Untersuchung einiger Stichprobenverfahren für Zeitreihen

Autor(en): **Nievergelt, E.**

Objekttyp: **Article**

Zeitschrift: **Mitteilungen / Vereinigung Schweizerischer Versicherungsmathematiker = Bulletin / Association des Actuaires Suisses = Bulletin / Association of Swiss Actuaries**

Band (Jahr): **62 (1962)**

PDF erstellt am: **29.06.2024**

Persistenter Link: https://doi.org/10.5169/seals-555169

Untersuchung einiger Stichprobenverfahren für Zeitreihen

Von E. Nievergelt, Bern [1])

Zusammenfassung

Es werden 3 Stichprobenverfahren für Zeitreihen auf ihre Wirksamkeit untersucht und einander gegenübergestellt. Besonders behandelt werden die Fragen der Schichtung, der Berechnung des Stichprobenumfangs bei optimaler Aufteilung auf die einzelnen Schichten und der Kostenminimalisierung.

1. Kapitel

Differenzenmethode ohne Schichtung der Grundgesamtheit

1. Einleitung

Eine Zeitreihe besteht aus einer Menge von Beobachtungen, welche chronologisch angeordnet sind. Beispiele: Index der Konsumentenpreise, Tagesumsatz eines Verkaufsunternehmens, tägliche Niederschlagsmenge, monatlich abgeschlossene Zahl von Versicherungspolicen, monatliche Anzahl verkaufter Fahrkarten usw.

Oft ist die periodische Ermittlung einer solchen Zahl mit grossen Umtrieben verbunden, namentlich wenn es sich um ein Total oder einen Mittelwert vieler einzelner Grössen handelt. In solchen Fällen ist es viel rationeller, die Information mit Hilfe eines permanenten Stichprobenverfahrens zu beschaffen.

Die einfachste Methode, in jedem Zeitabschnitt eine Anzahl Elemente zufallsmässig auszuwählen und diese zu erheben, ist organisatorisch meistens mit grossem Aufwand verbunden. Es gibt viel wirksamere Verfahren, welche weniger Arbeit verursachen. Von diesen werden einige ausgewählt und näher untersucht. Es handelt sich dabei

[1]) Dr. Erwin Nievergelt, Generaldirektion SBB, Bern.

um Verfahren zur periodischen Schätzung des Mittelwertes (oder des Totals) der Charakteristik einer Anzahl Elemente, welche über eine Anzahl Zeitabschnitte unverändert bleiben.

2. *Modell und Bezeichnungen*

Wir betrachten eine Reihe von Zeitabschnitten. Dies können z. B. Tage, Monate oder Jahre sein. Es wird jedoch nirgends vorausgesetzt, dass die Zeitabschnitte gleich lang sind.

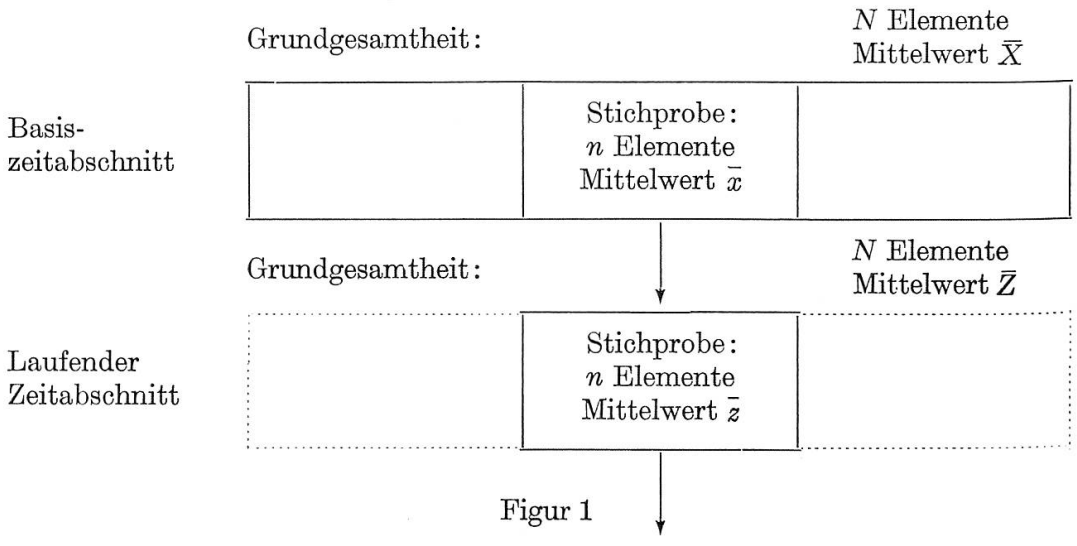
Die Grundgesamtheit, die aus N Elementen besteht, sei über alle Zeitabschnitte konstant. Jedem Element ist ein Merkmalswert zugeordnet. In einem bestimmten, dem *Basiszeitabschnitt*, heisst der Merkmalswert X, im *laufenden Zeitabschnitt* nennen wir ihn Z.

Dem i-ten Element sei der Merkmalswert X_i bzw. Z_i zugeordnet. Für den Basiszeitabschnitt seien die Merkmalswerte *aller Elemente* der Grundgesamtheit bekannt.

Aus der Grundgesamtheit wird zufallsmässig eine Stichprobe von n Elementen ausgewählt. Im laufenden Zeitabschnitt werden die Merkmalswerte nur für die Elemente der Stichprobe erhoben. Die Stichprobe ist über alle Zeitabschnitte *unveränderlich.*

Gesucht ist eine möglichst gute Schätzung für den unbekannten wahren Mittelwert $\overline{Z}$ des laufenden Zeitabschnittes, wobei die Kenntnis des Stichprobenmittelwertes $\bar{z}$ für den laufenden Zeitabschnitt, des Stichprobenmittelwertes $\bar{x}$ und des wahren Mittelwertes $\overline{X}$ für den Basiszeitabschnitt verwendet wird.

Schematisch lässt sich diese Methode folgendermassen darstellen:



Figur 1

Dieses Modell stellt einen Spezialfall des allgemeineren Modells dar, bei dem für den Basiszeitabschnitt die Merkmalswerte einer erweiterten Stichprobe von $n + n'$ Elementen bekannt sind. In den laufenden Zeitabschnitten werden dann die n' Elemente entweder gegen n' neue ausgetauscht oder weggelassen. Wir verweisen auf die Literatur.

Im Abschnitt 5 wird der Fall untersucht, bei dem die Merkmalswerte der Grundgesamtheit für *zwei* Basiszeitabschnitte bekannt sind.

