahren
– Übungsaufgaben –

Willi Mutschler
willi@mutschler.eu

Version: 20. Juli 2018

Aufgabe 1: Einschlusswahrscheinlichkeiten

Betrachten Sie eine kleine Grundgesamtheit mit $N = 4$ Elementen: $U = \{u_1, u_2, u_3, u_4\}$. Eine Stichprobe der Größe $n = 2$ soll gezogen werden.

(a) Betrachten Sie für den Moment den Fall einer einfachen Zufallsstichprobe bei der jede Stichprobe s mit derselben Wahrscheinlichkeit gezogen werden kann. Berechnen Sie (i) $M = |\mathcal{S}|$, (ii) π_k und (iii) die Summe aller Einschlusswahrscheinlichkeiten der Elemente $k \in U$.

(b) Zeigen Sie, dass bei (a) die Kovarianz zwischen den Einschlussindikatoren I_k und I_l negativ ist.

(c) Betrachten Sie nun folgendes Stichprobendesign: $\mathcal{S}_n = \{s_1, s_2, s_3\}$ mit $s_1 = \{u_1, u_3\}$, $s_2 = \{u_1, u_4\}$ und $s_3 = \{u_2, u_4\}$. Nehmen Sie folgende Wahrscheinlichkeiten an: $p(s_1) = 0.1$, $p(s_2) = 0.6$ und $p(s_3) = 0.3$. Berechnen Sie (i) alle Einschlusswahrscheinlichkeiten π_k, (ii) die Summe aller Einschlusswahrscheinlichkeiten und (iii) alle Einschlusswahrscheinlichkeiten π_{kl}.

(d) Berechnen Sie die Kovarianzmatrix der Einschlussindikatoren in (c).

Aufgabe 2: Schätzung mithilfe von Einschlusswahrscheinlichkeiten

Betrachten Sie eine kleine Grundgesamtheit U der Größe $N = 5$. Die Werte von Y in der Grundgesamtheit betragen $\{1, 2, 5, 12, 30\}$. Eine Stichprobe (ohne Zurücklegen) der Größe $n = 3$ soll gezogen werden. Für die Wahrscheinlichkeiten der M möglichen Stichproben gilt:

$$p(s_i) = \frac{i}{1 + \cdots + M}$$

(a) Berechnen Sie die Anzahl M aller möglichen Stichproben.

(b) Berechnen Sie alle Elemente von $\mathcal{S}_n$. Stellen Sie hierzu eine Matrix mit Inklusionsindikatoren auf.

(c) Berechnen Sie die Einschlusswahrscheinlichkeiten erster (π_k) und zweiter (π_{kl}) Ordnung.

(d) Berechnen Sie die Kovarianzen der Einschlussindikatoren I_k und I_l.

(e) Schätzen Sie den Mittelwert der Grundgesamtheit mithilfe des π-Schätzers: $\hat{\bar{y}}_\pi = \frac{1}{N} \sum_s \frac{y_k}{\pi_k}$ für alle M möglichen Stichproben.

(f) Zeigen Sie numerisch, dass $\hat{\bar{y}}_\pi$ eine unverzerrte Schätzfunktion für den Mittelwert ist.

(g) Schätzen Sie für die M Stichproben die Varianz des obigen π-Schätzers für den Mittelwert mithilfe der Schätzfunktion $\hat{V}(\hat{\bar{y}}_\pi) = \frac{1}{N^2} \sum \sum_s \Delta_{kl} \breve{y}_k \breve{y}_l$. Was fällt ihnen bei der ersten Stichprobe auf?

(h) Schätzen Sie für die M Stichproben die Varianz des obigen π-Schätzers für den Mittelwert mithilfe der Schätzfunktion $\hat{V}(\hat{\bar{y}}_\pi) = -\frac{1}{2N^2} \sum \sum_s \breve{\Delta}_{kl} (\breve{y}_k - \breve{y}_l)^2$

(i) Zeigen Sie numerisch, dass beide Varianzschätzer unverzerrt sind.

Aufgabe 3: Horvitz-Thompson-Schätzer für einfache Zufallsstichproben ohne Zurücklegen

Zeigen Sie, dass bei der einfachen Zufallsstichproben ohne Zurücklegen folgendes für den Horvitz-Thompson Schätzer gilt:

(a) $\hat{t}_\pi = N\bar{y}_s = \frac{1}{f}\sum_s y_k$ ist ein unverzerrter Schätzer für die Merkmalssumme.

(b) $V(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{y_U}$ ist die Varianz von $\hat{t}_\pi$

(c) $\hat{V}(\hat{t}_\pi) = N^2\frac{1-f}{n}S^2_{y_s}$ ist ein unverzerrter Schätzter für die Varianz.

Es gilt $f = n/N$, $S^2_{y_U} = \frac{1}{N-1}\sum_U(y_k - \bar{y}_U)^2$ und $S^2_{y_s} = \frac{1}{n-1}\sum_s(y_k - \bar{y}_s)^2$.

Aufgabe 4: Bernoulli Stichprobenziehungen

Betrachten Sie folgendes Verfahren um eine Stichprobe aus einer Grundgesamtheit mit N Elementen auszuwählen. Sei π_B eine Konstante derart, dass $0 < \pi_B < 1$. Gegeben sind N unabhängige Realisationen $\varepsilon_1, \ldots, \varepsilon_N$ aus einer auf dem Intervall $[0; 1]$ gleichverteilten Zufallsvariable. Die Stichprobe wird anhand folgender Regel erstellt: Falls $\varepsilon_k < \pi_B$ wird das Element k in die Stichprobe aufgenommen, ansonsten nicht.

(a) Welche Aussagen können Sie zum Stichprobendesign, Einschlussindikatoren, Einschlusswahrscheinlichkeiten und Stichprobengröße bei diesem Verfahren fällen?

(b) Wie lautet die Wahrscheinlichkeit, dass eine Stichprobe genau die Größe $n_s = n$ hat?

(c) Wie lautet der π-Schätzer für die Merkmalssumme? Geben Sie einen Ausdruck für die Varianz an.

(d) Vergleichen Sie dieses Design mit der einfachen Zufallsauswahl ohne Zurücklegen. Berechnen Sie hierzu den Designeffekt und setzen Sie $N\pi = n$. Interpretieren Sie ihr Ergebnis im Zusammenhang mit dem Variationskoeffizienten, $cv_{y_U} = S_{y_U}/\bar{y}_U$, der die Standardabweichung des Merkmals ins Verhältnis zum Merkmalsmittelwert setzt.

Aufgabe 5: Stichprobenmittel und -median

Betrachten Sie eine kleine Grundgesamtheit vom Umfang $N = 5$ mit den Merkmalswerten $Y_1 = 3, Y_2 = 1, Y_3 = 0, Y_4 = 1$ und $Y_5 = 5$. Betrachten Sie eine einfache Zufallsstichprobe ohne Zurücklegen vom Umfang $n = 3$.

(a) Für alle möglichen Stichproben berechnen Sie jeweils das Stichprobenmittel und zeigen Sie, dass das Stichprobenmittel erwartungstreu für das Mittel der Grundgesamtheit ist.

(b) Für alle möglichen Stichproben berechnen Sie jeweils den Stichprobenmedian. Bestimmen Sie, ob der Stichprobenmedian erwartungstreu für den Median in der Grundgesamtheit ist.

Aufgabe 6: Schraubenlieferungen

Von vier Lieferungen, $L_1, \ldots, L_4$, mit $N_1 = 500$, $N_2 = 200$, $N_3 = 200$ und $N_4 = 100$ Schrauben soll eine Lieferung auf Stichprobenbasis bezüglich der Genauigkeit der Schraubenlänge überprüft werden. Die Lieferungen sind nicht getrennt, die Schrauben jedoch markiert, so dass man erkennen kann, zu welcher der vier Lieferungen eine Schraube gehört. Es wird blindlings eine der $N_1 + N_2 + N_3 + N_4$ Schrauben ausgewählt und festgestellt, zu welcher Lieferung sie gehört. Diese Lieferung wird dann für die Überprüfung der Schrauben auf Stichprobenbasis ausgewählt.
Handelt es sich bei der Auswahl der Lieferung um eine einfache Zufallsauswahl vom Umfang $n = 1$? Geben Sie gegebenfalls die Wahrscheinlichkeiten p_j an, dass die Lieferung L_j ausgewählt wird $(j = 1, 2, 3, 4)$.

Aufgabe 7: Unterschriftenaktion

Die Studierenden der Fakultät Statistik der TU Dortmund starten eine Unterschriftenaktion zur Abschaffung der 8-Uhr Vorlesungen an der TU Dortmund. Die Unterschriften werden auf insgesamt 676 Blatt Papier gesammelt, wobei auf jedes Blatt 42 Unterschriften passen. Nicht alle Blätter enthalten jedoch die Maximalanzahl an Unterschriften. In einer einfachen Zufallsauswahl vom Umfang $n = 50$ Blatt Papier werden die Anzahlen der Unterschriften gezählt. Die folgende Tabelle enthält das Ergebnis:

y_i	42	41	36	32	29	27	23	19	16	15	14	11	10	9	7	6	5	4	3
f_i	23	4	1	1	1	2	1	1	2	2	1	1	1	1	1	3	2	1	1

Schätzen Sie die Gesamtanzahl an Unterschriften für diese Aktion und geben Sie ein 80% Konfidenzintervall (unter Annahme der Normalverteilung) für die Gesamtanzahl an.

