Verkehrsökonometrie für Bachelor-Studierende

Sommersemester 2021, Lösungsvorschläge zu Übung Nr. 10

Lösungsvorschlag zu Aufgabe 10.1: Umfrage zur Verkehrsmittelwahl

Allgemeines

Durch die Erhebung der Stichprobe werden 8 Ja-Anteilswerte (f_k, $k = 1(\text{PKW}, ..., 8(\text{Fuss}))$ simultan ermittelt. Unter der Annahme eines wahren Anteilswertes μ_k der Ja-Antworten für Modus k (der später durch die relativen Häufigkeiten angenähert wird) haben die entsprechenden Merkmale Y_{ki}, also Entscheidungen der Person i für Alternative k, die Ausprägungen

$$Y_{ki} = \begin{cases} 1 & \text{"ja" bzw. 1 mit Wahrsch. } \mu_k \\ 0 & \text{"nein" bzw. 0 mit Wahrsch. } 1-\mu_k \end{cases}$$

Es handelt sich also um binärverteilte Zufallsvariablen. Als Sonderfall einer (n, μ)- Binomialverteilung für $n = 1$ (Bernoulli-Verteilung) hat Y_{ki} den Erwartungswert μ_k und die Varianz $\mu_k(1-\mu_k)$.

Wir betrachten nun für jede Alternative k den bei Zufalls-Stichproben gültigen erwartungstreuen und (wenn es keine weiteren Informationen gibt) effizienten Schätzer

$$f_k = \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} Y_{ki},$$

wobei Y_{ki} das Merkmal Y_k im Stichprobenelement i bedeutet. Gemäß dem Skript gilt für Erwartungswert und Varianz

$$\begin{aligned} E(\hat{\mu}_k) &= \frac{1}{n}\sum_{i=1}^{n} E(Y_{ki}) = \frac{n}{n}\mu_k = \mu_k, \\ V(\hat{\mu}_k) &= V\left(\frac{1}{n}\sum_{i=1}^{n} V(Y_{ki})\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_{ki}) = \frac{\mu_k(1-\mu_k)}{n}. \end{aligned}$$

Zur Bestimmung der Konfidenzintervalle zu gegebener Fehlerwahrscheinlichkeit α benötigen wir aber nicht nur Erwartungswert und Varianz der Schätzer $\hat{\mu}_k$, sondern die Verteilung selbst (sonst könnte man ja nicht die benötigten Quantile bestimmen). Hier vereinfacht der Zentrale Grenzwertsatz (ZGWS), also die Approximation einer Summe von Zufallsvariablen durch eine Normalverteilung, die Berechnung enorm. Hinreichende Bedingungen für die Anwendbarkeit des ZGWS sind

(a) Unabhängigkeit der einzelnen Summanden Y_{ki},

(b) Endliche Varianz $V(Y_{ki})$

(c) Jeder Summand Y_{ki} der Summe $n\hat{\mu}_k = \sum_i Y_{ki}$ hat nur einen geringen Anteil an der Gesamtvarianz $n^2V(\hat{\mu}_k)$ der Summe. Häufig werden Einzelvariianzen kleiner als 1/30stel der Gesamtvarianz gefordert. Bei den unsymmetrischen diskreten Bernoulliverteilungen wird die Bedingung 3 häufig schärfer formuliert durch

$$n^2V(\hat{\mu}_k) = nV(Y_{ki}) = n\mu_k(1-\mu_k) > 9 \quad \forall k. \tag{1}$$

(Warum ist dies schärfer?)

Bedingung 1 ist aufgrund der geforderten Unabhängigkeit bzw. bei $N \gg n$ (Umfang der GG sehr viel größer als Stichprobenumfang) erfüllt. Ferner ist die Varianz $V(Y_{ki}) = \mu_k(1-\mu_k) \leq \frac{1}{4}$, also endlich. Nur Bedingung 3 ist möglicherweise nicht erfüllt. Zur Auswertung wird μ_k abgeschätzt durch die relative Häufigkeit bzw. den Anteilswert f_k in der Stichprobe.[1] Es genügt hier, die schärfste Bedingung zu prüfen, die sich aus der *kleinsten* Varianz ergibt. Hier ist es die Klasse $k = 3$ (Motorradfahrer), für die gilt

$$n\mu_3(1-\mu_3) \approx nf_3(1-f_3) = 1500 \cdot 0.01(1-0.01) = 14.8 > 9, \tag{2}$$

so dass der Schätzer $f_k = \hat{\mu}_k$ (die mit der Stichprobengröße n normierte Summe genügend vieler binomialverteilter Zufallsvariablen) als *normalverteilt* angenommen werden kann.

