MÖGLICHKEITEN ZUR EFFEKTIVITÄTSSTEIGERUNG ZWEIPHASIGER STICHPROBENINVENTUREN

J. Saborowski
Abteilung für Forstliche Biometrie und Informatik
Universität Göttingen

SUMMARY

Double sampling for stratification and for regression, respectively, are widespread tools for reduction of time and costs in forest inventory. However, especially in large scale inventories it is sometimes not reasonable to apply one regression model for the whole area. Therefore we recommend the application of a combined sampling technique: double sampling for stratification with single or double sampling for regression. This technique allows individual regression models in all strata. The theoretical foundation is given by a general expression of the variance in cases where the phase-one-sample is stratified into L strata and a nearly arbitrary phase-two-sample is drawn to estimate the strata means. This general case includes also the above mentioned combined sampling techniques. Optimum sample sizes are developped and a comparison with conventional double sampling for regression is carried out. Finally, variance and variance estimator are developped for a special double sampling technique for regression and *pps*-sampling, mentioned in the literature.

1. EINLEITUNG

Obwohl mehrphasige und mehrstufige Stichprobenverfahren in der Stichprobentheorie schon seit langem bekannte Techniken der Stichprobenerhebung sind, haben sie im forstlichen Bereich erst mit dem zunehmenden Einsatz von Luft- und Satellitenbildern in der Waldinventur einen besonderen Aufschwung erfahren. Wenngleich eine Kombination beider Verfahren möglich ist, müssen sie begrifflich klar auseinandergehalten werden. Mehrstufige Verfahren sollen gegenüber einfachen zufälligen oder systematischen Stichproben durch eine Zusammenfassung von meist räumlich zusammenhängenden Objekten (z.B. Bäumen) zu Klumpen (cluster) und Teilklumpen zu einer weniger zeitraubenden bzw. kostspieligen Stichprobenerhebung führen. Mehrphasige Stichprobeninventuren profitieren dagegen von einfach bzw. preiswert zu erhebenden Zusatzinformationen (Hilfsvariablen), die eine reduzierte Erhebung der eigentlichen, kostspieligeren Zielvariablen bei gleichem Stichprobenfehler erlauben. Über die Darstellungen in den einschlägigen Monographien zur Stichprobentheorie hinaus findet man umfangreiche Informationen über mehrstufige und mehrphasige Stichprobenverfahren, speziell im Zusammenhang mit forstlichen Anwendungen, in *FRAYER 1979, JEYARATNAM u.a. 1984 und SABOROWSKI 1990.*

Im deutschsprachigen Raum werden in jüngerer Zeit besonders zweiphasige Stichproben von verschiedenen Autoren auf ihre Tauglichkeit in Verbindung mit Luftbilddaten überprüft. *WOLFF 1990* und *KÄTSCH 1991* untersuchen zweiphasige Stichproben zur Regressionsschätzung, um mit Hilfe von Variablen, die aus dem Luftbild (1:6000) gewonnen werden (z.B. die photogrammetrischen Variablen Ober- und Mittelhöhe, Kronenanzahl etc. und unterschiedlichste Transformationen davon), terrestrische Erhebungen auf ein Mindestmaß beschränken zu können. Hohe multiple Korrelationen zwischen Luftbild- und terrestrischen Variablen belegen die Wirtschaftlichkeit der zweiphasigen Regressionsschätzung im Fall des zugrunde liegenden Datenmaterials aus dem Hils (Weserbergland). *KÖHL 1990* berichtet über eine Pilotinventur im Kanton Tessin, die die Anwendung zweiphasiger Regressionsstichproben wegen zu geringer Korrelationen zwischen Luftbildvolumen und terrestrischem Volumen nicht sinnvoll erscheinen läßt. Er sucht die Ursachen in einem möglicherweise ungünstigen Luftbildmaßstab (1:25000) und dem kleinflächig stark variierenden Holzvorrat im Gebirgswald. Alternativ wird jedoch ein zweiphasiges Stichprobenmodell zur Stratifizierung vorgeschlagen, in dem nach 5 Waldentwicklungsstufen stratifiziert wird. Es führt bei geringen Einbußen in der Genauigkeit zu erheblichen Kosteneinsparungen.

Zweiphasige Stichproben, die Zusatzinformationen erheben, um sinnvoll zu stratifizieren oder mit Hilfe von Regressionsschätzungen kostspielige Messungen durch preiswertere zu substituieren, scheinen demnach erfolgversprechende Konzepte für großräumige Inventuren zu sein. In dieser Arbeit sollen beide Strategien, Stratifizierung und Regressionsschätzung, miteinander kombiniert werden, um einen noch wirtschaftlicheren Einsatz der für eine Inventur zur Verfügung stehenden Mittel anzustreben. Wir beschränken uns hier auf die Darstellung und Diskussion der Ergebnisse. Einzelheiten der notwendigen Beweise werden im Anhang dargestellt.

Zu diesem Zweck werden zunächst in kurzer Form die beiden grundlegenden zweiphasigen Stichprobenverfahren skizziert, wie sie auch in den üblichen Lehrbüchern zu finden sind. Die Nomenklatur entspricht weitgehend derjenigen in *COCHRAN 1977*.

2. GRUNDTYPEN ZWEIPHASIGER STICHPROBENVERFAHREN

Wir unterstellen für dieses und die folgenden Kapitel, daß das Inventurgebiet aus genau N Stichprobeneinheiten $u_1, \ldots, u_N$ besteht, die wir uns z.B. als wie auch immer geformte Probeflächen, Probekreise oder auch Einzelbäume vorstellen können. Jede dieser Stichprobeneinheiten sei durch einige quantitative (wie z.B. Luftbildmittelhöhe, Beschirmungsgrad, terrestrisches Volumen etc.) oder auch qualitative Merkmale (Entwicklungsstufe, Mischungsgrad, Exposition, etc.) charakterisiert, z.B.

