

Der folgende Text war ursprünglich gedacht als ein Kapitel für ein Lehrbuch der induktiven Statistik. Er nimmt Bezug auf Kap. 10 und andere Kapitel in meinem Buch "Induktive Statistik, Formeln, Aufgaben, Klausurtraining", Oldenbourg Verlag, 6. Aufl. 1998. Den Text habe ich im Wesentlichen im Mai 2006 verfaßt und ihn dann aber aktuell überarbeitet.

Febr. 2010 Peter von der Lippe

# Kapitel 10: Stichprobentheorie

Peter von der Lippe

Gegenstand dieses Kapitels, das sich in vier Abschnitte gliedert sind (1) einige praktische Probleme der Durchführung von Stichprobenerhebungen sowie einfache Stichprobenpläne, (2) die geschichtete Stichprobe und (3) die Klumpenstichprobe, sowie (4) mehrstufige Verfahren. Der Begriff des Stichprobenplans ist in Def. 10.1 definiert worden.<sup>1</sup> Im Rahmen der statistischen Qualitätskontrolle wird im gleichen Sinne von Prüfanweisung, -plan oder -vorschrift gesprochen. Der Entwurf und die Beurteilung von Stichprobenplänen ist eine der Aufgaben der Stichprobentheorie, die sich im übrigen ausschließlich mit solchen Teilerhebungen beschäftigt, die ganz oder zum größten Teil auf einer Zufallsauswahl beruhen.<sup>2</sup> Eine nichtzufällige Auswahl wie z.B. die Quotenauswahl oder eine cut-off Auswahl (z.B. alle Betriebe ab 20 Beschäftigte werden betrachtet) ist nicht Gegenstand der statistischen Stichprobentheorie.

## 1. Einführung, Stichprobenumfang, Hochrechnung

### a) Probleme der Durchführung von Stichprobenerhebungen

Die praktischen Probleme im Zusammenhang mit Stichproben sind Gegenstand der Stichprobentheorie und beschäftigen auch die "Empirische Sozialforschung". Bei der Durchführung sind u.a. folgende Punkte zu berücksichtigen:

- die Konkretisierung des Untersuchungsziels,
- die Erhebungsplanung, d.h. die Entscheidung über Stichprobenplan, Auswahltechnik usw., die den Zielen und Rahmenbedingungen der Untersuchung Rechnung trägt,
- die Abgrenzung der Grundgesamtheit (GG), z.B. die Feststellung des Kreises der [potentiell] Berichtspflichtigen in der amtlichen Statistik,
- der Einsatz einer Auswahltechnik (Übers. 10.2), die die Zufälligkeit der Auswahl sicherstellt,

---

<sup>1</sup> Kenntnis meines Buches "Induktive Statistik" im Oldenbourg Verlag ist an einigen Stellen vorteilhaft. Was hier als Def. 10.1 eingeführt wird, ist dort Def. 8. 3, auch die Übersicht 10.1 findet sich dort als Übers. 8.5. Es wurde versucht, einige Erklärungen von Symbolen und Konzepten aus dem dortigen Kapitel 8 in diesen Text (und ggfls. in den Anhang zu diesem Text, vgl. Seite 35ff) zu übertragen. Das kann aber an einigen Stellen nicht ganz ausreichend sein. Insbesondere sind allgemeine Bemerkungen zur Punkt- und Intervallschätzung und zu Parametertests im Falle einer einfachen (uneingeschränkten) Stichprobe hier nicht alle vom Kapitel 8 des genannten Lehrbuchs in den hier vorliegenden Text übernommen worden, so dass ein gewisses Maß an Vorkenntnis auf diesem Gebiet schon vorausgesetzt werden muss.

<sup>2</sup> Streng genommen kann man nur in diesem Fall von einer "Stichprobe" oder einem "sample" sprechen. Der Ausdruck quota sample oder purposive sample ist mithin nicht sehr sinnvoll.

- die Planung der Erhebung, und damit u.a. auch die Abschätzung des erforderlichen Stichprobenumfangs,
- die Behandlung von Antwortausfällen (Nichtbeantwortung, *non reponse Problem*) bei der Nichtbeantwortung sind übrigens einige wichtige Unterscheidungen vorzunehmen:
  - ◆ eine Frage wird nicht beantwortet (item nonresponse) oder
  - ◆ die Einheit selbst nimmt nicht an der Befragung teil (unit nonresponse), sei es,
    - weil sie nicht mehr existiert oder weil sie
    - nicht mehr aufzufinden ist
    - die Mitarbeit verweigert
- die Nutzung von Kenntnissen über andere Merkmale (Y, Z, ...) als das Untersuchungsmerkmal (X) für die Ziehung der Stichprobe oder die Konstruktion von Schätzfunktionen (für die Parameter der Verteilung von X wie etwa  $\mu$  oder  $\sigma^2$ ) also für Stichprobenpläne und Hochrechnungen,
- die Durchführung von Probeerhebungen (bzw. Vorerhebungen) zur Überprüfung von praktischen Aspekten der Erhebungsarbeit (z.B. Optimierung des Fragebogens) oder zur Erlangung von Kenntnissen über die GG, die bei Anwendung bestimmter Stichprobenpläne erforderlich sind,
- die Behandlung von (z.B. systematischen) Fehlern neben dem (zufälligen) Auswahlfehler, also z.B. von Angabe- oder Meßfehlern, die bei Stichproben genauso wie bei Totalerhebungen auftreten können und natürlich auch bei der Stichprobenplanung zu berücksichtigen sind, wenngleich es keine spezifischen Stichprobenprobleme sind,
- die Aufbereitung des Datenmaterials und Organisation der Erhebungsarbeit (field work).

Auf einige der hier genannten Probleme wird im folgenden eingegangen.

Das Ziel der Beschäftigung mit Auswahltechniken und Stichprobenplänen (-modellen, Auswahlverfahren)<sup>3</sup> ist die Entwicklung kostengünstiger und effizienter (bei gleichem Stichprobenumfang eine größere Genauigkeit liefernde ) Verfahren, die evtl. vorhandenen Kenntnissen über die Beschaffenheit der Grundgesamtheit Rechnung tragen. "Beschaffenheit" bezieht sich auf z.B. Größe (Umfang) und Struktur der GG, räumliche Streuung der Einheiten und die Möglichkeiten, sie zu entnehmen und zu prüfen, Vorhandensein eines Auswahlrahmens (z.B. Adressenlisten etc.), Korrelation der Untersuchungsmerkmale mit anderen Merkmalen usw. Kenntnisse über die GG können ggfls. durch Probeerhebungen erworben werden.

## b) Stichprobenpläne

Wir definieren zunächst den Begriff "Stichprobenplan" und die "einfache" Stichprobe. In der dann folgenden Übersicht stellen wir einige Stichprobenpläne vor.

### Def. 10.1: Stichprobenplan, einfache Stichprobe

- a) Ein *Stichprobenplan* ist eine Festlegung über die Art der Entnahme von Elementen der Grundgesamtheit (GG).
- b) Der einfachste Stichprobenplan ist die sog. *einfache* (oder: reine) *Zufallsauswahl* (Stichprobe) bei *gleicher* Auswahlwahrscheinlichkeit jeder Einheit der GG (allgemein bedeutet

---

<sup>3</sup> Die Abgrenzung ist z.T. nicht einfach. So wird z. B. in der Literatur die systematische Auswahl als Technik der zufälligen Ziehung (wie die Verwendung von Zufallszahlen) oder aber auch als ein Stichprobenmodell (wie z.B. die geschichtete Stichprobe) behandelt.

Zufallsauswahl: bei a priori bekannter [aber evtl. verschiedener] Auswahlwahrscheinlichkeit und unabhängiger Ziehung; vgl. auch Bem. 3).

**Bemerkungen zu Def. 10.1:**

1. Die Größe der Stichprobenfehler hängt u.a. vom Stichprobenplan ab. In den Kapiteln 8 und 9 wird allein die einfache Stichprobe behandelt. Im Kap. 10 werden auch Stichprobenpläne bei Ausnutzung von Kenntnissen über die Beschaffenheit der GG behandelt (vgl. Übers. 10.1).
2. Beispiel der Abhängigkeit des Stichprobenfehlers  $\sigma_{\bar{x}}$  des arithmetischen Mittels vom Stichprobenplan. Er ist bei

a) uneingeschränkter Zufallsauswahl und

1. Ziehen ohne Zurücklegen von  $n < N$  Elementen aus einer Grundgesamtheit des Umfangs  $N$

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}}, \text{ bzw.}$$

2. Ziehen mit Zurücklegen, bzw. bei  $\frac{N-n}{N-1} \approx 1 - \frac{n}{N} \approx 1$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

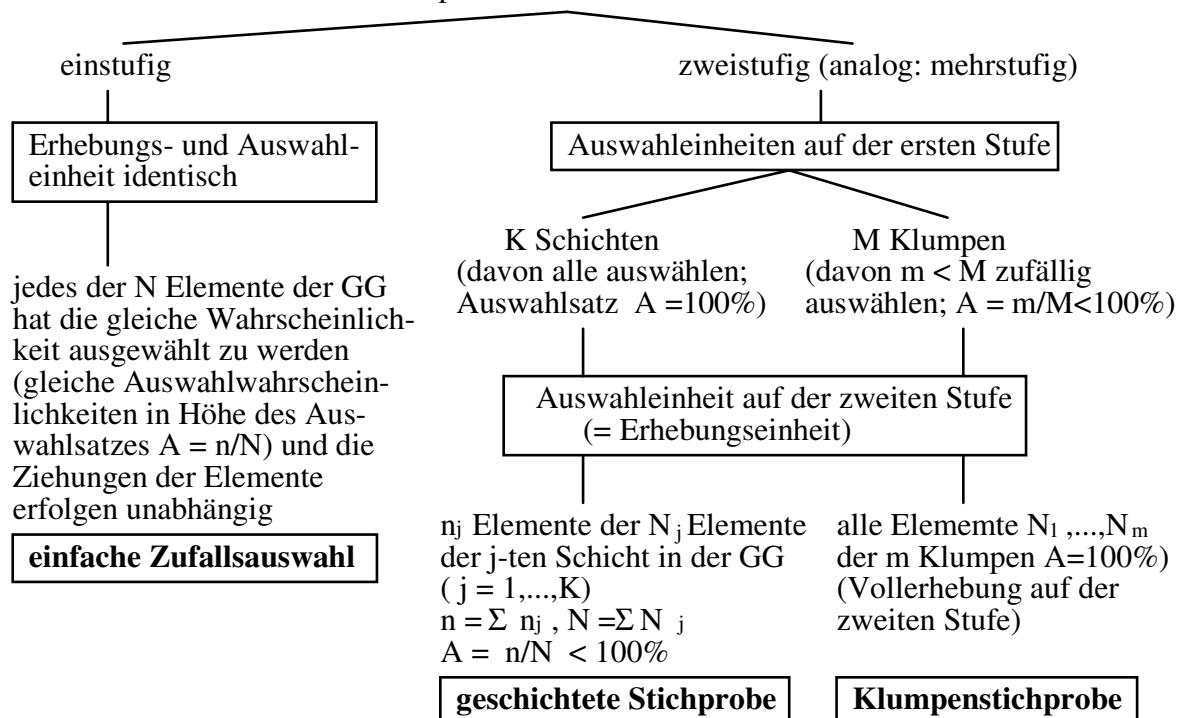
b) andere Stichprobenpläne:

1. Geschichtete Stichprobe vgl. Gl. 10.12f
2. Klumpenstichprobe vgl. Gl. 10.22 .

3. Die Ziehung eines Elements (einer Einheit) der GG stellt ein Zufallsexperiment dar. Die Ziehung der Einheit  $i$  oder der Einheit  $j$  sind die Elementarereignisse. Sind diese gleichwahrscheinlich (Laplace-Annahme), so spricht man von *uneingeschränkter Zufallsauswahl*. Sind darüber hinaus die Ziehungen (Wiederholungen des Zufallsexperiments) unabhängig, wie bei Ziehen mit Zurücklegen, so spricht man von *einfacher Zufallsauswahl*. Die Stichprobenwerte  $x_1, x_2, \dots, x_n$  sind dann Realisationen von unabhängigen identisch verteilten Zufallsvariablen  $X_1, X_2, \dots, X_n$ .

**Übersicht 10.1: Einige einfache Stichprobenpläne**

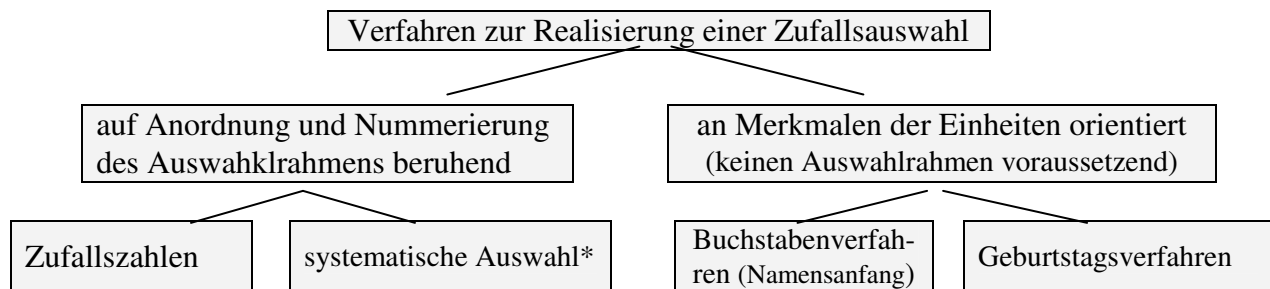
Zufallsauswahlverfahren (Stichprobe) bei endlicher Grundgesamtheit  
 Kennzeichen: a priori bekannte Auswahlwahrscheinlichkeit



### c) Techniken der Zufallsauswahl

Bisher wurden keine näheren Angaben darüber gemacht, *wie* eine Stichprobe aus einer realen Grundgesamtheit zu ziehen ist. Gedacht war dabei stets an irgendeinen Zufallsvorgang (z.B. Werfen einer Münze oder eines Würfels, Ziehen aus einer Urne). Bei einer Stichprobe aus einer konkreten, umfangreichen GG wird es jedoch meist nicht einfach sein, einen Mechanismus zu konstruieren, der eine Ziehung nach dem Zufallsprinzip sicherstellt.

#### Übersicht 10.2: Techniken der Zufallsauswahl



\*Berücksichtigung jeder  $k$ -ten Karteikarte nach der  $i$ -ten (Zufallsstart  $i$ ). Der Abstand  $k$  (als Anzahl der Karteikarten oder als Breite des Kartenstapels) wird durch den Auswahlssatz  $n/N$  definiert. Oder: Karteikarten deren fortlaufende Nummer auf  $i$  lautet (Schlußziffernverfahren).

Eine echte Zufallsauswahl zu realisieren kann sehr schwierig sein. Die Ziehung aus einer Urne (Losverfahren) ist nur bei einer kleinen GG praktikabel. Bei Verfahren, die eine Auswahlgrundlage (sampling frame), z.B. eine durchnummerierte Kartei benötigen ist darauf zu achten, dass diese nicht zyklisch (periodisch) angeordnet ist. Ein großer Vorteil von Flächenstichproben (oder allgemein Klumpenstichproben) ist es, dass keine vollständige Auflistung der Einheiten der GG (also kein "sampling frame") benötigt wird.

Orientiert man sich an Merkmalen der Untersuchungseinheiten (dabei i.d.R. Personen), z.B. am Namen oder Geburtstag, um eine möglichst zufällige Auswahl vorzunehmen, so ist es natürlich wesentlich, dass diese Merkmale nicht mit den Untersuchungsmerkmalen korrelieren.

### d) "Repräsentativität" einer Stichprobe

Man findet in der Literatur, insbesondere zum Marketing, auch Überlegungen darüber, wann eine konkrete Stichprobe mehr oder weniger „repräsentativ“ sei und es gibt sogar Vorschläge, wie dies zu messen sei. Dabei geht man meist davon aus, dass eine Stichprobe dann „repräsentativ“ sei, wenn die Verteilung bestimmter Merkmale in der Stichprobe mit der in der GG „möglichst gut“ übereinstimmt, also die Strukturen ähnlich sind. Auf dem gleichen Gedanken beruht auch das Quotenverfahren, wonach eine Stichprobe bei bestimmten Merkmalen eine ähnliche Struktur aufweisen sollte, wie die Grundgesamtheit. Gemeinsam an solchen Überlegungen ist, dass sie sich orientieren am *Ergebnis* hinsichtlich *einer konkreten* Stichprobe.

