Peter von der Lippe (Jan. 2012)

Genauigkeit von Schichtmittelwerten und Genauigkeit des Gesamtmittelwerts einer geschichteten Stichprobe

Eine kurze Notiz zu meinem Text

"Wie groß muss meine Stichprobe sein, damit sie repräsentativ ist?[1]

Das Problem

Es hat mich sehr gefreut, zu sehen, dass mein o.g. Text an verschiedenen Stellen wahrgenommen wurde. Das darin behandelte Problem, wie man den Stichprobenumfang zu planen hat, wenn man bei einer geschichteten Stichprobe

- nicht nur für Schätzungen (z.B. der Mittelwerte μ_1, ..., μ_K) *bezüglich der einzelnen Schichten* (Teilmassen) bestimmten Genauigkeitsforderungen (was i.d.R. in Gestalt von Vorgaben hinsichtlich der K Varianzen $\mathrm{Var}(\bar{x}_k) = \sigma^2_{\bar{x}_k}$ bei $k = 1, ..., K$ erfolgt) genügen möchte, sondern dies auch
- von der entsprechenden Schätzung für das entsprechende Gesamtaggregat (also für alle Teilmassen zusammen) fordert (in Gestalt von Vorgaben hinsichtlich der einen Varianz $\mathrm{Var}(\bar{x}) = \sigma^2_{\bar{x}}$ für die Gesamtmasse)

scheint ein sehr verbreitetes Problem zu sein. Bei der Deutschen Bundesbank trat es auf als Zusammenhang zwischen der Genauigkeit der Länderergebnisse einerseits und des europäischen Ergebnisses andererseits.[2] Legt man den Fokus auf die Länder (nationale Stichproben) sind die Formeln für die einfache Stichprobe angebracht, steht dagegen das europäische Ergebnis im Vordergrund ist nach Maßgabe der geschichteten Stichprobe zu verfahren. Nimmt man diese Formeln (für die geschichtete Stichprobe), so erhält man die in meinem o.g. Text dargestellten paradoxen Ergebnisse (Auswahlsatz umso kleiner größer die Schicht ist – ohne dass angenommen wurde, dass die großen Schichten mehr oder weniger homogen sind als die kleinen - und Gesamtstichprobenumfang abhängig von der Anzahl K der Schichten, also von dem Ausmaß der Differenzierung der Masse in Teilmassen).[3]

Bei der optimalen Aufteilung einer geschichteten Stichprobe wird $\mathrm{Var}(\bar{x}) = \sigma^2_{\bar{x}}$ minimiert. Die Varianzen $\mathrm{Var}(\bar{x}_k) = \sigma^2_{\bar{x}_k}$ sind demgegenüber ein daraus resultierendes Nebenprodukt, d.h. für sie ist nicht die Berücksichtigung bestimmter Vorgaben vorgesehen und sie sind somit nicht Gegenstand der Minimierung.

Es empfiehlt sich zunächst Zusammenhänge zwischen der einen Varianz $\sigma^2_{\bar{x}}$ und den K Varianzen $\mathrm{Var}(\bar{x}_k) = \sigma^2_{\bar{x}_k}$ darzustellen und zwar zunächst allgemein und dann speziell für die geschichtete Stichprobe. Mir fiel auf, dass mein o.g. Papier in diesem Punkt nicht sehr klar und ausführlich ist, so dass es nützlich ist zunächst hierauf etwas einzugehen.[4]

---

[1] Untertitel "Wie viele Einheiten müssen befragt werden? Was heißt 'Repräsentativität'?" Das Papier vom Februar 2011 ist als Download auf meiner Homepage und als Diskussionsbeitrag Nr. 187 des Fachbereichs Wirtschaftswissenschaften Univ. Duisburg Essen, Campus Essen verfügbar

[2] Bei der hier vorliegenden Notiz nehme implizit Bezug auf eine Ausarbeitung von zwei Mitarbeitern der Deutschen Bundesbank, die nach eigener Auskunft dabei in Kenntnis meines o.g. Papiers waren

[3] Mein ursprüngliches Anwendungsproblem war eine geschichtete Stichprobe von Arztpraxen, wobei mir auffiel, dass der Auswahlsatz aus der großen Schicht der Hausärzte relativ klein war, aber aus der kleinen Schicht etwa der Augenärzte sehr viel größer. Was für ärztliche Fachgruppen gilt, kann entsprechend auf Länder der Europäischen Union übertragen werden.

[4] Die hier vorliegende Notiz ist bewusst sehr einfach gehalten und mag helfen, manche Stellen in dem anderen Papier leichter zu verstehen.