Aufgabe 8: Verteilung des π Schätzers

Betrachten Sie den Datensatz `psid.csv`. Dieser beinhaltet einen Auszug aus dem Panel zur Analyse von Einkommensdynamiken der Universität Michigan. Nehmen Sie an, der vorhandenen Datensatz stellt ihre Grundgesamtheit mit $N = 1000$ Merkmalsträgern dar.

(a) Betrachten Sie die Variable „wage“, die den Lohn der Merkmalsträger enthält. Schätzen Sie, basierend auf einer einfachen Zufallsstichprobe ohne Zurücklegen mit $n = 20$, den Durchschnittslohn in der Grundgesamtheit. Wie hoch ist die geschätzte Varianz dieses Schätzers?

(b) Schreiben Sie eine Funktion, die für eine gegebene einfache Zufallsstichprobe ohne Zurücklegen den Schätzwert für die Varianz des π-Schätzers ausgibt.

(c) Approximieren Sie die Verteilungen der Schätzfunktionen für den Mittelwert sowie seine Varianz mithilfe eines sogenannten „einfachen nichtparametrischen Bootstraps“. Hierzu ziehen Sie $B = 10000$ unterschiedliche Stichproben und berechnen jeweils die Schätzwerte für Durchschnittslohn und Varianz. Speichern Sie diese ab und betrachten Sie die zugehörigen Verteilungen. Wie verhält sich diese zu den wahren Werten der Grundgesamtheit?

(d) Wiederholen Sie Schritt (b) mit $n = 250$ und $B = 100000$.

(e) Ihr Dozent möchte nun, dass Sie – wie in der Stichprobenpraxis üblich – Konfidenzintervalle für den Durchschnittslohn mithilfe der Normalverteilung berechnen. Was erwidern Sie ihm?

Aufgabe 9: Anteilsschätzung in einfachen Zufallsstichproben

(a) Leiten Sie analytisch die Schätzfunktionen für einen Anteilsschätzer in einfachen Zufallsstichproben ohne Zurücklegen her.

(b) Wiederholen Sie die vorherige Aufgabe zur Verteilung des π-Schätzers um den Anteil an Personen zu schätzen, die im Dienstleistungssektor tätig sind. Betrachten Sie hierzu die Variable „sector“, die den Wert 7 für eine Tätigkeit im Dienstleistungssektor annimmt.

Aufgabe 10: Bundestagswahl

Man interessiert sich für den Stimmenanteil, den eine Partei bei der nächsten Bundestagswahl erhalten wird. Wie viele Wahlberechtigte muss man befragen, wenn das Konfidenzintervall (unter Annahme der Normalverteilung) möglichst eine Länge von weniger als 10% des Schätzwertes für den unbekannten Stimmenanteil haben soll und bekannt ist, dass der Stimmenanteil der Partei (a) etwa bei 50%, (b) zwischen 15 und 20%, (c) etwa bei 5% liegen wird und eine Überdeckungswahrscheinlichkeit von $(1-\alpha) = 0.95$ gefordert werden soll. Warum können Sie bei ihren Berechnungen den Korrekturfaktor $\frac{N-n}{N-1}$ vernachlässigen?

Aufgabe 11: Effizienz systematischer Auswahl

Betrachten Sie die systematische Auswahl mit $N = an$, wobei a eine ganzzahlige Zahl darstellt.

(a) Zeigen Sie, dass die Varianz des π-Schätzers der Populationssumme umgeformt werden kann zu

$$V(\hat{t}_\pi) = \frac{N^2 S_{y_U}^2}{n}[(1-f) + (n-1)\delta]$$

mit $f = n/N = 1/a$. Interpretieren Sie den Ausdruck.

(b) Zeigen Sie, dass der Designeffekt im Vergleich zur einfachen Zufallsstichprobe ohne Zurücklegen gleich $1 + \frac{n-1}{1-f}\delta$ ist. Für welche Werte von δ ist die systematische Auswahl effizienter?

(c) Diskutieren Sie die Praxisrelevanz ihrer Ergebnisse.

Aufgabe 12: Reisanbaufläche

Die insgesamt bebaute Fläche a_i (in acre) und die Reisanbaufläche y_i (in acre) wurde für eine Stichprobe von 25 Dörfern aus insgesamt 892 Dörfern erhoben. Dabei wurden die Stichprobenelemente mit Zurücklegen und mit Wahrscheinlichkeiten, die proportional der insgesamt bebauten Fläche sind, gezogen. Die bebaute Gesamtfläche beträgt 568565 acres. Die Daten der Erhebung sind in der Datei `reisanbau.csv` zusammengestellt. Schätzen Sie die Reisanbaufläche der 892 Dörfer und geben Sie ein 95%-Konfidenzintervall (unter Annahme der Normalverteilung) für die Reisanbaufläche an!

Aufgabe 13: Unterschiedliche Auswahlwahrscheinlichkeiten mit R

Betrachten Sie die R Funktion `sample`. Diese ermöglicht es, sequentiell mit gegebenen Ein-Zug-Auswahlwahrscheinlichkeiten p_k eine Stichprobe mit unterschiedlichen Auswahlwahrscheinlichkeiten zu ziehen. Zeigen Sie mithilfe einer Simulationsstudie, dass die Funktion die Vorgabe $\pi_k = np_k$ für größenproportionale Stichproben nicht erfüllt. Führen Sie folgende Schritte durch:

- Setzen sie die Größe der Stichprobe $n = 3$ und die Größe der Grundgesamtheit $N = 5$.
- Benutzen Sie folgende ungleiche Auswahlwahrscheinlichkeiten: `p <- 4:8/sum(4:8)`
- Ziehen Sie $B = 10000$ Stichproben mithilfe des `sample(1:N,n,prob=p)` Befehls.
- Überprüfen Sie, wie oft die Elemente $1, ..., 5$ in den jeweiligen Stichproben vorkommen und berechnen Sie die relative Häufigkeit indem sie durch die Anzahl B dividieren. Hinweis: Verwenden Sie die `%in%` Funktion.
- Vergleichen Sie ihr Ergebnis mit `p*n`.

Wiederholen Sie obiges Vorgehen indem Sie anstelle des `sample` Befehls den Befehl `sampford` aus der Bibliothek `pps` verwenden.

Aufgabe 14: Größenproportionale Auswahl mit R

Laden Sie das Paket `samplingbook`. Dieser beinhaltet einen Datensatz, `data(influenza)`, in dem Daten zu Grippeerkrankungen der 424 Stadt- und Landkreise aus dem Jahr 2007 abrufbar sind. Die Variable `district` enthält die Namen der Stadt- bzw. Landkreise, die Variable `population` die Einwohnerzahl, und `cases` die Anzahl der Influenza-Erkrankungen aus dem Jahr 2007. Schätzen Sie nun anhand einer Stichprobe die Anzahl der Influenza-Fälle für ganz Deutschland.

(a) Verwenden Sie als Hilfsgröße die Einwohnerzahl der Landkreise und ziehen Sie eine größenproportionale Stichprobe der Landkreise vom Umfang $n = 20$. Betrachten Sie die gezogenen Kreise. Hinweis: Benutzen Sie hierzu den Befehl `pps.sampling`. Wählen Sie einen geeigneten Algorithmus.

(b) Schätzen Sie nun den Mittelwert der Influenza-Fälle mit verschiedenen Methoden der Varianzschätzung. Hinweis: Benutzen Sie hierzu die Funktion `htestimate`.

(c) Bestimmen Sie ein Konfidenzintervall für die Gesamtanzahl der Krankheitsfälle (unter Verwendung der Normalverteilung). Vergleichen Sie mit der tatsächlichen Anzahl an Krankheitsfällen.

Aufgabe 15: Größenproportionale Auswahl mit n=2

Aus einer Grundgesamtheit vom Umfang N werden nacheinander zwei Einheiten nach einem größenproportionalen Verfahren entnommen und **nicht** wieder zurückgelegt. Zeigen Sie:

$$\text{(a) } \pi_k = p_k \left(1 + \sum_{l=1, k\neq l}^{N} p_l (1-p_l)^{-1}\right) \text{ und (b) } \pi_{kl} = p_k p_l \left(\frac{1}{1-p_k} + \frac{1}{1-p_l}\right), \quad i \neq j,$$

wobei $p_k = x_k \left(\sum_{j=1}^{N} x_j\right)^{-1}$.