3. Linearer Ansatz

Für den unbekannten Mittelwert $\overline{Z}$ des laufenden Zeitabschnittes sucht man eine Schätzung, welche keinen systematischen Fehler aufweist und deren Varianz minimal ist. Eine solche Schätzung nennt man *Minimalschätzung ohne Bias.* Als Ansatz verwendet man das folgende lineare Polynom, in dem alle bekannten Mittelwerte vorkommen und dessen Koeffizienten nach den obigen Bedingungen zu bestimmen sind.

$$\bar{z}_m = a_1\overline{X} + a_2\bar{x} + c\bar{z}. \tag{3.1}$$

Für eine biasfreie Schätzung gilt

$$E\bar{z}_m = \overline{Z} = E(a_1\overline{X} + a_2\bar{x} + c\bar{z}) = a_1\overline{X} + a_2\overline{X} + c\overline{Z}.$$

Daraus folgt

$$a_1 + a_2 = 0, \qquad c = 1. \tag{3.2}$$

Somit führt die Forderung der Erwartungstreue auf den Ansatz

$$\bar{z}_m = a(\overline{X} - \bar{x}) + \bar{z}. \tag{3.3}$$

4. Minimalschätzung

Der unbekannte Koeffizient a ist so zu bestimmen, dass die Varianz von $\bar{z}_m$ minimal wird.

$$\sigma^2_{\bar{z}m} = a^2\sigma^2_{\bar{x}} - 2a\sigma_{\bar{x}\bar{z}} + \sigma^2_{\bar{z}}. \tag{4.1}$$

Durch Nullsetzen der Ableitung

$$\frac{d\sigma^2_{\bar{z}m}}{da} = 2a\sigma^2_{\bar{x}} - 2\sigma_{\bar{x}\bar{z}} \tag{4.2}$$

erhält man

$$a = \frac{\sigma_{\bar{x}\bar{z}}}{\sigma^2_{\bar{x}}}. \tag{4.3}$$

Da die 2. Ableitung $2\sigma_{\bar{x}}^2$ immer positiv ist, handelt es sich bei diesem Extremum um ein Minimum. Wir betrachten hier den Fall der Ziehung der Stichprobe «*ohne Zurücklegen*», welcher praktisch weitaus der häufigste ist. Die Varianz der Mittelwerte $\bar{x}$ und $\bar{z}$ ist dann

$$\sigma_{\bar{x}}^2 = \frac{S_x^2(N-n)}{nN}, \qquad \sigma_{\bar{z}}^2 = \frac{S_z^2(N-n)}{nN} \tag{4.4}$$

$$\text{mit} \quad S_x^2 = \frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N-1}, \quad S_z^2 = \frac{\sum_{i=1}^{N}(Z_i-\overline{Z})^2}{N-1}$$

und die Kovarianz

$$\sigma_{\bar{x}\bar{z}} = \frac{S_{xz}(N-n)}{nN} \tag{4.5}$$

$$\text{mit} \quad S_{xz} = \frac{\sum_{i=1}^{N}(X_i-\overline{X})(Z_i-\overline{Z})}{N-1}.$$

Für a erhält man

$$a = \frac{S_{xz}}{S_x^2} \tag{4.6}$$

oder, bei Einführung des Korrelationskoeffizienten

$$\varrho = \frac{S_{xz}}{S_x S_z}, \tag{4.7}$$

$$a = \varrho \frac{S_z}{S_x}. \tag{4.8}$$

Die Minimalschätzung ohne Bias lautet also

$$\bar{z}_m = \varrho \frac{S_z}{S_x}(\overline{X}-\bar{x}) + \bar{z}. \tag{4.9}$$

Der Stichprobenmittelwert $\bar{z}$ des laufenden Zeitabschnitts wird mit Hilfe der Differenz des wahren Mittelwertes $\overline{X}$ und des Stichprobenmittelwertes $\bar{x}$ des Basiszeitabschnittes korrigiert. Wir nennen diese Methode deshalb *Differenzenmethode*. Die Korrektur ist um so stärker, je grösser die Korrelation ϱ, die Streuung S_z und je kleiner die Streuung S_x ist.

Setzt man (4.4), (4.5) und (4.8) in (4.1) ein, so bekommt man die *Minimalvarianz*

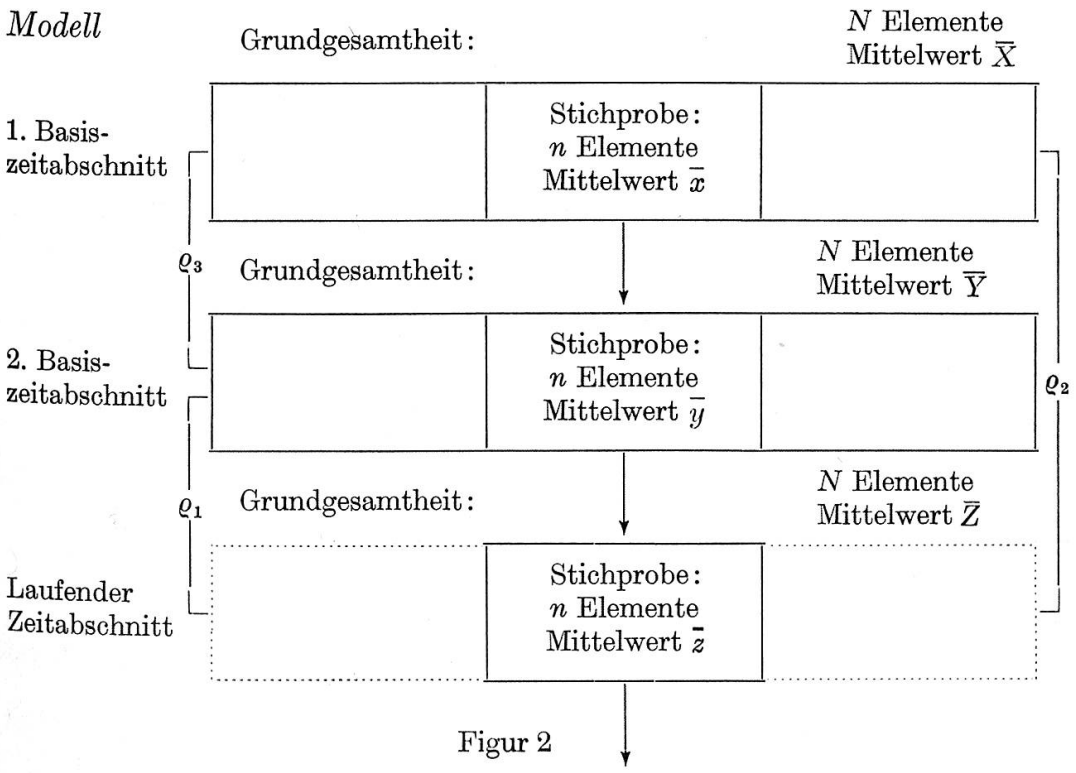
$$\sigma^2_{\bar{z}_m}\,(\min) = \frac{N-n}{nN} S_z^2 (1-\varrho^2)\,. \qquad (4.10)$$

Die Minimalvarianz hängt sehr stark ab von der Korrelation ϱ zwischen den Merkmalswerten des Basis- und des laufenden Zeitabschnitts. Ist der Absolutwert der Korrelation gleich Eins, so ist die Varianz Null, d.h. die Schätzung enthält überhaupt keinen Fehler. Ist die Korrelation gleich Null, so erhält man Formel (4.4), dies ist die Varianz des unkorrigierten Schätzwertes $\bar{z}$. Daraus folgt, dass Schätzungen nach der Formel (4.9) in der Regel besser, mindestens jedoch so gut als solche nach dem gewöhnlichen Stichprobenverfahren sind. Beträgt die Korrelation z. B. $\varrho = 0{,}95$, so ist der nach dieser Methode zu erwartende Fehler nur 30 % des gewöhnlichen Stichprobenfehlers.