Nun müssen wir noch das Prognoseintervall bzw. die Fehlerwahrscheinlichkeit *definieren*: Wir fordern, dass eine Stichprobe vom Umfang n in $1-\alpha = 95\%$ der Fälle im Prognoseintervall liegt und mit $\alpha = 5\%$ davon abweichen kann. Da die Stichprobe normalverteilt ist (s.o.), ist das Intervall durch das entsprechende Quantil der Normalverteilung gegeben. Da der Fehler sich symmetrisch für Abweichungen nach oben und unten aufteilt, können wir

- nicht nur bei bekannten μ_k das *Prognoseintervall*

  $$\hat{\mu}_k \in [\mu_k - \Delta\hat{\mu}_k, \mu_k + \Delta\hat{\mu}_k]$$

  bestimmen,

- sondern auch bei bekannten Schätzer $\hat{\mu}_k$ das *Vertrauensintervall* bzw *Konfienzintervall* für den wahren Wert μ_k,

$$\mu_k \in [\hat{\mu}_k - \Delta\hat{\mu}_k, \hat{\mu}_k + \Delta\hat{\mu}_k], \quad \Delta\hat{\mu}_k = z_{1-\frac{\alpha}{2}}\sigma_k, \quad \sigma_k \approx \hat{\sigma}_k = \sqrt{\frac{f_k(1-f_k)}{n}}. \tag{3}$$

Die Umkehrung der Schließrichtung vom Prognose- zum Konfidenzintervall sowie die Verwendung des Schätzers $\hat{\sigma}_k$ für σ_k ergibt nur einen "Fehler der Fehlerabschätzung", welcher irrelevant ist.

Für $\alpha = 5\%$ ergibt sich aus der Quantiltabelle $z_{0.975} \approx 1.9600$

[1] Die Varianz wird nur zur Bestimmung des Stichprobenfehlers benötigt. Insofern stellt die Näherung von μ_k durch f_k nur einen Fehler zweiter Ordnung ("Fehler der Fehlerabschätzung") dar, welcher vernachlässigt werden kann.

Eigentliche Lösung

(a) Mit der Gl. (3) ergeben sich folgende Stichprobenfehler:

k	1 PKW	2 PKW (Mitf.)	3 KR	4 Tram	5 Bus	6 S-Bahn	7 Rad	8 Fuß
f_k	0.43	0.03	0.01	0.20	0.12	0.03	0.13	0.05
$\Delta f_{0.95,k}$	0.025	0.0086	0.0050	0.020	0.016	0.0086	0.017	0.011

(b) Da der Stichprobenfehler für Anteilswerte μ proportional zu $\sqrt{\mu(1-\mu)}$ ist, ergibt sich der größte Fehler für den Anteil, welcher am nächsten an 50% liegt – hier für μ_1 mit dem Stichprobenergebnis $f_1 = 0.43$. Wird die Stichprobe so vergrößert, dass Δf_1 höchstens 0.02 beträgt, liegen alle anderen Stichprobenfehler unter diesem Wert. Unter der Annahme, das sich der Anteil μ_1 weiterhin in der Größenordnung 0.43 bewegt ergibt sich:

$$\begin{aligned}
& & \Delta f_1 &\leq 0.02 \\
\iff & z_{0.975}\sqrt{\frac{f_1(1-f_1)}{n}} &&\leq 0.02 \\
\iff & & n &\geq \frac{z^2_{0.975} f_1(1-f_1)}{0.02^2} \\
\iff & & n &\geq \frac{1.96^2 \cdot 0.43 \cdot 0.57}{0.02^2} \approx 2354. \qquad (4)
\end{aligned}$$

Lösungsvorschlag zu Aufgabe 10.2: Umfrage zur Radwege-Sanierung

(a) Durch die Meinungsumfrage des Vorjahrs sind die autobesitzabhängigen Meinungsanteile (der „Meinungs-Split") erhoben worden. Es handelt sich um die bedingten Häufigkeiten $f(\text{Meinung}|\text{Autobesitz})$, aus welchen man bei Kenntnis der relativen Häufigkeiten $f(\text{Autobesitz})$ die unbedingten Meinungsanteile exakt berechnen kann (deskriptive Statistik auf der Menge der Umfrageteilnehmer). Gibt f den Ja-Anteil in der Stichprobe an und bedeuten f_1 und f_2 die bedingten Ja-Anteile in den Untergruppen der Auto- und Nicht-Auto-Besitzer (Teilumfänge $n_1 = 400$ und $n_2 = 600$, so gilt