Es ist nicht überraschend, dass man solche Überlegungen in Arbeiten von Statistikern nicht findet und es ist wichtig, sich klar zu machen, dass eine solche Betrachtungsweise auch ziemlich unsinnig ist. Ein Vergleich mit der GG ist i.d.R. gar nicht möglich, weil diese ja meist unbekannt ist. Es liegt im Wesen der Zufälligkeit begründet, dass die einzelnen konkreten Stichproben sehr unterschiedlich ausfallen können. Ein und das gleiche Auswahlverfahren müßte so gesehen unterschiedlich „repräsentative“ Stichproben erzeugen, solche die zufällig „repräsentativ“ sind und solche, die es weniger oder überhaupt nicht sind. Das wäre nicht

akzeptabel, denn alle auf die gleiche Art gezogenen Stichproben sind von gleicher Qualität. Nicht das *Ergebnis* einer konkreten Stichprobe, sondern die *Voraussetzungen* unter denen alle möglicherweise zu ziehenden Stichproben gezogen werden ist relevant für die Beurteilung der Güte einer Stichprobenziehung.

Was hier mit der „Repräsentativität“ in einer dem Konzept der Zufallsauswahl gar nicht angemessenen Betrachtungsweise versucht wird zu messen ist das, was die Statistiker mit dem „Stichprobenfehler“ messen. Dieser Fehler ist ein Zufallsfehler, wenn eine Zufallsauswahl vorgenommen wird, und er betrifft nur den Auswahlfehler, nicht auch systematische Fehler, wie z.B. Angabefehler. Er bezieht sich auf die Gesamtheit aller unter den gleichen Bedingungen gezogenen Stichproben (d.h. auf die Stichprobenverteilung), nicht auf eine einzelne konkrete Stichprobe. Aussagen über den Stichprobenfehler sind Wahrscheinlichkeitsaussagen, die sich nicht auf den Einzelfall beziehen.

Niemand käme auf die Idee zu sagen, dass eine 3 und eine 4 ein „repräsentativerer“ Würfelwurf als eine 1 und 2 sei, nur weil das Ergebnis näher an  $\mu=3,5$  statt 1,5 liegt. Man kann auch leicht Beispiele konstruieren, in denen kein einziger Stichprobenmittelwert  $\bar{x}$  mit dem wahren Mittelwert  $\mu$  der GG übereinstimmt, bei denen es, so gesehen, also keine „repräsentative“ Stichprobe gibt.

Zusammenfassend kann man sagen. Nicht das Ergebnis eines einzelnen Zufallexperiments, sondern die Voraussetzungen unter denen alle seine Wiederholungen ablaufen, sind von Interesse und bestimmen dessen Qualität. „Repräsentativität“ messen zu wollen, indem man eine konkrete Stichprobe mit der Grundgesamtheit vergleicht, verkennt völlig das Wesen einer Zufallsauswahl.

### e) Notwendiger Stichprobenumfang bei einfacher Zufallsauswahl

Der für eine Stichprobe von geforderter Genauigkeit und Sicherheit mindestens erforderliche Stichprobenumfang  $n^*$  ergibt sich aus einer Umformung der Formeln für das Schwankungsintervall (direkter Schluß). Man erhält dann die Formeln der Übersicht 10.3. Die Größe  $n^*$  hängt ab von:

- der Genauigkeit und der gewünschten
- Sicherheit und der
- Homogenität der GG

hinsichtlich eines für die Untersuchung besonders repräsentativen Merkmals. Generell gilt damit (was auch sehr plausibel ist)

Ein größerer (geringer) Stichprobenumfang wird erforderlich, wenn eine größere (geringere) Genauigkeit und Sicherheit gewünscht wird und die GG wenig (mehr) homogen ist.

1. **Genauigkeit** ist definiert als **absoluter Fehler**  $e$ . Das ist die halbe Länge des (symmetrischen zweiseitigen) Schwankungsintervalls (direkter Schluß) gem. **Übers. 8.8**<sup>4</sup> also (vgl. Übersicht umseitig:

Löst man diese Gleichungen nach  $n$  auf, so erhält man die in Übers. 10.3 zusammengestellten Formeln für den notwendigen Stichprobenumfang, der  $n^*$  genannt wird.

Der **relativer Fehler** ist der absolute Fehler im Verhältnis zum zu schätzenden Parameter  $\theta$ , also:  $e^* = \frac{e}{\theta}$ . Für die Spezialfälle  $\theta = \mu$  und  $\theta = \pi$  gilt somit:  $e^* = e/\mu$  und  $e^* = e/\pi$ . Im

<sup>4</sup> Die Übers. 8.8 auf die hier Bezug genommen wird, findet sich im Anhang zu diesem Text (vgl. Seite 36)

homograden Fall kann es verwirrend sein, wenn man bedenkt, dass – weil es hier um Anteile geht – der absolute Fehler  $e$  auch schon einen Anteil (Prozentsatz) darstellt.

	Ziehen mit Zurücklegen (ZmZ) oder $N \rightarrow \infty$	Ziehen ohne Zurücklegen (ZmZ) oder: der relativ geringe Umfang $N$ ist zu berücksichtigen
heterograd (Mittelwerte)	Fall 1 $e = z_\alpha \frac{\sigma}{\sqrt{n}}$	Fall 5 $e = z_\alpha \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
homograd (Anteile)	Fall 2 $e = z_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}$	Fall 6 $e = z_\alpha \sqrt{\frac{\pi(1-\pi)}{n}} \sqrt{\frac{N-n}{N-1}}$

2. **Sicherheit** ist die Wahrscheinlichkeit  $1 - \alpha$ , der ein kritischer Wert  $z_\alpha$  zugeordnet ist. Genauigkeit und Sicherheit sind konkurrierende Forderungen: Denn mit zunehmender Sicherheit steigt der kritische Wert  $z_\alpha$  und damit auch der Fehler  $e$ . Umgekehrt wird ein Fehler  $e$  durch den Umfang der GG und der Stichprobe und zudem durch die Sicherheitswahrscheinlichkeit beeinflusst.
3. Zur Beurteilung der **Homogenität der Grundgesamtheit** wird die Varianz  $\sigma^2$  bzw.  $\pi(1 - \pi)$  der GG betrachtet. Der Stichprobenfehler  $\sigma_{\bar{x}}$  bzw.  $\sigma_p$  ist direkt proportional zu  $\sigma$  bzw.  $\sqrt{\pi(1 - \pi)}$ . Es ist unmittelbar einsichtig, dass die Streuung der GG (gemessen an der Varianz) ein Bestimmungsfaktor für den notwendigen (mindestens erforderlichen) Stichprobenumfang ist. Sind im Extremfall alle  $N$  Elemente der GG identisch (also  $\sigma=0$ ) so genügt eine Stichprobe vom Umfang  $n=1$  um die GG vollständig zu kennen.

Häufig ist die Varianz der GG nicht bekannt. Mit  $\sigma^{*2}$  bzw.  $\pi^*(1 - \pi^*)$  soll angedeutet werden, dass diese Größen geschätzt sind. Eine konservative Schätzung des notwendigen Stichprobenumfangs erhält man im homograden Fall mit  $\pi^*(1 - \pi^*) = 1/4$ , da dies der maximale Wert der Varianz einer Zweipunktverteilung ist.

### Def. 10.2: Notwendiger Stichprobenumfang

Der zur Schätzung eines Mittel-, bzw. Erwartungswerts  $\mu$  oder eines Anteilswerts bzw. einer Wahrscheinlichkeit  $\pi$  bei einer gewünschten Sicherheit  $1 - \alpha$  und Genauigkeit  $e$  (absoluter Fehler) mindestens erforderliche Stichprobenumfang  $n^*$  heißt notwendiger Stichprobenumfang. Übers. 10.3 enthält die Abschätzungen des Stichprobenumfangs  $n^*$  aufgrund der Formeln für die Intervallschätzung (Übers. 8.8) und damit unter Berücksichtigung der Grenzwertsätze.

#### Exkurse:

- Es gibt auch Formeln für den erforderlichen Stichprobenumfang um z.B.
  - eine Varianz mit vorgegebener Genauigkeit und Sicherheit abschätzen zu können oder, um
  - im Zwei-Stichproben-Fall einen hypothetischen Unterschied (etwa  $\mu_1 - \mu_2 = \Delta$ ) mit einer bestimmten Irrtumswahrscheinlichkeit in den Stichproben zu erkennen.
- Würde man demgegenüber den für die Intervallschätzung von  $\mu$  für eine Sicherheit von  $1 - \alpha$  erforderlichen Stichprobenumfang bei einem absoluten Fehler in Höhe von  $e = \varepsilon$  mit der Tschebyscheffschen Ungleichung abschätzen, also ohne Benutzung der Grenzwertsätze?

wertsätze , so erhalte man anstelle von Gl. 10.1 in Übers. 10.2 den Wert  $n^* \geq \frac{1}{\alpha} \cdot \frac{\sigma^{*2}}{e^2}$

was natürlich erheblich größer ist als  $n^*$  gem. Gl. 10.1 also  $n^* \geq z_\alpha^2 \cdot \frac{\sigma^{*2}}{e^2}$ .

**Übersicht 10.3**

**Notwendiger Stichprobenumfang bei einfacher Zufallsauswahl<sup>a)</sup>**  
 (Fallunterscheidung wie in Übers. 8.8)

	heterograd	homograd
ohne Endlichkeitskorrektur (ZmZ)	(10.1) $n^* \geq \frac{z_\alpha^2 \sigma^{*2}}{e^2} = \frac{z_\alpha^2 V^2}{e^2}$ (mit $V = \sigma^*/\mu$ Variationskoeffizient)	(10.2) <sup>b)</sup> $n^* \geq \frac{z_\alpha^2 \pi^* (1 - \pi^*)}{e^2} = \frac{z_\alpha^2 (1 - \pi^*)}{e^2 \pi^*}$
mit Endlichkeitskorrektur (ZoZ)	(10.3) <sup>c)</sup> $n^* \geq \frac{K}{e^2 + \frac{K}{N}}$ , mit $K = z_\alpha^2 \sigma^{*2}$ und mit entsprechender Formel unter Verwendung von $e^*$	(10.4) <sup>d)</sup> $n^* \geq \frac{K'}{e^2 + \frac{K'}{N}}$ mit $K' = z_\alpha^2 \pi^* (1 - \pi^*)$

a) Zur geschichteten Stichprobe vgl. auch Gl. 10.15 und 10.16

b) für den „ungünstigsten Fall“  $\pi^*(1 - \pi^*) = 0,25$  ergibt sich  $n^* \geq \frac{z_\alpha^2}{4e^2}$

c) Wenn  $N - 1 \approx N$  sonst  $n^* \geq NK / [e^2(N - 1) + K]$

d) für den "ungünstigsten Fall"  $\pi^*(1 - \pi^*) = 0,25$  :  $n^* \geq \frac{z_\alpha^2}{4e^2 + \frac{z_\alpha^2}{N}} = \frac{Nz_\alpha^2}{N4e^2 + z_\alpha^2}$

Das liegt daran, dass  $1/\alpha > z_\alpha^2$ , denn:

$1 - \alpha$	$z_\alpha$	$z_\alpha^2$	$\frac{1}{\alpha}$	$\frac{1}{\alpha} : z_\alpha^2$
0,90	1,6449	2,7057	10	3,696
0,95	1,9600	3,8416	20	5,206
0,99	3,2910	10,8307	100	9,233

Bei einer geforderten Sicherheit von 90% (99%) wäre der danach erforderliche Stichprobenumfang 3,7 - mal (9,2 - mal) so groß wie gem. Übers. 10.3.

**e) Hochrechnung**

Hochrechnung<sup>5</sup> ist das Problem, von einem Punktschätzer  $\bar{x} = \hat{\mu}$  bzw.  $p = \hat{\pi}$ , also von einem Mittelwert bzw. Anteilswert auf eine Merkmalssumme  $N\mu$  oder eine Gesamthäufigkeit (einen Bestand)  $N\pi$  zu schließen.

<sup>5</sup> In der „Alltagssprache“, z.B. bei Fernsehsendungen am Wahlabend wird „Hochrechnung“ im Sinne der "Punktschätzung" gebraucht, nicht in dem oben definierten Sinne einer „Rückvergrößerung“ eines Bildes von den kleineren Dimensionen einer Stichprobe auf die größeren Dimensionen der Grundgesamtheit.

**Def. 10.3: Merkmalssumme  $X_S$ , bzw. Bestand  $X_B$ , freie Hochrechnung**

Ist  $\mu$  der Mittelwert einer endlichen Grundgesamtheit des Umfangs  $N$ , so heißt

$$X_S = \sum_{i=1}^N x_i = N\mu \quad \text{Merkmalssumme (Gesamtmerkmalsbetrag, Totalwert).}$$

Im homograden Fall gilt entsprechend für die eine Gesamthäufigkeit (Anzahl der Erfolge in der endlichen GG) bzw. für einen **Bestand** (etwa von Personen)  $X_B = N\pi$ , wenn  $\pi$  der Anteil ist (die Merkmalswerte  $x_1, x_2, \dots, x_n$  der GG sind dann Meßwerte, die 0 oder 1 betragen). Unter Hochrechnung versteht man die Schätzung von  $X_S$  bzw.  $X_B$  aufgrund der Stichprobenschätzer  $\hat{\mu} = \bar{x}$ , bzw.  $\hat{\pi} = p$ . Schätzt man  $X$  mit

$$(10.5a) \quad \hat{X}_S = N\bar{x} \quad (\text{heterograd, Merkmalssumme}) \text{ bzw.}$$

$$(10.5b) \quad \hat{X}_B = Np \quad (\text{homograd, Bestand, absolute Anzahl})$$

aufgrund des Stichprobenmittel- bzw. -anteils werts so spricht man von **freier Hochrechnung**.

**Bemerkungen zu Def. 10.3**

1. Der Gl. 10.5a liegt der Gedanke zugrunde, dass man die Merkmalssumme  $n\bar{x}$  der *Stichprobe* durch Multiplikation mit dem reziproken Auswahlatz, also durch Multiplikation mit  $N/n$ , auf die größere Dimension der Grundgesamtheit also zum Betrag  $X_S$  „vergrößert“ (analog die Überlegung zu Gl. 10.5b), d.h.

$$X_S = N\mu \cong \hat{X}_S = \frac{N}{n} (n\bar{x}) = N\bar{x} \quad \text{im heterograden Fall, bzw.}$$

$$X_B = N\pi \cong \hat{X}_B = \frac{N}{n} (np) = Np \quad \text{im homograden Fall}$$

( $\cong$  soll heißen: wird geschätzt mit)

2. Dem entspricht die plausible Annahme, dass evtl. die Merkmalssumme  $X_S$  (bzw. der Bestand  $X_B$ ) im gleichen Maße über- oder unterschätzt wird wie der Mittelwert (bzw. der Anteilswert) d.h. die Annahme

$$\frac{X_S}{\hat{X}_S} = \frac{\mu}{\bar{x}} \quad \text{bzw.} \quad \frac{X_B}{\hat{X}_B} = \frac{\pi}{p} \quad \text{führt zu Gl. 10.5a bzw. 10.5b.}$$

3. Da somit offenbar  $\hat{X}_S$  bzw.  $\hat{X}_B$  eine Lineartransformation von  $\bar{x}$  bzw.  $p$  ist, gelten z.B. bei ZoZ nach Übersicht 8.8 (Fall 5 bzw. Fall 6)

$$E(\hat{X}_S) = N \cdot E(\bar{X}) = N\mu = X_S \quad \text{bzw.} \quad E(\hat{X}_B) = N \cdot E(p) = N \cdot \pi = X_B$$

$$V(\hat{X}_S) = N^2 \cdot V(\bar{X}) = \frac{N^2}{n} \frac{N-n}{N-1} \sigma^2 \quad \text{bzw.} \quad V(\hat{X}_B) = N^2 \cdot V(p) = N^2 \frac{(N-n)}{(N-1)} \frac{\pi(1-\pi)}{n}$$

(ansonsten  $\hat{V}(\hat{X}_B)$ ;  $\hat{V}(\hat{X}_S)$  wenn die Varianzen  $\sigma^2$  bzw.  $\pi(1-\pi)$  nicht bekannt sind).