$u_1, \ldots, u_N$	N Stichprobeneinheiten
$y_1, \ldots, y_N$	quantitative Zielvariable
$x_1, \ldots, x_N$	quantitative Hilfsvariable
$q_1, \ldots, q_N$	qualitative Hilfsvariable

Weiter sei

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

und durch die qualitative Hilfsvariable q werde das Inventurgebiet in insgesamt L Straten aus je N_h $(h = 1, \ldots, L)$ Stichprobeneinheiten eingeteilt $(N = \sum_{h=1}^{L} N_h)$, in denen

$$\bar{Y}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} y_{hj} \qquad \text{und} \qquad S_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}(y_{hj} - \bar{Y}_h)^2$$

Mittelwerte und Streuungen der Straten bezeichnen.

$$\rho = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{Y})^2 \sum_{i=1}^{N}(x_i - \bar{X})^2}} \qquad \rho_h = \frac{\sum_{j=1}^{N_h}(y_{hj} - \bar{Y}_h)(x_{hj} - \bar{X}_h)}{\sqrt{\sum_{j=1}^{N_h}(y_{hj} - \bar{Y}_h)^2 \sum_{j=1}^{N_h}(x_{hj} - \bar{X}_h)^2}}$$

sind die Korrelationen zwischen x und y im gesamten Inventurgebiet bzw. in den einzelnen Straten.

a) *Zweiphasige Stichprobe zur Stratifizierung*

Hier wird zunächst (Phase 1) eine einfache Zufallsstichprobe vom Umfang n' ohne Zurücklegen aus den N Einheiten des Inventurgebietes gezogen. Sie wird wie üblich der Einfachheit halber von 1 bis n' durchnumeriert

$$u_1, u_2, \ldots, u_{n'}$$

An diesen Stichprobeneinheiten wird das Merkmal q ermittelt

$$q_1, q_2, \ldots, q_{n'}$$

(z.B. die Entwicklungsstufe aus dem Luftbild), um sie den L unterschiedlichen Straten zuzuordnen. Die Anzahl n'_h von Stichprobeneinheiten, die so dem Stratum h zugeordnet werden, ist demnach eine zufällige Größe, während $n' = \sum_{h=1}^{L} n'_h$ natürlich ein fest vorgegebener Stichprobenumfang ist. Der Anteil $w_h = n'_h/n'$ eines Stratums h an der Stichprobe schätzt den Anteil $W_h = N_h/N$ dieses Stratums am gesamten Inventurgebiet erwartungstreu.

Aus den n'_h Stichprobeneinheiten jedes Stratums h wird nun eine weitere Stichprobe (Phase 2) zufällig und ohne Zurücklegen ausgewählt, deren Umfang n_h vor der Erhebung anteilig festgelegt werden muß. D.h. genauer, daß die Verhältnisse

$$\nu_1, \ldots, \nu_L \qquad \text{mit} \qquad \nu_h = \frac{n_h}{n'_h}$$

festgelegt werden, so daß $0 < \nu_h \leq 1$. n_h kann also höchstens gleich n'_h sein. Lediglich an diesen Teilstichproben wird das (kostspielige) Merkmal y ermittelt, also z.B. das terrestrische Volumen.

Dann kann $\bar{Y}$ durch

$$\bar{y}_{2st} = \sum_{h=1}^{L} w_h \cdot \bar{y}_h = \sum_{h=1}^{L} \frac{n'_h}{n'} \cdot \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$$

geschätzt werden, und es gilt $E\,\bar{y}_{st} = \bar{Y}$ und

$$Var\,\bar{y}_{2st} = \frac{1}{n'}\Big(1 - \frac{n'}{N}\Big)S^2 + \sum_{h=1}^{L} \frac{W_h}{n'} S_h^2 \Big(\frac{1}{\nu_h} - 1\Big) \tag{2.1}$$

$\bar{y}_{st}$ ist also ein erwartungstreuer Schätzer für den Gesamtmittelwert mit der Varianz (2.1). Letztere kann durch

$$v(\bar{y}_{2st}) = \frac{N-1}{N} \sum_{h=1}^{L} \Big(\frac{n'_h - 1}{n' - 1} - \frac{n_h - 1}{N - 1}\Big) \frac{w_h s_h^2}{n_h} + \frac{N - n'}{N(n' - 1)} \sum_{h=1}^{L} w_h (\bar{y}_h - \bar{y}_{2st})^2$$

ebenfalls erwartungstreu geschätzt werden.

b) *Zweiphasige Stichprobe zur Regressionsschätzung*

An den n' Stichprobeneinheiten der Phase 1 wird bei diesem Verfahren an Stelle von q die quantitative Hilfsvariable x ermittelt

$$u_1, u_2, \ldots, u_{n'}$$

$$x_1, x_2, \ldots, x_{n'}$$

In der Phase 2 wird an einer Teilstichprobe vom Umfang $n < n'$, die ebenfalls ohne Zurücklegen aus den n' Einheiten der Phase 1 gezogen wird, die Zielvariable y erhoben. Diese Teilstichprobe dient dann zur Schätzung der Parameter eines linearen Regressionsmodells mit Hilfe der Methode der kleinsten Quadrate, so daß für jede der n' Stichprobeneinheiten der Phase 1 eine Schätzung

$$\hat{y}_i = \bar{y} + b \cdot (x_i - \bar{x})$$

des Merkmalswertes y_i erfolgen kann, in der

$$b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

verwendet wird. $\bar{x}$ und $\bar{y}$ sind die Mittelwerte über die n Einheiten der Phase 2. Dann liegt es nahe $\bar{Y}$ durch

$$\bar{y}_{2lr} = \frac{1}{n'} \sum_{i=1}^{n'} \hat{y}_i = \bar{y} + b(\bar{x}' - \bar{x})$$

zu schätzen. $\bar{x}'$ ist der Mittelwert über alle n' Einheiten der Phase 1. Es gilt $E\,\bar{y}_{2lr} \approx \bar{Y}$ und

$$Var\,\bar{y}_{2lr} \approx \frac{S^2(1 - \rho^2)}{n} + \frac{S^2\rho^2}{n'} - \frac{S^2}{N} \tag{2.2}$$

wenn $1/n$ klein gegenüber 1 d.h. n groß ist. $\bar{y}_{21r}$ hat einen Bias, der von der Ordnung 1 ist, d.h. der wie $1/n$ gegen 0 konvergiert. Er ist besonders klein, wenn die Annahme eines linearen Zusammenhangs zwischen y und der oder den quantitativen Hilfsvariablen zutrifft. Der Stichprobenfehler kann durch

$$v(\bar{y}_{21r}) = \frac{s^2(1-r^2)}{n} + \frac{s^2 r^2}{n'} - \frac{s^2}{N}$$

geschätzt werden. Hierbei sind s^2 und r die empirische Varianz von y und der empirische Korrelationskoeffizient auf der Basis der n Einheiten aus Phase 2.

Wird in $\bar{y}_{21r}$ statt der einfachen eine multiple Regression verwendet, so sind statt ρ und r die entsprechenden multiplen Größen zu verwenden. Unter Umständen muß der Stichprobenfehler (2.2) gemäß *COCHRAN 1977, (12.57)* korrigiert werden.

Die zweiphasige Regressionsschätzung hängt stark von der Ausprägung des linearen Zusammenhangs zwischen der Zielgröße und den Hilfsvariablen ab. Insbesondere wenn in Straten, in die das Inventurgebiet zerlegt werden kann, unterschiedliche Zusammenhänge bestehen, die im gepoolten Datenmaterial nicht erkennbar sind, wie z.B. einfache Niveauverschiebungen (Abb.), d.h. ungleiche Konstanten im linearen Regressionsmodell, können individuelle Regressionsschätzungen in diesen Straten zu einer effektiveren Inventur beitragen.