1. Varianzzerlegung für die Gesamtstichprobe

a) allgemein

Man kann in $\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}$ bzw. $\hat{\sigma}_{\bar{x}}^2 = \frac{\hat{\sigma}_x^2}{n} = \frac{\hat{\sigma}^2}{n}$ die Größe $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x})^2$ in Varianzkomponenten zerlegen. Wir tun dies der Einfachheit halber mit

(1) $\quad s^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x})^2 = \frac{1}{n}\sum\sum x_{ik}^2 - \bar{x}^2$

Der Ausdruck $\frac{1}{n}\sum\sum x_{ik}^2$ (das zweite Anfangsmoment) kann bei Verwendung der K inneren Varianzen $s_k^2 = \frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)^2 = \frac{1}{n_k}\sum x_{ik}^2 - (\bar{x}_k)^2$ und der daraus als Mittelwert gebildeten internen Varianz

(2) $\quad s_{int}^2 = \sum_{k=1}^{K}\frac{n_k}{n}s_k^2 = \frac{1}{n}\sum_k\sum_i x_{ik}^2 - \sum_{k=1}^{K}\frac{n_k}{n}\bar{x}_k^2$

ersetzt werden, so dass man erhält

(3) $\quad s^2 = s_{int}^2 + \sum_{k=1}^{K}\frac{n_k}{n}\bar{x}_k^2 - \bar{x}^2 = s_{int}^2 + s_{ext}^2$

mit der externen Varianz

(3a) $\quad s_{ext}^2 = \sum_{k=1}^{K}\frac{n_k}{n}(\bar{x}_k - \bar{x})^2 = \sum_k\frac{n_k}{n}\bar{x}_k^2 - \bar{x}^2$ da gilt $\bar{x} = \sum\frac{n_k}{n}\bar{x}_k$.

Gl. 3 ist das Stichproben-Analogon zu Gl. 21 im Papier "Wie groß muss meine Stichprobe sein?" (kurz WGS-Papier)

$$\sigma^2 = \sum_{k=1}^{K}\frac{N_k}{N}(\mu_k - \mu)^2 + \sum_{k=1}^{K}\frac{N_k}{N}\sigma_k^2 = V_{ext} + V_{int},$$

was ein definitorischer Zusammenhang für Größen der Grundgesamtheit ist. Damit erhält man für den Stichprobenfehler

(4) $\quad \hat{\sigma}_{\bar{x}}^2 = \frac{\hat{\sigma}^2}{n} \approx \frac{s^2}{n} = \frac{1}{n}(s_{int}^2 + s_{ext}^2) = \frac{1}{n}\sum\frac{n_k}{n}s_k^2 + \frac{1}{n}\sum\frac{n_k}{n}(\bar{x}_k - \bar{x})^2.$

b) bei geschichteter Stichprobe

In diesem Fall erhält man einen erheblich geringeren Wert für s^2; denn jetzt wird $\bar{x}$ gebildet als Linearkombination der Schichtmittelwerte gem. Gl. 10 des WGS-Papiers

$\bar{x} = \frac{N_1}{N}\cdot\bar{x}_1 + \ldots + \frac{N_K}{N}\cdot\bar{x}_K = \sum_k\frac{N_k}{N}\cdot\bar{x}_k$, so dass für die Varianz der Linearkombination gilt

(5) $\quad \hat{\sigma}_{\bar{x}}^2 = \sum_k\left(\frac{N_k}{N}\right)^2\hat{\sigma}_{\bar{x}_k}^2 = \sum_k\left(\frac{N_k}{N}\right)^2\frac{\hat{\sigma}_k^2}{n_k} \approx \sum_k\left(\frac{N_k}{N}\right)^2\frac{s_k^2}{n_k} = \sum_k\omega_k^2\frac{s_k^2}{n_k}$

Bei proportionaler Aufteilung $N_k/N = n_k/n$ für alle k (oder wenn $\bar{x}$ als Linearkombination mit den Gewichten statt gebildet wird) vereinfacht sich das zu

(6) $\quad \hat{\sigma}_{\bar{x}}^2 = \frac{\hat{\sigma}^2}{n} = \frac{1}{n}\sum\frac{n_k}{n}\hat{\sigma}_k^2 \approx \frac{1}{n}\sum\frac{n_k}{n}s_k^2.$

womit der Unterschied zu (4) sehr deutlich wird.