Aufgabe 16: Wasserverschmutzung

In einer Studie zur Wasserverschmutzung wird eine Stichprobe von Seen in einer Studienregion mit 320 Seen durch die folgende Prozedur gezogen: Ein Rechteck der Länge l und Breite b wird um das Studiengebiet auf einer Karte eingezeichnet. Ein Paar von gleichverteilten Zufallszahlen zwischen 0 und 1 wird erzeugt. Die erste Zufallszahl des Paares wird mit der Länge l und die zweite mit der Breite b multipliziert, um Lagekoordinaten innerhalb der Studienregion zu bestimmen. Wenn die Lagekoordinaten in einem See sind, wird dieser See ausgewählt. Das Auswahlverfahren wird solange durchgeführt bis vier Seen ausgewählt worden sind. Der erste See in der Stichprobe wurde bei diesem Auswahlverfahren zweimal ausgewählt, die beiden anderen Seen nur einmal. Die Schadstoffkonzentration für die drei Seen in der Stichprobe betragen 2, 5 und 10 (ppm). Die Größe der Seen (in km^2) sind 1.2, 0.2 und 0.5. Insgesamt sind 80 km^2 der Studienregion durch Seen bedeckt.

Geben Sie einen unverzerrten Schätzer für die durchschnittliche Schadstoffkonzentration pro See in der Grundgesamtheit an sowie eine Schätzung für die Varianz der Schätzers der mittleren Schadstoffkonzentration!

Aufgabe 17: Weizenproduktion

In einer Population von $N = 3$ Farmen werden $n = 2$ Farmen mit Auswahlwahrscheinlichkeiten proportional zu ihrer Größe gezogen, um die Gesamtproduktion von Weizen zu schätzen. In der folgenden Tabelle sind die Werte der Population gegeben.

Untersuchungseinheit (Farm) k	1	2	3
Auswahlwahrscheinlichkeit p_k	0.3	0.2	0.5
Weizenproduktion (in t)	11	6	25

Betrachten Sie jede *geordnete* pps-Stichprobe mit Zurücklegen vom Umfang $n = 2$ und berechnen Sie den Horvitz-Thompson-Schätzer und den Hansen-Hurwitz-Schätzer für jede Stichprobe. Zeigen Sie die Unverzerrtheit der beiden Schätzer und berechnen Sie die Standardabweichung der beiden Schätzer in der Population.

Aufgabe 18: Varianzzerlegung

Zeigen Sie, dass gilt:

$$(N-1)S_{y_U}^2 = \sum_{h=1}^{H}(N_h - 1)S_{yU_h}^2 + \sum_{h=1}^{H} N_h(\bar{y}_{U_h} - \bar{y}_U)^2$$

Aufgabe 19: Kommunalwahl (1)

Da bei der letzten Kommunalwahl der Anteil der Wähler der *Opportunistischen Partei* (OP) in verschiedenen Bevölkerungskreisen unterschiedlich war, entschließt man sich für eine Wahlprognose zu einem geschichteten Auswahlverfahren aus drei Bevölkerungsschichten. Neben der Frage, ob bei der nächsten Wahl die OP gewählt wird, wird zusätzlich das Alter der befragten Personen erhoben. Die Ergebnisse der Erhebung sind in nachfolgender Tabelle zusammengefasst:

Schicht h	W_h	$\hat{P}_h$	$\bar{y}_h$	s_h^2
1	0.2	0.40	28	90
2	0.5	0.15	45	85
3	0.3	0.25	60	80

Hierbei bezeichnet $\hat{P}_h$ den Anteil der Personen, die die OP wählen wollen, $\bar{y}_h$ das Durchschnittsalter und s_h^2 die empirische Varianz des Alters in der Schicht h, $h = 1, 2, 3$.

(a) Schätzen Sie den zu erwartenden Anteil, den die OP bei der nächsten Wahl erhält, und das Durchschnittsalter. Geben Sie zudem, falls möglich, für beide Parameter jeweils 95%-Konfidenzintervalle an.

(b) Berechnen Sie für beide erhobenen Merkmale eine optimale Aufteilung des Stichprobenumfangs mit den obigen Ergebnissen als Vorinformation.

Aufgabe 20: Bauernhöfe

In der folgenden Tabelle sind die Bauernhöfe eines Landes entsprechend ihrer Größe geschichtet. Ferner wird die durchschnittliche mit Weizen bebaute Fläche angegeben.

Größe (ha)	Anzahl der Bauernhöfe	Durchschnittliche mit Weizen bebaute Fläche	Standard-abweichung
0 – 40	394	5.4	8.3
41 – 80	461	16.3	13.3
81 – 120	391	24.3	15.1
121 – 160	334	34.5	19.8
161 – 200	169	42.1	24.5
201 – 240	113	50.1	26.0
> 240	148	63.8	35.2

Es soll eine Stichprobe vom Umfang $n = 100$ Bauernhöfe gezogen werden. Wie groß sind die Stichprobenumfänge der einzelnen Schichten bei

(a) proportionaler Aufteilung?

(b) optimaler Aufteilung?

Vergleichen Sie die Genauigkeit dieser Verfahren mit der Genauigkeit bei einer einfachen Zufallsstichprobe. (*Hinweis*: Betrachten Sie das Verhältnis der Varianzen der Mittelwertschätzer.)

Aufgabe 21: Erhebungskosten

Es soll eine geschichtete Zufallsstichprobe gezogen werden. Es wird vermutet, dass die Erhebungskosten gleich der Summe $\sum c_h n_h$ sind. Ferner wird erwartet, dass bei der Untersuchung die wichtigen Schätzwerte etwa folgende Größe haben:

Schicht	W_h	S_h	c_h
1	0.4	10	4 Euro
2	0.6	20	9 Euro

(a) Berechnen Sie diejenigen Werte für n_1/n und n_2/n, die bei vorgegebener Varianz $\mathrm{Var}(\hat{\bar{Y}}_U)$ die Erhebungskosten minimieren.

(b) Wie groß muss der Stichprobenumfang sein, wenn $\mathrm{Var}(\hat{\bar{Y}}_U) = 1$ sein soll? Vernachlässigen Sie den Korrekturfaktor.

(c) Wie groß sind die Erhebungskosten?

Aufgabe 22: Erfrischungsräume

In einem Unternehmen sind 62% der Beschäftigten männliche Fach- oder Hilfskräfte, 31% sind weibliche Schreibkräfte, 7% der Angestellten sind mit leitenden Aufgabe beschäftigt. Die Unternehmensleitung will mit einer Stichprobe vom Umfang $n = 400$ den Anteil der Beschäftigten schätzen, die die firmeneigenen Erfrischungsräume nutzen. Nach groben Schätzungen werden sie von 40 bis 45% der männlichen Fach- und Hilfsarbeiter, von 20 bis 30% der weiblichen Schreibkräfte und von 5 bis 10% der leitenden Angestellten benutzt.

(a) Wie würden Sie den Stichprobenumfang zwischen den drei Schichten aufteilen?

(b) Wenn die wahren Anteilswerte 48, 21 und 4% sind, wie groß ist die Standardabweichung des geschätzten Anteils $\hat{P}$.

(c) Wie groß ist die Standardabweichung des geschätzten Anteils $\hat{P}$, wenn nur eine einfache Zufallsstichprobe vom Umfang $n = 400$ gezogen wird?

Aufgabe 23: Nachträgliches Schichten

Aus der Gesamtheit der Teilnehmer einer Lehrveranstaltung einer Technischen Universität wird eine einfache Zufallsstichprobe entnommen und die Merkmale "Geschlecht" (m/w), "Körpergröße" (in cm) und "Jeansträger" (ja/nein) erhoben. Man erhält nachfolgende Ergebnisse:

Geschlecht	m	m	w	m	w	w	m	w	w	w	m	m
Körpergröße	182	179	165	192	175	165	182	170	171	172	182	193
Jeansträger	n	j	n	j	j	j	j	n	n	n	n	n

Schichten Sie nachträglich nach dem Geschlecht ($N_1 = 63$ m/$N_2 = 57$ w) und berechnen Sie

(a) erwartungstreue Schätzer für die durchschnittliche Körpergröße und den Anteil der Jeansträger,

(b) die Varianzschätzer zu den Schätzern aus (a) und

(c) vergleichen Sie diese Ergebnisse mit denen für eine einfache Zufallsstichprobe. Hat sich die nachträgliche Schichtung gelohnt?

Aufgabe 24: Qualitätsmängel

Bei der CD-Produktion der Firma *Schall & Rausch* treten leider auch Qualitätsmängel auf, die während der Herstellung nicht bemerkt werden. Zur Qualitätsprüfung vor Auslieferung wird deshalb aus den 1000 bereits verpackten Kartons eines Produktionsabschnittes eine einfache Zufallsstichprobe von 10 Kartons entnommen und hierin jeweils alle 20 CD's auf ihre Qualität überprüft. Die Ergebnisse dieser Überprüfungen sind in nachfolgender Tabelle zusammengestellt:

Karton	1	2	3	4	5	6	7	8	9	10
Anzahl defekter CD's	3	1	1	0	2	0	1	2	2	1

Schätzen Sie den Anteil defekter CD's während des Produktionsabschnittes erwartungstreu.