5. *Zwei Basiszeitabschnitte*

Da sich bereits mit einem Basiszeitabschnitt wesentliche Verbesserungen erzielen lassen, liegt der Gedanke nahe, zwei Basiszeitabschnitte voll zu erheben und die Resultate zur Korrektur der Stichprobenschätzung des laufenden Abschnitts zu verwenden.

Modell



Figur 2

Die Herleitung der Formeln für die Minimalschätzung und die Minimalvarianz ist ganz analog wie im Fall eines Basiszeitabschnitts. Wir geben hier nur die Resultate.

Linearer Ansatz

$$\bar{z}_m = a_1\overline{X} + a_2\bar{x} + b_1\overline{Y} + b_2\bar{y} + c\bar{z}. \tag{5.1}$$

Biasfreie Schätzung

$$\bar{z}_m = a(\overline{X}-\bar{x}) + b(\overline{Y}-\bar{y}) + \bar{z}. \tag{5.2}$$

Varianz

$$\sigma^2_{\bar{z}_m} = a^2\sigma^2_{\bar{x}} + b^2\sigma^2_{\bar{y}} + \sigma^2_{\bar{z}} + 2ab\,\sigma_{\bar{x}\bar{y}} - 2a\,\sigma_{\bar{x}\bar{z}} - 2b\,\sigma_{\bar{y}\bar{z}}. \tag{5.3}$$

Minimalschätzung

$$\bar{z}_m = \frac{S_z}{S_x}\,\frac{\varrho_2-\varrho_1\varrho_3}{1-\varrho_3^2}\,(\overline{X}-\bar{x}) + \frac{S_z}{S_y}\,\frac{\varrho_1-\varrho_2\varrho_3}{1-\varrho_3^2}\,(\overline{Y}-\bar{y}) + \bar{z}. \tag{5.4}$$

Minimalvarianz

$$\sigma^2_{\bar{z}_m}(\min) = \frac{N-n}{nN}\,S_z^2\,\frac{1-\varrho_1^2-\varrho_2^2-\varrho_3^2+2\varrho_1\varrho_2\varrho_3}{1-\varrho_3^2}; \tag{5.5}$$

dabei ist ϱ_1 die Korrelation zwischen den Y_i und Z_i,
ϱ_2 die Korrelation zwischen den X_i und Z_i,
ϱ_3 die Korrelation zwischen den X_i und Y_i.

In den Formeln (5.4) und (5.5) muss noch der Fall $\varrho_3 = \pm 1$ untersucht werden.

Ist $\varrho_3 = \pm 1$, so liegen die Punkte (x_i, y_i) in einem X, Y-Koordinatensystem auf der Geraden

$$y = \pm cx + d.$$

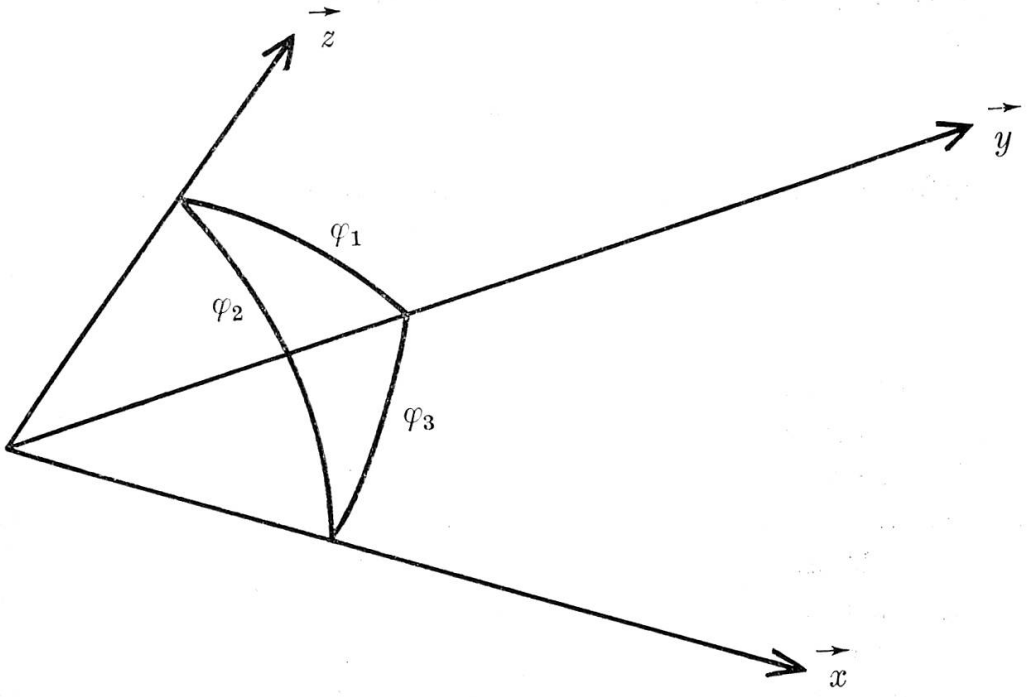
Daraus folgt

$$\varrho_1 = \frac{\sigma_{yz}}{\sigma_y\sigma_z} = \frac{\sigma_{\pm cx+d,\,z}}{\sigma_{\pm cx+d}\,\sigma_z} = \frac{\pm c\,\sigma_{xz}}{c\,\sigma_x\sigma_z} = \pm\varrho_2.$$

Durch Vertauschung der Variabeln erhält man den Satz:

$$\begin{aligned} \varrho_2 &= \pm\varrho_3 \quad \text{wenn} \quad \varrho_1 = \pm 1, \\ \varrho_1 &= \pm\varrho_3 \qquad\quad\;\; \varrho_2 = \pm 1, \\ \varrho_1 &= \pm\varrho_2 \qquad\quad\;\; \varrho_3 = \pm 1. \end{aligned} \tag{5.6}$$

Dieses Ergebnis kann auch geometrisch interpretiert werden. Im n-dimensionalen Euklidischen Raum lässt sich der Korrelationskoeffizient als cos des Zwischenwinkels zweier Vektoren deuten.