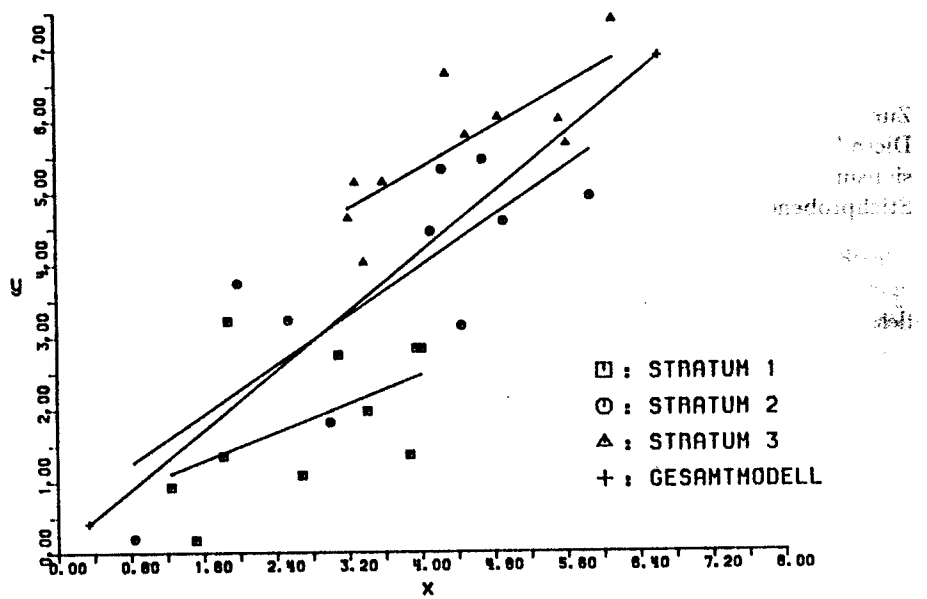

Abb.: *Drei simulierte zweidimensional normalverteilte Stichproben (x_i, y_i) vom Umfang $n = 10$ mit $\rho = 0.8$, $S_x^2 = 1.1$, $S_y^2 = 0.7$. Die Erwartungswerte sind (3,2), (4,4) und (5,6).*

Aus dieser Idee entstanden die kombinierten Schätzverfahren des vierten Kapitels. Im dritten Kapitel wird ein übergeordnetes Ergebnis formuliert, mit dessen Hilfe die Stichprobenfehler für die später folgenden kombinierten Verfahren abgeleitet werden können.

3. ALLGEMEINERE ZWEIPHASIGE STICHPROBEN ZUR STRATIFIZIERUNG

Wir unterstellen einmal, daß mit der Bezeichnung $\hat{\bar{Y}}_h$ irgendein nicht näher spezifizierter Schätzer gemeint ist, der bei gegebener (Phase 1)-Stichprobe $u_1, \ldots, u_{n'}$ ein erwartungstreuer Schätzer für den Mittelwert des Merkmals y der in dieser Stichprobe enthaltenen, zu Stratum h gehörigen Einheiten ist, d.h.

$$E(\hat{\bar{Y}}_h \mid Phase\,1) = \frac{1}{n'_h} \sum_{j=1}^{n'_h} y_{hj}$$

Mit

$$\bar{y}_{2st,allg} = \sum_{h=1}^{L} w_h \hat{\bar{Y}}_h$$

gilt dann nämlich

$$E\,\bar{y}_{2st,allg} = \bar{Y} \tag{3.1}$$

$$Var\,\bar{y}_{2st,allg} = \frac{1}{n'}\left(1 - \frac{n'}{N}\right) S^2 + E \sum_{h=1}^{L} w_h^2\, Var(\hat{\bar{Y}}_h \mid Phase\,1) \tag{3.2}$$

Dabei ist $Var(\hat{\bar{Y}}_h \mid Phase\,1)$ die (bedingte) Varianz dieses Schätzers bei gegebener Phase 1. Diese Darstellung zeigt, auf welche Weise man den Stichprobenfehler einer so allgemeinen zweiphasigen Stichprobe erhält. Im Standardfall a) von Kapitel 2 ist z.B.

$$\hat{\bar{Y}}_h = \bar{y}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$$

$$\begin{aligned} Var(\hat{\bar{Y}}_h \mid Phase\,1) &= \frac{1}{n_h}\left(1 - \frac{n_h}{n'_h}\right) \frac{1}{n'_h - 1} \sum_{j=1}^{n'_h} (y_{hj} - \bar{y}'_h)^2 \\ &= \frac{1}{n_h}\left(1 - \frac{n_h}{n'_h}\right) \cdot s'^2_h \end{aligned}$$

Die Berechnung des Erwartungswertes in (3.2) liefert dann nach einigen Schritten die bekannte Formel (2.1). Ähnlich werden die Ergebnisse des nächsten Kapitels aus (3.1) und (3.2) hergeleitet.

4. KOMBINATION MIT EINFACHER REGRESSIONSSCHÄTZUNG

In diesem Kapitel wird $\hat{\bar{Y}}_h$ als einfacher (einphasiger) Regressionsschätzer angenommen. Dazu muß an den n' Stichprobeneinheiten $u_1, \ldots, u_{n'}$ der ersten Phase zusätzlich zum Merkmal q auch eine quantitative Hilfsvariable x aufgenommen werden.

$$u_1, \ldots, u_{n'}$$
$$q_1, \ldots, q_{n'}$$
$$x_1, \ldots, x_{n'}$$

q dient dann wieder der Zuordnung der Stichprobeneinheiten zu L gegebenen Straten. Aus den Stichprobeneinheiten $u_{h1}, \ldots, u_{hn'_h}$ jedes Stratums h wird anschließend eine Teilstichprobe vom Umfang n_h ohne Zurücklegen ausgewählt, wobei auch hier wieder der Stichprobenanteil $\nu_h = n_h/n'_h$ vor der Erhebung fixiert werden muß. An dieser Teilstichprobe wird das Merkmal y erhoben, so daß sich hier für jedes Stratum individuelle Regressionsparameter b_h schätzen lassen. Mit

$$\hat{\bar{Y}}_h = \bar{y}_h + b_h(\bar{x}'_h - \bar{x}_h)$$

kann der Mittelwert $\bar{Y}$ dann durch

$$\bar{y}_{2st,lr} = \sum_{h=1}^{L} w_h \cdot \hat{\bar{Y}}_h$$

geschätzt werden. Dabei sind $\bar{x}'_h$ und $\bar{x}_h$ die Mittelwerte aus den n'_h Stichprobeneinheiten des Stratums h und aus der daraus gezogenen Teilstichprobe vom Umfang n_h, und

$$b_h = \frac{\sum_{j=1}^{n_h}(y_{hj} - \bar{y}_h)(x_{hj} - \bar{x}_h)}{\sum_{j=1}^{n_h}(x_{hj} - \bar{x}_h)^2}$$