Damit sind auch Konfidenzintervalle für Merkmalssummen ( $X_S$ ) bzw. Merkmalsbeträge ( $X_B$ ) zu berechnen. Wie man sieht, ergeben die mit  $N$  multiplizierten Grenzen eines Konfidenzintervalls für  $\mu$  bei unbekannter Varianz, nämlich  $\bar{x} \pm z \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$ , so dass die

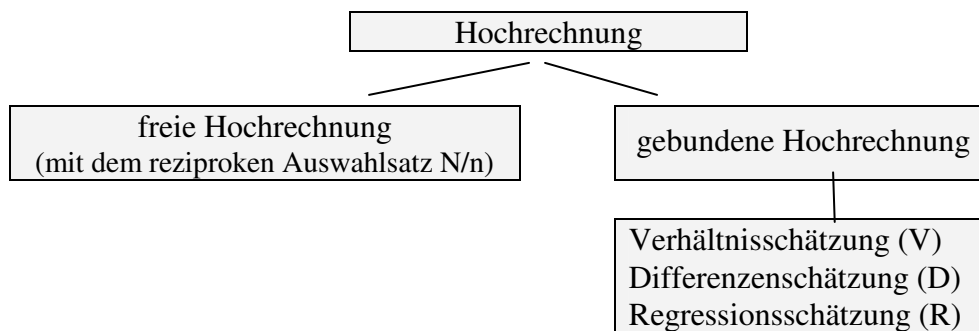
Grenzen eines Konfidenzintervalls für  $X_S$  lauten

$$\hat{X}_S \pm z_\alpha \cdot N \cdot \frac{\hat{\sigma}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}} = z \cdot \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{N(N-n)}.$$

#### Def 10.4: gebundene Hochrechnung

Eine Hochrechnung, bei der Informationen aus der Stichprobe oder aus anderen Untersuchungen über andere Variablen ( $Y, Z, \dots$ ) zur Schätzung von  $X_S$  bzw.  $X_B$  herangezogen werden, heißt gebundene Hochrechnung.

#### Übersicht 10.4: Hochrechnungsverfahren



#### Bemerkungen zu Def 10.4

##### 1. Beispiele

$X$  = Rechnungsbetrag unbezahlter Rechnungen und  $Y$  = ...bezahlter Rechnungen oder  
 $X$  = Ernteertrag und  $Y$  = Anbaufläche oder  
 $X$  = Umsatz und  $Y$  = Arbeitszeit usw.

2. Angenommen, es sei der Mittelwert  $\mu_x$  der Variable  $X$  in der GG oder die Merkmalssumme  $X_S = N\mu_x$  zu schätzen und der Mittelwert  $\mu_y$  der Variable  $Y$  in der GG sei bekannt, dann kann man  $\mu_x$  schätzen durch

(a)  $\hat{\mu}_x^{(V)} = f\mu_y$  mit  $f = \bar{x}/\bar{y}$  als Schätzer für  $F = \mu_x/\mu_y$  oder mit

(b)  $\hat{\mu}_x^{(R)} = \bar{x} + d(\mu_y - \bar{y})$  mit  $d$  = Regressionskoeffizient ( $d = s_{xy}/s_y^2$ )

wenn die Stichprobenregressionsfunktion lautet  $N \cdot \hat{\mu}_x = \hat{X}_S = c + dy$ , so dass man hiermit  $\mu_x$  schätzt als  $c + d\mu_y$

(c)  $\hat{\mu}_x^{(D)} = \bar{x} + k(\mu_y - \bar{y})$  mit  $k = \text{const.}$ ,

was die üblichen Ansätze der (a) Verhältnis- (b) Regressions- und (c) Differenzschätzung sind. Abgesehen von  $\hat{\mu}_x^{(D)}$  sind die Schätzungen nicht erwartungstreu und auf die Eigenschaften soll hier nicht weiter eingegangen werden. Sehr einfach zu durchschauen sind allerdings die Eigenschaften der Differenzschätzung von  $\mu_x$  bzw.  $X_S$  mit  $\hat{\mu}_x^{(D)}$  bzw.  $N \cdot \hat{\mu}_x^{(D)}$ , denn  $E(\hat{\mu}_x^{(D)}) = E(\bar{X}) = \mu_x$  (da auch  $E(\bar{Y}) = \mu_y$ ) und nach den Sätzen über die Linearkombination von Zufallsvariablen  $(\bar{X}, \bar{Y})$  gilt

$$V(\hat{\mu}_x^{(D)}) = \frac{N-n}{Nn} (\hat{\sigma}_x^2 + k^2 \hat{\sigma}_y^2 - 2k \hat{\sigma}_{xy}) \text{ mit } \hat{\sigma}_{xy} = \text{Stichprobenkovarianz.}$$

Diese Verfahren laufen darauf hinaus, die einfache Punktschätzung von  $\mu_x$  mit  $\bar{x}$  zu korrigieren unter Ausnutzung des Umstandes, dass  $X$  und  $Y$  miteinander korreliert sind.

3. In der englischen Literatur wird meist kein eigener Begriff für das mit den Def. 10.3 und 10.4 beschriebene Problem der Hochrechnung verwendet. Man spricht von „estimating totals“ (Hochrechnung) und „estimating means“ (Punktschätzung) und auch bei der Behandlung bestimmter Verfahren der gebundenen Hochrechnung wie z.B. „ratio estimates“ (Verhältnisschätzung) oder wie „regression estimates“ fällt auf, dass es eigentlich von untergeordneter Bedeutung ist, ob  $X$  oder  $\mu_X$  geschätzt wird. Wir folgen hier lediglich der in der deutschen Literatur (noch) üblichen Praxis einen Begriff „Hochrechnung“ zu benutzen, der eigentlich überflüssig ist.

## 2. Geschichtete Stichprobe (stratified sample)

### a) Beschreibung des Stichprobenplans und Notation

In der Praxis bereitet die Durchführung einer uneingeschränkten Zufallsauswahl i.a. Schwierigkeiten; aus diesem Grunde werden häufig alternative Stichprobenpläne angewendet. Sie können leistungsfähiger sein, als die uneingeschränkte Zufallsauswahl in dem Sinne, dass sie bei gleichem (Gesamt-)umfang  $n$  zu einer geringeren Varianz der Schätzer für den Mittel- bzw. Anteilswert führen. Dies ist möglich, weil die GG eine Struktur besitzt (z.B. die Zerlegung in homogene Schichten), der mit dem Auswahlverfahren Rechnung getragen wird (was jedoch auch gewisse Kenntnisse über die Struktur der GG voraussetzt, die bei uneingeschränkter Zufallsauswahl nicht erforderlich sind).

#### Def. 10.5: Schichten, geschichtete Stichprobe

Teilt man eine GG vom Umfang  $N$  in  $K$  disjunkte Teilmengen mit den Umfängen  $N_k$ ,  $k=1, \dots, K$  mit

$$(10.6) \quad N = N_1 + N_2 + \dots + N_K = \sum_{k=1}^K N_k$$

und zieht aus den  $K$  Teilmengen Zufallsstichproben mit den Umfängen  $n_k$ ,  $k=1, \dots, K$ , mit

$$(10.7) \quad n = n_1 + n_2 + \dots + n_K = \sum_{k=1}^K n_k,$$

so heißen die  $K$  disjunkten Teilmengen *Schichten*, und das Auswahlverfahren *geschichtete Zufallsauswahl*. Die erhaltene Stichprobe vom Umfang  $n$  ist eine *geschichtete Stichprobe*.

#### Bemerkungen zu Def. 10.5

1. Zur Bildung der Schichten wird eine sogenannte Schichtungsmerkmal bestimmt, das mit dem zu untersuchenden Merkmal in Zusammenhang steht, und nach dem die Einteilung der Schichten erfolgt. Die Schichtenerhebung ist eine Zerlegung (Partition) aufgrund eines oder mehrerer Schichtungsmerkmale.
2. Schichten sollten so gebildet werden, dass die Schichten in sich möglichst homogen und untereinander möglichst unterschiedlich sind, d.h. die Varianzen  $\sigma_k^2$  sollten klein und die Differenzen  $|\mu_k - \mu|$  sollten groß sein.

In der Stichprobentheorie ist es üblich, Größen der GG mit Großbuchstaben und solche der Stichprobe mit Kleinbuchstaben, zu bezeichnen. Im Interesse der Einheitlichkeit der Darstellung sollen hier, wie in anderen Kapiteln, Parameter der GG mit griechischen Buchstaben ge-

kennzeichnet werden und zwischen Groß- und Kleinschreibung unterschieden werden, wenn es erforderlich ist, den Unterschied zwischen einer Zufallsvariable und ihrer Realisation zu betonen.

**Übersicht 10.5: Notation zur geschichteten Stichprobe (im heterograden Fall)**

Begriff	GG	Stichprobe
Umfang	N	n
Anzahl der Schichten	K	K
Umfang der k-ten Schicht	$N_k, k = 1, \dots, K$	$n_k, k = 1, \dots, K$
Schichtmittelwert	$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} X_{ki}$	$\bar{X}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} X_{kj}$
Gesamtmittelwert	$\mu = \frac{1}{N} \sum_{k=1}^K N_k \cdot \mu_k$	$\bar{X} = \frac{1}{N_k} \sum_{k=1}^K \bar{X}_k$ (= Gl. 10.9a unten)
Schichtvarianz	$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} (X_{ki} - \mu_k)^2$	$S_k^2 = \frac{1}{n_k} \sum_{j=1}^{n_k} (X_{kj} - \bar{X}_k)^2$ (Stichprobenvarianz)
Schätzer	$\hat{\mu} = \bar{X} = \frac{1}{N} \sum N_k \cdot \bar{x}_k$ (Schätzer für Gesamtmittelwert)	$\sigma_k^2 = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} (X_{kj} - \bar{X}_k)^2$ (Schätzer der Varianz $\sigma_k^2$ )
Gesamtvarianz	$\sigma^2 = \frac{1}{N} \sum N_k \sigma_k^2 + \frac{1}{N} \sum N_k (\mu_k - \mu)^2$	Varianz von x nicht von Interesse wohl der $V(\bar{X})$ , vgl. dazu Übers. *****

**b) Mittel- und Anteilswertschätzung**

Bei Schätzfunktionen  $\hat{\theta}$  für  $\theta$ , wie etwa P für  $\pi$  (homograden Fall) oder  $\bar{X}$  für  $\mu$  (heterograden Fall) gilt die Aggregationsformel

$$(10.8) \quad \hat{\theta} = \sum_{k=1}^K \frac{N_k}{N} \hat{\theta}_k,$$

wonach  $\hat{\theta}$  eine Linearkombination der K unabhängigen Zufallsvariablen  $\hat{\theta}_k$  ist, so dass gilt

$(10.9) \quad E(\hat{\theta}) = \sum_{k=1}^K \frac{N_k}{N} E(\hat{\theta}_k) \quad \text{und} \quad (10.10) \quad V(\hat{\theta}) = \sum_{k=1}^K \left(\frac{N_k}{N}\right)^2 V(\hat{\theta}_k),$
---

denn eine geschichtete Stichprobe bedeutet K unabhängige Stichproben aus K Schichten zu ziehen. Je nach Art der Stichprobenfunktion  $\hat{\theta}$  und der Stichprobenziehung erhält man anstelle von Gl. 10.6 bis 10.7 spezielle Formeln, wie etwa im *heterograden Fall* (Mittelwert, die in der folgenden Übersicht zusammengestellten Formeln):

**Übersicht 10.6**

Parameter der Stichprobenverteilung von  $\bar{X}$  bei geschichteter Stichprobe für Intervallschätzungen und Tests bezüglich  $\mu$  bzw.  $\pi$

Gl. (10.9) $E(\hat{\theta})$	Gl. (10.10) $V(\hat{\theta})$
<p><b>ZmZ und ZoZ (10.9a)</b></p> $E(\bar{X}) = \sum_{k=1}^K \frac{N_k}{N} E(\bar{X}_k) = \mu,$ <p>da <math>\sum_{k=1}^K \frac{N_k}{N} \mu_k = \mu</math></p> <p>für den Erwartungswert spielt es keine Rolle ob mit oder ohne Zurücklegen gezogen wird</p>	<p><b>ZmZ (10.10a)</b></p> $V(\bar{X}) = \sum_{k=1}^K \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k}, \text{ bzw. } \hat{V}(\bar{X}) = \sum_{k=1}^K \frac{N_k^2}{N^2} \frac{\hat{\sigma}_k^2}{n_k}$ <p><b>ZoZ (10.10b)</b></p> $V(\bar{X}) = \sum_{k=1}^K \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k} \frac{N_k - n_k}{N_k - 1} \text{ bzw. da geschätzt wird}$ $\hat{V}(\bar{X}) = \sum_{k=1}^K \frac{N_k}{N^2} \frac{\hat{\sigma}_k^2}{n_k} (N_k - n_k) \text{ denn mit } \hat{\sigma}_k^2 \text{ anstelle von } \sigma_k^2$ <p>lauten die K finite multipliers <math>(N_k - n_k)/N_k</math></p>

Aus den Gleichungen 10.10 folgt auch, dass der Standardfehler  $\sigma_{\bar{x}} = \sqrt{V(\bar{X})}$  umso kleiner ist und damit die Schätzung umso besser ist, je homogener die K Schichten (je kleiner die K Varianzen  $\sigma_k^2$ ) sind

Im **homograden Fall** (Betrachtung des Anteilwertes  $\pi$ ) ergeben sich analoge Formeln mit  $\sigma_k^2 = \pi_k(1 - \pi_k)$  und weil der Parameter  $\pi$  durch p geschätzt so ist die zu Gl. 10.10b analoge Formel:

$$(10.10c) \quad \hat{V}(P) = \hat{\sigma}_p^2 = \sum_k \left( \frac{N_k}{N} \right)^2 \frac{p_k q_k}{n_k - 1} \frac{N_k - n_k}{N_k}$$

**c) Aufteilung (Allokation) der Stichprobe**

Der gesamte Stichprobenumfang n kann auf unterschiedliche Art auf die einzelnen K Stichproben aus den K Schichten aufgeteilt werden. Bei dieser Bestimmung der Umfänge  $n_k$  ( $k = 1, 2, \dots, K$  mit  $\sum_{k=1}^K n_k = n$ ) sind verschiedene Vorgehensweisen üblich:

**1. Proportionale Aufteilung**

$$(10.11) \quad \frac{n_k}{n} = \frac{N_k}{N}, \quad \text{für alle } k = 1, 2, \dots, K$$

Daraus folgt (10.11a): 
$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n_k}{N_k} = \frac{n}{N}$$

(gleiche Auswahlätze bei allen K Schichten).

Die Aufteilung der Gesamtstichprobe auf die K Schichten erfolgt proportional zu deren Anteil an dem Gesamtumfang N der Grundgesamtheit (Gl. 10.10). Das impliziert, dass die Auswahlätze bei allen K Schichten identisch sind (Gl. 10.10a) und dass wegen

$$\hat{\mu} = \bar{X} = \frac{1}{N} \sum N_k \cdot \bar{X}_k = \frac{1}{n} \sum n_k \cdot \bar{X}_k$$

der Schätzer  $\hat{\mu}$  für den Mittelwert  $\mu$  der GG identisch ist mit dem Stichprobenmittel.

### 2. Nichtproportionale Aufteilung

Eine Möglichkeit der nichtproportionalen Aufteilung ist die sogenannte optimale Aufteilung, die auf J. Neyman zurückgeht. Die Idee der optimalen Aufteilung ist die Minimierung der Schätzervarianz  $V(\hat{\theta})$  zum einen unter Beibehaltung eines vorgegebenen Stichprobenumfangs  $n$  und zum anderen unter Einhaltung vorgegebener Kosten  $C$ . Eine Minimierung von  $V(\hat{\theta})$  etwa von  $V(\bar{X})$  unter der Nebenbedingung  $n = \sum_{k=1}^K n_k$  liefert:

(10.11a) 
$$\frac{n_k}{n} = \frac{N_k \sigma_k}{\sum_{k=1}^K N_k \sigma_k} .$$

Man beachte, dass bei Gleichheit der Varianzen  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$  die optimale Aufteilung identisch ist mit der proportionalen Aufteilung (nach Jerzy Neyman).

#### Herleitung von Gl. 10.11a

Die Herleitung wird für den Fall ZmZ durchgeführt.