$$\hat{\mu} = f = \frac{1}{n}(n_1 f_1 + n_2 f_2 = 0.6415$$

(b) Bei einer reinen Zufallsinformation wird die in der GG bekannte Information des wahren Autobesitzanteils von $\vartheta_1 = 40\,\%$ *nicht* berücksichtigt. Damit gilt analog zur ersten Aufgabe (Die Bedingung $nf(1-f) > 9$ für die Anwendbarkeit des ZGWS ist erfüllt) $\hat{\mu} = f$ und $E(f) = \mu$ sowie

$$\mu \in [f - \Delta f, f + \Delta f], \quad \Delta f = z_{0.975}\sqrt{\frac{f(1-f)}{n}} = 0.0298$$

(c) Man befragt unter den Autobesitzern $n_1 = n\vartheta_1 = 1000 \cdot 0.4 = 400$ zufällig ausgewählte Personen, und $n_2 = n - n_1 = 600$ zufällig ausgewählte Personen ohne Auto. Daraus erhält man zwei Ja-Anteile f_1 und f_2, die als Schätzer für die wahren Ja-Anteile μ_1 und μ_2 in den beiden Bevölkerungsschichten dienen. Die Schätzer gehorchen jeweils einer Wahrscheinlichkeitsverteilung mit

$$E(f_1) = \mu_1, \quad V(f_1) = \frac{f_1(1-f_1)}{n_1},$$
$$E(f_2) = \mu_2, \quad V(f_2) = \frac{f_2(1-f_2)}{n_2},$$

Der erwartungstreue Schätzer des Anteilwertes μ der Grundgesamtheit ergibt sich *derselbe* Schätzer wie bei der reinen Zufallsauswahl:

$$\hat{\mu} = f = \vartheta_1 f_1 + \vartheta_2 f_2, \quad \vartheta_1 = 0.4, \quad \vartheta_2 = 0.6$$

Zur Berechnung der Varianz kann man aber hier, im gegensatz zur reinen Zufallsauswahl, ausnutzen, dass die Teilumfänge n_1 und n_2 bzw. die Anteile ϑ_1 und ϑ_2 der beiden Ausprägungen des Shcichtungsmerkmals (Autobesitz) a priori fest sind und man somit f als gewichtete Summe zweier unabhängiger Zufallsvariablen ausdrücken kann. Mit der Varianz-Rechenregel

$$V(aX + bY) = a^2 V(X) + b^2 V(Y)$$

(X, Y sind zwei unabhängige Zufallsvariablen und a und b reelle Konstnten) erhält man

$$\hat{\sigma}^2_{\text{Quotenauswahl}} = V(\hat{\mu}) = \vartheta_1^2 V(f_1) + \vartheta_2^2 V(f_2)$$

und mit Einsetzen der obigen Formal $V(f_k) = f_k(1-f_k)/n_k$ letztendlich

$$\hat{\sigma}^2_{\text{Quotenauswahl}} = 10^{-4}, \quad \hat{\sigma}_{\text{Quotenauswahl}} = 0.010.$$

Damit ergibt sich als maximaler Stichprobenfehler:

$$\Delta f_{\text{Quotenauswahl}} = z_{0.975} \cdot \hat{\sigma}_{\text{Quotenauswahl}} = 1.96 \cdot 0.010 = \underline{\underline{0.0196}}.$$

Dies ist deutlich geringer als der entsprechende Wert $\Delta f_{\text{Zufallsauswahl}} = 0.0298$.

(d) Obwohl keine geschichtete Stichprobe erhoben wurde, kann die Stichprobe mit einer geeigneten *Entzerrung* wie eine geschichtete Stichprobe ausgewertet werden, da das die Schichten begründende Merkmal mit erhoben wurde (Autobesitz).
Zufallsbedingt sind die 40% Autobesitzer durch Befragung von 443 Personen in der Stichprobe ($n = 1000$) überrepräsentiert, also $f_1 = 0.443 > \vartheta_1 = 0.4$. Die 60% Nicht-Autobesitzer sind hingegen durch die 557 Personen der Stichprobe unterrepräsentiert, $f_2 = 0.557 < \vartheta_1 = 0.6$. Um dies zu entzerren, werden die Antworten der Nichtautobesitzer mit $E_1 = \vartheta_1/f_1 < 1$ unter- und die Antworten der Autobesitzer mit $E_2 = \vartheta_2/f_2 > 1$ übergewichtet.