Bei gegebener Phase 1 lassen sich die bekannten Ergebnisse über Regressionsschätzer verwenden, so daß hier gilt

$$E\left(\hat{\bar{Y}}_h \,|Phase\,1\right) \approx \frac{1}{n'_h}\sum_{j=1}^{n'_h} y_{hj} \qquad Var\left(\hat{\bar{Y}}_h|Phase\,1\right) \approx \frac{1-\nu_h}{n_h} s'^2_h (1 - r'^2_h)$$

Der bekannte Bias der Regressionsschätzer und von deren Varianzapproximation (siehe *COCHRAN 1977*) überträgt sich dann natürlich auch auf den kombinierten zweiphasigen Schätzer $\bar{y}_{2st,lr}$. Es gilt

$$E\,\bar{y}_{2st,lr} \approx \bar{Y}$$

$$Var\,\bar{y}_{2st,lr} \approx \frac{1}{n'}\left(1 - \frac{n'}{N}\right)S^2 + \sum_{h=1}^{L}\frac{W_h}{n'}\left(\frac{1}{\nu_h} - 1\right)S_h^2(1 - \rho_h^2) \tag{4.1}$$

und der Stichprobenfehler $Var\,\bar{y}_{2stlr}$ kann durch

$$v(\bar{y}_{2stlr}) = v(\bar{y}_{2st}) - \sum_{h=1}^{L}\frac{w_h}{n'}\left(\frac{1}{\nu_h} - 1\right)s_h^2 r_h^2 \tag{4.2}$$

geschätzt werden. In den Formeln (4.1) und (4.2) bedeuten ρ_h die Korrelation zwischen x und y über alle N_h Einheiten und r_h die aus den n_h Stichprobeneinheiten geschätzte Korrelation in Stratum h.

Offensichtlich hat dieses zweiphasige Verfahren, das die Stratifizierung mit individuellen Regressionsschätzern in den Straten kombiniert, einen gegenüber dem bekannten zweiphasigen Verfahren aus Kapitel 2.a) um den Betrag

$$Var\,\bar{y}_{2st} - Var\,\bar{y}_{2st,lr} \approx \sum_{h=1}^{L} \frac{W_h}{n'} \left(\frac{1}{\nu_h} - 1\right) S_h^2 \rho_h^2 \tag{4.3}$$

geringeren Stichprobenfehler, dessen Größe von der Straffheit des Zusammenhangs zwischen x und y abhängt. Es darf dabei aber nicht vergessen werden, daß diese Verringerung des Stichprobenfehlers bei sonst übereinstimmenden Stichprobenumfängen n' und n_h durch den zusätzlichen Meßaufwand erkauft wird, der mit der Erfassung der Variablen x verbunden ist. Dennoch ist zu erwarten, daß bei hinreichend hoher Korrelation und Kostendifferenz für die Erhebung der Merkmale x und y bei gleichen Gesamtkosten ein geringerer Stichprobenfehler möglich ist. Mit der Frage der für diese Zielsetzung optimalen Stichprobenumfänge wollen wir uns nun befassen.

5. OPTIMALE STICHPROBENUMFÄNGE FÜR $\bar{y}_{2st,lr}$

Als optimale Stichprobenumfänge bezeichnen wir diejenige Festlegung von n' und $\nu_h = n_h/n'_h$, die bei vorgegebenen Kosten den Stichprobenfehler (siehe (4.1)) oder umgekehrt bei vorgegebenem Stichprobenfehler die Kosten der Inventur minimiert. Die Kosten einer zweiphasigen Inventur, wie sie in Kapitel 4. beschrieben wurde, setzen sich aus den Kosten c' für die Zuordnung einer der n' Stichprobeneinheiten zu einem der L Straten und den Kosten $c_h(x)$ bzw. $c_h(y)$ für die Erhebung des Merkmals x bzw. y im Stratum h zusammen.

$$\begin{aligned} C =& c'n' + \sum_{h=1}^{L} c_h(x) n'_h + \sum_{h=1}^{L} c_h(y) n_h \\ =& n' \left(c' + \sum_{h=1}^{L} c_h(x) w_h + \sum_{h=1}^{L} c_h(y) \nu_h w_h\right) \end{aligned}$$

Sind die einzelnen Kosten unabhängig von den Straten, so gilt demnach

$$C = n' \left(c' + c(x) + c(y) \sum_{h=1}^{L} \nu_h w_h\right)$$

Beide Formulierungen der Gesamtkosten sind aber zufällige Größen, da auch die Stichprobenumfänge n'_h und n_h zufällig sind. Deshalb arbeitet man statt dessen in der Planungsphase mit den zu erwartenden Kosten

$$E\,C = n' \left(c' + \sum_{h=1}^{L} c_h(x) W_h + \sum_{h=1}^{L} c_h(y) \nu_h W_h\right)$$

bzw.

$$EC = n'\Big(c' + c(x) + c(y)\sum_{h=1}^{L}\nu_h W_h\Big)$$

Mit Hilfe der Cauchy-Schwarzschen Ungleichung erhält man als Lösung der Minimierungsaufgabe

$$\nu_h = \sqrt{\frac{c' + \sum_{h=1}^{L} W_h c_h(x)}{c_h(y)} \cdot \frac{S_h^2(1-\rho_h^2)}{S^2 - \sum_{h=1}^{L} W_h S_h^2(1-\rho_h^2)}}$$

und n' ergibt sich aus den vorgegebenen zu erwartenden Kosten gemäß

$$n' = \frac{EC}{c' + \sum_{h=1}^{L} W_h\big(c_h(x) + c_h(y)\nu_h\big)}$$

oder aus der vorgegebenen $Var\,\bar{y}_{2st,lr}$ gemäß

$$n' = \frac{S^2 + \sum_{h=1}^{L} W_h\big(\frac{1}{\nu_h} - 1\big)S_h^2(1-\rho_h^2)}{Var\,\bar{y}_{2st,lr} + (S^2/N)}$$

Die durch diese Wahl von n und der ν_h erreichte minimale Varianz ergibt sich ebenfalls aus der Cauchy-Schwarzschen Ungleichung als

$$(Var\,\bar{y}_{2st,lr})_{min} = \frac{1}{EC}\Big(\sqrt{c' + \sum W_h c_h(x)}\cdot \cdot\sqrt{S^2 - \sum W_h S_h^2(1-\rho_h^2)} + \sum W_h\sqrt{c_h(y)S_h^2(1-\rho_h^2)}\Big)^2 - \frac{S^2}{N}$$

6. KOMBINATION MIT ZWEIPHASIGER REGRESSIONSSCHÄTZUNG

Im Unterschied zu dem in Kapitel 4 beschriebenen Verfahren wird hier die Hilfsvariable x erst nach der Stratifizierung der Stichprobe aus Phase 1 und nur an Unterstichproben vom Umfang $n_h = \nu_h n_h'$ aus den Straten erhoben.