Die Lagrange Funktion  $F = V(\bar{X}) + \lambda \left( \sum_{k=1}^K n_k - n \right) = \sum_{k=1}^K \left( \frac{N_k}{N} \right)^2 \frac{\sigma_k^2}{n_k} + \lambda \left( \sum_{k=1}^K n_k - n \right)$  mit dem Lagrange-Multiplikator  $\lambda$  ist nach den Stichprobenumfängen  $n_1, n_2, \dots, n_k$  partiell abzuleiten und Null zu setzen. Man erhält die folgenden  $K$  Gleichungen:

$$\begin{aligned} \frac{\partial F}{\partial n_1} &= - \left( \frac{N_1}{N} \right)^2 \frac{\sigma_1^2}{n_1^2} + \lambda = 0 \Rightarrow n_1 \sqrt{\lambda} = \frac{N_1}{N} \sigma_1 \\ &\vdots \\ (*) \quad &\vdots \\ \frac{\partial F}{\partial n_K} &= - \left( \frac{N_K}{N} \right)^2 \frac{\sigma_K^2}{n_K^2} + \lambda = 0 \Rightarrow n_K \sqrt{\lambda} = \frac{N_K}{N} \sigma_K \end{aligned}$$

Die Summe der  $K$  - Gleichungen des Gleichungssystem (\*) liefert

$$(**) \quad \sum_{k=1}^K n_k \sqrt{\lambda} = n \sqrt{\lambda} = \frac{1}{N} \sum_{k=1}^K N_k \sigma_k$$

Dividiert man die  $k$ -te Gleichung von Gl. \* durch Gl. \*\*, so erhält man GL. 10.11a. Damit sind Stichprobenumfänge für  $k = 1, 2, \dots, K$ , eindeutig bestimmt.

Man beachte, dass  $V(\bar{X})$  kein Maximum hat (die Umfänge  $n_k$  können beliebig klein sein).<sup>6</sup>

Durch die optimale Aufteilung des Stichprobenumfangs  $n$  in die Stichprobenumfänge  $n_1, \dots, n_K$  der  $K$  Schichten gem. Gl. 10.11a wird die Varianz des Schätzers  $\hat{\theta}$  bei Einhaltung des vorgegebenen Stichprobenumfangs von  $n$  Einheiten minimiert.

<sup>6</sup> Man kann leicht zeigen, dass die dargestellte Vorgehensweise eine Minimierung der Varianz, nicht eine Maximierung liefert.

Minimierung der Varianz  $V(\bar{X})$  unter der Nebenbedingung gegebener Gesamtkosten führt zu

$$(10.11b) \quad \frac{n_k}{n} = \frac{\frac{1}{\sqrt{c_k}} N_k \sigma_k}{\sum_{k=1}^K \frac{1}{\sqrt{c_k}} N_k \sigma_k}$$

Im folgenden wird  $V(\hat{\theta})$  für den Fall ZmZ unter der Nebenbedingung, der Einhaltung vorgegebener Gesamtkosten  $C = c_0 + \sum_{k=1}^K c_k n_k$  minimiert. Es bedeuten dabei  $c_0$  fixe Kosten und  $c_k$  variable Kosten pro Einheit in der  $k$ -ten Schicht,  $k = 1, \dots, K$ . Es wird also angenommen, dass die Kosten einer Erhebung linear von den Umfängen von  $n_k$  abhängen (mit den konstanten Koeffizienten  $c_0, c_1, \dots, c_k$ ).

Analog zur vorausgegangenen Herleitung wird die Lagrange-Funktion

$$(*) \quad F = \sum_{k=1}^K \left( \frac{N_k}{N} \right)^2 \cdot \frac{\sigma_k^2}{n_k} + \lambda \left( c_0 + \sum_{k=1}^K c_k \cdot n_k - C \right)$$

nach  $n_1, \dots, n_k$  partiell abgeleitet und gleich Null gesetzt. Man erhält die folgenden  $K$  Gleichungen:

$$\frac{\partial F}{\partial n_k} = - \left( \frac{N_k}{N} \right)^2 \cdot \frac{\sigma_k^2}{n_k^2} + \lambda \cdot c_k = 0; \quad k=1, 2, \dots, K. \quad \text{Es folgt daraus: } n_k \sqrt{\lambda} = \frac{1}{\sqrt{c_k}} \cdot \frac{1}{N} \cdot N_k \cdot \sigma_k$$

und somit gilt:

$$\frac{n_k}{n} = \frac{\frac{1}{\sqrt{c_k}} \cdot N_k \cdot \sigma_k}{\sum_{k=1}^K \frac{1}{\sqrt{c_k}} \cdot N_k \cdot \sigma_k}, \quad k=1, 2, \dots, K.$$

Im Vergleich zu (10.11a) hängen die optimalen Stichprobenumfänge hier zusätzlich von den variablen Kosten ab.

Der gewünschte Stichprobenumfang  $n$  fließt ebenfalls in die Bestimmung von  $n_k$  ein, wird jedoch nicht als Nebenbedingung berücksichtigt.

Zu beachten ist, dass bei der optimalen Aufteilung die Varianzen  $\sigma_k^2$  der Schichten bekannt sein müssen. Zudem kann der Fall eintreten, dass  $n_k$  größer als  $N_k$  ist.

### d) Vergleich der optimalen mit der proportionalen Aufteilung

Sind alle  $\sigma_k$  (bzw.  $\sigma_k$  und  $c_k$ ) gleich, ist die proportionale Aufteilung zugleich die optimale.

Gl. 10.11 bzw. 10.11a eingesetzt in Gl. 10.10a ergibt:

$$(10.12) \quad V(\bar{X})_{\text{opt}} = \frac{1}{n} \left( \sum \frac{N_k}{N} \sigma_k \right)^2 = \frac{1}{n} \left[ \sum \frac{n_k}{n} \left( \frac{\sum N_k \sigma_k}{N} \right) \right]^2 = \frac{\bar{\sigma}^2}{n}$$

(wegen  $\frac{n_k}{n} \frac{\sum N_k \sigma_k}{N} = \frac{N_k \sigma_k}{N}$  nach Gl. 10.11a); was offenbar nicht größer ist als

$$(10.13) \quad V(\bar{X})_{\text{prop}} = \frac{1}{n} \sum \frac{N_k}{N} \sigma_k^2 = \frac{1}{n} \sum \frac{n_k}{n} \sigma_k^2.$$

Der Klammerausdruck in Gl. 10.12 (zweiter Teil, runde Klammer) kann als mittlere Schicht-Standardabweichung  $\bar{\sigma}$  gedeutet werden, so dass gilt:

$$(10.13a) \quad V(\bar{X})_{\text{prop}} - V(\bar{X})_{\text{opt}} = \frac{1}{n} \left[ \sum \frac{N_k}{N} \sigma_k^2 - \bar{\sigma}^2 \right] = \frac{1}{n} \sum \frac{N_k}{N} (\sigma_k - \bar{\sigma})^2 \geq 0.^7$$

Bei diesem Größenvergleich liegt der gleiche Zusammenhang vor wie bei  $E(X^2) > [E(X)]^2$ .

Man sieht an dieser Beziehung und schon an Gl. 10.11a, dass die proportionale Aufteilung optimal ist, wenn alle Varianzen innerhalb der Schichten gleich sind. Ansonsten erhält man bei proportionaler Aufteilung eine höhere Varianz des Schätzers  $\bar{X}$ .

Auf die entsprechenden Formeln im Fall ZoZ soll hier verzichtet werden.

Eine andere i.d.R. nicht proportionale Aufteilung von  $n$  in  $n_k^*$  ( $n = \sum n_k^*$ ) wäre die Aufteilung mit vorgegebener (gewünschter) Genauigkeit  $e_k$  in jeder Schicht (Gl. 10.15a).

### e) Schichtungseffekt

Während soeben zwei Varianten der geschichteten Stichprobe (optimale und proportionale Aufteilung) untereinander verglichen wurden, gilt es jetzt die geschichtete Stichprobe mit der uneingeschränkten Zufallsauswahl zu vergleichen.

Bekanntlich gilt bei einfacher Zufallsstichprobe (uneingeschränkte Zufallsauswahl und unabhängige Züge, also ZmZ)

$$\sigma_{\bar{X}}^2 = V(\bar{X})_{\text{einf}} = \frac{\sigma^2}{n},$$

wobei die Varianz  $\sigma^2$  der Grundgesamtheit wie folgt in externe und interne Varianz

$$\sigma^2 = \sum_{k=1}^K \frac{N_k}{N} (\mu_k - \mu)^2 + \sum_{k=1}^K \frac{N_k}{N} \sigma_k^2 = V_{\text{ext}} + V_{\text{int}} = nV(\bar{X})_{\text{einf}}.$$

zu zerlegen ist. Im Falle einer geschichteten Stichprobe mit proportionaler Aufteilung gilt demgegenüber wegen Gl. 10.13:

$$nV(\bar{X})_{\text{prop}} = \sum_{k=1}^K \frac{N_k}{N} \sigma_k^2 = V_{\text{int}} \leq \sigma^2 = V_{\text{ext}} + V_{\text{int}} = nV(\bar{X})_{\text{einf}}.$$

Sobald eine externe Varianz auftritt (zwischen den Schichten große Unterschiede sind) ist der Standardfehler der Schätzung bei Schichtung und proportionaler Aufteilung kleiner als bei einfacher Stichprobe (es entsteht ein "Schichtungsgewinn").

Der Schichtungseffekt ist umso größer je homogener die Schichten sind (je kleiner  $V_{\text{int}}$  ist).

Bei der Herleitung dieses Zusammenhangs, wonach wegen

$$(10.13b) \quad V(\bar{X})_{\text{prop}} = \frac{V_{\text{int}}}{n} \leq V(\bar{X})_{\text{einf}} = \frac{V_{\text{int}}}{n} + \frac{V_{\text{ext}}}{n}$$

ein Schichtungseffekt stets dann eintritt, wenn es eine externe Varianz gibt, ist jedoch zu beachten, dass dieses Ergebnis für die geschichtete Stichprobe hergeleitet wurde unter der Annahme einer **proportionalen** Aufteilung.<sup>8</sup>

<sup>7</sup> Gleichheit, wenn alle  $K$  Standardabweichungen  $\sigma_k$  gleich sind.

<sup>8</sup> Auf die optimale Aufteilung wird in Abschn. f eingegangen.

Es ist deshalb zu beachten, dass im allgemeinen Fall einer geschichteten Stichprobe (also die Aufteilung nicht proportional ist)

- ein Schichtungseffekt auch dann auftreten kann, wenn es keine externe Varianz gibt (wenn also  $V_{\text{ext}} = 0$  ist) und
- wenn aber eine externe Varianz auftritt die geschichtete Stichprobe in jedem Fall besser ist als die uneingeschränkte Zufallsauswahl.

Beide Möglichkeiten werden im nachfolgenden Beispiel 10.1 demonstriert.

### Beispiel 10.1

Die Grundgesamtheit bestehe aus zwei Schichten mit jeweils  $N_i = 4$  ( $i = 1, 2$ ) Einheiten mit den Merkmalswerten  $x_{ij}$  ( $j = 1, \dots, 4$ ):

Schicht 1: 0, 4, 6, 10 ( $\mu_1 = 5, \sigma_1^2 = 13$ )

Schicht 2: 4, 4, 6, 6 ( $\mu_2 = 5, \sigma_2^2 = 1$ )

Man ziehe Stichproben vom Umfang

a)  $n = 3$  und

b)  $n = 4$

und jeweils eine uneingeschränkte einfache Stichprobe und eine geschichtete Stichprobe mit verschiedener Aufteilung mit Zurücklegen und bestimme jeweils die Varianz der Stichprobenverteilung von  $\bar{X}$ .

### Lösung 10.1

#### a) Stichproben vom Umfang $n = 3$

Bemerkenswert an dem Beispiel ist das Fehlen einer externen Varianz in der GG, weil  $\mu_1 = \mu_2 = \mu = 5$ . Für die Varianz erhält man  $\sigma^2 = \sigma_{\text{int}}^2 = 7 = \frac{1}{2}(13+1)$  und danach ist

$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{7}{3} = 2,33$  im Falle der einfachen Stichprobe. Für die geschichtete Stichprobe erhält man für  $\sigma_{\bar{x}}^2$  jeweils die folgenden Werte:

$$(1) \quad n_1 = 1, n_2 = 2 \rightarrow \sigma_{\bar{x}}^2 = \left(\frac{N_1}{N}\right)^2 \frac{\sigma_1^2}{n_1} + \left(\frac{N_2}{N}\right)^2 \frac{\sigma_2^2}{n_2} = \frac{27}{3} = 3,375$$

$$(2) \quad n_1 = 2, n_2 = 1 \rightarrow \sigma_{\bar{x}}^2 = \frac{15}{8} = 1,875$$

Für die optimale Aufteilung erhält man gem. 10.11a die Anteile  $n_1/n = \sqrt{13}/(1 + \sqrt{13}) = 0,783$  (also  $n_1 = 2,35$ ) und  $n_2 = 3 \cdot (1 - 0,783) = 0,65$ . Die zweite Aufteilung kommt also der optimalen Aufteilung näher. Bemerkenswert ist, dass man einen Schichtungsgewinn erzielt obgleich keine externe Varianz vorliegt. Man beachte auch, dass bei einer ungünstigen Aufteilung (wie Nr. (1)) die Schichtung zu einem Verlust an Genauigkeit führen kann ( $3,375 > 2,33$ ).

#### b) Stichproben vom Umfang $n=4$

Anders als im Fall a) kann hier auch eine proportionale Aufteilung ( $n_1 = n_2 = 2$ ) studiert werden. Für die optimale Aufteilung erhält man  $n_1 = 3,13 \approx 3$  und  $n_2 \approx 1$  und bei proportionaler Aufteilung, wie nach Gl. 10.12 zu erwarten war, keinen Schichtungsgewinn. Für

die Varianz  $\sigma_{\bar{x}}^2$  der Stichprobenverteilung von  $\bar{x}$  erhält man bei uneingeschränkter Zufallsauswahl  $\sigma_{\bar{x}}^2 = 7/4 = 1,75$  und bei geschichteter Stichprobe folgende Werte in Abhängigkeit von der Aufteilung:

$$\begin{aligned} n_1 = 1, n_2 = 3 &\rightarrow \sigma_{\bar{x}}^2 = 3,33 \\ n_1 = 2, n_2 = 2 &\rightarrow \sigma_{\bar{x}}^2 = 1,75 \\ n_1 = 3, n_2 = 1 &\rightarrow \sigma_{\bar{x}}^2 = 1,33 < 1,75. \end{aligned}$$

Weitere Überlegungen aufgrund des Beispiels 10.1

Um die Verringerung der Varianz der Stichprobenverteilung auf 1,33 deutlich zu machen, mag es nützlich sein, auch einmal die Stichprobenverteilung im Falle des Ziehens ohne Zurücklegen zu betrachten. Man erhält bei Schichtung mit der Aufteilung  $n_1 = 3$  und  $n_2 = 1$  genau  $\binom{4}{3}\binom{4}{1} = 16$  Stichproben im Vergleich zu  $\binom{8}{4} = 70$  Stichproben bei uneingeschränkter Zufallsauswahl. Extreme Stichproben mit den Werten 0, 4, 4, 4 ( $\bar{x} = 3$ ) oder 10, 6, 6, 6 ( $\bar{x} = 14,5$ ) sind z.B. bei der geschichteten Stichprobe mit der Aufteilung  $n_1/n_2 = 3/1$  nicht möglich.

In den 70 Stichproben sind die genannten 16 Stichproben enthalten, darüberhinaus aber auch noch je eine Stichprobe mit der extremen Aufteilung  $n_1 = 4, n_2 = 0$  und  $n_1 = 0$  und  $n_2 = 4$  sowie 16 und 36 Stichproben der beiden anderen Aufteilungen  $n_1 = 1, n_2 = 3$  und  $n_1 = 2, n_2 = 2$  (also  $70 = 1 + 16 + 36 + 16 + 1$ ).