Aufgabe 25: Kommunalwahl (2)

Es soll der Stimmenanteil der OP (Opportunistische Partei) bei der bevorstehenden Kommunalwahl vorhergesagt werden. Man wählt deshalb $n = 200$ Wahlberechtigte zufällig aus, und fragt sie nach ihrer Einstellung. 80 Wahlberechtigte erklären, die OP wählen zu wollen; 60 dieser 80 Befragten hatten bereits bei der letzten Wahl die OP gewählt; von den 120 Befragten, die die OP nicht wählen wollen, hat bei der letzten Wahl keiner die OP gewählt.

(a) Schätzen Sie den gesuchten Anteilswert und geben Sie die Varianz der Schätzung an.

(b) Wie würden Sie den gesuchten Anteilswert schätzen, wenn bei der letzten Gemeinderatswahl die OP 25% der Wählerstimmen errungen hätte? Berechnen Sie den Differenzen- und Quotientenschätzer. Geben Sie auch die Varianz der jeweiligen Schätzung an.

Solutions

Aufgabe 1: (a) Die Anzahl an Stichproben mit Element k ist $\binom{N-1}{n-1}$. Hinzufügen von Element k zu diesen Stichproben ergibt Stichprobengröße n. k wird aber auch zur Grundgesamtheit hinzugefügt, diese hat dann N Elemente. Also laut LaPlace Definition von Wahrscheinlichkeiten gilt für die Einschlusswahrscheinlichkeit erster Ordnung:

$$\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-1-n+1)!}}{\frac{N!}{n!(N-n)!}} = \frac{\frac{(N-1)!}{(n-1)!}}{\frac{N!}{n!}} = \frac{n}{N}$$

Für die Einschlusswahrscheinlichkeiten zweiter Ordnung gilt analog:

$$\pi_{kl} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{n!}{(n-2)!}}{\frac{N!}{(N-2)!}} = \frac{n(n-1)}{N(N-1)}$$

(b) Es gilt: $\pi_k = \pi_l = \frac{n}{N}$ und $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$. Dann folgt für die Kovarianz:

$$\begin{aligned} Cov(I_k, I_l) &= \pi_{kl} - \pi_k \pi_l = \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N} \\ &= -\frac{n}{N}\left(\frac{1-n}{N-1} + \frac{n}{N}\frac{N-1}{N-1}\right) = \frac{-n}{N}\left(\frac{1-n+n/N(N-1)}{N-1}\right) \\ &= \frac{-n}{N}\left(\frac{1-n/N}{N-1}\right) < 0 \end{aligned}$$

da $n/N > 0$, $1 - n/N > 0$ und $N - 1 > 0$.

Der R-Code könnte folgendermaßen aussehen:

```
## Aufgabe Einschlusswahrscheinlichkeiten
# Matrix mit Einschlusswahrscheinlichkeiten
Iks <- function(x,y) as.numeric(is.element(x,y))
N <- 4
n <- 2

# a)
S <- combn(1:N,n)
M <- choose(N,n)
ps <- rep(1/M,M)
ind <- apply(S,2,function(z) Iks(1:N,z)); ind
pi_k = colSums(t(ind)*ps);
round(pi_k,2)
sum(pi_k)

#c)
M <- 3
S <- cbind(c(1,3),c(1,4),c(2,4))
ps <- c(0.1,0.6,0.3)
ind <- apply(S,2,function(z) Iks(1:N,z)); ind
pi_k <- colSums(t(ind)*ps)
round(pi_k,2)
sum(pi_k)

#d)
pi_kl <- matrix(NA,N,N)
for (k in 1:N){
  for (l in 1:N) {
```

```
      pi_kl[k,l] <- sum(apply(S,2,function(z) Iks(k,z)*Iks(l,z))*ps)
    }
  }
  Delta_kl <- matrix(NA,N,N)
  for (k in 1:N){
    for (l in 1:N) {
      Delta_kl[k,l] <- pi_kl[k,l] - pi_kl[k,k]*pi_kl[l,l]
    }
  }
```

Aufgabe 2: Der R-Code könnte folgendermaßen aussehen:

```
#########################################################################
#### Aufgabe Schaetzung mithilfe von Einschlusswahrscheinlichkeiten
#########################################################################
# Matrix mit Einschlusswahrscheinlichkeiten
Iks <- function(x,y) as.numeric(is.element(x,y))
Y <- c(1,2,5,12,30)
N <- length(Y)
n <- 3
ps <- 1:M/sum(1:M); round(ps,3)
#a)
M <- choose(N,n);M
#b)
S <- combn(N,n);S
ind <- apply(S,2,function(z) Iks(1:N,z)); ind
#c)
pi_k <- colSums(t(ind)*ps);round(pi_k,3)
pi_kl <- matrix(NA,N,N)
for (k in 1:N){
  for (l in 1:N){
    pi_kl[k,l] <- sum(apply(S,2,function(z) Iks(k,z)*Iks(l,z))*ps)
  }
}
round(pi_kl,3)
#d)
Delta_kl <- matrix(NA,N,N)
for (k in 1:N){
  for (l in 1:N) {
    Delta_kl[k,l] <- pi_kl[k,l] - pi_kl[k,k]*pi_kl[l,l]
  }
}
round(Delta_kl,2)
#e)
ybar.hat <- 1/N*apply(S,2,function(z) sum(Y[z]/pi_k[z]))
round(ybar.hat,2)
#f)
mean(Y) #wahrer Wert
sum(ybar.hat*ps) #unverzerrter Schaetzer ergibt wahren Wert
#g)
# Funktion die die Varianz des Horvitz-Thompson Schaetzers fuer jede Stichprobe
    schaetzt
vhatHT <- function(s){
  n <- length(s)
  sl <- rep(NA,n)
  sk <- sl
  for (j1 in 1:n){
    k <- s[j1]
    for (j2 in 1:n) {
      l <- s[j2]
      sl[j2] <- 1/pi_kl[k,l]*(pi_kl[k,l]/(pi_k[k]*pi_k[l])-1)*Y[k]*Y[l]
    }
    sk[j1] <- sum(sl)
```

```
  }
  sum(sk)/N^2
}
vHT <- apply(S,2,vhatHT)
round(vHT,2)
# Erster Wert ist negativ! Dies kann passieren beim Varianz Schaetzer von Horvitz-
    Thompson

#h) Alternativ Yates-Grundi Schaetzer
vhatYG <- function(s){
  n <- length(s)
  sl <- rep(NA,n)
  sk <- sl
  for (j1 in 1:n){
    k <- s[j1]
    for (j2 in 1:n) {
      l <- s[j2]
      sl[j2] <- Delta_kl[k,l]/pi_kl[k,l]*(Y[k]/pi_k[k]-Y[l]/pi_k[l])^2
    }
    sk[j1] <- sum(sl)
  }
  sum(sk)*(-1)/(2*N^2)
}
vYG <- apply(S,2,vhatYG)
round(vYG,2)

#i)
sum((ybar.hat-mean(Y))^2*ps) # wahrer Wert der Varianz des Schaetzers
sum(vHT*ps)
sum(vYG*ps)
```

Aufgabe 3: (a) Der π Schätzer für die Merkmalssumme vereinfacht sich zu:

$$\hat{t}_\pi = \sum_U I_k \frac{y_k}{\pi_k} = \sum_U I_k \frac{y_k}{n/N} = \frac{N}{n} \sum_U I_k y_k = \frac{N}{n} \sum_s y_k = N\bar{y}_s$$

Dieser ist unverzerrt, da

$$E\left(\frac{N}{n}\sum_s y_k\right) = \frac{N}{n} E\left(\sum_U I_k y_k\right) = \frac{N}{n}\sum_U y_k E(I_k) = \frac{N}{n}\sum_U y_k \frac{n}{N} = \sum_U y_k$$