$u_{h1}, \ldots\ldots\ldots, u_{hn_h'}$ (*Stichprobe aus Stratum h*)

$u_{h1}, \ldots, u_{hn_h}$ (*Teilstichprobe*)

$x_{h1}, \ldots, x_{hn_h}$ (*Hilfsvariable*)

Die Zielvariable y wird wiederum an einer Teilstichprobe von $u_{h1}, \ldots, u_{hn_h}$ des Umfangs $n_h^* = \nu_h^* n_h$ ermittelt. Sämtliche Stichproben und Teilstichproben werden wie bisher ohne Zurücklegen gezogen, und die Stichprobenfraktionen ν_h und ν_h^* liegen zwischen 0 und 1 mit der Zusatzforderung, daß n_h und n_h^* größer als 1 sind, damit Varianzschätzungen möglich sind. Insgesamt haben wir es nun mit einem dreiphasigen Verfahren zu tun, dessen erste Phase der Stratifizierung dient, während in den

Phasen zwei und drei je Stratum eine zweiphasige Regressionsschätzung bei gegebener Stichprobe der Phase 1 durchgeführt wird.

Als Schätzer $\hat{\bar{Y}}_h$, gemäß Kapitel 3 wird hier der zweiphasige Regressionsschätzer

$$\hat{\bar{Y}}_h = \bar{y}_h^* + b_h^*(\bar{x}_h - \bar{x}_h^*)$$

verwendet. Dabei sind $\bar{x}_h$ und $\bar{x}_h^*$ die Mittelwerte aus den n_h Stichprobeneinheiten des Stratums h, an denen die Hilfsvariable ermittelt wurde, und aus der daraus gezogenen Teilstichprobe vom Umfang n_h^* (entsprechend $\bar{y}_h^*$), und

$$b_h^* = \frac{\sum_{j=1}^{n_h^*}(y_{hj} - \bar{y}_h^*)(x_{hj} - \bar{x}_h^*)}{\sum_{j=1}^{n_h^*}(x_{hj} - \bar{x}_h^*)^2}$$

Mit den für große n_h^* gültigen Formeln

$$E(\hat{\bar{Y}}_h \mid Phase\,1) \approx \frac{1}{n_h'}\sum_{j=1}^{n_h'} y_{hj} \qquad Var(\hat{\bar{Y}}_h \mid Phase\,1) \approx \frac{s_h'^{\,2}(1-r_h'^{\,2})}{n_h^*} + \frac{s_h'^{\,2}r_h'^{\,2}}{n_h} - \frac{s_h'^{\,2}}{n_h'}$$

(siehe (2.2)) ist dann der Mittelwertschätzer

$$\bar{y}_{2st2,lr} = \sum_{h=1}^{L} w_h \cdot \hat{\bar{Y}}_h$$

((3.1) und (4.1) entsprechend) annähernd erwartungstreu, und seine Varianz ergibt sich nach weiteren Schritten als

$$Var\,\bar{y}_{2st2,lr} \approx \frac{1}{n'}\Big(1 - \frac{n'}{N}\Big)S^2 + \sum_{h=1}^{L}\frac{W_h}{n'}\left(\frac{S_h^2(1-\rho_h^2)}{\nu_h^*\nu_h} + \frac{S_h^2\rho_h^2}{\nu_h} - S_h^2\right) \tag{6.1}$$

Sie kann durch

$$v(\bar{y}_{2st,2lr}) = v(\bar{y}_{2st}) + \sum_{h=1}^{L}\frac{w_h}{n'}\Big(\frac{1}{\nu_h^*} - 1\Big)\frac{s_h^{*2}(1-r_h^{*2})}{\nu_h} \tag{6.2}$$

geschätzt werden. In Formel (6.2) ist r_h^* die empirische Korrelation zwischen x und y über die n_h^* Einheiten der Phase 3, s_h^{*2} die Streuung der y-Werte derselben Einheiten. Vergleicht man (6.1) mit (2.1) und (4.1), so ergibt sich

$$Var\,\bar{y}_{2st,2lr} - Var\,\bar{y}_{2st} \approx \sum_{h=1}^{L}\frac{W_h}{n'}\Big(\frac{1}{\nu_h^*} - 1\Big)\frac{S_h^2(1-\rho_h^2)}{\nu_h} \geq 0$$

$$Var\,\bar{y}_{2st,2lr} - Var\,\bar{y}_{2st,lr} \approx \sum_{h=1}^{L}\frac{W_h}{n'}\left(\Big(\frac{1}{\nu_h^*} - 1\Big)\frac{S_h^2(1-\rho_h^2)}{\nu_h} + \Big(\frac{1}{\nu_h} - 1\Big)S_h^2\rho_h^2\right) \geq 0$$

Zusammen mit (4.3) ist damit eine Anordnung der drei Verfahren aus 2.a), 4. und 6. nach Stichprobenfehlern möglich. Sie gilt, wenn in allen drei Verfahren mit gleichem n' und gleichen ν_h gearbeitet wird. In der Praxis wird man demgegenüber natürlich immer anstreben, die aufwendigere Erhebung des Merkmals y soweit wie möglich durch die Hilfsvariable x zu substituieren, um so ohne Vergrößerung des Stichprobenfehlers mit geringeren Kosten auszukommen. D.h. es muß in jeder konkreten Situation geprüft werden, ob z.B. $\bar{y}_{2st,lr}$ gegenüber $\bar{y}_{2st}$ mit kleinerem ν_h und dadurch verringerten Kosten durch Nutzung von Hilfsvariablen dennoch einen vergleichbar kleinen Stichprobenfehler erreicht.

Insbesondere aber können $\bar{y}_{2st,lr}$ und $\bar{y}_{2st,2lr}$ in einer Situation wie sie in der Abbildung von Kapitel 2 dargestellt ist kostengünstigere Lösungen ermöglichen als y_{2lr}.

Für $\nu_h^* = 1$ d.h. $n_h^* = n_h$ stimmen im übrigen die Varianzen von $\bar{y}_{2st,2lr}$ und $\bar{y}_{2st}$ überein. Dies ist auch plausibel, denn in diesem Fall ist $\hat{\bar{Y}}_h = \bar{y}_h$ und damit $\bar{y}_{2st,2lr} = \bar{y}_{2st}$.