Die Varianz der Stichprobenverteilung von  $\bar{x}$  bei geschichteter Stichprobe und ZoZ ist

$$V(\bar{X})^{(G)} = \frac{N_1^2}{N^2} \frac{\sigma_1^2}{n_1} \frac{N_1 - n_1}{N_1 - 1} + \frac{N_2^2}{N^2} \frac{\sigma_2^2}{n_2} \frac{N_2 - n_2}{N_2 - 1} = \frac{1}{12} \left[ \frac{13}{n_1} (4 - n_1) + \frac{1}{n_2} (4 - n_2) \right]$$

im Unterschied zur uneingeschränkten (ungeschichteten) Zufallsauswahl

$$V(\bar{X})^{(U)} = \frac{\sigma^2}{n} \frac{N - n}{N - 1} = \frac{7}{4} \frac{8 - 4}{8 - 1} = 1.$$

Von Interesse mögen die jeweils 16 Stichproben bei den folgenden Aufteilungen sein:

Aufteilung  $n_1 = 1, n_2 = 3; n = 4$  und die Kenngrößen (statistics)

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{i1}, \bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{i2}, \bar{x} = \frac{\bar{x}_1 + \bar{x}_2}{2}, \bar{x}^* = \frac{\sum_{i=1}^{n_1} x_{i1} + \sum_{i=1}^{n_2} x_{i2}}{n}$$

Elemente aus Schicht		Anzahl der Stichproben	$\bar{x}_1$	$\bar{x}_2$	$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2}{2}$	$\bar{x}^* = \frac{\sum x_{i1} + \sum x_{i2}}{4}$
1	2					
0	4, 4, 6	2	0	4,67	2,33	3,5
0	4, 6, 6	2	0	5,33	2,67	4
4	4, 4, 6	2	4	4,67	4,33	4,5
4	4, 6, 6	2	4	5,33	4,67	5
6	4, 4, 6	2	6	4,67	5,33	5
6	4, 6, 6	2	6	5,33	5,67	5,5
10	4, 4, 6	2	10	4,67	7,33	6
10	4, 6, 6	2	10	5,33	7,67	6,5

Es gilt  $E(\bar{X}) = E(\bar{X}^*) = 5$ ,  $V(\bar{X}) = 59/18 = 3,278$  und die Varianz  $V(\bar{X}^*) = 0,875$

Man beachte, dass der Schätzer  $\bar{X}$  als gewogenes Mittel aus den einzelnen Stichprobenmittelwerten  $\bar{x}_1$  und  $\bar{x}_2$  gewonnen wird und nicht einfach als ungewogenes Mittel aus den insgesamt  $n_1 + n_2 = 4$  Stichprobenwerten bestimmt werden. Ein so gebildetes Mittel soll  $\bar{x}^*$  genannt werden.

Hinsichtlich einer Stichprobenfunktion  $\bar{X}^* = \frac{n_1}{n} \bar{X}_1 + \frac{n_2}{n} \bar{X}_2$  anstatt  $\bar{X} = \frac{N_1}{N} \bar{X}_1 + \frac{N_2}{N} \bar{X}_2$  wäre nämlich die Stichprobenverteilung bei dieser Aufteilung identisch mit derjenigen von  $\bar{X}^*$  bei der folgenden Aufteilung  $n_1 = 3$ ,  $n_2 = 1$ :

Elemente aus Schicht		Anzahl der Stichprobe	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}$	$\bar{x}^*$
1	2					
0, 4, 6	4	2	3,33	4	3,67	3,5
0, 4, 6	6	2	3,33	6	4,67	4
0, 4, 10	4	2	4,67	4	4,33	4,5
0, 4, 10	6	2	4,67	6	5,33	5
0, 6, 10	4	2	5,33	4	4,67	5
0, 6, 10	6	2	5,33	6	5,67	5,5
4, 6, 10	4	2	6,67	4	5,33	6
4, 6, 10	6	2	6,67	6	6,33	6,5

Es gilt jetzt  $V(\bar{X}) = 11/18 = 0,611$  und  $V(\bar{X}^*) = 0,875$

Im Falle der extremen Aufteilungen  $n_1 = 4$ ,  $n_2 = 0$  und  $n_1 = 0$ ,  $n_2 = 4$  sowie bei proportionaler Aufteilung sind natürlich die Stichprobenverteilungen von  $\bar{X}$  und  $\bar{X}^*$  identisch.

Schichtung ist auch eine Möglichkeit um zu verhindern, dass zufällig kein Element einer bestimmten Schicht in einer Stichprobe vertreten ist. Mit diesem Ziel wird häufig auch in der amtlichen Statistik eine geschichtete Stichprobe gezogen. Als Grenzfall der Schichtung könnte man das **cut off "sample"** (= Abschneidegrenze, Konzentrationsprinzip) betrachten, bei dem  $n_1$  gegen 0 und  $n_2$  gegen  $N_2$  strebt (also die Auswahlätze gegen 0 und 100% streben). Eine solche "Auswahl" kann dann aber keine Zufallsauswahl sein, die Wahrscheinlichkeitsrechnung ist nicht anwendbar und man kann bei 0% und 100% auch generell nicht von "Auswahl" sprechen.

## f) weitere Bemerkungen zum Schichtungseffekt

- Mit Gl. 10.13b wurde gezeigt, dass ein Schichtungsgewinn bei geschichteter Stichprobe mit proportionaler Aufteilung entstehen kann wenn  $V_{\text{ext}}$  gering ist. Bei optimaler Aufteilung vergrößert sich der Schichtungseffekt (d.h. die Differenz  $nV(\bar{X})_{\text{ext}} - nV(\bar{X})_{\text{prop}} > 0$ ) noch einmal und zwar in dem Maße, indem die  $K$  Standardabweichungen der Schichten  $\sigma_k$  um  $\bar{\sigma}$  streuen, denn  $nV(\bar{X})_{\text{prop}} - nV(\bar{X})_{\text{opt}} = \sum \frac{n_k}{n} (\sigma_k - \bar{\sigma})^2$ .
- Unter dem Schichtungseffekt versteht man, dass die Varianz des Schätzers  $\bar{X}$  bei einer geschichteten Stichprobe,  $V(\bar{X})_{\text{gesch}} = \sum_{k=1}^K \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k}$  (gem. Gl. 10.10a) kleiner ist, als bei uneingeschränkter Zufallsauswahl und ZmZ (einfache Stichprobe), die dann den Wert

$$V(\bar{X})_{\text{einf}} = \frac{1}{n} \sigma^2 = \frac{1}{n} \left[ \sum \frac{N_k}{N} (\mu_k - \mu)^2 + \sum \frac{N_k}{N} \sigma_k^2 \right] = \frac{1}{n} (V_{\text{ext}} + V_{\text{int}}) \text{ annimmt.}$$

Offenbar kann man Gl. 10.10a auch wie folgt schreiben

$$V(\bar{X})_{\text{gesch}} = \frac{1}{n} \sum \left( \frac{N_k n}{N n_k} \right) \left( \frac{N_k}{N} \sigma_k^2 \right)$$

Bei proportionaler Aufteilung ist der Ausdruck in der ersten Klammer nach dem Summenzeichen eins, d.h. für alle  $k=1, \dots, K$  gilt  $\frac{N_k n}{N n_k} = 1$ .

Tritt keine externe Varianz auf ( $\mu_k = \mu \quad \forall k$ ), so ist eine geschichtete Stichprobe nicht besser, als die einfache Stichprobe. Bei einer anderen Aufteilung, als der proportionalen kann – wie gezeigt – sehr wohl ein Schichtungseffekt auch dann auftreten, wenn es keine externe Varianz gibt. So ist z.B. bei optimaler Aufteilung

$$V(\bar{X})_{\text{opt}} = \frac{1}{n} \sum \frac{\bar{\sigma}}{\sigma_k} \left( \frac{N_k}{N} \sigma_k^2 \right) = \frac{1}{n} \bar{\sigma}^2$$

kleiner (oder gleich, wenn alle  $\sigma_k = \bar{\sigma}$ )  $V(\bar{X})_{\text{einf}}$  unter der Voraussetzung  $V_{\text{ext}} = 0$ , also .

3. Es wurde außerdem behauptet und im Beispiel 10. demonstriert, dass eine ungünstige Aufteilung der geschichteten Stichprobe auch ein Ergebnis liefern kann, das hinsichtlich der Genauigkeit ( $\sigma_{\bar{x}}$ ) schlechter ist, als das einer einfachen Zufallsauswahl. Es muß dann gelten

$$\sum \frac{N_k n}{N n_k} \left( \frac{N_k}{N} \sigma_k^2 \right) > \sum \frac{N_k}{N} (\mu_k - \mu)^2 + \sum \frac{N_k}{N} \sigma_k^2 \text{ und somit}$$

$$(10.14) \quad \sum \frac{N_k}{N} \sigma_k^2 \left( \frac{N_{k \cdot n} - 1}{N_{n \cdot k}} \right) > \sum \frac{N_k}{N} (\mu_k - \mu)^2$$

Im Beispiel 10.1 ist wegen fehlender externer Varianz die rechte Seite der Ungleichung Null. Für die linke Seite gilt bei  $n_1 = 1, n_2 = 2, n = 3$

$$\frac{13}{2} \left( \frac{3}{2} - 1 \right) + \frac{1}{2} \left( \frac{3}{4} - 1 \right) = \frac{25}{8} > 0$$

und bei  $n_1 = 2, n_2 = 1, n = 3$

$$\frac{13}{2} \left( \frac{3}{4} - 1 \right) + \frac{1}{2} \left( \frac{3}{2} - 1 \right) = -\frac{13}{8} + \frac{1}{4} = -\frac{11}{8} < 0.$$

Im ersten Fall ist in der Tat die Varianz  $\sigma_{\bar{x}}^2$  bei geschichteter Stichprobe größer als bei einfacher Zufallsauswahl, so dass genau ein "Schichtungsverlust" eintritt, während im zweiten Fall ein Schichtungsgewinn zu verzeichnen ist.

Aus der Ungleichung 10.14, die die Bedingung für einen "Schichtungsverlust" beschreibt, ist auch ersichtlich, dass dieser *nicht* eintreten kann

- ◆ bei proportionaler Aufteilung (linke Seite der Ungl. ist Null) und damit auch bei der
- ◆ optimalen Aufteilung, weil  $\sigma_{\bar{x}}$  dann nicht größer sein kann als bei proportionaler Aufteilung.

## g) Notwendiger Stichprobenumfang bei geschichteter Stichprobe

**1. Proportionale Aufteilung:** Für den absoluten Fehler  $e$  im heterograden Fall ZmZ gilt nach Gl. 10.10a  $e^2 = z^2 V(\bar{X})_{\text{prop}}$ , was nach Gl. 10.13 und aufgelöst nach  $n$  den folgenden Mindeststichprobenumfang  $n^*$  ergibt:

$$(10.15) \quad n^* \geq \frac{z^2}{e^2} \cdot V_{\text{int}} = \frac{z^2}{e^2} \sum \frac{N_k}{N} \sigma_k^2,$$

im Unterschied zur einfachen Stichprobe  $n^* \geq \frac{z^2}{e^2} \cdot \sigma^2 = \frac{z^2}{e^2} (V_{\text{int}} + V_{\text{ext}})$

Sobald eine externe Varianz auftritt ( $V_{\text{ext}} > 0$ ) ist also der bei der gleicher Genauigkeit und Sicherheit (gemessen durch  $e$  und  $z$ ) erforderliche Stichprobenumfang geringer, wenn man eine geschichtete Stichprobe mit proportionaler Aufteilung zieht als im Fall einer einfachen Stichprobe.

Zu einer Abschätzung des notwendigen Stichprobenumfangs **im homograden Fall** und mit Endlichkeitskorrektur kann man wie folgt gelangen:<sup>9</sup>

Setzt man  $n_k$  gem. Gl. 10.11a in Gl. 10.10c ein, so erhält man wegen  $e = z\hat{\sigma}_p$

$$e^2 = z^2 \sum_k \left( \frac{N_k}{N} \right)^2 \frac{p_k q_k}{n \frac{N_k}{N} - 1} \frac{N_k - n \frac{N_k}{N}}{N_k} \approx z^2 \sum_k \alpha_k^2 \frac{p_k q_k}{n \alpha_k} \left( 1 - \frac{n}{N} \right) \quad \text{mit } \alpha_k = \frac{N_k}{N} \quad \text{und der An-}$$

nahme, dass gilt  $n_k - 1$  für alle  $k = 1, 2, \dots, K$  und damit

$$e^2 \approx z \left( \frac{1}{n} - \frac{1}{N} \right) \sum_{k=1}^K \alpha_k p_k q_k \quad \text{was dann nach } n \text{ aufgelöst für den mindesumfang } n^* \text{ ergibt}$$

$$(10.15a) \quad n^* \geq \frac{z^2 \sum \alpha_k p_k q_k}{e^2 + \frac{1}{N} \sum \alpha_k p_k q_k}$$

Es ist leicht zu sehen, dass bei  $\pi_k$  statt  $p_k = \hat{\pi}_k$  und entsprechend  $1 - \pi_k$  statt  $q_k = 1 - \hat{\pi}_k$  sowie  $\alpha_k = N_k/N$  und  $N \rightarrow \infty$  (also Ziehen mit Zurücklegen) daraus Gl. 10.14 erhält zumal der Varianz  $\sigma_k^2$  im heterograden Fall (Mittelwerte, stetig abgestufte Werte von  $x$ ) die Größe  $\pi_k (1 - \pi_k)$  entspricht, denn dann ist

$$n = n^* \geq \frac{z^2}{e^2} \sum_k \frac{N_k}{N} \cdot \pi_k (1 - \pi_k) = \frac{z^2}{e^2} \sum_k \frac{N_k}{N} \cdot \sigma_k^2.$$

Es ist nun interessant zu sehen, wie sich dies von der einfachen Zufallsauswahl unterscheidet.

Dabei ist zu beachten, dass gilt  $\pi = \sum_k \frac{N_k}{N} \pi_k$

<sup>9</sup> entnommen aus einer Unterrichtung des ZI der KVB, die ich im Rahmen meiner Tätigkeit im wiss. Beirat für das (Ärzte)Praxispanel verfasst habe.

Mindeststichprobenumfang  $n^* \geq$ 

	einfache Zufallsauswahl	geschichtete Stichprobe
mit Zurücklegen	$\frac{z^2}{e^2} \cdot \pi(1-\pi)$	$\frac{z^2}{e^2} \sum_k \frac{N_k}{N} \cdot \pi_k (1-\pi_k)$
ohne Zurücklegen	$\frac{K'}{e^2 + \frac{K'}{N}}$ mit $K' = z^2 \pi(1-\pi)$	$\frac{z^2 K^*}{e^2 + \frac{K^*}{N}}$ mit $K^* = \sum_k \frac{N_k}{N} \cdot \pi_k (1-\pi_k)$

Man sieht dass eine gewisse Symmetrie besteht. Worauf es ankommt ist schließlich, inwieweit

$V_{\text{gesch}} = \sum_k \frac{N_k}{N} \cdot \pi_k (1-\pi_k)$  abweicht von (genauer: kleiner ist als)

$$V_{\text{einfach}} = \pi(1-\pi) = \sum_k \frac{N_k}{N} \cdot \pi_k \left(1 - \sum_k \frac{N_k}{N} \pi_k\right).$$

Wenn die Varianzen innerhalb der Schichten unterschiedlich sind, dann würde man zugleich auch einen Gewinn (mehr "Genauigkeit", also ein kleinerer Fehler bei gleichem Stichprobenumfang bzw. ein kleinerer Stichprobenumfang bei gleicher Genauigkeit) wenn man eine optimale Aufteilung vornehmen würde. Dazu ein paar Beispiele (wir rechnen der Einfachheit halber mit dem Fall "mit Zurücklegen"):

**Variante 1**

$N = 1000, N_1 = 300, N_2 = 700 \pi_1 = 0,2$  und  $\pi_2 = 0,6$

Man erhält dann<sup>10</sup>

$$V_{\text{einfach}} = \pi(1-\pi) = 0,48 \cdot 0,52 = 0,2496 \text{ weil } \pi = 0,3 \cdot 0,2 + 0,7 \cdot 0,6 = 0,48$$

$$V_{\text{gesch}} = \sum_k \frac{N_k}{N} \cdot \pi_k (1-\pi_k) = 0,3 \cdot 0,2 \cdot 0,8 + 0,7 \cdot 0,6 \cdot 0,4 = 0,216$$

mit  $z^2 = 4$  und  $e^2 = (0,1)^2 = 0,01$  also  $z^2/e^2 = 400$  erhält man für die einfache Stichprobe einen Mindeststichprobenumfang von  $400 \cdot 0,2496 = 99,84$  und bei der geschichteten Stichprobe  $400 \cdot 0,216 = 86,4$ . Der Unterschied wäre erheblich größer (als 86 zu 100), wenn sich die Varianzen innerhalb der beiden Schichten stärker unterschieden.

