Stichprobenverfahren
– Klausur –

Dr. Willi Mutschler

10. August 2017

Bitte ausfüllen:

Name, Vorname:	Geburtsdatum:
Matrikelnummer:	Unterschrift:

Hinweise:
Die Klausur besteht aus acht Aufgaben, die alle zu bearbeiten sind.

Sie dürfen einen Taschenrechner und ein einseitig handschriftlich beschriebenes DIN A4 Blatt als Hilfsmittel verwenden. Dieses wird am Ende eingesammelt, jedoch nicht bewertet.

Die Bearbeitungszeit beträgt 90 Minuten.

Bitte schreiben Sie Ihren Namen auf jedes Blatt!

Ergebnis:

Punkte	Note	Unterschrift

Aufgabe 1: Einfache Zufallsstichprobe ohne Zurücklegen

Betrachten Sie eine Grundgesamtheit mit zu untersuchendem Merkmal Y mit folgenden Werten:

Y :	1	2	4	3	5	7	6	8	9

Der Mittelwert und die Varianz von Y in der Grundgesamtheit sind $\bar{Y}_U = 5$ und $S^2_{y_U} = 7.5$.

Betrachten Sie eine einfache Zufallsstichprobe ohne Zurücklegen der Größe $n = 6$.

(a) Berechnen Sie die Anzahl aller möglichen Stichproben.

(b) Wie ist die Horvitz-Thompson Schätzfunktion für den Mittelwert der Grundgesamtheit in diesem Fall definiert? Zeigen Sie, dass diese unverzerrt ist.

(c) Berechnen Sie die Varianz des Horvitz-Thompson Mittelwertschätzers für obige Daten.

Antwort:

Aufgabe 1: 5 Punkte

(a) $M = \binom{9}{6} = 84$ [1 Punkt]

(b) Hier Einschlusswahrscheinlichkeit:$E(I_k) = \pi_k = \frac{n}{N}$, somit HT-Schätzer:

$$\hat{\bar{y}} = \frac{1}{N}\sum_s \frac{y_k}{\pi_k} = \frac{1}{N}\sum_s \frac{y_k}{\frac{n}{N}} = \frac{1}{n}\sum_s y_k = \bar{y}_s$$

Erwartungstreu, da:

$$E(\hat{\bar{y}}) = E[\frac{1}{n}\sum_U I_k y_k] = \frac{1}{n}\sum_U y_k E(I_k) = \frac{1}{n}\sum_U y_k \frac{n}{N} = \frac{1}{N}\sum_U y_k = \bar{y}_U$$

[3 Punkte]

(c) $V(\hat{\bar{y}}) = \frac{1-\frac{n}{N}}{n} S^2_{y_U} = 0.4167$ [1 Punkt]

Aufgabe 2: Geschichtete Zufallsstichprobe

Betrachten Sie eine Grundgesamtheit mit zu untersuchendem Merkmal Y sowie Attribut X, welche folgende Werte annehmen:

Y :	1	2	4	3	5	7	6	8	9
X :	1	1	1	2	2	2	3	3	3

Der Mittelwert und die Varianz von Y in der Grundgesamtheit sind $\bar{Y}_U = 5$ und $S^2_{y_U} = 7.5$.

Betrachten Sie eine geschichtete Zufallsstichprobe der Größe $n = 6$, wobei X die Schichtvariable darstellt. Nehmen Sie hierzu an, dass innerhalb der Schichten eine einfache Zufallsstichprobe mit identischer Größe $n_h = 2$ gezogen wird.
Hinweis: Für die Schichtmittel und Schichtvarianzen gilt:

$$\bar{Y}_{U_1} = 7/3, \qquad \bar{Y}_{U_2} = 5, \qquad \bar{Y}_{U_3} = 23/3,$$
$$S^2_{y_{U_1}} = 7/3, \qquad S^2_{y_{U_2}} = 4, \qquad S^2_{y_{U_3}} = 7/3.$$

(a) Berechnen Sie die Anzahl aller möglichen Stichproben.

(b) Wie ist die Horvitz-Thompson Schätzfunktion für den Mittelwert der Grundgesamtheit in diesem Fall definiert? Zeigen Sie, dass diese unverzerrt ist.

(c) Berechnen Sie die Varianz des Horvitz-Thompson Mittelwertschätzers für obige Daten.

(d) Was versteht man unter dem Schichtungsprinzip? Wie sehe die optimale Einteilung in drei Schichten für die obigen Daten aus?