7. OPTIMALE STICHPROBENUMFÄNGE FÜR $\bar{y}_{2st,2lr}$

Wie in Kapitel 5 wird auch hier wieder nach Festlegungen von n', ν_h und ν_h^* gesucht, die bei vorgegebenen zu erwartenden Kosten die Varianz von $\bar{y}_{2st,2lr}$ minimieren. c', $c_h(x)$ und $c_h(y)$ sind schon von dort bekannt, und die Gesamtkosten bzw. deren Erwartungswert sind

$$\begin{aligned} C =& c'n' + \sum_{h=1}^{L} c_h(x) n_h + \sum_{h=1}^{L} c_h(y) n_h^* \\ =& n' \Big(c' + \sum_{h=1}^{L} c_h(x) \nu_h w_h + \sum_{h=1}^{L} c_h(y) \nu_h^* \nu_h w_h \Big) \end{aligned}$$

$$EC = n' \Big(c' + \sum_{h=1}^{L} c_h(x) \nu_h W_h + \sum_{h=1}^{L} c_h(y) \nu_h^* \nu_h W_h \Big)$$

Die gesuchten optimalen Stichprobenfraktionen sind

$$\nu_h = \sqrt{\frac{c'}{c_h(x)} \cdot \frac{\rho_h^2 S_h^2}{S^2 - \sum_{h=1}^{L} W_h S_h^2}} \qquad \nu_h^* = \frac{1}{\nu_h} \sqrt{\frac{c_h(x)}{c_h(y)} \cdot \frac{1-\rho_h^2}{\rho_h^2}}$$

und

$$n' = \frac{EC}{c' + \sum_{h=1}^{L} c_h(x) \nu_h W_h + \sum_{h=1}^{L} c_h(y) \nu_h^* \nu_h W_h}$$

falls EC, bzw.

$$n' = \frac{S^2 + \sum_{h=1}^{L} W_h \left(\frac{S_h^2 (1-\rho_h^2)}{\nu_h^* \nu_h} + \frac{S_h^2 \rho_h^2}{\nu_h} - S_h^2 \right)}{Var\, \bar{y}_{2st,2lr} + (S^2/N)}$$

falls $Var\,\bar{y}_{2st,2lr}$ vorgegeben wird. Die mit dieser Wahl der Stichprobenumfänge erzielte minimale Varianz ist

$$(Var\,\bar{y}_{st,2lr})_{min} = \frac{1}{EC}\Big(\sqrt{c'(S^2 - \sum W_h S_h^2)} + \\ + \sum W_h \sqrt{c_h(x) S_h^2 \rho_h^2} + \sum W_h \sqrt{c_h(y) S_h^2 (1-\rho_h^2)}\Big)^2 - \frac{S^2}{N}$$

8. VERGLEICH VON $\bar{y}_{2lr}$ UND $\bar{y}_{2st,lr}$

Der Vergleich der Varianzen der bekannten zweiphasigen Regressionsschätzung (n'-mal Hilfsvariable x, davon n-mal auch Zielgröße y) einerseits

$$\begin{aligned} Var\,\bar{y}_{lr} &\approx \frac{S^2(1-\rho^2)}{n} + \frac{S^2\rho^2}{n'} - \frac{S^2}{N} \\ &= \frac{S^2(1-\rho^2)}{n} + \Big(\frac{1}{n'} - \frac{1}{N}\Big)S^2 + \frac{S^2(\rho^2-1)}{n'} \\ &= \frac{1}{n'}\Big(1 - \frac{n'}{N}\Big)S^2 + \Big(\frac{1}{n} - \frac{1}{n'}\Big)S^2(1-\rho^2) \end{aligned}$$

und der mit einer einfachen Regressionsschätzung kombinierten zweiphasigen Stratifizierung andererseits mit

$$\sum_{h=1}^{L} n'_h = n' \qquad \sum_{h=1}^{L} n_h = n$$

also gleichen Stichprobenumfängen für die Merkmale x und y und der Vereinfachung:

$$\text{identische Stichprobenanteile} \quad \nu_h = \nu = \frac{n}{n'} \quad \text{in allen Straten}$$

d.h.

$$\begin{aligned} Var\,\bar{y}_{2st,lr} &\approx \frac{1}{n'}\Big(1 - \frac{n'}{N}\Big)S^2 + \sum_{h=1}^{L} \frac{W_h}{n'}\Big(\frac{1}{\nu} - 1\Big)S_h^2(1-\rho_h^2) \\ &= \frac{1}{n'}\Big(1 - \frac{n'}{N}\Big)S^2 + \Big(\frac{1}{n} - \frac{1}{n'}\Big)\sum_{h=1}^{L} W_h S_h^2(1-\rho_h^2) \end{aligned}$$

zeigt, daß in diesem Fall

$$\begin{aligned} &Var\,\bar{y}_{2lr} > Var\,\bar{y}_{2st,lr} \qquad \Leftrightarrow \\ &S^2(1-\rho^2) > \sum_{h=1}^{L} W_h S_h^2(1-\rho_h^2) \quad =: \overline{S_h^2(1-\rho_h^2)} \end{aligned}$$

Bei den hier vorausgesetzten Stichprobenumfängen sind natürlich im Normalfall höhere Kosten durch die zusätzliche Stratifizierung zu berücksichtigen. Aber je größer der Unterschied zwischen $S^2(1-\rho^2)$ und dem über alle Straten gebildeten Mittelwert

$\overline{S_h^2(1-\rho_h^2)}$ ist, umso eher wird das kombinierte Verfahren bei gleichen Kosten einen geringeren Stichprobenfehler haben, bzw. den gleichen Stichprobenfehler mit geringeren Kosten erreichen. Darüberhinaus wird z.B. bei der Wiederholung einer zu einem früheren Zeitpunkt durchgeführten Inventur häufig eine gewisse Anzahl als Stratifizierungsmerkmal geeigneter Variablen bereits vorhanden sein, so daß die Stratifizierung keine weiteren Kosten verursacht.

In dem in der Abbildung von Kapitel 2 dargestellten Fall ergibt sich für die Gesamtpopulation

$$S^2(1-\rho^2) = 3.82(1-0.77^2) = 1.56$$

und für die einzelnen Straten

S_h^2	ρ_h	$S_h^2(1-\rho_h^2)$
1.04	0.52	0.76
2.76	0.79	1.04
0.93	0.79	0.35

Mit $W_h = N_h/N = 1/L$ folgt daraus

$$\overline{S_h^2(1-\rho_h^2)} = (0.76 + 1.04 + 0.35)/3 = 0.72$$

Bei etwa gleich großen Korrelationen wie in der Gesamtpopulation haben sich hier die deutlich kleineren Varianzen S_h^2 in den Straten gegenüber $S^2 = 3.82$ durchgesetzt.

9. ZWEIPHASIGE REGRESSIONSSCHÄTZUNG FÜR pps-AUSWAHL

Als Versuch zur Effektivitätsverbesserung der zweiphasigen Stichprobe zur Regressionsschätzung (siehe 2.b)) ist auch das Verfahren von *KÄTSCH 1991* zu werten, das dort als "Listenstichprobe" bezeichnet wird. Dieses Verfahren verwendet die zweiphasigen Regressionsschätzungen

$$\hat{y}_i = \bar{y} + b \cdot (x_i - \bar{x}) \qquad i = 1, \ldots, n'$$

nach der notwendigen Normierung als Auswahlwahrscheinlichkeiten für eine weitere pps-Auswahl von Probekreisen, in denen erneut die Werte des (terrestrischen) y-Merkmals festgestellt werden. Wir betrachten hier wieder nur eine unabhängige Variable, während dort ein multiples Regressionsmodell verwendet wird. Ein Vergleich mit der konkurrierenden reinen zweiphasigen Regressionsschätzung wurde bei *Kätsch* nicht durchgeführt, da kein Schätzer für die Varianz bekannt war.