Warum ist der Stichprobenfehler des Gesamtmittels $\bar{x}$ (also $\hat{\sigma}_{\bar{x}}^2$) bei geschichteter Stichprobe (nach Gl. 6) so viel kleiner als bei einer einfachen Stichprobe (Gl. 4)? Der Grund liegt darin, dass extreme Stichproben nicht gezogen werden, wenn die externe Varianz groß ist. Wir verdeutlichen das an einem kleinen Beispiel mit zwei Schichten mit $N_1=N_2=2$ Elementen. Die x-Werte in Schicht 1 sind

$x_A = 2$ und $x_B = 4$ (so dass $\bar{x}_1 = 3$ ist) und die x-Werte in Schicht 2 sind $x_c = 20$ und $x_D = 40$ (so dass $\bar{x}_2 = 30$ ist). Wirft man alle Einheiten in einen Topf (nach Art der einfachen Stichprobe), so sind 6 Stichproben möglich AB, AC, AD, BC, BD, und CD. Zieht man aus Schicht 1 und 2 jeweils eine Stichprobe von nur einem Element ($n_1 + n_2 = 2$) so hat man auch Stichproben vom Umfang $n = 2$, aber die extremen Zusammenstellungen AB und CD mit den Mittelwerten 3 und 30, die bei der einfachen Stichprobe möglich sind, können jetzt nicht gezogen werden, es bleiben nur mittlere Kombinationen mit moderaten Mittelwerten, zwischen 11 bei AC und 22 bei BD.

2. Fehler des Gesamtmittelwerts und Fehler der Schichtmittelwerte

Aus (5) ergibt sich, dass der Stichprobenfehler des Gesamtmittelwerts ($\hat{\sigma}^2_{\bar{x}}$) eine gewogene Summe der der Stichprobenfehler der Schichtmittelwerte ($\hat{\sigma}^2_{\bar{x}_k}$) ist. Die Summe der Gewichte ist jedoch nicht 1. Der Mittelwert der K Größen N_k/N ist $1/K$ weil $\Sigma N_k = N$ und für die Varianz der Größen N_k gilt

$$(7) \quad \mathrm{var}\left(\frac{N_k}{N}\right) = \frac{1}{K}\sum_k \left(\frac{N_k}{N} - \frac{1}{K}\right)^2 = \frac{1}{K}\sum \left(\frac{N_k}{N}\right)^2 - \left(\frac{1}{K}\right)^2 \geq 0.$$

Ferner ist wegen $\Sigma N_k = N$ auch $\mathrm{var}\left(\frac{N_k}{N}\right) < \frac{K-1}{K^2} = \frac{1}{K}\left(1 - \frac{1}{K}\right) < 1$, so dass man erhält

$$(8) \quad \frac{1}{K} \leq \sum_k \left(\frac{N_k}{N}\right)^2 = K \cdot \mathrm{var}\left(\frac{N_k}{N}\right) + \frac{1}{K} < 1 - \frac{1}{K}.$$

Bestimmt man die Stichprobenumfänge für die K schichten so, dass die Genauigkeit (Stichprobenfehler) bei allen Schichten – unabhängig von den Schichtumfängen – gleich groß ist, $\hat{\sigma}^2_{\bar{x}_k} = c$ für $k = 1, \ldots, K$, dann vereinfacht sich (5) zu

$$(5a) \quad \hat{\sigma}^2_{\bar{x}} = c\sum_k \left(\frac{N_k}{N}\right)^2,$$

so dass der Fehler des Gesamtmittelwerts ($\hat{\sigma}^2_{\bar{x}}$) kleiner als der annahmegemäß jeweils gleich große Fehler der Schichtmittelwerte ($\hat{\sigma}^2_{\bar{x}_k} = c$) ist. Nach Gl. 8 gilt nämlich[5]

$$(5b) \quad c\frac{1}{K} \leq \hat{\sigma}^2_{\bar{x}} \leq c\frac{K-1}{K}.$$

Eine solche Bedingung $\hat{\sigma}^2_{\bar{x}_k} = c$ verlangt, die Stichprobenumfänge $n_1, n_2, \ldots$ der Schichten so zu setzen, dass die Gleichungen $\hat{\sigma}^2_{\bar{x}_k} = \frac{\hat{\sigma}^2_k}{n_k} = c$ gelten. Aus Summation dieser Gleichungen folgt, dass der Gesamtstichprobenumfang n nach

$$(9) \quad n = \frac{\sum_k \hat{\sigma}^2_k}{c}$$

zu bestimmen ist – was nach (5a) $\hat{\sigma}^2_{\bar{x}_k} = \frac{\sum_k \hat{\sigma}^2_k}{n}\sum\left(\frac{N_k}{N}\right)^2$ ergibt – und nach

$$(10) \quad \frac{n_k}{n} = \frac{\hat{\sigma}^2_k}{\sum \hat{\sigma}^2_k}$$

aufzuteilen ist. Man vergleiche das mit der optimalen Aufteilung (Gl. 18 in WGS)

$$\frac{n_k}{n} = \frac{N_k\,\sigma_k}{\sum N_k\,\sigma_k} = \frac{N_k\,\sigma_k}{N\bar{\sigma}} \text{ bzw. } \frac{n_k}{n} = \frac{N_k\,\hat{\sigma}_k}{\sum N_k\,\hat{\sigma}_k}.$$

[5] Ich verdanke diese Erkenntnis den beiden Mitarbeitern der Bundesbank.