(b) Die Varianz lässt sich umformen zu:

$$
\begin{aligned}
V(\hat{t}_\pi) &= \sum\sum_U (\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum_U \pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 + \sum\sum_{U,k\neq l}(\pi_{kl}-\pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \\
&= \sum_U \frac{n}{N}\left(1-\frac{n}{N}\right)\left(\frac{y_k}{n/N}\right)^2 + \sum\sum_{U,k\neq l}\left(\frac{n}{N}\frac{n-1}{N-1} - \frac{n}{N}\frac{n}{N}\right)\frac{y_k}{n/N}\frac{y_l}{n/N} \\
&= N^2\frac{n}{N}\left(1-\frac{n}{N}\right)\frac{1}{n^2}\sum_U y_k^2 + N^2\left(\frac{n}{N}\frac{n-1}{N-1} - \frac{n}{N}\frac{n}{N}\right)\frac{1}{n^2}\sum\sum_{U,k\neq l} y_k y_l \\
&= N^2\frac{1}{n}\left(\frac{1}{N} - \frac{n}{N^2}\right)\sum_U y_k^2 + N^2\frac{1}{n}\left(\frac{(n-1)}{N(N-1)} - \frac{n}{N^2}\right)\sum\sum_{U,k\neq l} y_k y_l \\
&= N^2\frac{1}{n}\left(\frac{N-n}{N^2}\right)\sum_U y_k^2 + N^2\frac{1}{n}\left(\frac{N^2(n-1)-nN(N-1)}{N^2N(N-1)}\right)\sum\sum_{U,k\neq l} y_k y_l \\
&= N^2\frac{1}{n}\left(\frac{N-n}{N^2}\right)\sum_U y_k^2 + N^2\frac{1}{n}\left(\frac{nN^2-N^2-nN^2+nN}{N^2N(N-1)}\right)\sum\sum_{U,k\neq l} y_k y_l \\
&= N^2\frac{1}{n}\left(\frac{N-n}{N-1}\frac{N-1}{N^2}\right)\sum_U y_k^2 + N^2\frac{1}{n}\left(\frac{-(N-n)}{N^2(N-1)}\right)\sum\sum_{U,k\neq l} y_k y_l \\
&= N^2\frac{1}{n}\frac{N-n}{N-1}\left(\left(\frac{1}{N} - \frac{1}{N^2}\right)\sum_U y_k^2 - \frac{1}{N^2}\sum\sum_{U,k\neq l} y_k y_l\right) \\
&= N^2\frac{1}{n}\frac{N-n}{N-1}\left(\frac{1}{N}\sum_U y_k^2 - \frac{1}{N^2}\sum_U y_k^2 - \frac{1}{N^2}\sum\sum_{U,k\neq l} y_k y_l\right) \\
&= N^2\frac{1}{n}\frac{N-n}{N-1}\left(\frac{1}{N}\sum_U y_k^2 - \frac{1}{N^2}\sum\sum_U y_k y_l\right) \\
&= N^2\frac{1-f}{n}S_{y_U}^2
\end{aligned}
$$

(c) Der Varianzschätzer ist unverzerrt, da:

$$
\begin{aligned}
E\left(\hat{V}(\hat{t}_\pi)\right) &= E\left(\sum\sum_s \check{\Delta}\check{y}_k\check{y}_l\right) = E\left(\sum\sum_U I_k I_l \check{\Delta}\check{y}_k\check{y}_l\right) = \sum\sum_U E(I_kI_l)\check{\Delta}\check{y}_k\check{y}_l \\
&= \sum\sum_U \pi_{kl}\frac{\Delta_{kl}}{\pi_{kl}}\check{y}_k\check{y}_l = \sum\sum_U \Delta_{kl}\check{y}_k\check{y}_l = V(\hat{t}_\pi)
\end{aligned}
$$

Aufgabe 4: (a) Für die Einschlusswahrscheinlichkeit gilt $Pr(\varepsilon_k < \pi_B) = \pi_k = \pi_B$. Für $k \neq l$ gilt, dass das Ereignis „k und l werden beide ausgewählt“ unabhängig ist, also I_k und I_l unabhängig und identisch verteilt sind. Somit ist der Einschlussindikator I_k Bernoulli verteilt mit Parameter π_B. Es gilt: $E(I_k) = \pi_B$, $V(I_k) = \pi_B(1-\pi_B) = \Delta_{kk}$ und für $k \neq l$: $Cov(I_k, I_l) = \pi_B^2 - \pi_B\pi_B = 0 = \Delta_{kl}$. Die Stichprobengröße ist zufällig und Binomial verteilt mit Parametern N und π_B, mit $E(n_s) = N\pi_B$ und $V(n_s) = N\pi_B(1-\pi_B)$. Somit ist das Stichprobendesign gegeben durch:

$$
p(s) = \underbrace{\pi_B \cdot \ldots \cdot \pi_B}_{n_s} \cdot \underbrace{(1-\pi_B)\cdot\ldots\cdot(1-\pi_B)}_{N-n_s} = \pi_B^{n_s}(1-\pi_B)^{N-n_s}
$$

(b) $Pr(n_s = n) = \binom{N}{n}\pi_B^n(1-\pi_B)^{N-n}$

(c) $\hat{t}_\pi = \frac{1}{\pi_B}\sum_s y_k$ mit Varianz $V_{BE}(\hat{t}_\pi) = \frac{1-\pi_B}{\pi_B}\sum_U y_k^2$

(d) $\sum_U y_k^2$ lässt sich umformen zu: $\sum_U y_k^2 = (N-1)S_{Y_U}^2 + N(\bar{y}_U)^2 = \left[1 - \frac{1}{N} + \frac{1}{(cv_{y_U})^2}\right] NS_{y_U}^2$. Um einen fairen Vergleich zu gewährleisten, setzen wir $E(n_s) = N\pi = n$, dann ist der Designeffekt gegeben durch:

$$deff = \frac{V_{BE}(\hat{t}_\pi)}{V_{SI}(\hat{t}_\pi)} = 1 - \frac{1}{N} + \frac{1}{(cv_{y_U})^2}.$$

Oft liegt der Variationskoeffizient zwischen $0.5 \leq cv_{y_U} \leq 1$, was einem Designeffekt von ungefähr 2 bis 5 entsprechen würde. Somit lässt sich zusammenfassen, dass das sogenannte Bernoulli Sampling (BE) oft weniger präzise für den π-Schätzer ist als die einfache Zufallsstichprobe ohne Zurücklegen (SI). Der Grund liegt in der zusätzlichen Variabilität in der Stichprobengröße. Dies kann man berücksichtigen und beispielsweise einen anderen unverzerrten Schätzer verwenden, z.B. $\hat{t}_{alt} = \frac{n}{n_s}\hat{t}_\pi$.

Aufgabe 5: Es gibt $M = \binom{N}{n} = \binom{5}{3} = 10$ mögliche Stichproben. In der Grundgesamtheit ist das Mittel gleich 2 und der Median gleich 1.

Stichprobe		Mittelwert	Median
1	1 2 3	4/3	1
2	1 2 4	5/3	1
3	1 2 5	9/3	3
4	1 3 4	4/3	1
5	1 3 5	8/3	3
6	1 4 5	9/3	3
7	2 3 4	2/3	1
8	2 3 5	6/3	1
9	2 4 5	7/3	1
10	3 4 5	6/3	1
$\sum$		20	16

Das Stichprobenmittel (20/10) ist erwartungstreu für den Merkmalsdurchschnitt, aber Stichprobenmedian ist nicht erwartungstreu für den Median der Grundgesamtheit.

Aufgabe 6: Es handelt sich um keine einfache Zufallsstichprobe vom Umfang $n = 1$, denn: $p_1 = \frac{500}{1000} =$, $p_2 = p_3 = 0.2$ und $p_4 = 0.1$, d.h. die Auswahlwahrscheinlichkeiten sind verschieden. Das Auswahlverfahren ist eine einfache Zufallsauswahl genau dann, wenn die Umfänge der Lieferungen alle gleich groß sind.

Aufgabe 7: $N = 676, n = 50$. Es gilt: $\hat{t}_\pi = N\bar{y}_s = N\frac{1}{n}\sum_s f_k y_k = \frac{1471}{50} = 676 \cdot 29.42 = 19887.92$. $V(\hat{t}_\pi) = N^2\frac{1-n/N}{n}S_{y_s}^2 = 676^2\frac{1-50/676}{50}\frac{1}{49}(54497 - 50 \cdot 29.42^2) = 1937990$. 80% Konfidenzintervall: $\hat{t}_\pi \pm 1.28 \cdot \sqrt{1937990} = [18103.84; 21672]$.

Aufgabe 8: Der R-Code könnte folgendermaßen aussehen:

```
1 psid <- read.csv2("psid.csv")
```

```
N <- nrow(psid)
n <- 20
Y <- psid$wage
f <- n/N
Ybar <- mean(Y); Ybar
V.Ybar <- (1-f)/n*var(Y); V.Ybar

f_var <- function(y,N){
  n <- length(y)
  f <- n/N
  return((1-f)/n*var(y))
}

B <- 100000
m <- rep(NA,B)
v <- rep(NA,B)
for (b in 1:B) {
  y <- sample(Y,n)
  m[b] <- mean(y)
  v[b] <- f_var(y,N)
}

vm <- v/10^6
V.Ybarm <- V.Ybar/10^6

plot(density(m),lwd=2,xlab='Schaetzwert',main='')
arrows(Ybar,5e-06,Ybar,0,length=0.1,angle=25)
text(Ybar,7.5e-06,expression(bar(y)[U]))

plot(density(vm),lwd=2,xlab='Schaetzwert in Millionen',main='')
arrows(V.Ybarm+1000,0.002,V.Ybarm,0,length=0.1,angle=25)
text(V.Ybarm+1000,0.0026,expression(V(bar(y)[U])))
```

Bei der Normalverteilung ist das Konfidenzintervall im Allgemeinen $\hat{\theta} \pm u_{1-\alpha/2}[V(\hat{\theta})]^{1/2}$. Wir treffen damit implizit Aussagen über Asymptotische Eigenschaften des Schätzers auf Grundlage eines Zentralen Grenzwertsatzes. Auch für abhängig identisch verteilte Zufallsvariablen (wie beim Ziehen ohne Zurücklegen) gibt es solch einen Satz, allerdings gilt, dass falls die Verteilung von Y stark schief ist (wie im Beispiel der Fall), dann benötigen wir einen sehr hohen Stichprobenumfang für die Konvergenz zur Normalverteilung. In diesem Fall ist es folglich sinnvoller sich die Konfidenzintervalle zu simulieren/bootstrapen.