Figur 3

$$\varrho_1 = \cos\varphi_1, \quad \varrho_2 = \cos\varphi_2, \quad \varrho_3 = \cos\varphi_3; \tag{5.7}$$
$$\vec{x} = \{x_1 - \bar{x},\ x_2 - \bar{x},\ \ldots,\ x_n - \bar{x}\} \quad \text{usw.}$$

Strebt nun ϱ_3 gegen 1, so geht φ_3 gegen Null, und die beiden Vektoren $\vec{x}$ und $\vec{y}$ schliessen sich zusammen. Dann wird $\varphi_1 = \varphi_2$ und damit $\varrho_1 = \varrho_2$.
Strebt ϱ_3 gegen -1, so geht φ_3 gegen 180° und die beiden Vektoren bekommen entgegengesetzt gleiche Richtung. Dann wird $\varphi_2 = 180° - \varphi_1$ und damit $\varrho_1 = -\varrho_2$.

Führen wir in (5.4) und (5.5) den Grenzprozess $\varrho_3 \to 1$, $\varrho_1 \to \varrho_2$ durch, so erhalten wir die Minimalschätzung bei $\varrho_3 = 1$

$$z_m = \frac{S_z}{S_x}\frac{\varrho_2}{2}(\overline{X} - \bar{x}) + \frac{S_z}{S_y}\frac{\varrho_2}{2}(\overline{Y} - \bar{y}) + \bar{z} \tag{5.8}$$

und die Minimalvarianz bei $\varrho_3 = 1$

$$\sigma^2_{z_m}(\min) = \frac{N-n}{nN} S_z^2 (1 - \varrho_2^2). \tag{5.9}$$

Bei $\varrho_3 = -1$ erhält der Koeffizient von $(\overline{Y} - \bar{y})$ in (5.8) ein negatives Vorzeichen, während die Minimalvarianz gleich bleibt.

Aus Figur 3 ist ferner folgendes ersichtlich: Gibt man etwa ϱ_1 und ϱ_2 vor, so ist ϱ_3 nicht bestimmt, aber gewissen Einschränkungen unterworfen.

2. Kapitel

Differenzenmethode mit Einteilung der Grundgesamtheit in Schichten

1. Allgemeines und Bezeichnungen

Ist die Grundgesamtheit aus Elementen zusammengesetzt, die in bezug auf den Merkmalswert heterogen sind, d. h. merkliche Unterschiede aufweisen, so lässt sich der Stichprobenumfang durch eine zweckmässige Einteilung der Grundgesamtheit in Schichten erheblich reduzieren. Massgebend für die Einteilung sind die Grösse des Merkmalswertes, dessen Streuung und die Korrelation innerhalb einer Schicht.

Alle Formeln, Bezeichnungen und Aussagen des 1. Kapitels gelten auch für eine Schicht. Eine Grösse, die sich auf die h-te Schicht bezieht, erhält den Index h.

Die Anzahl der Schichten sei L. Es wird hier nur der Fall *eines* Basiszeitabschnittes untersucht. Die Herleitung der Formeln für den Fall zweier Basiszeitabschnitte ist wörtlich gleich.

2. Mittelwert

Entsprechend Formel (4.9) des ersten Kapitels ist die Minimalschätzung ohne Bias für den Mittelwert der h-ten Schicht

$$\bar{z}_{mh} = \varrho_h \frac{S_{zh}}{S_{xh}} (\overline{X}_h - \bar{x}_h) + \bar{z}_h . \tag{2.1}$$

Da $N_h \bar{z}_{mh}$ eine biasfreie Schätzung für den Totalwert Z_h der h-ten Schicht ist, erhält man mit

$$\bar{z}_m = \frac{\sum_{h=1}^{L} N_h \bar{z}_{mh}}{N} \tag{2.2}$$

eine biasfreie Schätzung für den Mittelwert $\overline{Z}$ der Grundgesamtheit.

3. *Varianz*

Die Minimalvarianz von $\bar{z}_{mh}$ ist gemäss Formel (4.10) des 1. Kapitels

$$\sigma^2_{\bar{z}_{mh}}(\min) = \frac{N_h - n_h}{n_h N_h} S^2_{zh}(1-\varrho^2_h). \tag{3.1}$$

Da $\bar{z}_{mh}$ unabhängig ist von $\bar{z}_{mk}$ für $h \neq k$, ist nach (2.2) $\bar{z}_m$ eine Linearkombination von unabhängigen Zufallsvariabeln. Die Varianz von $\bar{z}_m$ ist deshalb

$$\sigma^2_{\bar{z}_m} = \frac{\sum_{h=1}^{L} N_h^2 \sigma^2_{\bar{z}_{mh}}(\min)}{N^2}. \tag{3.2}$$

4. *Optimale Aufteilung bei festem Stichprobenumfang n*

Wir wollen nun die Frage untersuchen, wie die Stichprobe auf die einzelnen Schichten verteilt werden muss, damit die Varianz (3.2) minimal wird.

Wir suchen die Extrema von $\sigma^2_{\bar{z}_m}$ unter der Nebenbedingung

$$\sum n_h - n = 0. \tag{4.1}$$

Um die Formeln zu vereinfachen, führen wir folgende Bezeichnung ein:

$$G_h^2 = S^2_{zh}(1-\varrho^2_h). \tag{4.2}$$

Die Minimalvarianz von $\bar{z}_{mh}$ ist dann

$$\sigma^2_{\bar{z}_{mh}}(\min) = \frac{N_h - n_h}{n_h N_h} G_h^2 \tag{4.3}$$

und die Varianz von $\bar{z}_m$

$$\sigma^2_{\bar{z}_m} = \frac{1}{N^2} \sum_{h}^{L} \frac{N_h(N_h - n_h)}{n_h} G_h^2. \tag{4.4}$$

Um die Extrema zu finden, leiten wir die Funktion von Lagrange

$$\Phi = \frac{1}{N^2} \sum \frac{N_h(N_h - n_h)}{n_h} G_h^2 + \lambda(\sum n_h - n) \tag{4.5}$$

nach n_h ab und erhalten

$$\frac{\partial \Phi}{\partial n_h} = -\frac{N_h^2 G_h^2}{N^2 n_h^2} + \lambda. \tag{4.6}$$

Nullsetzen und Auflösen nach n_h ergibt

$$n_h = \frac{N_h G_h}{N\sqrt{\lambda}}. \tag{4.7}$$

Durch Summation über alle Schichten unter Berücksichtigung von (4.1) erhalten wir

$$\sqrt{\lambda} = \frac{\sum N_h G_h}{Nn} \tag{4.8}$$

und somit

$$n_h = \frac{N_h G_h}{\sum N_h G_h} n = \frac{N_h S_{zh}\sqrt{1-\varrho_h^2}}{\sum N_h S_{zh}\sqrt{1-\varrho_h^2}} n. \tag{4.9}$$

Wir ersetzen n_h in (4.4) durch (4.9) und bekommen die optimale Varianz für $\bar{z}_m$

$$\sigma^2(\text{opt}) = \frac{1}{N^2}\left(\frac{(\sum_h N_h G_h)^2}{n} - \sum_h N_h G_h^2\right). \tag{4.10}$$

Die Varianz von $\bar{z}_m$ wird somit am kleinsten, wenn die Stichprobe nach (4.9) auf die einzelnen Schichten verteilt wird. Aus (4.9) ist ersichtlich, dass der Stichprobenanteil n_h einer Schicht linear mit der Anzahl Elemente N_h einer Schicht und der Streuung S_{zh} innerhalb einer Schicht wächst, jedoch mit zunehmender Korrelation abnimmt.