Antwort:

Aufgabe 2: 8 Punkte

(a) $M = \left[\binom{3}{2}\right]^3 = 27$ [1 Punkt]

(b) Hier Einschlusswahrscheinlichkeit:$E(I_{hk}) = \pi_{hk} = \frac{n_h}{N_h}$, somit HT-Schätzer:

$$\hat{\bar{y}} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_{s_h} = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{n_h} y_k$$

Erwartungstreu, da:

$$E(\hat{\bar{y}}) = E[\frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} I_{hk} y_k] = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} E(I_{hk}) y_k = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{1}{n_h}\sum_{k=1}^{N_h} \frac{n_h}{N_h} y_k = \frac{1}{N}\sum_{h=1}^{H}\sum_{k=1}^{N_h} y_k = \bar{y}_U$$

[4 Punkte]

(c) $V(\hat{\bar{y}}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \frac{1-\frac{n_h}{N_h}}{n_h} S_{y_{U_h}}^2 = 0.1605$ [1 Punkt]

(d) Die Schichten sollten so gewählt werden, dass die Variablen (oder Merkmalsträger) innerhalb einer Schicht so ähnlich wie möglich sind. Die einzelnen Schichten sollten sich untereinander so weit wie möglich unterscheiden.

Eine optimale Aufteilung wäre demnach:

Y :	1	2	4	3	5	7	6	8	9
X :	1	1	2	1	2	3	2	3	3

[2 Punkte]

Aufgabe 3: Cluster Zufallsstichprobe

Betrachten Sie eine Grundgesamtheit mit zu untersuchendem Merkmal Y sowie Attribut X, welche folgende Werte annehmen:

Y :	1	2	4	3	5	7	6	8	9
X :	1	1	1	2	2	2	3	3	3

Der Mittelwert und die Varianz von Y in der Grundgesamtheit sind $\bar{Y}_U = 5$ und $S^2_{y_U} = 7.5$.

Betrachten Sie nun eine Zufallsstichprobe mithilfe von Clustern, wobei X die Clusterzugehörigkeitsvariable darstellt. Die Stichprobe soll aus $n_I = 2$ Clustern bestehen, wobei diese mit identischen Einschlusswahrscheinlichkeiten gezogen werden.

(a) Was lässt sich allgemein über die Stichprobengröße n bei einer Cluster-Stichprobe aussagen? Gilt dies auch in dem vorliegenden Fall?

(b) Wie ist die Horvitz-Thompson Schätzfunktion für den Mittelwert der Grundgesamtheit in diesem Fall definiert? Zeigen Sie, dass diese unverzerrt ist.

(c) Berechnen Sie die Varianz des Horvitz-Thompson Mittelwertschätzers für obige Daten.

(d) Was versteht man unter dem Clusterprinzip? Wie sehe die optimale Aggregierung der Merkmalsträger in drei Cluster für die obigen Daten aus?

Antwort:

Aufgabe 3: 10 Punkte

(a) Normalerweise ist n eine Zufallsgröße, da die Anzahl an Merkmalsträgern innerhalb der ausgewählten Cluster unterschiedlich sein kann. Hier haben jedoch alle Cluster eine identische Größe, somit ist $n = 2 \cdot 3 = 6$ fix. [2 Punkte]

(b) Hier Einschlusswahrscheinlichkeit:$E(I_{Ii}) = \pi_{Ii} = \frac{n_I}{N_I}$, somit HT-Schätzer:

$$\hat{\bar{y}} = \frac{1}{N}\hat{t}_\pi = \frac{1}{N}\sum_{s_I}\frac{t_i}{\pi_{Ii}} = \frac{1}{N}N_I\sum_{s_I}\frac{t_i}{n_I} = \frac{1}{N}N_I\sum_{s_I}\sum_{U_i}\frac{y_k}{n_I}$$

Erwartungstreu, da:

$$E(\hat{\bar{y}}) = E[\frac{1}{N}N_I\sum_{s_I}\sum_{U_i}\frac{y_k}{n_I}] = \frac{N_I}{N}\sum_{s_I}\sum_{U_i}y_k\frac{E(I_{Ik})}{n_I} = \frac{N_I}{N}\sum_{s_I}\sum_{U_i}y_k\frac{n_I/N_I}{n_I} = \frac{1}{N}\sum_{s_I}\sum_{U_i}y_k = \frac{1}{N}\sum_{U}y_k = \bar{y}_U$$

[4 Punkte]

(c) Es gilt (mit den Hilfswerten aus dem Aufgabentext zu Aufgabe 2): $\bar{t}_{U_I} = \frac{7}{3} + 5 + \frac{23}{3} = 15$ und $S^2_{tU_I} = \frac{1}{N_I - 1}\sum_{U_I}(t_i - \bar{t}_{U_I})^2 = \frac{1}{3-1}\left((3 \cdot \frac{7}{3} - 15)^2 + (3 \cdot 5 - 15)^2 + (3 \cdot \frac{23}{3} - 15)^2\right) = 64$. Somit $V(\hat{\bar{y}}) = \frac{1}{N^2}N_I^2\frac{1}{n_I}(1 - \frac{n_I}{N_I})S^2_{tU_I} = 1.1852$ [2 Punkte]

(d) Cluster sollten so gewählt werden, dass innerhalb eines Clusters die Variablen (oder Merkmalsträger) so heterogen wie möglich sind. Die einzelnen Cluster sollten sich aber so wenig wie möglich voneinander unterscheiden.