Aufgabe 9: (a) Der π-Schätzer in einfachen Zufallsstichproben für einen Anteil ist gerade der π-Schätzer für den Durchschnitt. Also: $\hat{P} = \hat{\bar{y}}_\pi = \frac{1}{n}\sum_s y_k$. Die Varianz ist gegeben durch:

$$V(\hat{P}) = \frac{1-f}{n} S^2_{y_U} = \frac{N-n}{Nn}\frac{1}{N-1}\left(\sum_U y_k^2 - \bar{y}\right) = \frac{1}{n}\frac{N-n}{N-1}\left(\frac{1}{N}\sum_U y_k^2 - \left(\frac{1}{N}\sum_U y_k\right)^2\right)$$
$$= \frac{1}{n}\frac{N-n}{N-1}\left(\frac{1}{N}\sum_U y_k\right)\left(1 - \sum_U y_k\right) = \frac{1}{n}\frac{N-n}{N-1}P(1-P)$$

Dies kann unverzerrt geschätzt werden mit

$$\hat{V}(\hat{P}) = \frac{1-f}{n} S^2_{y_s} = \frac{1-f}{n}\frac{n}{n-1}\left(\frac{1}{n}\sum_s y_k\right)\left(1 - \frac{1}{n}\sum_s y_k\right) = \frac{1-f}{n-1}\hat{P}(1-\hat{P})$$

(b) Der R-Code könnte folgendermaßen aussehen:

```r
psid <- read.csv2('psid.csv')
N <- nrow(psid)
B <- 10000
n <- 30
f <- n/N
e <- rep(NA,B)
v <- rep(NA,B)
Y <- psid$sector==7
for (i in 1:B){
  y <- sample(Y,n)
  e[i] <- mean(y)
  v[i] <- (1-f)/n*var(y)
}
plot(density(e))
# grob normal verteilt, insbesondere bei hoeheren n
plot(density(v))
# linksschief, definitiv nicht normalverteilt
```

Aufgabe 10: Sei $\hat{P} = \hat{\bar{y}}_\pi = \frac{1}{n}\sum_s y_k$ der erwartungstreue Schätzer für den Anteil mit Varianz $V(\hat{P}) = \frac{1}{n}\frac{N-n}{n-1}P(1-P)$. Wir können den Korrekturfaktor vernachlässigen, da $N \approx 80$ Millionen, also $V(\hat{P}) \approx \frac{1}{n}P(1-P)$. Dann gilt für das Konfidenzintervall (unter Annahme der Normalverteilung): $P \pm \sqrt{V(\hat{P})}u_{1-\alpha/2}$. Länge des Konfidenzintervalls ist $2\sqrt{V(\hat{P})}$, dies soll gleich $0.1P$ sein, also:

$$2\sqrt{V(\hat{P})}u_{1-\alpha/2} = 2\sqrt{\frac{P(1-P)}{n}}u_{1-\alpha/2} = 0.1P$$
$$\Leftrightarrow n = \frac{P(1-P)u_{1-\alpha/2}^2}{0.05^2 P^2}$$

Mit $\alpha = 0.05$ gilt dann

(a) $P = 0.5 \Rightarrow n \approx 1537$

(b) $P = 0.175 \Rightarrow n \approx 7244$

(c) $P = 0.05 \Rightarrow n \approx 29196$

Aufgabe 11: (a) Laut Vorlesung gilt für $N = an$:

$$\begin{aligned} V(\hat{t}_\pi) &= N \cdot SSB = N(SST - SSW) = \frac{N(N-a)}{N-1}SST\left(\frac{N-1}{N-a} - 1 + 1 - \frac{N-1}{N-a}\frac{SSW}{SST}\right) \\ &= N(N-a)\frac{SST}{N-1}\left(\frac{N-1-N+a}{N-a} + \delta\right) \\ &\overset{N=an}{=} Na(n-1)S_{y_U}^2\left(\frac{a-1}{a(n-1)} + \delta\right) \\ &= \frac{N^2}{n}S_{y_U}^2\left(1 - \frac{1}{a} + (n-1)\delta\right) \end{aligned}$$

Je homogener die Elemente in einer systematischen Stichprobe sind, desto weniger effizient ist der Schätzer.

(b) $V_{SI} = N^2 \frac{1-f}{n} S^2_{y_U}$ und $V_{SY} = \frac{N^2 S^2_{y_U}}{n}[(1-f) + (n-1)\delta] = N^2 \frac{1-f}{n} S^2_{y_U} + \frac{N^2 S^2_{y_U}}{n}(n-1)\delta$. Der Designeffekt ist dann:

$$deff = \frac{V_{SY}}{V_{SI}} = 1 + \frac{n-1}{1-f}\delta$$

Systematisches Sampling ist effizienter als die einfache Zufallsstichprobe falls $\delta < 0$.

(c) In der Praxis müssen wir versuchen (falls möglich) die Grundgesamtheit so anzuordnen, dass die Merkmalswerte y_k innerhalb der systematischen Stichproben so heterogen wie möglich sind (z.B. durch Nachbarschaften, linearen Trend,...)

Aufgabe 12: Der R Code könnte folgendermaßen aussehen:

```
reisanbau <- read.csv2('reisanbau.csv')
N <- 892
n <- 25
X.dot <- 568565
sum(reisanbau$Reisflaeche/reisanbau$Flaeche)
y.sum <- X.dot * sum(reisanbau$Reisflaeche/reisanbau$Flaeche)/n
y.sum
var(reisanbau$Reisflaeche/reisanbau$Flaeche)
var.y.sum <- X.dot^2 * var(reisanbau$Reisflaeche/reisanbau$Flaeche)/n
var.y.sum
sqrt(var.y.sum)

lower <- y.sum - sqrt(var.y.sum)*qnorm(0.975)
upper <- y.sum + sqrt(var.y.sum)*qnorm(0.975)
cbind(lower, upper)
```

Aufgabe 13: Der R-Code könnte folgendermaßen aussehen:

```
library(pps)
B <- 10000; N <- 5; n <- 3
e_sample <- matrix(NA,B,n)
e_sampford <- matrix(NA,B,n)
p <- 4:8/sum(4:8);p

for (i in 1:B){
  e_sample[i,] <- sample (1:N,n,prob=p)
  e_sampford[i,] <- sampford(p,n)
}
pi_emp_sample <- rep(NA,N)
pi_emp_sampford <- rep(NA,N)

for (i in 1:N){
  pi_emp_sample[i] <- sum(apply(e_sample,1, function(z) i%in%z))
  pi_emp_sampford[i] <- sum(apply(e_sampford,1, function(z) i%in%z))
}

rbind(p*n,round(pi_emp_sample/B,3),round(pi_emp_sampford/B,3))
```

Aufgabe 14: Der R-Code könnte folgendermaßen aussehen:

```
library(samplingbook)
```

```r
data(influenza)
summary(influenza)

# 1) pps.sampling
pps <- pps.sampling(z=influenza$population,n=20,method='sampford')
pps
sample <- influenza[pps$sample,]
sample

# 2) htestimate
pps <- pps.sampling(z=influenza$population,n=20,method='midzuno')
sample <- influenza[pps$sample,]
N <- nrow(influenza)

# Exakte Varianzberechnung
PI <- pps$PI
htestimate(sample$cases, N=N, PI=PI, method='ht')
htestimate(sample$cases, N=N, PI=PI, method='yg')
# Approximierte Varianzschaetzung
pk <- pps$pik[pps$sample]
htestimate(sample$cases, N=N, pk=pk, method='hh')
pik <- pps$pik

# Konfidenzintervale basierend auf der Normalverteilung
est.ht <- htestimate(sample$cases, N=N, PI=PI, method='ht')
est.ht$mean*N
lower <- est.ht$mean*N - qnorm(0.975)*N*est.ht$se
upper <- est.ht$mean*N + qnorm(0.975)*N*est.ht$se
c(lower,upper)
# Wahrer Wert an Grippeerkrankungen
sum(influenza$cases)
```