5. Der Stichprobenumfang bei vorgegebener Genauigkeit

Betrachten wir in (4.10) die optimale Varianz als gegeben, so kann man den erforderlichen Stichprobenumfang n berechnen.

$$n = \frac{(\sum N_h G_h)^2}{N^2\sigma^2(\text{opt}) + \sum N_h G_h^2}. \tag{5.1}$$

Unter der Annahme, dass der standardisierte Mittelwert

$$\frac{\bar{z}_m - \bar{Z}}{\sigma(\text{opt})}$$

die normale Verteilungsfunktion $\Phi(t)$ besitzt, gilt:

$$|\bar{z}_m - \bar{Z}| \leqq t\sigma(\text{opt}) \tag{5.2}$$

mit der Wahrscheinlichkeit $2\Phi(t) - 1$.

Für $t = 1{,}96$ ist $2\Phi(t) - 1 = 0{,}95$,
für $t = 2{,}58$ ist $2\Phi(t) - 1 = 0{,}99$.

Wir bezeichnen $$e = t\sigma\,(\text{opt}) \tag{5.3}$$

als den *absoluten* und

$$\varepsilon = \frac{t\sigma\,(\text{opt})}{\bar{Z}} \tag{5.4}$$

als den *relativen Fehler* des geschätzten Mittelwertes $\bar{z}_m$.

Mit diesen Bezeichnungen erhält man den erforderlichen Stichprobenumfang n bei vorgegebenem maximalem relativen Fehler ε und der Sicherheit t

$$n = \frac{t^2 \left(\sum_h^L N_h G_h\right)^2}{\varepsilon^2 N^2 \bar{Z}^2 + t^2 \sum_h^L N_h G_h^2}. \tag{5.5}$$

Wir wiederholen die Voraussetzungen, unter denen Formel (5.5) gilt:

a) $\bar{z}_{mh}$ ist eine Minimalschätzung ohne Bias nach Formel (2.1).

b) Die Aufteilung auf die einzelnen Schichten ist optimal nach Formel (4.9).

c) $\bar{z}_m$ ist normal verteilt mit dem Mittelwert $\bar{Z}$ und der Streuung $\sigma\,(\text{opt})$.

6. *Schätzung der unbekannten Parameter und Bemerkungen zur Anwendung der Methode*

Bei der Planung für die Anwendung der Methode stösst man zuerst auf die Schwierigkeit, dass in Formel (5.5) die Parameter $G_h(S_{zh}, \varrho_h)$ und $\bar{Z}$ nicht bekannt sind. Dieselben sind deshalb durch eine Vorerhebung mit Hilfe einer Stichprobe von n' Elementen zu schätzen. Nachdem man die Grundgesamtheit in Schichten eingeteilt hat, zieht man in jeder Schicht eine genügend grosse Zahl n'_h Elemente, erhebt ihre Merkmalswerte für den laufenden Zeitabschnitt und findet folgende biasfreie Schätzungen:

für S^2_{zh} $$s^2_{zh} = \frac{\sum_i^{n'_h} (z_{hi} - \bar{z}_h)^2}{n'_h - 1}, \tag{6.1}$$

$$\text{für} \quad \varrho_h \qquad r_h = \frac{\sum_i^{n'_h} (x_{hi} - \bar{x}_h)(z_{hi} - \bar{z}_h)}{\sqrt{\sum_i^{n'_h} (x_{hi} - \bar{x}_h)^2 \sum_i^{n'_h} (z_{hi} - \bar{z}_h)^2}}, \tag{6.2}$$

$$\text{für} \quad \bar{Z} \qquad \bar{z} = \frac{\sum_h^L N_h \frac{\sum_i^{n'_h} z_{hi}}{n'_h}}{N}. \tag{6.3}$$

Nach der Wahl von ε und t berechnet man mit Formel (5.5) den minimalen Stichprobenumfang. Dieser wird dann nach Formel (4.9) optimal auf die einzelnen Schichten verteilt, womit die Planung beendet ist.

In jeder Schicht werden nun die berechneten n_h Elemente zufallsmässig ausgewählt und deren Merkmalswerte in jedem Zeitabschnitt erhoben. Die Schätzungen für die gesuchten Mittelwerte findet man mit den Formeln (2.1) und (2.2).

Hierzu sind folgende *Bemerkungen* zu machen:

a) Die Streuung S_{xh} könnte man genau berechnen, da man im Basiszeitabschnitt die Merkmalswerte der Grundgesamtheit kennt. Wie nähere Untersuchungen zeigen, wird jedoch die Schätzung genauer, wenn man in Formel (2.1) auch für S_{xh} die Stichprobenschätzung s_{xh} einsetzt.

b) Der Stichprobenumfang wurde unter der Voraussetzung berechnet, dass in Formel (2.1) die Parameter ϱ_h, S_{xh}, S_{zh} genau bekannt sind.

Da diese Werte aber aus der Stichprobe geschätzt werden müssen, ist in Wirklichkeit die Varianz nach Formel (3.1) etwas zu klein. Eine Approximationsformel für die genaue Varianz lässt sich zwar herleiten, ist aber ausserordentlich kompliziert. Da sich jedoch eine Funktion in der Nähe ihres Minimums nur wenig verändert, kann durch eine geringe Erhöhung des Stichprobenumfangs dieser Fehler ausgeglichen werden.

c) Der Stichprobenumfang und dessen optimale Aufteilung wird nur für *einen* Zeitabschnitt berechnet. Für die andern Zeitabschnitte ist die Aufteilung nicht mehr optimal. Auch diese Tatsache spricht für eine gewisse Erhöhung des berechneten Stichprobenumfangs.