Eine optimale Aufteilung wären demnach Cluster mit gleichem Mittelwert:

Y :	1	2	4	3	5	7	6	8	9
X :	1	3	2	2	1	3	3	2	1

[2 Punkte]

Aufgabe 4: Einfache Zufallsstichprobe mit Zurücklegen

Aus einer Grundgesamtheit vom Umfang $N = 10$ wird eine Stichprobe vom Umfang $n = 3$ mit Zurücklegen gezogen. Die Auswahlwahrscheinlichkeiten p_i, $i = 1, \ldots, 10$, sind dabei unterschiedlich. Folgende Werte mit zugehörigen Auswahlwahrscheinlichkeiten wurden gezogen:

$$y_1 = 3,\ p_1 = 0.06, \quad y_2 = 10,\ p_2 = 0.20, \quad y_3 = 7,\ p_3 = 0.10$$

(a) Schätzen Sie die Populationssumme mit dem Hansen-Hurwitz-Schätzer und schätzen Sie die Varianz dieses Schätzers!

(b) Schätzen Sie die Populationssumme mit dem Horvitz-Thompson-Schätzer und schätzen Sie die Varianz dieses Schätzers!

Antwort:

Aufgabe 4: 9 Punkte

(a) $\hat{t}_{HH} = \frac{1}{n}\sum_{k=1}^{n}\frac{y_k}{p_k} = \frac{1}{3}(50 + 50 + 70) = 56,67$ und

$$\hat{V}(\hat{t}_{HH}) = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{y_{k_i}}{p_{k_i}} - \hat{t}_{HH}\right)^2 = 44,44$$

[3 Punkte]

(b) $\pi_k = 1 - (1 - p_k)^m$, also: $\pi_1 = 0.169, \pi_2 = 0.488, \pi_3 = 0.271$. Dann

$$\hat{t}_{HT} = \sum_{k=1}^{n}\frac{y_k}{\pi_k} = 64,03$$

$\pi_{kl} = Pr(k\&lins) = Pr(kins) + Pr(lins) - Pr(kinsoderlins) = \pi_k + \pi_l - [1 - (1 - p_k - p_l)^m]$. Also: $\pi_{12} = 0.0626, \pi_1 3 = 0.033, \pi_{23} = 0.102$. Somit:

$$\hat{V}(\hat{t}_{HT}) = \sum\sum_{s}\pi_{kl}^{-1}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)y_k y_l = -449.6851$$

[6 Punkte]

Aufgabe 5: Optimale Schichtgröße

Es soll geschätzt werden, wie viel Geld die Haushalte einer Großstadt auf ihren laufenden Konten haben. Die Haushalte werden nach der Höhe ihres Einkommens in zwei Schichten gegliedert. Es wird vermutet, dass der Kontostand eines Haushalts mit größerem Einkommen neunmal so groß wie der Kontostand eines Haushalts mit kleinerem Einkommen ist. Ferner wird angenommen, dass S_h proportional der Quadratwurzel des Schichtmittels ist.
Es gibt 4000 Haushalte in der Schicht mit größerem und 20000 Haushalte in der Schicht mit kleinerem Einkommen. Es soll eine Stichprobe vom Umfang $n = 1000$ gezogen werden.

Teilen Sie den Gesamtstichprobenumfang optimal (ohne Berücksichtigung einer Kostenfunktion) auf die beiden Schichten auf!

Antwort:

Aufgabe 5: [5 Punkte]
$N_1 = 20000, N_2 = 4000, n = 1000, 9\bar{y}_{s_1} = \bar{y}_{s_2}$.
Damit: $S_{y_1} = k\sqrt{\bar{y}_{s_1}}$ und $S_{y_2} = k\sqrt{\bar{y}_{s_2}} = k3\sqrt{\bar{y}_{s_1}}$. Einsetzen in

$$n_h^{opt} = n\frac{N_h S_{y_h}}{N_1 S_{y_1} + N_2 S_{y_2}}$$

ergibt: $n_1^{opt} = 375$ und $n_2^{opt} = 625$.