Aufgabe 15: (a)

$$
\begin{aligned}
\pi_k &= Pr(u_k \text{ in } 1) + Pr(u_k \text{ in } 2 \text{ und nicht in } 1) \\
&= p_k + Pr(u_k \text{ nicht in } 1) \cdot Pr(u_k \text{ in } 2 | u_k \text{ nicht in } 1) \\
&= p_k + \sum_{l=1, l\neq k}^{N} Pr(u_l, l \neq k, \text{ in } 1) \cdot Pr(u_k \text{ in } 2 | u_l, l \neq k \text{ in } 1) \\
&= p_k + \sum_{l=1, l\neq k}^{N} \left(\frac{x_l}{\sum_{j=1}^{N} x_j} - \frac{x_k}{\sum_{j=1}^{N} x_j - x_l} \right) \\
&= p_k + \sum_{l=1, l\neq k}^{N} \left(p_l \cdot \frac{\sum_{j=1}^{N} x_j}{\sum_{j=1}^{N} x_j} \cdot \frac{x_k}{\sum_{j=1}^{N} x_j - x_l} \right) \\
&= p_k + \sum_{l=1, l\neq k}^{N} \left(p_l \cdot p_k \cdot \frac{1}{1-p_l} \right) \\
&= p_k \left(1 + \sum_{k\neq l} p_l (1-p_l)^{-1} \right)
\end{aligned}
$$

(b)

$$\begin{aligned}\pi_{kl} &= Pr(u_k \text{ in } 1)Pr(u_l \text{ in } 2|u_k \text{ in } 1) + Pr(u_l \text{ in } 1)Pr(u_k \text{ in } 2|u_l \text{ in } 1)\\ &= p_k \cdot \frac{x_l}{\sum_{j=1}^{N} x_j - x_k} + p_l \cdot \frac{x_k}{\sum_{j=1}^{N} x_j - x_l}\\ &= p_k \frac{p_l}{1-p_k} + p_l \frac{p_k}{1-p_l}\\ &= p_k p_l (\frac{1}{1-p_k} + \frac{1}{1-p_l})\end{aligned}$$

Aufgabe 16: $N = 320$, $n = 4$, $p_1 = 1.2/80 = p_2$, $p_3 = 0.2/80$, $p_4 = 0.5/80$. Der Hansen-Hurwitz Schätzer für den Durchschnitt ist:

$$\hat{\bar{y}}_{HH} = \frac{1}{Nn} \sum_{k=1}^{N} \frac{y_k}{p_k} = 3.021$$

Die geschätzte Varianz lautet:

$$\hat{V}(\hat{\bar{y}}_{HH}) = \frac{1}{N^2 \frac{1}{n(n-1)}} \sum_{k=1}^{N} \left(\frac{y_k}{p_k} - \frac{1}{n} \sum_{j=1}^{N} \frac{y_j}{p_j} \right)^2 = 2.326$$

Aufgabe 17: Für Hansen-Hurwitz: $y_1/p_1 = 36.67$, $y_2/p_2 = 30$ und $y_3/p_3 = 50$. Für Horvitz-Thompson: $\pi_k = 1 - (1 - p_k)^m$, also $\pi_1 = 0.51$, $\pi_2 = 0.36$ und $\pi_3 = 0.75$. Damit dann $y_1/\pi_1 = 36.67$, $y_2/\pi_2 = 30$ und $y_3/\pi_3 = 50$. Zusammenfassend:

Stichprobe	Auswahl-Wkeit	Stipro-Werte	HH	HT
1,1	0.09	(11,11)	36,67	21.57
2,2	0.04	(6,6)	30	16.67
3,3	0.25	(25,25)	30	33.33
1,2	0.06	(11,6)	33.33	38.24
2,1	0.06	(6,11)	33.33	38.24
1,3	0.15	(11,25)	43.33	54.90
3,1	0.15	(25,11)	43.33	54.90
2,3	0.10	(6,25)	40	50
3,2	0.10	(25,6)	40	50

Der Mittelwert von HH ist 42 mit Std-Abweichung 5.89, während für HT der Mittelwert 42 mit Std-Abweichung 12.10 ist.

Aufgabe 18:

$$\begin{aligned}(N-1)S_{y_U}^2 &= \sum_{h=1}^{H}\sum_{k=1}^{N_h}(y_{h_k} - \bar{y}_U)^2 = \sum_{h=1}^{H}\sum_{k=1}^{N_h}\left[(y_{h_k} - \bar{y}_{U_h}) + (\bar{y}_{U_h} - \bar{y}_U)\right]^2\\ &= \sum_{h=1}^{H}\sum_{k=1}^{N_h}(y_{h_k} - \bar{y}_{U_h})^2 + \sum_{h=1}^{H}\sum_{k=1}^{N_h}(\bar{y}_{U_h} - \bar{y}_U)^2 + 2\sum_{h=1}^{H}\sum_{k=1}^{N_h}(y_{h_k} - \bar{y}_{U_h})(\bar{y}_{U_h} - \bar{y}_U)\end{aligned}$$

Da $2\sum_{h=1}^{H}\sum_{k=1}^{N_h}(y_{h_k}-\bar{y}_{U_h})(\bar{y}_{U_h}-\bar{y}_U) = 2\sum_{h=1}^{H}(\bar{y}_{U_h}-\bar{y}_U)\underbrace{\sum_{k=1}^{N_h}(y_{h_k}-\bar{y}_{U_h})}_{=0} = 0$, folgt:

$$\begin{aligned}(N-1)S_{y_U}^2 &= \sum_{h=1}^{H}\sum_{k=1}^{N_h}(y_{h_k}-\bar{y}_{U_h})^2 + \sum_{h=1}^{H}\sum_{k=1}^{N_h}(\bar{y}_{U_h}-\bar{y}_U)^2 \\ &= \sum_{h=1}^{H}(N_h-1)S_{y_{U_h}}^2 + \sum_{h=1}^{H}N_h(\bar{y}_{U_h}-\bar{y}_U)^2\end{aligned}$$

Aufgabe 19: (a) Wahlprognose: $\hat{P} = \sum_{h=1}^{3} W_h\hat{P}_h = 0.23$.
Durchschnittsalter: $\hat{\bar{y}}_\pi = \sum_{h=1}^{H} W_h\bar{y}_h = 46.1$.
Da n_h nicht gegeben ist, sind die Konfidenzintervalle nicht berechenbar.

(b) Für die Optimale Aufteilung gilt $n_h = n\frac{N_h S_{y_{U_h}}}{\sum_{i=1}^{N} W_i S_{y_{U_i}}}$.

- Wahlprognose: Wir benötigen die Varianz der Personen, die die OP wählen, in den jeweiligen Schichten. Es gilt:

$$S_{y_{U_h}}^2 = \frac{1}{N_h-1}\sum_{U_h}(y_k-\bar{y}_{u_h})^2 = \left(\frac{1}{N_h-1}\sum y_k^2\right) - \bar{y}_{U_h}^2 = \left(\frac{N_h}{N_h-1}\frac{1}{N_h}\sum_{U_h} y_k^2\right) - \left(\frac{1}{N_h}\sum_{U_h} y_k\right)^2$$

  Da y_k nur Werte 0 oder 1 annimmt, gilt $\frac{1}{N_h}\sum_{U_h} y_k^2 = \frac{1}{N_h}\sum_{U_h} y_k = \hat{P}_h$. Außerdem ist $N_h/(N_h-1) \approx 1$. Folglich:

$$S_{y_{U_h}}^2 = \hat{P}_h - \hat{P}_h^2 = \hat{P}_h(1-\hat{P}_h)$$

  Somit errechnen wir: $n_1 = n\cdot 0.2411$, $n_2 = n\cdot 0.4393$ und $n_3 = n\cdot 0.3169$.
- Durchschnittsalter: Schätze $S_{y_{U_h}}$ durch s_h, dann ergibt sich: $n_1 = n\cdot 0.2065$, $n_2 = n\cdot 0.5016$ und $n_3 = n\cdot 0.2920$.

Aufgabe 20: (a) Bei proportionaler Aufteilung gilt $n_h = nW_h$. Hier: $n = 100$ und Gesamtanzahl an Bauernhöfen: 2010

Schicht	1	2	3	4	5	6	7	$\sum$
W_h	0.196	0.229	0.195	0.166	0.084	0.056	0.074	1
n_h	20	23	19	17	8	6	7	100

(b) Bei optimaler Aufteilung gilt $\frac{n_h}{n} = \frac{N_h s_h}{\sum_i N_i s_i}$.

Schicht	1	2	3	4	5	6	7
n_h (ungerundet)	9.6	17.9	17.3	19.3	12.1	8.6	15.2
n_h (gerundet)	10	18	17	19	12	9	15

Genauigkeit:
Bei Schichtschätzung: $\hat{V}(\hat{\bar{y}}) = \sum_{h=1}^{H}\frac{N_h^2}{N^2}\frac{1-f_h}{n_h}S_{y_{s_h}}^2$. Somit $\hat{V}_{prop} = 3.2620$ und $\hat{V}_{opt} = 2.7254$.
Bei der einfachen Zufallsstichprobe gilt $\hat{V}_{EZ} = \frac{1}{n}\left(1-\frac{n}{N}\right)S_{y_s}^2$. Weiter gilt die Varianzzerlegungsformel $(N-1)S_{y_s}^2 = \sum_{h=1}^{H}(N_h-1)S_h^2 + \sum_{h=1}^{H}N_h(\bar{y}_{s_h}-\bar{y}_s)^2$. Nun ist $\bar{y}_s =$

$\frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h = 26.3168$ und $S^2_{y_s} = 1243156/(N-1) = 618.7935$, dh. $\hat{V}_{EZ} = 5.880078$. Effizienzvergleich:

$$\frac{\hat{V}_{EZ}}{\hat{V}_{prop}} = 1.8026 \qquad \frac{\hat{V}_{EZ}}{\hat{V}_{opt}} = 2.1575$$

Aufgabe 21: (a) $\frac{n_h}{n} = \frac{W_h s_h/\sqrt{c_h}}{\sum_i W_i s_i/\sqrt{c_i}}$, also: $n_1/n = 1/3$ und $n_2/n = 2/3$.