Für die praktische Durchführung ist es zu empfehlen, jedesmal den maximalen Stichprobenfehler mit den Formeln (3.1) und (3.2) zu überprüfen.

d) Wie praktische Versuche gezeigt haben, hängt die Wirksamkeit des Verfahrens sehr stark von der Art der Einteilung der Grundgesamtheit in Schichten ab. Eine Methode, welche die beste Art der Einteilung ergibt, ist bis jetzt noch nicht gefunden worden. Immerhin lassen sich folgende Richtlinien angeben: Damit die Schätzungen für die Parameter S_{zh}, S_{xh} und vor allem ϱ_h genügend genau sind, darf die Anzahl n_h der Elemente der Stichprobe in jeder Schicht nicht zu klein sein. Man darf deshalb die Anzahl der Schichten nicht zu gross wählen. Die Schichteinteilung sollte so erfolgen, dass die Korrelation ϱ_h innerhalb einer Schicht möglichst hoch ist, denn davon hängt der Stichprobenfehler am stärksten ab.

7. Minimalisierung der Gesamtkosten bei vorgegebener Genauigkeit

Bei der Beschaffung der Information über ein Element der Stichprobe entstehen zwei Arten von Kosten: *a)* einmalige Kosten; *b)* laufende Kosten.

Die *einmaligen Kosten* sind die zusätzlichen Kosten, die durch die Neuerfassung eines Elementes entstehen. Sie sind z. B. bedingt durch die Instruktion des Personals, das die betreffenden Daten zusammenzustellen und zu liefern hat.

Die *laufenden Kosten* fallen jedesmal an, wenn das betreffende Element erhoben wird. Es sind z. B. die Arbeitskosten, die durch das Zusammenstellen und Liefern der verlangten Daten verursacht werden.

Wir setzen nun voraus, dass sowohl die einmaligen wie die laufenden Kosten für alle Einheiten *derselben Schicht* gleich sein sollen. Ist dies nicht der Fall, so kann es durch eine Verfeinerung der Schichteinteilung erreicht werden. Es sind

C_{1h} einmalige Kosten pro Element in der h-ten Schicht,
C_{2h} laufende Kosten pro Element in der h-ten Schicht,
ω Anzahl der laufenden Zeitabschnitte.

Die einmaligen Kosten der h-ten Schicht sind

$$N_h C_{1h}.$$

Die laufenden Kosten derselben Schicht sind

$$N_h C_{2h} + \omega n_h C_{2h}.$$

Somit ergeben sich folgende *Gesamtkosten*

$$C = \sum_{h=1}^{L} N_h C_{1h} + \sum_{h=1}^{L} N_h C_{2h} + \omega \sum_{h=1}^{L} n_h C_{2h}. \tag{7.1}$$

Wir berechnen nun die *optimale Aufteilung* der Stichprobe auf die einzelnen Schichten im Sinne einer *Minimalisierung der Gesamtkosten bei vorgegebener Varianz* der Schätzung $\bar{z}_m$ (4.4)

$$\sigma^2_{\bar{z}m} = \frac{1}{N^2} \sum_{h}^{L} \frac{N_h (N_h - n_h)}{n_h} G_h^2 = d^2. \tag{7.2}$$

Die Funktion von Lagrange lautet (7.3)

$$\Phi = \sum N_h C_{1h} + \sum N_h C_{2h} + \omega \sum n_h C_{2h} + \lambda \left[\frac{1}{N^2} \sum \frac{N_h (N_h - n_h)}{n_h} G_h^2 - d^2 \right]$$

Die Ableitung von Φ nach n_h

$$\frac{\partial \Phi}{\partial n_h} = \omega C_{2h} - \lambda \frac{N_h^2 G_h^2}{N^2 n_h^2} \tag{7.4}$$

wird nullgesetzt und nach n_h aufgelöst

$$n_h = \sqrt{\lambda} \frac{N_h G_h}{N \sqrt{\omega C_{2h}}}. \tag{7.5}$$

$\sqrt{\lambda}$ erhält man, wenn man (7.5) in die Nebenbedingung (7.2) einsetzt

$$\sqrt{\lambda} = \frac{N \sqrt{\omega} \sum (N_h G_h \sqrt{C_{2h}})}{N^2 d^2 + \sum N_h G_h^2}. \tag{7.6}$$

Substitution von (7.6) in (7.5) ergibt

$$n_h = \frac{N_h G_h \sum (N_h G_h \sqrt{C_{2h}})}{\sqrt{C_{2h}} (N^2 d^2 + \sum N_h G_h^2)}. \tag{7.7}$$

Den Gesamtstichprobenumfang erhält man durch Summation von (7.7) über alle Schichten.

$$n = \frac{\sum (N_h G_h \sqrt{C_{2h}}) \sum (N_h G_h / \sqrt{C_{2h}})}{N^2 d^2 + \sum N_h G_h^2}. \tag{7.8}$$

Die Division von (7.7) durch (7.8) liefert schliesslich

$$n_h = \frac{N_h G_h / \sqrt{C_{2h}}}{\sum (N_h G_h / \sqrt{C_{2h}})} n \tag{7.9}$$

Es ist interessant festzustellen, dass bei Annahme von schichtunabhängigen Kosten

$$C_{2h} = C_2$$

Formel (7.9) in (4.9) und Formel (7.8) in (5.1) übergeht. Minimalisierung der Gesamtkosten bedeutet in diesem Fall Minimalisierung des Stichprobenumfangs.

Aus Formel (7.9) halten wir fest, dass der Stichprobenanteil einer Schicht umgekehrt proportional zur Quadratwurzel aus den Erhebungskosten pro Element dieser Schicht ist.

Die *Minimalkosten* erhält man durch Einsetzen von (7.7) in (7.1)

$$C\,(\min) = \sum^{L} N_h C_{1h} + \sum^{L} N_h C_{2h} + \omega \frac{\left[\sum^{L} (N_h G_h \sqrt{C_{2h}})\right]^2}{N^2 d^2 + \sum^{L} N_h G_h^2}. \tag{7.10}$$

Es ist noch zu bemerken, dass die Varianz der Schätzung $\bar{z}_m$ nur für denjenigen Zeitabschnitt vorgegeben werden kann, auf Grund dessen die Planung durchgeführt wird. Um die Genauigkeit überwachen zu können, ist es zu empfehlen, die Varianz für jeden Zeitabschnitt jeweils nach Formel (7.2) zu berechnen.

3. Kapitel

Quotientenmethode und Vergleich der verschiedenen Methoden

1. Die Quotientenmethode

Da zu vermuten ist, dass sich der Stichprobenmittelwert zum Mittelwert der Grundgesamtheit im Basiszeitabschnitt ungefähr gleich verhält wie im laufenden Zeitabschnitt, kann man den Mittelwert $\bar{Z}$ schätzen durch

$$\bar{z}_q = \frac{\bar{z}}{\bar{x}} \bar{X} \tag{1.1}$$

Es seien

$$r = \frac{z}{\bar{x}}, \qquad R = \frac{\bar{Z}}{\bar{X}}. \tag{1.2}$$

Man benützt r oft als Schätzung für R, obwohl die Schätzung einen, wenn auch kleinen, systematischen Fehler hat. Deshalb ist auch die Schätzung (1.1), im Gegensatz zur Differenzenmethode, nicht biasfrei. Das Verhältnis des Bias $B = Er - R$ zur Streuung σ_r nimmt jedoch mit $\sqrt{n}$ ab, so dass der Bias schon bei mässig grossen Stichproben vernachlässigt werden kann.