Die Schätzer $\hat{\mu}_1$ für die wahren Ja-Anteile μ_1 unter den Autobesitzern bzw. der Anteile $\hat{\mu}_2$ für die wahren Ja-Anteile μ_2 der Nichtbesitzer sind wie üblich die relativen Häufigkeiten in den jeweiligen Schichten:

$$\begin{aligned} \hat{\mu}_1 = f(\text{ja}|\text{Auto}) &= \frac{102}{443} = 0.230 \quad \text{und} \\ \hat{\mu}_2 = f(\text{ja}|\text{kein Auto}) &= \frac{530}{557} = 0.952. \end{aligned}$$

Der entzerrende Schätzer $\hat{\mu}_\text{E}$ ist gegeben durch

$$\begin{aligned} \hat{\mu}_\text{E} &= \frac{1}{n}\sum_i Y_i E_{k(i)} && (5) \\ &= f_1 E_1 \hat{\mu}_1 + f_2 E_2 \hat{\mu}_2 && (6) \\ &= \vartheta_1 \hat{\mu}_1 + \vartheta_2 \hat{\mu}_2. && (7) \end{aligned}$$

Erwartungswert

$$\begin{aligned} E(\hat{\mu}_\text{E}) &= \vartheta_1 E(\hat{\mu}_1) + \vartheta_2 E(\hat{\mu}_2) \\ &= \vartheta_1 \mu_1 + \vartheta_2 \mu_2 = \mu \end{aligned}$$

Der entzerrende Schätzer ist also – ebenso wie das einfache arithmetische Mittel – erwartungstreu.

Varianz

Innerhalb der jeweiligen Gruppen können die Schätzer $\hat{\mu}_1$ und $\hat{\mu}_2$ separat als einfaches arithmetisches Mittel der durch Zufallsauswahl gewonnenen Elemente angesehen werden, also

$$\begin{aligned} V(\hat{\mu}_1) &= \frac{\mu_1(1-\mu_1)}{n_1} \approx \frac{0.230 \cdot (1-0.230)}{443} = 4.00 \cdot 10^{-4}, \\ V(\hat{\mu}_2) &= \frac{\mu_2(1-\mu_2)}{n_2} \approx \frac{0.952 \cdot (1-0.952)}{557} = 8.20 \cdot 10^{-5}. \end{aligned}$$

Die Varianz des eigentlichen entzerrenden Schätzers $\hat{\mu}_\text{E} = \vartheta_1 \hat{\mu}_1 + \vartheta_2 \hat{\mu}_2$ wird durch die übliche Regel für die Varianz einer Linearkombination von Zufallsvariablen mit verschwindender Kovarianz zwischen $\hat{\mu}_1$ und $\hat{\mu}_2$ ermittelt:[2]

$$V(\hat{\mu}) = \vartheta_1^2 V(\hat{\mu}_1) + \vartheta_2^2 V(\hat{\mu}_2) = 9.31\,10^{-5}$$

[2] Es seien X und Y Zufallsvariable und a, b Zahlen sowie $Z = aX + bY$. Dann gilt $V(Z) = a^2 V(X) + b^2 V(Y) + 2ab\text{Cov}(X, Y)$.

Da nach dem ZGWS $\hat{\mu}_1$ und $\hat{\mu}_2$ normalverteilt sind, gilt dies auch für die Linerakombination $\hat{\mu}$. Daher lautet das Konfidenzintervall, in welchem der wahre Wert μ mit 95% Wahrscheinlichkeit liegt,

$$\mu \in f(\text{ja}) \pm z_{0.975}\sigma = 0.6630 \pm 0.0189\,.$$

Mit dieser Zufallsstichprobe, bei deren Auswertung die (aus anderen Quellen stammende) sehr exakt vorgegebene 40%:60% Verteilung berücksichtigt wurde, erhält man also einen genaueren Schätzer, als mit einer von vorne herein nach dem Anteil des Autobesitzes geschichteten Stichprobe, vgl. Aufgabenteil c). Diese Abweichung (ungenauere Untersuchung der Nicht-Autobesitzerschicht, 600→557) zahlt sich hier aus, weil der Term $0.952(1-0.952)$ eine verhältnismäßig hohe Genauigkeit für den Schätzer $f(\text{kein Auto})$ nach sich zieht, so dass es sich lohnt, den Stichprobenumfang für diese eigentlich größere und damit wichtigere Schicht zugunsten der Autobesitzer zu verringern.