(b) $n_1 = 1/3n$, $n_2 = 2/3n$, vernachlässige f_1 und f_2:

$$\begin{aligned} V(\hat{\bar{y}}_U) &= \frac{1}{n_1}\frac{N_1}{N^2}(1-f_1)S^2_{y_1} + \frac{1}{n_2}\frac{N_2}{N^2}(1-f_2)S^2_{y_2} \\ &\approx \frac{1}{n_1}W_1^2 s_1^2 + \frac{1}{n_2}W_2^2 s_2^2 = \frac{1}{n}(3W_1^2 S_1^2 + \frac{3}{2}W_2^2 S_2^2) \stackrel{!}{=} 1 \\ &\Leftrightarrow n = 264, \text{ d.h. } n_1 = 264/3 = 88, n_2 = n - n_1 = 176 \end{aligned}$$

(c) Erhebungskosten: $C = 88 \cdot 4 + 176 \cdot 9 = 1936$

Aufgabe 22: (a) Es gilt $n = 400$, $W_1 = 0.62$ (männliche Fach-oder Hilfsarbeiter), $W_2 = 0.31(weibliche Schreibkräfte)$ und $W_3 = 0.07$ (Angestellte mit leitenden Aufgaben). Vermutungen: $P_1 \in [0.4, 0.45]$, $P_2 \in [0.2, 0.3]$ und $P_3 \in [0.05, 0.1]$. Deshalb Annahme:

- $P_1 = 0.425 \Rightarrow s_1^2 \approx P_1(1-P_1) = 0.2444$
- $P_2 = 0.25 \Rightarrow s_2^2 \approx P_2(1-P_2) = 0.1875$
- $P_3 = 0.075 \Rightarrow s_3^2 \approx P_3(1-P_3) = 0.0694$

Optimale Aufteilung: $n_1 = n\frac{W_1 s_1}{sum_{i=1}^{3} W_i s_i} = 267, n_2 = 117, n_3 = 16$

(b) $P_1 = 0.48$, $P_2 = 0.21$, $P_3 = 0.04$:

$$\begin{aligned} \sqrt{V(\hat{P})} &= \left(\sum_{h=1}^{3} W_h^2 \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) s_h^2\right)^{1/2} \\ &= \left(\sum_{h=1}^{3} W_h^2 \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) \frac{N_h}{N_h - 1} P_h(1-P_h)\right)^{1/2} \\ &\approx \left(\sum_{h=1}^{3} W_h^2 \frac{1}{n_h} P_h(1-P_h)\right)^{1/2} \end{aligned}$$

Mit $P_1 = 0.48 \Rightarrow s_1^2 \approx P_1(1-P_1) = 0.2496$, $P_2 = 0.21 \Rightarrow s_2^2 \approx P_2(1-P_2) = 0.1659$ und $P_3 = 0.04 \Rightarrow s_3^2 \approx P_3(1-P_3) = 0.0384$. Daraus folgt, dass $n_1 = 276, n_2 = 112, n_3 = 12$. Somit $\sqrt{V(\hat{P})} = 0.02248$.

(c) $S^2 = \frac{1}{N-1}\sum_{h=1}^{H}(N_h - 1)S_h^2 + \frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{y}_h - \bar{y}_U)^2 \approx \sum_{h=1}^{H} W_h S_h^2 + \sum_{h=1}^{H} W_h(P_h - P)^2$ mit $P = \sum_{h=1}^{H} W_h P_h = 0.3655$. Die Varianz bei der einfachen Zufallsstichprobe ist ungefähr $V(\hat{P}) \approx 1/nS^2 = 0.02408^2$. Somit ist die gesuchte Standardabweichung 0.02408.

Aufgabe 23: $H = 2, N_1 = 63, N_2 = 57, n_1 = 6, n_2 = 6$

(a) Körpergröße:
$\bar{y}_1 = 185, \bar{y}_2 = 169,67$. $\hat{\bar{y}} = \sum_{h=1}^{2} \frac{N_h}{N} \bar{y}_{s_h} = 177.72$
Jeansträger:
$\bar{y}_1 = 3/6$, $\bar{y}_2 = 2/6$, $\hat{P} = \frac{1}{120}(63\frac{3}{6} + 57\frac{2}{6}) = 0.4208$

(b) Körpergröße:
$s_1^2 = 35.2$, $s_2^2 = 15.86$, d.h. $\hat{V}(\hat{\bar{y}}) = 6.225$
Jeansträger:
$s_1^2 = 0.3, s_2^2 = 0.267$, d.h. $\hat{V}(\hat{\bar{y}}) = 0.0201$

(c) Körpergröße:
$\bar{y} = 177.3$, $\hat{V}(\hat{\bar{y}}) = 6.55$
Jeansträger: $\hat{P} = 0.4167$, $\hat{V}(\hat{P}) = 0.01988$

Aufgabe 24:

Aufgabe 25: $N > 80$ Millionen, Auswahlsatz vernachlässigbar. Definiere

$$y_k = \begin{cases} 1 & \text{, falls } u_k \text{ die OP jetzt wählt} \\ 0 & \text{, sonst} \end{cases}$$

und

$$x_k = \begin{cases} 1 & \text{, falls } u_k \text{ die OP früher gewählt hat} \\ 0 & \text{, sonst} \end{cases}$$

(a) Keine Vorinformation, also freie Schätzung: $\hat{P} = \frac{80}{200} = 0.4$ und

$$\begin{aligned} \hat{V}(\hat{P}) = \frac{1}{n}(1 - n/N)S_y^2 &\approx \frac{1}{n(n-1)} \left(\sum_{k=1}^{n} y_k^2 - n\bar{y}_s^2 \right) \\ &= \frac{1}{n(n-1)} \left(n\hat{P} - n\hat{P}^2 \right) = \frac{1}{n-1}\hat{P}(1-\hat{P}) = 0.0012 \end{aligned}$$

(b) Vorinformationen: In der Grundgesamtheit: $\bar{x}_U = 0.25$. In der Stichprobe: $\bar{x}_s = \frac{60}{200} = 0.3$

- Differenzenschätzung: $\hat{P} = \bar{y}_s - \bar{x}_s + \bar{x}_U = 0.4 - 0.3 + 0.25 = 0.35$

$$\begin{aligned} \hat{V}(\hat{P}) \approx \frac{1}{n}(S_{y_s}^2 + S_{x_s}^2 - 2S_{xy_s}) &= \frac{1}{n(n-1)} \sum_{k=1}^{n} (y_k - x_k - \bar{y}_s + \bar{x}_s)^2 \\ &= \frac{1}{n(n-1)} \left(\sum_{k=1}^{n} (y_k - x_k)^2 - n(\bar{y}_s + \bar{x}_s)^2 \right) \\ &= \frac{1}{n(n-1)} \left(20 - 200 * 20^2/200^2 \right) = 0.00045 \end{aligned}$$

  da 20 mal $(y_k = 1, x_k = 0)$ während $(y_k = 0, x_k = 1)$ kein mal auftritt.
- Verhältnisschätzung: $\hat{P} = \bar{x}_U \frac{\bar{y}_s}{\bar{x}_s} = 0.25 \frac{0.4}{0.33}$ Approximierte Varianz ist

$$A\hat{V}(\hat{P}) = \frac{1}{n}(1 - n/N) \left(S_{y_s}^2 + \hat{r} S_{x_s}^2 - 2\hat{r} S_{xy_s} \right) \approx \frac{1}{n} \left(S_{y_s}^2 + \hat{r} S_{x_s}^2 - 2\hat{r} S_{xy_s} \right)$$

mit

$$\hat{r} = 0.4/0.3$$
$$S^2_{y_s} = \frac{n}{n-1}\hat{P}(1-\hat{P}) = 0.2412$$
$$S^2_{x_s} = \frac{n}{n-1}\hat{P}_x(1-\hat{P}_x) = 200/199 \cdot 0.3 \cdot 0.7 = 0.2111$$
$$S_{xy_s} = \frac{1}{n-1}(\sum x_i y_i - n\bar{x}\bar{y}) = \frac{n}{n-1}(\bar{x} - \bar{x}\bar{y}) = \frac{n}{n-1}\hat{P}_x(1-\hat{P}) = 200/199 \cdot 0.3 \cdot 0.6 = 0.1809$$

Damit ist die approximierte Varianz gleich 0.00067