**Variante 2**

$N = 1000, N_1 = 300, N_2 = 700 \pi_1 = 0,1$  und  $\pi_2 = 0,5$

Jetzt ist  $V_{\text{einfach}} = \pi(1-\pi) = 0,38 \cdot 0,62 = 0,2356$  und

$$V_{\text{gesch}} = 0,3 \cdot 0,1 \cdot 0,9 + 0,7 \cdot 0,5 \cdot 0,5 = 0,202.$$

Die Stichprobenumfänge erhält man wieder bei Multiplikation mit  $z^2/e^2 = 400$ . Es ergibt sich 80,8 also 81 und 94,24 also 94.

**Variante 3**

$N = 1000, N_1 = 300, N_2 = 700 \pi_1 = 0,1$  und  $\pi_2 = 0,9$

$$V_{\text{einfach}} = \pi(1-\pi) = 0,66 \cdot 0,34 = 0,2244 \rightarrow 400 \cdot 0,2244 = 89,76 \approx 90 \text{ statt } 100 \text{ bei } \pi(1-\pi) = \frac{1}{4}$$

$$V_{\text{gesch}} = 0,3 \cdot 0,1 \cdot 0,9 + 0,7 \cdot 0,9 \cdot 0,1 = 0,09 \rightarrow 400 \cdot 0,09 = 36.$$

Hier wird der Unterschied natürlich sehr beträchtlich. Bei proportionaler Aufteilung muss die im Verhältnis 3 zu 7 (wie die Grundgesamtheit) aufgeteilt werden, also  $36 \cdot 0,3 = 10,8$  also 11 zu 25. Es ist nicht schwer jetzt für die drei Varianten die etwas komplizierten Formeln ohne Zurücklegen anzuwenden (bei denen also auch die Größe  $N$  mit ins Spiel kommt).

<sup>10</sup> Wir betrachten hier stets die  $n$ -fachen Varianzen des Mittelwerts ( $x$ -quer) und rechnen mit  $z = 4$  (statt korrekt bei 95% Sicherheit mit  $z=1,96$ ).

**2. Bei optimaler Aufteilung** erhält man durch Einsetzen von Gl. 10.11a in Gl. 10.10a

$V(\bar{X}) = \sum_{k=1}^K \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k}$  für  $e^2 = z^2 V(\bar{X})$  im Fall ZmZ nach einigen Umformungen und Auflösung nach  $n$  die Gleichung für den Mindestumfang  $n^*$

$$(10.15b) \quad n^* \geq \frac{z^2}{e^2} \left( \sum \frac{N_k}{N} \sigma_k \right)^2 = \frac{z^2}{e^2} \bar{\sigma}^2$$

was weniger ist als die entsprechende Gl. 10.15  $n^* \geq \frac{z^2}{e^2} \sum \frac{N_k}{N} \sigma_k^2$  bei proportionaler Aufteilung

weil  $\sum \frac{N_k}{N} \sigma_k^2 > \left( \sum \frac{N_k}{N} \sigma_k \right)^2$ .

Es ist auch eine i.d.R. nicht-proportionale Aufteilung der Stichprobe in der Weise möglich, dass für jede Schicht eine vorgegebene Genauigkeit (gemessen am absoluten Fehler  $e_k$ ) eingehalten wird, was bei  $n_k^* \geq \frac{z^2}{e_k^2} \cdot \sigma_k^2 \quad \forall k$  gewährleistet ist.

### 3. Klumpenstichprobe (cluster sample) und zweistufige Auswahl

#### a) Beschreibung des Auswahlverfahrens

Aus organisatorischen und wirtschaftlichen Gründen ist es oft vorteilhaft eine sog. "Klumpenstichprobe" zu ziehen, weil die Erhebungseinheiten bereits "gebündelt" als Klumpen (cluster) vorliegen. Ein wichtiger Spezialfall der Klumpenstichprobe ist die Flächenstichprobe (area sample). Personen, die in einer Gemeinde wohnen, stellen einen Klumpen dar.

Die praktischen Vorteile sind vor allem:

- Verringerung von Reisekosten (bei einem area sample) durch Bündelung der Feldarbeit
- es ist keine vollständige Auswahlgrundlage (sampling frame) erforderlich (z.B. eine Einwohnerkartei für das gesamte Bundesgebiet) und entsprechende Verzeichnisse sind ggfls nur für die ausgewählten Klumpen zu beschaffen
- anderes als bei einer Schichtung sind kaum Kenntnisse über die GG erforderlich (es ist meist noch nicht einmal nötig den Umfang der GG oder der einzelnen Klumpen zu kennen).

Den Vorteilen stehen aber auch Nachteile gegenüber. Wie im folgenden gezeigt wird, kann die Präzision (Standardfehler der Schätzung) bei diesem Stichprobendesign geringer sein als bei einer uneingeschränkten Zufallsauswahl.

#### Def. 10.6

- Teilt man die GG vom Umfang  $N$  in  $M$  disjunkte Teilmengen mit den Umfängen  $N_i$  ( $i = 1, \dots, M$ ) so dass  $N_1 + N_2 + \dots + N_M = \sum_{i=1}^M N_i$  und wählt man hieraus  $m < M$  Teilmengen aus, so heißen die Teilmengen Klumpen und das Auswahlverfahren Klumpenauswahl.

- Werden von den  $M$  Klumpen  $m$  zufällig ausgewählt und diese  $m$  Klumpen mit *allen* ihren Einheiten ausgezählt so spricht man von einem *einstufigen* Klumpenstichprobe<sup>11</sup>
- Werden dagegen von den  $M$  Klumpen  $m$  zufällig ausgewählt und jeweils  $n_j$  Einheiten beim ausgewählten  $j$ -ten Klumpen zufällig ausgewählt (Auswahlsätze  $n_j/N_j$  kleiner als 100%) und untersucht so spricht man von einer *zweistufigen* Klumpenauswahl.

### Bemerkungen zu Def. 10.6

Ein Klumpen ist eine natürliche (vorgefundene) Ansammlung von Untersuchungseinheiten, die in sich möglichst heterogen sein sollte (eine verkleinerte GG) und die Klumpen sollten untereinander möglichst homogen sein. Hinsichtlich Klumpen und Schichten werden also gegensätzliche Forderungen aufgestellt.

Gegensätzlich ist auch die Art der Auswahl: es werden alle  $K$  Schichten berücksichtigt aber innerhalb der Schichten die Einheiten ausgewählt und es wird bei der einstufigen Klumpenauswahl nur  $m < M$  Klumpen berücksichtigt, aber innerhalb der ausgewählten Klumpen alle Elemente.

### b) Notation, Fragestellungen

Es empfiehlt sich auch zur Behandlung der zweistufigen Klumpenauswahl folgende Subskripte zu vereinbaren

$i = 1, 2, \dots, M$  für die Klumpen der GG

$j = 1, 2, \dots, m$  für die ausgewählten Klumpen der Stichprobe

und auch unterschiedliche Subskripte für die Elemente in den Klumpen vorzusehen

$k = 1, 2, \dots, N_i$  der Klumpen der GG

$l = 1, 2, \dots, n_j$  der ausgewählten Elemente des ausgewählten Klumpens

obgleich natürlich mit  $i$  und  $j$  eventl. der gleiche Klumpen und mit  $k$  und  $l$  das gleiche Element gemeint sein kann.

Ferner ist es (ebenfalls zum Verständnis der zweistufigen Auswahl) vorteilhaft nicht nur die Schätzung des arithmetischen Mittels sondern auch des Totalwerts (der Merkmalssumme) zu betrachten:  $X = \sum_{i=1}^M X_i = \sum_{i=1}^M \sum_{k=1}^{N_i} x_{ik} = N\mu$  ist der Totalwert der GG als Summe der Klumpentotalwerte. Große Buchstaben  $X$  bezeichnen Merkmalssummen, kleine einzelne Merkmalsbeiträge. Das ist jedoch vor allem dann unschön, aber schwer vermeidbar, wenn Erwartungswerte gebildet werden. Was die Symbolik betrifft so entstehen auch dadurch Probleme, weil unterschiedliche Arten von Mittelwerten definiert werden müssen. Ein mittlerer Totalwert von  $M$  Klumpen  $\bar{X}_{(M)} = \frac{1}{M} \sum X_i$  oder von  $m$  ausgewählten Klumpen  $\bar{X}_{(m)} = \frac{1}{m} \sum X_j$  ist zu unterscheiden von den Mittelwerten etwa  $\mu = \frac{1}{N} \sum X_i = \frac{1}{N} \sum N_i \mu_i$  (Summe über alle Klumpen).

Ein Mittelwert  $\hat{\mu}$  ist nicht einfach definiert als  $\frac{1}{m} \sum_{j=1}^m X_j$  (Summe über die ausgewählten Klumpen) sondern gem. Gl. 10.18.

Zur Verbesserung der Übersichtlichkeit ist in allen Fällen, in denen keine Zweifel aufkommen sollten, auch auf eine Indizierung der Summen verzichtet worden. Die Darstellung wird auch

<sup>11</sup> Klumpenauswahl im engeren Sinne, in der Art, wie in Übers. 10.1. In Gegenüberstellung zur geschichteten Stichprobe beschrieben; Auswahlsatz auf der zweiten Stufe 100%. Der Stichprobenumfang  $n$  liegt dann mit der Anzahl  $m$  der ausgewählten Klumpen fest mit  $n = \sum N_j$  ( $j = 1, 2, \dots, m$ ).

dadurch kompliziert, weil Totalwerte aufgrund aller Einzelwerte eines Klumpens oder auch nur aufgrund der Merkmalswerte ausgewählten Einheiten eines Klumpens gebildet sein können.

Auf Beweise ist im folgenden stets verzichtet worden und es wurde statt dessen versucht, Zusammenhänge anhand von Zahlenbeispielen zu verifizieren und damit auch zu interpretieren.

## c) Einfache Klumpenstichprobe

### 1. Schätzung des Totalwerts

Untersucht werden zwei Schätzfunktionen

$$(10.16) \quad \hat{X} = \sum_{j=1}^m \frac{M}{m} X_j = \frac{M}{m} \sum_{j=1}^m X_j \quad \text{mit } X_j = \sum_{k=1}^{N_j} x_{jk} \quad \text{und}$$

$$(10.16a) \quad \hat{X}^* = \frac{N}{n} \sum_{j=1}^m X_j$$

mit  $n = \sum N_j$  und  $X_j = N_j \mu_j$  da im einstufigen Verfahren die Klumpen voll ausgezählt werden. Die Schätzfunktion  $\hat{X}$  ist erwartungstreu und hat die Varianz

$$(10.17) \quad V(\hat{X}) = M \left( \frac{M}{m} - 1 \right) \frac{1}{M-1} \sum_{i=1}^M (X_i - \bar{X}_{(M)})^2 = \frac{M}{m} \frac{M-m}{M-1} \sum_{i=1}^M (X_i - \bar{X}_{(M)})^2$$

Diese "wahre" Fehlervarianz ist zu unterscheiden von der geschätzten Fehlervarianz, für die gilt

$$(10.17a) \quad \hat{V}(\hat{X}) = M \left( \frac{M}{m} - 1 \right) \frac{1}{m-1} \sum_{j=1}^m (X_j - \bar{X}_{(m)})^2 = \frac{M}{m} \frac{M-m}{m-1} \sum_{i=1}^M (X_i - \bar{X}_{(M)})^2$$

Die Schätzfunktion  $\hat{X}^*$  ist dagegen nur näherungsweise erwartungstreu und hat aber in der Regel eine erheblich kleinere Varianz, nämlich

$$(10.17b) \quad V(\hat{X}^*) \approx M \left( \frac{M}{m} - 1 \right) \frac{1}{M-1} \sum_{i=1}^M N_i^2 (\mu_i - \mu)^2 = \frac{M}{m} \frac{M-m}{M-1} \sum_{i=1}^M N_i^2 (\mu_i - \mu)^2. \quad \text{Das leicht}$$

sehen mit  $\sum_{i=1}^M (X_i - \bar{X}_{(M)})^2 = \sum_{i=1}^M (N_i \mu_i - N \mu)^2 = \sum_{i=1}^M (N_i^2 \mu_i^2 - 2 N_i N \mu \mu_i + N^2 \mu^2)$  im Vergleich zu

$$\sum_{i=1}^M N_i^2 (\mu_i - \mu)^2 = \sum_{i=1}^M \left( N_i^2 \mu_i^2 - 2 N_i N \mu \mu_i + \mu^2 \sum_i N_i^2 \right).$$

Es ist klar, dass  $N^2 = (N_1 + \dots + N_M)^2 > N_1^2 + \dots + N_M^2$

Aus den Schätzfunktionen für den Totalwert  $X$  werden die Schätzfunktionen für den Mittelwert  $\mu$  abgeleitet, denn  $\mu = X/n$ . Häufig wird zur Vereinfachung angenommen, dass alle Klumpen den gleichen Umfang haben, also  $N_i = \bar{N}$  für alle  $i = 1, \dots, M$  so dass  $N = M \bar{N}$  und

$n = m \bar{N}$ . Dann sind die Schätzfunktionen  $\hat{X}$  und  $\hat{X}^*$  identisch, da dann  $\frac{N}{n} = \frac{M \bar{N}}{m \bar{N}} = \frac{M}{m}$ .

### Beispiel 10.2

Die GG bestehe aus den Klumpen (5, 10, 15), (6, 11, 16), (7, 12, 12, 17) mit dem Totalwert  $X=111$  und ungleichen Klumpengrößen  $N_1 = N_2 = 3$  und  $N_3 = 4$ . Es werden  $m = 2$  der  $M = 3$  Klumpen ausgewählt und total erhoben. Die Stichprobenverteilung ist dann

Klumpen	$\hat{X}$	$\hat{X}^*$
1 und 2	$\frac{3}{2} \cdot 30 + \frac{3}{2} \cdot 33 = 94,5$	$\frac{10}{6}(30 + 33) = 105$
1 und 3	$\frac{3}{2} \cdot 30 + \frac{3}{2} \cdot 48 = 117$	111,43
2 und 3	$\frac{3}{2} \cdot 33 + \frac{3}{2} \cdot 48 = 121,5$	115,71

Da jede Stichprobe gleichwahrscheinlich ist, gilt

$E(\hat{X}) = 111$  und  $V(\hat{X}) = E(\hat{X} - X)^2 = E(\hat{X}^2) - [E(\hat{X})]^2 = 12460,5 - 12321 = 139,5$ . Man erhält diesen Wert auch mit Gl. 10.17 da  $X_1 = 30, X_2 = 33, X_3 = 48$  und  $\frac{1}{3}(30 + 33 + 48) = \frac{111}{3} = 37$ ,

wie man sieht ist  $\bar{X}_{(M)} = 37 \neq \mu = \frac{111}{10} = 11,1$ .

Demgegenüber ist die Schätzfunktion  $\hat{X}^*$  nicht erwartungstreu. Man sieht leicht, dass  $E(\bar{X}^*) = \frac{1}{3}(105 + 111,43 + 115,71) = 110,714$  statt 111. Die Varianz ist näherungsweise nach

$$\text{Gl. 10.17b } V(\hat{X}^*) \approx 3 \cdot \left(\frac{3}{2} - 1\right) \frac{1}{2} \left[3^2(10 - 11,1)^2 + 3^2(11 - 11,1)^2 + 4^2(12 - 11,1)^2\right] = 17,955$$

Die Varianz von  $\hat{X}^*$  ist meist kleiner als die von  $\hat{X}$  (so auch im Beispiel) und der Präzisionsgewinn bei Verwendung von  $\hat{X}^*$  statt  $\hat{X}$  ist groß, wenn

- die Klumpenumfänge  $N_i$  sehr unterschiedlich sind
- die Klumpenmittelwerte  $\bar{X}_i = \mu_i$  weniger streuen als die Einzelwerte.