Für die Varianz von $\bar{z}_q$ leiten wir nun eine Näherungsformel her.

Es sei $f(u,v)$ eine reelle Funktion mit stetigen ersten partiellen Ableitungen im Gebiet D, welches den Punkt $P\ (U,V)$ enthält. Dann gilt nach Taylor

$$f(u,v) = f(U,V) + f'_u(u-U) + f'_v(v-V) + R_2, \tag{1.3}$$

wobei die partiellen Ableitungen im Punkt P berechnet werden. Setzen wir

$$\bar{x} = u, \qquad \bar{z} = v, \qquad \overline{X} = U, \qquad \overline{Z} = V, \tag{1.4}$$

$$f(u,v) = \overline{X}\frac{\bar{z}}{\bar{x}} = f(\bar{x},\bar{z}) = \bar{z}_q,$$

mit den partiellen Ableitungen

$$f'_u = -\overline{X}\frac{\bar{z}}{\bar{x}^2}, \qquad f'_v = \frac{\overline{X}}{\bar{x}}, \tag{1.5}$$

so wird aus (1.3)

$$\bar{z}_q = \overline{Z} - R(\bar{x}-\overline{X}) + (\bar{z}-\overline{Z}) + R_2. \tag{1.6}$$

Bei Vernachlässigung der Restglieder R_2 2. und höherer Ordnung bekommt man die approximative Varianz

$$\sigma^2_{\bar{z}_q} = R^2\sigma^2_{\bar{x}} - 2R\sigma_{\bar{x}\bar{z}} + \sigma^2_{\bar{z}} \tag{1.7}$$

oder, unter Benützung der Formeln (4.4), (4.5) und (4.7) des 1. Kapitels,

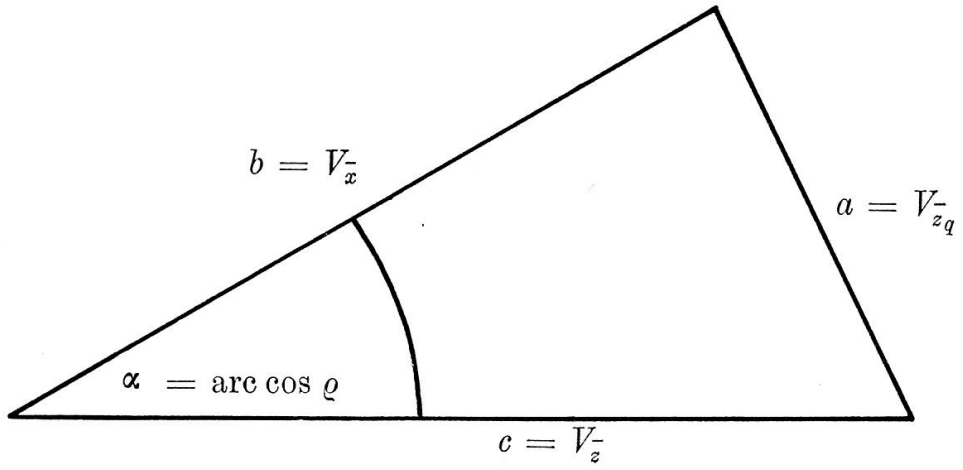
$$\sigma^2_{\bar{z}_q} = \frac{N-n}{nN}(R^2S_x^2 - 2R\varrho S_xS_z + S_z^2). \tag{1.8}$$

Eine Näherungsformel für den Bias von $\bar{z}_q$ kann mit der gleichen Methode hergeleitet werden, nur müssen bei der Taylorentwicklung (1.3) auch noch die quadratischen Glieder angeschrieben werden. Um den Einfluss der Korrelation bei der Quotientenmethode etwas besser diskutieren zu können, dividieren wir (1.7) durch $\overline{Z}^2$ und erhalten die *relative Varianz* von $\bar{z}_q$

$$V^2_{\bar{z}_q} = V^2_{\bar{x}} - 2\varrho V_{\bar{x}}V_{\bar{z}} + V^2_{\bar{z}}, \tag{1.9}$$

ausgedrückt durch die relativen Varianzen und Variationskoeffizienten von $\bar{x}$ und $\bar{z}$. Geometrisch lässt sich diese Formel durch den cosinus-Satz darstellen.

$$a^2 = b^2 - 2bc\cos\varphi + c^2. \tag{1.10}$$



Figur 4

$$a = V_{\bar{z}_q}, \qquad b = V_{\bar{x}}, \qquad c = V_{\bar{z}}, \qquad \cos\alpha = \varrho. \tag{1.11}$$

Schliesst sich der Winkel α, d. h. strebt ϱ gegen 1, so geht der Variationskoeffizient von $\bar{z}_q$ gegen $|V_{\bar{z}} - V_{\bar{x}}|$, öffnet sich α bis 180°, d. h. strebt ϱ gegen -1, so erreicht $V_{\bar{z}_q}$ den Maximalwert $V_{\bar{x}} + V_{\bar{z}}$.

Der Fehler der Schätzung $\bar{z}_q$ wird also am kleinsten, wenn ϱ nahe bei 1 ist und $V_{\bar{x}}$ möglichst gleich gross wie $V_{\bar{z}}$.

2. Vergleich der Fälle eines und zweier Basiszeitabschnitte bei der Differenzenmethode

Ein wertvolles Kriterium für den Vergleich der Wirksamkeit der beiden Methoden ist der Vergleich der Varianzen ihrer Minimalschätzungen. Werden beide Methoden auf dasselbe Problem angewandt, so ist

$$\varrho = \varrho_2. \tag{2.1}$$

Dividiert man die Minimalvarianzen für den Fall zweier und eines Basiszeitabschnittes durcheinander, also (5.5) durch (4.10) (1. Kapitel), so erhält man

$$Q(\varrho_1, \varrho_2, \varrho_3) = \frac{1 - \varrho_1^2 - \varrho_2^2 - \varrho_3^2 + 2\varrho_1\varrho_2\varrho_3}{(1-\varrho_2^2)(1-\varrho_3^2)}. \tag{2.2}$$

Diese Funktion wird durch eine dreidimensionale Hyperfläche im vierdimensionalen Raum dargestellt. Die vier Hyperebenen $\varrho_2 = \pm 1$,

$\varrho_3 = \pm 1$ und damit die Singularität $\varrho_1 = \varrho_2 = \varrho_3 = 1$ werden aus der Untersuchung ausgeschlossen. Die Berechnung der Extremalwerte zeigt, dass die Funktion Q ihr Maximum auf der Hyperfläche

$$\varrho_1 = \varrho_2 \varrho_3 \tag{2.3}$$

annimmt. Das Maximum ist

$$Q(\varrho_1 = \varrho_2 \varrho_3) = 1. \tag{2.4}$$

Die Methode mit einem Basiszeitabschnitt ist also nur unter der Bedingung (2.3) gleich genau wie diejenige mit zwei solchen, in allen andern Fällen liefert sie durchschnittlich weniger gute Schätzungen. Wie folgendes Beispiel zeigt, stellt sich jedoch die Frage, ob die Mehrarbeit, welche die Vollerhebung eines 2. Zeitabschnitts mit sich bringt, den oft nur kleinen Gewinn an Genauigkeit rechtfertigt.