Aus der (Phase 1)-Stichprobe mit Umfang n' wird also mit den Auswahlwahrscheinlichkeiten

$$p_i = \frac{\hat{y}_i}{\sum_{j=1}^{n'} \hat{y}_j} \qquad i = 1, \ldots, n'$$

eine Listenstichprobe vom Umfang n_0 mit den Merkmalswerten $y_1, \ldots, y_{n_0}$ gezogen. In ihr können dann auch Stichprobeneinheiten enthalten sein, die schon in Phase 2 im Rahmen einer einfachen Zufallsstichprobe ohne Zurücklegen ausgewählt wurden. Diese Listenstichprobe ermöglicht die Schätzung des Merkmalsmittelwerts $\bar{Y}$ durch

$$\bar{y}_{2lr,pps} = \frac{1}{n'} \cdot \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{y_i}{p_i} = \frac{1}{n'} \cdot \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{y_i}{\hat{y}_i} \cdot \sum_{j=1}^{n'} \hat{y}_j$$

Der (bedingte) Erwartungswert des Mittelwertschätzers $\bar{y}_{2lr,pps}$ bei gegebener zweiphasiger Stichprobe ist offenbar

$$\frac{1}{n'}\sum_{i=1}^{n'} y_i$$

d.h. der Mittelwert der Einheiten der (Phase 1)-Stichprobe. Damit ist der (unbedingte) Erwartungswert gerade der gesuchte Mittelwert im Inventurgebiet, und $\bar{y}_{2lr,pps}$ ist erwartungstreu. Die (bedingte) Varianz bei gegebener zweiphasiger Stichprobe ist durch die Formel

$$\frac{1}{n'^2}\cdot\frac{1}{n_0}\sum_{i=1}^{n'} p_i\Big(\frac{y_i}{p_i}-Y'\Big)^2$$

gegeben ($Y'=\sum_{i=1}^{n'} y_i$), d.h. durch die Varianz der pps-Mittelwertschätzung für eine Grundgesamtheit vom Umfang n'.

Bei der Varianzschätzung darf aber nicht nur die Streuung der reinen Listenstichprobe berücksichtigt werden, denn auch die vorausgegangene zweiphasige Stichprobe trägt natürlich zum Stichprobenfehler bei. Die Varianz von $\bar{y}_{2lr,pps}$ setzt sich daher nach einer bekannten Formel der Stichprobentheorie additiv aus zwei Komponenten zusammen

$$Var\,\bar{y}_{2lrpps} = Var\,E\,(\bar{y}_{2lr,pps} \mid Phase\,1u.2) \;+\; E\,Var\,(\bar{y}_{2lrpps} \mid Phase\,1u.2)$$

aus der Varianz des bedingten Erwartungswertes und dem Erwartungswert der bedingten Varianz von $\bar{y}_{2lr,pps}$. Daraus folgt also

$$Var\,\bar{y}_{2lr,pps} = Var\,\frac{1}{n'}\sum_{i=1}^{n'} y_i \;+\; E\,\frac{1}{n'^2}\cdot\frac{1}{n_0}\sum_{i=1}^{n'} p_i\Big(\frac{y_i}{p_i}-Y'\Big)^2$$

Der zweite Summand kann dann durch den Varianzschätzer für die pps-Listenstichprobe geschätzt werden, der erste durch

$$\frac{1}{n'}\Big(1-\frac{n'}{N}\Big)\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

Denn dies ist die Varianzschätzung für den Stichprobenmittelwert einer ohne Zurücklegen gezogenen Zufallsstichprobe vom Umfang n' (darum handelt es sich bei $\frac{1}{n'}\sum_{i=1}^{n'} y_i$) mit Hilfe einer ebensolchen Stichprobe vom Umfang n, die hier aus Phase 2 vorliegt. Insgesamt erhält man so den erwartungstreuen Varianzschätzer

$$v(\bar{y}_{2lr,pps}) = \frac{1}{n'}\Big(1-\frac{n'}{N}\Big)\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 + \frac{1}{n'^2}\cdot\frac{1}{n_0(n_0-1)}\sum_{i=1}^{n_0}\Big(\frac{y_i}{p_i}-\bar{y}_{2lr,pps}\Big)^2$$

für dieses Verfahren.

LITERATUR

COCHRAN, W.G. (1977): Sampling techniques. *Wiley, New York*

FRAYER, W.E. (1979): Multi-level sampling designs for resource inventories. *USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins*

JEYARATNAM, S.; Bowden, D.C.; Graybill, F.A.; Frayer, W.E. (1984): Estimation in multiphase designs for stratification. *Forest Science 30,2, S.484-491*

KÄTSCH, C. (1991): Zweiphasige Stichprobenverfahren für Zwecke der Betriebsinventur auf der Basis einfacher Luftbildauswertung. *Dissertation, Göttingen*

KÖHL, H.M. (1990): Zweiphasige Stichprobenverfahren zur Holzvorratsschätzung. *Jahrestagung Deutscher Verband Forstlicher Forschungsanstalten - Sektion Biometrie -, Göttingen*

SABOROWSKI, J. (1990): Schätzung von Varianzen und Konfidenzintervallen aus mehrstufigen Stichproben. *Schriften aus der Forstlichen Fakultät der Universität Göttingen 99, J.D. Sauerländer's, Frankfurt*

WOLFF, B. (1990): Verbesserung der Effizienz terrestrischer Kontrollstichproben durch ein zweiphasiges Stichprobenverfahren. *Jahrestagung Deutscher Verband Forstlicher Forschungsanstalten - Sektion Biometrie -, Göttingen*

ANHANG

Beweis (3.1):

Bei gegebener (Phase 1)-Stichprobe ist der bedingte Erwartungswert von $\bar{y}_{2st,allg}$ durch

$$E(\bar{y}_{st,allg} \mid Phase\,1) = E\left(\sum_{h=1}^{L} w_h \hat{\bar{Y}}_h \,\middle|\, Phase\,1\right) = \sum_{h=1}^{L} w_h E(\hat{\bar{Y}}_h \mid Phase\,1)$$

$$= \sum_{h=1}^{L} w_h \frac{1}{n'_h} \sum_{j=1}^{n'_h} y_{hj} = \frac{1}{n'} \sum_{h=1}^{L} \sum_{j=1}^{n'_h} y_{hj} = \frac{1}{n'} \sum_{j=1}^{n'} y_j$$

gegeben, so daß

$$E\,\bar{y}_{2st,allg} = E\,E(\bar{y}_{2st,allg} \mid Phase\,1) = E\,\frac{1}{n'} \sum_{j=1}^{n'} y_j = \bar{Y}$$

q.e.d.