## 2. Schätzfunktionen für den Mittelwert $\mu$

Der folgende Punktschätzer

$$(10.18) \quad \hat{\mu} = \frac{1}{N} \hat{X} = \frac{M}{N \cdot m} \sum_{j=1}^m X_j = \frac{M}{N \cdot m} \sum N_j \mu_j$$

auf der Basis des Schätzers  $\hat{X}$  (gem. Gl. 10.16) ist erwartungstreu (ohne Beweis und Demonstration am Beispiel) und im Falle von  $N_i = \bar{N}$  ist er identisch mit

$$(10.18a) \quad \hat{\mu}^* = \frac{1}{N} \hat{X}^* = \frac{1}{n} \sum N_j \mu_j \quad \text{mit } \sum N_j = n. \text{ Ist } N_i = \bar{N} \text{ so ist}$$

$$(10.19) \quad \bar{X} = \hat{\mu} = \frac{1}{m} \sum \mu_j \quad \text{der Schätzer für } \mu = \frac{1}{M} \sum \mu_i$$

Die Varianz dieses Schätzers ist bei einer Auswahl von  $m$  aus  $M$  Klumpen (ZoZ) gegeben mit

$$(10.20) \quad V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma_b^2}{m} \frac{M - m}{M - 1}$$

mit  $\sigma_b^2 =$  Varianz zwischen (between) den Klumpen. Sie ist definiert als

$$(10.20 a) \quad \sigma_b^2 = \frac{1}{N} \sum_{i=1}^M N_i (\mu_i - \mu)^2 = \frac{1}{M} \sum_{i=1}^M (\mu_i - \mu)^2 = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{\bar{N}} \sum_{k=1}^{\bar{N}} x_{ik} - \mu \right)^2$$

Die mit Gl. 10.20 bestimmte (wahre) Varianz von  $\bar{X}$  bei endlicher Grundgesamtheit ist nicht zu verwechseln mit der geschätzten (empirischen) Varianz von  $\bar{X}$ , nämlich  $\hat{V}(\hat{X}) = \frac{1}{N^2} \hat{V}(\hat{X})$

Im Fall  $N_i = \bar{N}$  (gleiche Klumpenumfänge) erhält man

$$(10.20b) \quad \hat{V}(\bar{X}) = \frac{M-m}{Mm} \hat{\sigma}_b^2 = \frac{1}{M} \left( \frac{M}{m} - 1 \right) \hat{\sigma}_b^2 \quad \text{mit} \quad \hat{\sigma}_b^2 = \frac{1}{m-1} \sum_{j=1}^m (\mu_j - \hat{\mu})^2.$$

Zum Verständnis der Klumpenstichprobe ( und auch zum Vergleich der Präzision dieses Stichprobendesigns mit der einfachen Stichprobe) ist es notwendig, den Ausdruck

$\hat{\sigma}_b^2 = \frac{1}{M} \sum_{i=1}^M (\mu_i - \hat{\mu})^2$  weiter zu analysieren (oder den Ausdruck  $\hat{\sigma}_b^2 = \frac{1}{m-1} \sum_{j=1}^m (\mu_j - \hat{\mu})^2$  also den Schätzer für  $\sigma_b^2$ ). Das soll im Folgenden geschehen und es gibt hierzu auch eine kurze Darstellung im Anhang.

Da  $\mu_i = \frac{1}{\bar{N}} \sum_{k=1}^{\bar{N}} X_{ik}$  und  $\sum_{i=1}^M \mu_i = \bar{N} \mu$  erhält man für  $\sigma_b^2$  nach Gl. 10.20a

$$(10.21) \quad \begin{aligned} \sigma_b^2 &= \frac{1}{N} \sum_{i=1}^M N_i (\mu_i - \mu)^2 = \frac{1}{M} \sum_{i=1}^M (\mu_i - \mu)^2 = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{\bar{N}} \sum_{k=1}^{\bar{N}} x_{ik} - \mu \right)^2 \\ &= \frac{1}{M\bar{N}^2} \sum_{i=1}^M \left( \sum_{k=1}^{\bar{N}} x_{ik} - \bar{N}\mu \right)^2 = \frac{1}{M\bar{N}^2} \sum_{i=1}^M \left[ \sum_{k=1}^{\bar{N}} (x_{ik} - \mu) \right]^2 \end{aligned}$$

Die Interpretation einer Summe, deren M Summanden quadrierte Summen sind, mag Schwierigkeiten bereiten und wird im Anhang noch demonstriert.

Die Summe der quadrierten Summen läßt sich zerlegen in

- $M\bar{N} = N$  Größen  $(x_{ik} - \mu)^2$ , die in ihrer Summe die N-fache ( $M\bar{N}$ -fache) Gesamtvarianz ( $\sigma^2$ ) der Variable X in der (endlichen) Grundgesamtheit ergeben und
- $M\bar{N}(\bar{N}-1)$  Glieder der Art  $(x_{ik} - \mu)(x_{il} - \mu)$  mit  $k \neq l$ , und  $k, l = 1, 2, \dots, N_i = \bar{N}$ . Das

Mittel dieser Produkte  $\frac{1}{M\bar{N}(\bar{N}-1)} \sum_{i=1}^M \left[ \sum_{k=1}^{\bar{N}} \sum_{l=1}^{\bar{N}} (x_{ik} - \mu)(x_{il} - \mu) \right] = \sigma_{kl} = \rho \sigma^2$ ;  $k \neq l$  (die

Doppelsumme in den eckigen Klammern hat  $\bar{N}(\bar{N}-1)$  Summanden) ist die durchschnittliche Kovarianz ( $\sigma_{kl}$ ) der Beobachtungen innerhalb der Klumpen und  $\rho$  heißt (durchschnittlicher) **Intraklasskorrelationskoeffizient** (intra-class correlation). Er ist ein Maß der Homogenität der Klumpen.

Man beachte, dass die Intraklasskorrelation  $\rho$  nicht den Zusammenhang zwischen zwei Variablen mißt, sondern die Unterschiedlichkeit der Elemente hinsichtlich einer Variable X (bzw. der Abweichung des X-Wertes vom Gesamtmittel  $\mu$ )

## d) Mittelwertschätzung bei einstufiger Klumpenauswahl und $N_i = \bar{N}$

### 1. Punktschätzung $\bar{X}$

$$(10.15) \quad \bar{X} = \hat{\mu} = \frac{\sum_j N_j \mu_j}{\sum_j N_j} = \frac{\sum_j \sum_k X_{jk}}{m\bar{N}} = \frac{\sum_j \mu_j}{m},$$

( $j = 1, 2, \dots, m$ ; Stichprobenumfang  $n = m\bar{N}$ ,  $i = 1, \dots, M$  und  $k = 1, 2, \dots, N_i = \bar{N}$ )

Wegen der Vollerhebung innerhalb eines Klumpens ist das Klumpenmittel  $\mu_j$  ohne Stichprobenfehler zu schätzen (allerdings nur bei den Klumpen, die in die Auswahl gelangen,  $j = 1, 2, \dots, m$ ).

Der Schätzwert  $\hat{\mu}$  beruht auf den  $m$  ausgewählten Klumpen im Unterschied zum wahren Mittelwert  $\mu$  der endlichen GG für den gilt  $\mu = \frac{\sum N_i \mu_i}{\sum N_i} = \frac{\sum \mu_i}{M}$  mit  $i = 1, 2, \dots, M$  und mit  $N = \sum N_i = M \cdot \bar{N}$ .

## 2. Varianz von $\bar{X}$

Die Varianz von  $\bar{X}$  bei einer Auswahl von  $m$  aus  $M$  Klumpen (ZoZ) ist wie dargestellt

$$(10.20) \quad V(\bar{X}) = \sigma_{\bar{x}}^2 = \frac{\sigma_b^2}{m} \frac{M-m}{M-1}$$

wobei für die Varianz  $\sigma_b^2$  zwischen (between) den Klumpen gilt

$$(10.21) \quad \sigma_b^2 = \frac{1}{M\bar{N}^2} \sum_{i=1}^M \left[ \sum_{k=1}^{\bar{N}} (x_{ik} - \mu) \right]^2$$

## 3. Interpretation

Es soll nun gezeigt werden, dass die Qualität der Klumpenstichprobe von der Intraklasskorrelation  $\rho$  abhängt. Offenbar ist  $M\bar{N}^2 = M\bar{N} + M\bar{N}(\bar{N} - 1)$ , so dass man aus Gl. 10.20a erhält

$$(10.21) \quad \sigma_b^2 = \frac{\sigma^2}{\bar{N}} + \frac{M\bar{N}(\bar{N} - 1)}{M\bar{N}^2} \rho \sigma^2 = \frac{\sigma^2}{\bar{N}} [1 + (\bar{N} - 1)\rho] = \frac{\sigma^2}{\bar{N}} V_{BL}.$$

Der Faktor in den eckigen Klammern wird auch (angenäherter) **Varianzaufblähungsfaktor** ( $V_{BL}$ ) genannt. Ist  $V_{BL} < 1$  weil  $\rho < 0$  ist ( $V_{BL} > 1$  wegen  $\rho > 0$ ), so ist die Klumpenstichprobe wirksamer als die einfache Stichprobe, denn Gl. 10.17 eingesetzt in Gl. 10.16 liefert:

$$(10.22) \quad V(\bar{X}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \frac{M-m}{M-1} [1 + (\bar{N} - 1)\rho] \approx \frac{\sigma^2}{n} \left(1 - \frac{m}{M}\right) V_{BL}.$$

für die Varianz der Stichprobenverteilung bei einer Klumpenstichprobe (wegen  $n = m\bar{N}$ ) im Vergleich zur Varianz

$$(10.23) \quad V(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{\sigma^2}{n} \cdot \frac{M\bar{N} - m\bar{N}}{M\bar{N} - 1} = \frac{\sigma^2}{n} \cdot \frac{M-m}{M-1/\bar{N}} \approx \frac{\sigma^2}{n} \left(1 - \frac{m}{M}\right)$$

bei einfacher Zufallsauswahl.

Wird in Gl. 10.22 und 10.23 jeweils  $\sigma^2$  durch  $\hat{\sigma}^2$  geschätzt, so ist im Nenner der Endlichkeitskorrektur  $M$  (bzw.  $N$ ) statt  $M-1$  (bzw.  $N-1$ ) zu schreiben, so dass  $\approx$  durch  $=$  zu ersetzen ist

$$\text{und } V(\bar{X}) = \frac{\sigma^2}{n} \left(1 - \frac{m}{n}\right) V_{BL}.$$

Man beachte, dass  $V_{BL}$  nur dann annähernd das Verhältnis der beiden Varianzen darstellt, wenn  $1/\bar{N} \approx 0$  ist. Das korrekte Verhältnis der Varianzen ist nämlich größer. Es ist

$$(10.23a) \quad V_{BL}^* = V_{BL} \frac{(M-m)(N-1)}{(M-1)(N-n)} = V_{BL} \frac{1 - \frac{1}{N}}{1 - \frac{1}{M}}$$

Der Näherungswert  $V_{BL}$  ist gleich dem exakten Faktor  $V_{BL}^*$  ist, wenn man  $M - 1 = M$  und  $N - 1 = N$  setzt. Der Ausdruck "Varianzaufblähungsfaktor" sollte nicht zu dem Mißverständnis führen, dass die Varianz der Stichprobenverteilung von  $\bar{X}$  bei einer Klumpenstichprobe stets größer sein muß als bei uneingeschränkter Zufallsauswahl. Der Faktor  $V_{BL}$  (und damit auch  $V_{BL}^*$ ) kann sehr wohl auch kleiner als 1 sein, er kann sogar Null sein, wie das folgende Beispiel zeigt, d.h. man „gewinnt“ durch eine Klumpenstichprobe, im Extremfall sogar soviel, dass eine sichere „Schätzung“ möglich ist, (nämlich dann, wenn die GG aus lauter identischen Klumpen besteht).

Da  $\rho$  ein Korrelationskoeffizient ist, sind die extremen Werte  $\rho = -1$  und  $\rho = +1$  denkbar, so dass der Intraklass-Korrelation  $\rho$  folgende Faktoren erhalten kann

$$\rho = -1 \rightarrow V_{BL} = 2 - \bar{N}, \text{ weil aber } V_{BL} \text{ nicht negativ sein kann, darf } \rho \text{ den Wert } -1/(\bar{N} - 1) \text{ nicht unterschreiten}$$

$$\rho = +1 \rightarrow V_{BL} = \bar{N}.$$

Es gilt: ist  $V_{BL} < 1$  so ist die Klumpenstichprobe effizienter, als die einfache Stichprobe (ZoZ) ist  $V_{BL} > 1$  so ist sie weniger effizient.

Aus  $V_{BL} = 1 - (\bar{N} - 1)\rho$  folgt dabei für die Intraklasskorrelation

ist $\rho < 0$	dann ist auch $V_{BL} > 1$
$\rho = 0$	$V_{BL} = 1$
$\rho > 0$	$V_{BL} < 1$ (Klumpenstichprobe effizienter)

Positive Intraklasskorrelation (= Homogenität der Klumpen) ist also vorteilhaft

Man kann sich leicht überzeugen, dass der Fall  $\rho = +1$  dann gegeben ist, wenn die Klumpen jeweils aus lauter gleichen Elemente bestehen, wenn es also innerhalb der Klumpen von jeweils  $\bar{N}$  Elementen keine Streuung gibt. Dann ist stets  $\sigma^2 = \sigma_b^2 = \sigma_{kl}$  und die Klumpenstichprobe ist wegen  $V_{BL} = \bar{N} > 1$  der uneingeschränkten Stichprobe unterlegen und zwar um so mehr, je größer die Klumpen sind.

Im Fall  $\rho = -1$  (bei  $\bar{N} = 2$ ) ist es offensichtlich, dass dann wegen  $\sigma_b^2 = 0$  die Varianz  $\sigma_x^2 = 0$  ist, so dass  $V_{BL} = 0$  (vgl. Beispiel 10.3).

### Beispiel 10.3

Mit dem folgenden Rechenbeispiel sollen die Überlegungen, die zu den Gl. 10.21 - 10.23 führten verifiziert werden. Weil hier extreme Konstellationen untersucht werden, eignet sich das Beispiel auch zur Interpretation der Gleichungen.

Man berechne  $\sigma^2$ ,  $\sigma_b^2$ ,  $\sigma_{kl}$ ,  $\rho$  und  $V_{BL}$  für die folgenden Situationen

- die Grundgesamtheit besteht aus  $M = 5$  identischen Klumpen mit jeweils  $\bar{N} = 3$  Einheiten  
 $x_{i1} = 5, x_{i2} = 10, x_{i3} = 15$  ( $\mu_1 = \dots = \mu_5 = \mu = 10$ )
- die GG besteht aus 5 Klumpen mit jeweils  $\bar{N} = 3$  Einheiten (5, 10, 15),

(6, 11, 16), (7, 12, 17), (8, 13, 18), (9, 14, 19) mit  $\mu_1 = 10, \mu_2 = 11, \dots, \mu_5 = 14, \mu = 12$

[In runden Klammern erscheinen jeweils die einzelnen Einheiten des Klumpens]

c) die 5 Klumpen seien (1, 2, 3), (4, 5, 6), (7, 8, 9), (10, 11, 12), (13, 14, 15) mit  $\mu_1 = 2, \mu_2 = 5$  usw. und  $\mu = 8$ .

Man berechne nun  $V(\bar{X})$  gem. (10.20) (Varianz der Stichprobenverteilung von  $\bar{X}$  bei einer Klumpenstichprobe) und  $V(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$ , die Varianz bei einer uneingeschränkten einfachen Stichprobe mit  $m = 2, n = 2 \cdot 3 = 6, N = 15, M = 5$  und man interpretiere das Ergebnis!

Lösung 10.3

Abgesehen von der Berechnung der Intraklass-Kovarianz  $\sigma_{kl}$  sind die Berechnungen einfach, so dass eine Tabelle der Ergebnisse ausreichen mag:

Fall	$\sigma^2$	$\sigma_b^2 = \sum(\mu_i - \mu)^2 / M$	$\sigma_{kl} = \rho\sigma^2$	$\rho$	$V_{BL}$	$V_{BL}^*$
a	250/15 = 16,667	0 (da alle Cluster-Mittel gleich)	-250/30 = -8,333	-0,5	0	0
b	280/15 = 18,667	$\frac{1}{5}(4+1+0+1+4) = 2$	-190/30 = -6,333	-0,339	0,321	0,375
c	280/15 = 18,667	$\frac{1}{5}(36+9+0+9+36) = 18$	530/30 = 17,667	+0,946	2,893	3,375

$$V_{BL}^* = V_{BL} \cdot \begin{bmatrix} 1-1/15 \\ -1-1/5 \end{bmatrix}$$

Für Fall a) sei die Berechnung der Intraklasskovarianz  $\sigma_{kl}$  beispielhaft vorgeführt. Es sind für jedes (etwa das i-te) Cluster die Produkte  $(x_{ik} - \mu)(x_{il} - \mu)$  zu bilden und zu addieren. Man erhält hier  $2 \cdot (5-10)(10-10) + 2(5-10)(15-10) + 2(10-10)(15-10) = -50$ . Die Summe über alle fünf identische Klumpen ist also -250. Der Nenner  $M\bar{N}(\bar{N}-1)$  ist dann  $5 \cdot 3 \cdot 2 = 30$ . Der Varianzaufblähungsfaktor ist  $V_{BL} = 1 + (\bar{N}-1)\rho = 1 + 2 \cdot (-0,5) = 0$ , d.h. in diesem Extremfall ist die Stichprobenverteilung von  $\bar{X}$  eine „Einpunktverteilung“ mit  $\sigma_{\bar{X}}^2 = 0$ . Sowohl  $\rho$  als auch  $V_{BL}$  haben in diesem konstruierten Beispiel ihre minimalen Werte.