Für $\varrho_1 = \varrho_2 = \varrho_3 = 0{,}8$ wird $Q = 0{,}8025$

und somit das Verhältnis der Streuungen

$$\sqrt{Q} = 0{,}9.$$

Der Fehler würde sich in dem Fall durch Hinzunahme eines 2. Basiszeitabschnitts um durchschnittlich 10% reduzieren.

Auf Spezialfälle $|\varrho_2| = 1$ oder $|\varrho_3| = 1$ gehen wir hier nicht mehr ein, sie lassen sich leicht mit der Regel von L'Hospital-Bernoulli behandeln. Für mathematisch Interessierte sei vermerkt, dass Q beim Grenzprozess $\varrho_1 \to 1$, $\varrho_2 \to 1$, $\varrho_3 \to 1$ alle Werte zwischen 0 und 1 annehmen kann, je nach dem Weg, auf dem die Singularität erreicht wird.

3. Vergleich der Differenzen- mit der Quotientenmethode

Da in der Praxis R oft nahe bei 1 und S_z nahe bei S_x liegt, beschränken wir uns bei diesem Vergleich auf den *Spezialfall* der Quotientenmethode

$$R = 1, \qquad S_x = S_z. \tag{3.1}$$

Nach Formel (1.8) wird die Varianz von $\bar{z}_q$ dann

$$\sigma^2_{\bar{z}_q} = \frac{N-n}{nN} S_z^2 \, 2(1-\varrho). \tag{3.2}$$

Diese Varianz vergleichen wir nun mit den Varianzen der Schätzungen nach der Differenzenmethode und nach dem gewöhnlichen Stichprobenverfahren.

	Schätzung	Varianz
I Gewöhnliche Stichprobe	$\bar{z} = \frac{1}{n}\sum^{n} z_i$	$\sigma^2_{\bar{z}} = \frac{N-n}{Nn} S_z^2$
II Differenzenmethode (ein Basiszeitabschnitt)	$\bar{z}_m = \varrho \frac{S_z}{S_x}(\overline{X}-\bar{x}) + \bar{z}$	$\sigma^2_{\bar{z}_m} = \frac{N-n}{Nn} S_z^2 (1-\varrho^2)$
III Quotientenmethode (Spezialfall)	$\bar{z}_q = \frac{\bar{z}}{\bar{x}} \overline{X}$	$\sigma^2_{\bar{z}_q} = \frac{N-n}{Nn} S_z^2 \, 2(1-\varrho)$

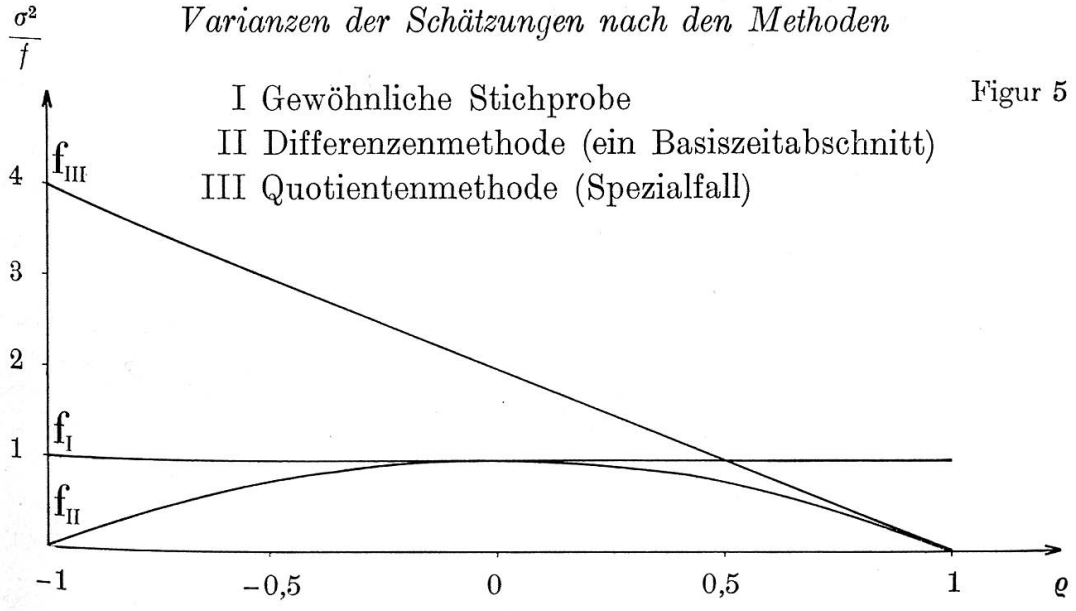
Da die 3 Varianzen alle den gemeinsamen Faktor

$$f = \frac{N-n}{Nn} S_z^2$$

haben, genügt es, die drei Funktionen

$$f_{\mathrm{I}} = 1, \qquad f_{\mathrm{II}} = 1-\varrho^2, \qquad f_{\mathrm{III}} = 2(1-\varrho)$$

miteinander zu vergleichen, was in Fig. 5 graphisch geschieht.



Figur 5

Die Differenzenmethode liefert immer durchschnittlich bessere Resultate als die beiden andern Methoden. Mit der Quotientenmethode muss man sehr vorsichtig sein, ihre Anwendung ist nur zu empfehlen bei hoher positiver Korrelation.

Literatur

Hansen, Hurwitz, Madow: Sample Survey Methods and Theory, Ch. 11, Sec. 8, New York-London 1956.

Résumé

Trois procédés d'échantillonnage de séries chronologiques sont confrontés et examinés du point de vue de leur efficacité. En particulier, l'auteur aborde le problème de la stratification et de la fixation de l'ampleur de l'échantillon dans l'hypothèse d'une disposition optimum de la stratification et d'un coût minimum.

Summary

Three sampling methods for time series are tested on their efficiency and mutually compared. The questions of stratification, calculation of the extent of samples at an optimal stratification and the minimizing of costs are treated in particular.

Riassunto

In rapporto alla loro efficacia vengono esaminati e confrontati fra loro 3 procedimenti di saggi per periodi cronologici. Particolarmente trattate sono le questioni della stratificazione, del calcolo dell'ampiezza del saggio in caso di distribuzione ottima nei diversi strati e della minimalizzazione dei costi.