Aufgabe 6: Anteilswerte und Modelbasierte Schätzung

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) Warum können Sie bei dieser Aufgabe den Auswahlsatz vernachlässigen?

Vernachlässigen Sie den Auswahlsatz im Folgenden.

(b) Schätzen Sie den gesuchten Anteilswert und geben Sie den Standardfehler der Schätzung an.

(c) 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. Eine Berechnung des Standardfehlers ist nicht notwendig.

Antwort:

Aufgabe 6: [6 Punkte]

(a) Da es sich um eine Wahl handelt, bei der die Wahlbevölkerung im Millionenbereich ist, also $N > 50$ Millionen, kann man den Auswahlsatz $f = n/N$ vernachlässigen. [1 Punkt]

(b) $\hat{P} = \frac{80}{200} = 0.4$ und

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

Somit ist der Standardfehler $\sqrt{\hat{V}(\hat{P})} = 0.0347$ [3 Punkte]

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

- Differenzenschätzung: $\hat{P} = \bar{y}_s - \bar{x}_s + \bar{x}_U = 0.4 - 0.3 + 0.25 = 0.35$
- Verhältnisschätzung: $\hat{P} = \bar{x}_U \frac{\bar{y}_s}{\bar{x}_s} = 0.25\frac{0.4}{0.3} = 1/3 = 0.33$

[2 Punkte]

Aufgabe 7: Einschlusswahrscheinlichkeiten

Da der Einschlussindikator I_k Bernoulli verteilt ist, gelten für ein beliebiges Stichprobendesign $p(s)$ für alle $k, l = 1, ..., N$ folgende Zusammenhänge:

$$(i)\ E(I_k) = \pi_k, \qquad (ii)\ V(I_k) = \pi_k(1-\pi_k) \qquad (iii)\ Cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l,$$

Für die Stichprobengröße n_s gilt überdies $n_s = \sum_U I_k$.

Zeigen Sie, dass bei einem Stichprobendesign mit fixer Stichprobengröße, $n_s = n$, folgendes gilt:

(a) $\sum_U \pi_k = n$

(b) $\sum \sum_{k \neq l} {}_U \pi_{kl} = n(n-1)$

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

Hinweis: Betrachten Sie $E(n_s)$ und $V(n_s)$.

Antwort:

Aufgabe 7: [9 Punkte]

(a) $n = E(n_s) = E(\sum_U I_k) = \sum_U E(I_k) = \sum_U \pi_k$ [2 Punkte]

(b) $0 = V(n_s) = \sum \sum_U Cov(I_k, I_l) = \left(\sum \sum_{U, k \neq l} \pi_{kl} - \pi_k\pi_l\right) + \sum_U \pi_k(1-\pi_k)$. Somit:

$$\sum \sum_{U, k \neq l} \pi_{kl} = \sum_{U, k \neq l} \pi_k\pi_l + \sum_U \pi_k\pi_k - \sum_U \pi_k = \sum \sum_U \pi_k\pi_l - \sum_U \pi_k = \sum_U \pi_k \sum_U \pi_l - \sum_U \pi_k = n \cdot n - n = n(n-1)$$

[4 Punkte]

(c) $\sum_{\substack{l \in U \\ l \neq k}} \pi_{kl} = \sum_{l \in U} \pi_{kl} - \pi_{kk} = \sum_{l \in U} E(I_k I_l) - \pi_k = E(I_k \sum_U I_l) - \pi_k = nE(I_k) - \pi_k = (n-1)\pi_k$. [3 Punkte]

Aufgabe 8: Nichtnegativitätsbedingung

Der Yates-Grundi-Sen Schätzer für die Varianz des π-Schätzers der Merkmalssumme ist:

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

wobei $p(s)$ ein Stichprobendesign mit fixer Stichprobengröße ist und $\pi_{kl} > 0$ für alle $k, l = 1, ..., N$.

(a) Benennen Sie eine Bedingung, so dass der Schätzer immer nichtnegativ ist.

(b) Überprüfen Sie diese Bedingung für die einfache Zufallsstichprobe ohne Zurücklegen.

Antwort:

Aufgabe 8: [4 Punkte]

(a) Es muss gelten, dass $\Delta_{kl} = Cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l < 0$ für alle $k \neq l \in U$. Für $k = l$ ist $\check{y}_k - \check{y}_l = 0$ immer gegeben. [1 Punkt]

(b) Für $k \neq l$:

$$\begin{aligned}\Delta_{kl} = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 &< 0\\ \frac{n-1}{N-1} &< \frac{n}{N}\\ N \cdot n - N &< N \cdot n - n\\ n &< N\end{aligned}$$

ist immer erfüllt. [3 Punkte]