Beweis (3.2):

Es ist

$$Var\,\bar{y}_{2st,allg} = Var\,E(\bar{y}_{2st,allg} \mid Phase\,1) + E\,Var(\bar{y}_{2st,allg} \mid Phase\,1)$$

mit

$$Var\,(\bar{y}_{2st,allg} \mid Phase\,1) = \sum_{h=1}^{L} w_h^2\, Var\,(\hat{\bar{Y}}_h \mid Phase\,1)$$

und (siehe Beweis (3.1))

$$Var\,E\,(\bar{y}_{2st,allg} \mid Phase\,1) = Var\,\frac{1}{n'}\sum_{j=1}^{n'} y_j = \frac{1}{n'}\Big(1-\frac{n'}{N}\Big)S^2$$

q.e.d.

Beweis (4.1):

Nach (3.2) und der bekannten Varianz des Regressionsschätzers folgt

$$Var\,(\bar{y}_{2st,lr}) \approx \frac{1}{n'}\Big(1-\frac{n'}{N}\Big)S^2 + E\sum_{h=1}^{L} w_h^2 \frac{1-\nu_h}{n_h} s_h'^{\,2}(1-r_h'^{\,2})$$

Der Erwartungswert wird wieder durch Einschaltung eines bedingten Erwartungswertes, diesmal bei gegebenen n_h', gebildet. Es ist also nicht die gesamte (Phase 1)-Stichprobe gegeben sondern lediglich die Anzahl n_h' der aus Stratum h ausgewählten Einheiten, so daß der bedingte Erwartungswert in jedem Stratum über alle einfachen Zufallsstichproben vom festen Umfang n_h' gebildet wird. Mit n_h' ist auch ν_h und damit n_h fest.

$$\begin{aligned}
E\sum_{h=1}^{L} w_h^2 \frac{1-\nu_h}{n_h} s_h'^{\,2}(1-r_h'^{\,2}) &= E\sum_{h=1}^{L} E\left(w_h^2 \frac{1-\nu_h}{n_h} s_h'^{\,2}(1-r_h'^{\,2}) \,\middle|\, n_h', h=1,\ldots,L\right) \\
&\approx E\sum_{h=1}^{L} w_h^2 \frac{1-\nu_h}{n_h} S_h^2(1-\rho_h^2) \qquad \textit{(nach Cochran 1977, S.195)} \\
&= E\sum_{h=1}^{L} \frac{w_h}{n'} \frac{1-\nu_h}{\nu_h} S_h^2(1-\rho_h^2) = \sum_{h=1}^{L} \frac{W_h}{n'}\Big(\frac{1}{\nu_h}-1\Big)S_h^2(1-\rho_h^2)
\end{aligned}$$

q.e.d.

Beweis (4.2):

Die Varianz von $\bar{y}_{2st,lr}$ in (4.1) unterscheidet sich von $Var\,\bar{y}_{2st}$ (siehe (2.1)) um den Ausdruck

$$-\sum_{h=1}^{L} \frac{W_h}{n'}\Big(\frac{1}{\nu_h}-1\Big)S_h^2\rho_h^2 \qquad (*)$$

Nach *Cochran 1977 (S.195)* gilt auch

$$E\left(\sum_{h=1}^{L} \frac{w_h}{n'}\Big(\frac{1}{\nu_h}-1\Big)s_h^2 r_h^2 \,\middle|\, n_h', h=1,\ldots,L\right) \approx \sum_{h=1}^{L} \frac{w_h}{n'}\Big(\frac{1}{\nu_h}-1\Big)S_h^2\rho_h^2$$

so daß

$$-\sum_{h=1}^{L}\frac{w_h}{n'}\Big(\frac{1}{\nu_h}-1\Big)s_h^2 r_h^2$$

annähernd erwartungstreu ist für (*) und damit

$$v(\bar{y}_{2st,lr}) = v(\bar{y}_{2st}) - \sum_{h=1}^{L}\frac{w_h}{n'}\Big(\frac{1}{\nu_h}-1\Big)s_h^2 r_h^2$$

annähernd erwartungstreu für (4.1).

q.e.d.

Beweis (6.1):
Nach (3.2) und der Varianz des zweiphasigen Regressionsschätzers (2.2) folgt für große n_h^*

$$Var\,(\bar{y}_{2st,2lr}) \approx \frac{1}{n'}\Big(1-\frac{n'}{N}\Big)S^2 + E\sum_{h=1}^{L} w_h^2\Big(\frac{s_h'^{\,2}(1-r_h'^{\,2})}{n_h^*} + \frac{s_h'^{\,2}r_h'^{\,2}}{n_h} - \frac{s_h'^{\,2}}{n_h'}\Big)$$

und mit der gleichen Argumentation wie im Beweis von (4.1) ergibt sich

$$\begin{aligned} &E\,w_h^2\Big(\frac{s_h'^{\,2}(1-r_h'^{\,2})}{n_h^*} + \frac{s_h'^{\,2}r_h'^{\,2}}{n_h} - \frac{s_h'^{\,2}}{n_h'}\,\Big|\, n_h', h=1,\dots,L\Big) \\ &\quad = w_h^2\Big(\frac{S_h^2(1-\rho_h^2)}{n_h^*} + \frac{S_h^2\rho_h^2}{n_h} - \frac{S_h^2}{n_h'}\Big) = \frac{w_h}{n'}\Big(\frac{S_h^2(1-\rho_h^2)}{\nu_h^*\nu_h} + \frac{S_h^2\rho_h^2}{\nu_h} - S_h^2\Big)\end{aligned}$$

q.e.d.

Beweis (6.2):
Der Vergleich von (6.1) und (2.1) zeigt, daß

$$Var\,\bar{y}_{2st,2lr} \approx Var\,\bar{y}_{2st} + \sum_{h=1}^{L}\frac{W_h}{n'}\Big(\frac{1}{\nu_h^*}-1\Big)\frac{S_h^2(1-\rho^2)}{\nu_h}$$

und die Summe auf der rechten Seite kann annähernd erwartungstreu durch

$$\sum_{h=1}^{L}\frac{w_h}{n'}\Big(\frac{1}{\nu_h^*}-1\Big)\frac{s_h^{*2}(1-r_h^{*2})}{\nu_h}$$

geschätzt werden, denn wie im Beweis zu (4.2) gilt

$$E\Big(\sum_{h=1}^{L}\frac{w_h}{n'}\Big(\frac{1}{\nu_h^*}-1\Big)\frac{s_h^{*2}(1-r_h^{*2})}{\nu_h}\,\Big|\, n_h', h=1,\dots,L\Big) \approx \sum_{h=1}^{L}\frac{w_h}{n'}\Big(\frac{1}{\nu_h^*}-1\Big)\frac{S_h^2(1-\rho_h^2)}{\nu_h}$$

q.e.d.