Die Konsequenzen für die Stichprobenverteilung von  $\bar{X}$  sind dann

Fall	Klumpenstichprobe $V(\bar{X}) = \frac{\sigma_b^2}{m} \frac{M-m}{M-1} = \frac{\sigma_b^2}{2} \cdot \frac{3}{4}$	uneingeschränkte (einfache) Stichprobe $V(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} = \frac{\sigma^2}{6} \cdot \frac{9}{14}$
a	0, da $\sigma_b^2 = 0$	25/14 = 1,7857 > 0
b	3/4, da $\sigma_b^2 = 2$	2 > 3/4
c	27/4 = 6,75, da $\sigma_b^2 = 18$	2 < 6,75

Bemerkungen zu den Ergebnissen:

Weil die Varianz  $\sigma^2$  in den Fällen b und c gleich ist, überrascht es nicht, dass in beiden Fällen für die uneingeschränkte Zufallsauswahl  $V(\bar{X}) = 2$  ist;

- dass Fall a zu einer sicheren „Schätzung“ mit  $\sigma_{\bar{x}}^2 = V(\bar{X}) = 0$  führt ist sehr plausibel, denn in diesem konstruierten Extremfall ist die Totalerhebung auch nur eines Klumpens identisch mit einer Totalerhebung der gesamten GG;
- im Fall b entsteht eine negative Intraklasskorrelation  $\rho$ , weil die fünf Klumpen alle jeweils die Werte enthalten, die in der Umgebung des Gesamtmittels  $\mu$  liegen
- sind dagegen die Klumpen so konstruiert (wie im Falle c), dass einige Klumpen mit allen ihren Merkmalswerten unterhalb  $\mu$  und andere oberhalb von  $\mu$  liegen, so entsteht eine positive Korrelation  $\rho$  und damit eine Verschlechterung der Schätzung durch eine Klumpenstichprobe (also  $V_{BL} > 1$ ).

### Beispiel 10.4

Man konstruiere Beispiele für Intraklassenkorrelationen von  $\rho = +1$  und  $\rho = -1$  und untersuche die Konsequenzen ( $M = 5$ ,  $\bar{N} = 2$ ).

### Lösung 10.4

- a) Ein Beispiel für  $\rho = +1$  wäre die folgende GG mit 5 Klumpen (5, 5), (6, 6), (7, 7), (8, 8), (9, 9) offenbar ist  $\sigma^2 = \sigma_b^2 = 2$  und die Klumpenstichprobe vom Umfang  $m\bar{N} = n = 2 \cdot 2 = 4$  ist schlechter als eine gleich umfangreiche uneingeschränkte Zufallsauswahl. Man erhält für die Klumpenstichprobe  $V(\bar{X}) = \frac{2}{2} \cdot \frac{5-2}{5-1} = \frac{3}{4}$  aber für die uneingeschränkte Zufallsauswahl  $V(\bar{X}) = \frac{2}{4} \cdot \frac{10-4}{10-1} = \frac{1}{3}$ .

Die Wahrscheinlichkeit eine extreme Auswahl z.B. die Werte 5, 5, 6, 6 zu erhalten ist im ersten Fall  $\binom{5}{2}^{-1} = \frac{1}{10}$ , im zweiten Fall dagegen nur  $\frac{1}{\binom{10}{4}} = \frac{1}{210}$  also nur 0,476% statt

10%. Deshalb überrascht es nicht, dass in diesem extremen Fall mit einer Klumpenstichprobe ein schlechteres Ergebnis als mit einer uneingeschränkten Zufallsauswahl erzielt wird.

- b) Ein Beispiel für  $\rho = -1$  wäre die GG (5,9), (6,8), (7,7), (8,6), (9,5). Hier ist  $\sigma_b^2 = 0$ ,  $\sigma^2 = 2$  und  $\sigma_{kl} = -2$ . Offenbar erlaubt die Ziehung eines beliebigen Klumpens die exakte Bestimmung von  $\mu$  weil  $\mu_1 = \dots = \mu_5$ , so dass es nicht verwundert, dass  $V_{BL} = 0$ .

### **e) Zweistufige (Klumpen-)Auswahl**

Auf diesen komplizierten Stichprobenplan soll hier nur kurz eingegangen werden, denn Einzelheiten sind der Spezialliteratur zur Stichprobentheorie zu entnehmen.

Die Varianz  $V(\bar{X})$  läßt sich etzt zerlegen in eine von  $\sigma_b^2$  und eine von  $\sigma_w^2$  (Varianz innerhalb [within] der Klumpen) abhängige Komponente. Anders als in Gl. 10.19 sind die Klumpenmit-

telwerte  $\mu_i$  durch  $\bar{x}_j$  zu schätzen. Man erhält wieder mit  $N_j = \bar{N}$  als Punktschätzer für  $\mu$  (der Gesamtmittelwert der Stichprobe)  $\bar{x} = \sum \bar{x}_j / m$ .

Der Stichprobenumfang ist jetzt  $n = \sum n_j \leq \sum N_j = m\bar{N}$ , weil auf der zweiten Stufe ausgewählt wird ( $n_j/N_j \leq 1$  für  $j = 1, 2, \dots, m$ ). Für die Varianz von  $\bar{X}$  erhält man mit  $\hat{\sigma}_b^2$  und  $\hat{\sigma}_{w_i}^2$ ,  $i = 1, \dots, M$

$$(10.24) \quad \hat{V}(\bar{X}) = \frac{1}{N^2} \left[ M^2 \left( \frac{1}{m} - \frac{1}{M} \right) \hat{\sigma}_b^2 + \frac{M}{m} \sum_{i=1}^m N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) \hat{\sigma}_{w_i}^2 \right]$$

Es soll jetzt kurz auch in diesem Fall zunächst die Schätzung des Totalwerts und dann die des Mittelwerts betrachtet werden. Analog zu Gl. 10.16 erhält man mit

$$\hat{X} = \sum_{j=1}^m \frac{M}{m} \frac{N_j}{n_j} X_j \quad \text{mit} \quad X_j = \sum_{l=1}^{n_j} x_{j,l} \quad (\text{was anders ist als in Gl. 10.16})$$

einen erwartungstreuen Schätzer für den Totalwert (Merkmalssumme) der GG, der wie folgt

definiert ist  $X = \sum_{i=1}^M \sum_{k=1}^{N_i} x_{ik} = \sum_{i=1}^M N_i \mu_i = N\mu$  da  $\mu = X/N$  erhält man auch mit

$$\bar{X} = \hat{\mu} = \hat{X} / N = \frac{1}{N} \frac{M}{m} \sum_{j=1}^m \frac{N_j}{n_j} \sum_{l=1}^{n_j} x_{j,l} \quad \text{einen erwartungstreuen Schätzer für } \mu.$$

Bei gleich großen Klumpen vereinfacht sich  $\hat{\mu}$  wie folgt

$$\bar{X} = \hat{\mu} = \frac{1}{m} \sum_{j=1}^m \bar{X}_j \quad \text{mit} \quad \bar{X}_j = \frac{1}{n} \sum x_{j,l}$$

für die Varianz des Totalwerts erhält man

$$(10.25) \quad V(\hat{X}) = \frac{M^2}{M-1} \left( \frac{M}{m} - 1 \right) \sigma_b^2 + \frac{M}{m} \sum_{i=1}^M \frac{N_i^2}{N_i - 1} \left( \frac{N_i}{n_i} - 1 \right) \sigma_{w_i}^2$$

mit  $\sigma_b^2 = \frac{1}{M} \sum (X_i - \bar{X})^2$  wobei  $X_i$  die Merkmalssumme des  $i$ -ten Klumpens und  $\bar{X}$  die durchschnittliche Merkmalssumme ist  $\left( \bar{X} = \frac{1}{M} \sum X_i = \frac{1}{M} \sum N_i \mu_i \neq \mu = \frac{1}{N} \sum N_i \mu_i \right)$  und

$$\sigma_{w_i}^2 = \frac{1}{N_i} \sum_{k=1}^{N_i} (x_{ik} - \mu_i)^2.$$

$\sigma_b^2$  ist wieder die Varianz zwischen /between) den Primäreinheiten (Klumpen) und  $\sigma_{w_i}^2$  sind die  $M$  Varianzen jeweils innerhalb (within) der  $i$ -ten Primäreinheit (also die Varianzen zwischen den Sekundäreinheiten). Auf die Aufgabe des Varianzschätzers soll hier verzichtet werden. Die eingeführten Begriffe werden anhand eines Beispiels demonstriert.

Da für den Schätzer  $\bar{X}$  des Mittelwerts  $\mu$  gilt  $\bar{X} = \frac{1}{N} \hat{X}$  ist

$$(10.26) \quad V(\bar{X}) = \frac{1}{N^2} \left[ \frac{M^2}{M-1} \left( \frac{M-m}{m} \right) \sigma_b^2 + \frac{M}{m} \sum_{i=1}^m \frac{N_i^2}{N_i - 1} \frac{N_i - n_i}{n_i} \sigma_{w_i}^2 \right]$$

In vielen Lehrbüchern findet sich auch die folgende Formel

$$V(\bar{X}) = \frac{1}{N^2} \left[ M^2 \left( \frac{1}{m} - \frac{1}{M} \right) \sigma_b^{*2} + \frac{M}{m} \sum_{i=1}^m N_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) \sigma_{wi}^{*2} \right]$$

$$V(\bar{X}) = \frac{1}{N^2} \left[ \frac{M^2}{M} \frac{M-m}{m} \sigma_b^{*2} + \frac{M}{m} \sum_{i=1}^m \frac{N_i^2}{N_i} \frac{N_i - n_i}{n_i} \sigma_{wi}^{*2} \right]$$

Es besteht kein Unterschied zur Gl. 10.30 weil die Varianzen dann wie folgt definiert sind

$$\sigma_b^{*2} = \frac{1}{M-1} \sum (X_i - \bar{X})^2 = \frac{M}{M-1} \sigma_b^2 \quad \text{und} \quad \sigma_{wi}^{*2} = \frac{1}{N_i-1} \sum_{k=1}^{N_i} (x_{ik} - \mu_i)^2 = \frac{N_i}{N_i-1} \sigma_{wi}^2$$

Beispiel 10.5

Fortführung von Bsp. 10.3 (Variante b, abgewandelt). Die GG besteht aus M=3 Klumpen mit jeweils  $N_i = \bar{N} = 3$  Elementen wie folgt (5, 10, 15) (6, 11, 16) und (7, 12, 17). Angenommen es werden m= 2 Klumpen ausgewählt und innerhalb des Klumpens jeweils 2 Elemente. Man bestimme nun die Stichprobenverteilung des Mittelwerts.

Lösung 10.5

Es gibt jetzt  $\binom{3}{2} = 3$  mögliche Stichproben auf der ersten Stufe (Auswahl von Klumpen) und dabei jeweils  $\binom{3}{2} \binom{3}{2} = 9$  Stichproben auf der zweiten Stufe (Auswahl der Elemente aus den Klumpen), zusammen also 27 gleichwahrscheinliche Stichproben. Die Merkmalssumme der GG beträgt 99 (weil  $\mu=11$ ). Für die ersten sechs Stichproben der Stichprobenverteilung gilt wenn die Klumpen: 1 und 2 ausgewählt sind:

Elemente aus Klumpen		Schätzung der Merkmalssumme des Klumpens			Schätzung des Mittelwerts
1	2	1	2	insges.	
5,10	6,11	3/2 · 15 = 22,5	3/2 · 17 = 25,5	3/2 · 48 = 72	72/9 = 8
	6,16	22,5	33	83,25	9,25
	11,6	22,5	40,5	94,5	10,5
5,15	6,11	30	25,5	83,25	9,25
	6,16	30	33	94,5	10,5
	11,16	30	40,5	105,75	11,75

Mit Berücksichtigung der übrigen 21 Stichproben erhält man als Ergebnis die Stichprobenverteilung des Totalwerts  $\hat{X}$  bzw. des arithmetischen Mittels  $\bar{X}$ .

$\hat{X}$	$\bar{X}$	P(...)	$\hat{X}$	$\bar{X}$	P(...)
72	8	1/27	103,5	11,5	3/27
76,5	8,5	1/27	105,75	11,75	2/27
81	9	1/27	110,25	12,25	2/27
83,25	9,25	2/27	114,75	12,75	2/27
87,75	9,75	2/27	117	13	1/27
92,25	10,25	2/27	121,5	13,5	1/27
94,5	10,5	3/27	126	14	1/27
99	11	3/27			

Wie man sieht ist die Stichprobenverteilung symmetrisch und es gilt  $E(\hat{X}) = 99$  sowie  $E(\bar{X}) = 11 = \mu$ . Für die Varianzen erhält man  $V(\hat{X}) = 182,25$  und  $V(\bar{X}) = 2,25$

Für die Varianzen  $\sigma_b^2$  und  $\sigma_{wi}^2$  erhält man in diesem Beispiel

$$\sigma_b^2 = \frac{1}{3}[(30-33)^2 + (33-33)^2 + (36-33)^2] = 6 \text{ weil } M=3 \text{ und } \bar{X}=33=1/3(30+33+33)$$

i	$\sigma_{wi}^2$
1	$\frac{1}{3}[(5-10)^2 + (10-10)^2 + (15-10)^2] = \frac{50}{3}$
2	$\frac{1}{3}[(6-11)^2 + (11-11)^2 + (16-11)^2] = \frac{50}{3}$
3	50/3

so dass man erhält für die Varianzkomponente auf der ersten Stufe (Auswahl von Klumpen)

$$\frac{M^2}{M-1} \left( \frac{M}{m} - 1 \right) \sigma_b^2 = \frac{9}{2} \left( \frac{3}{2} - 1 \right) 6 = 13,5 \text{ und für die Varianzkomponente auf der zweiten Stufe}$$

$$\text{(Auswahl von Einheiten innerhalb der Klumpen)} \quad \frac{M}{m} \sum \frac{N_i^2}{N_i - 1} \left( \frac{N_i}{n_i} - 1 \right) \sigma_{wi}^2 = 168,75 \text{ erhält}$$

## 4. Weitere Stichprobenpläne

### a) Ungleiche Auswahlwahrscheinlichkeiten: PPS-Verfahren

Prinzip: Berücksichtigung der Größe (des Merkmalsbetrags  $x_i$ ) der Einheiten ( $i = 1, 2, \dots, N$ ) einer endlichen GG bei Auswahl und Hochrechnung (Auswahl mit der Wahrscheinlichkeit  $w_i$  proportional zur Größe  $x_i$  der Einheit  $i$  [probability proportional to size PPS] statt mit  $w_i = 1/N$  für alle  $i$  bei einfacher Stichprobe). Die Schätzfunktion für  $\mu$  lautet mit  $w_i$  gem. Gl. 10.28:

$$(10.27) \quad \bar{X} = \frac{1}{n} \sum_{j=1}^n \frac{x_j}{Nw_j} \quad (j=1, 2, \dots, n) \text{ mit}$$

$$(10.28) \quad w_i = \frac{x_i}{\sum x_i} = \frac{x_i}{N\mu} \quad (i=1, 2, \dots, N)$$

im Unterschied zur einfachen Stichprobe mit  $\bar{x} = \frac{1}{n} \sum x_j$

Sie erlaubt eine exakte „Schätzung“ von  $\mu$  (sogar mit  $n=1$ ) wenn  $w_i = p_i$ , also  $w_i$  identisch ist mit dem Anteil von  $x_i$  am Gesamtmerkmalsbetrag  $\sum x_i$ . Im allgemeinen sind die Größen  $x_i$  jedoch nicht bekannt und damit auch nicht die Anteile  $p_i$ . Meist wird  $w_i$  mit einem anderen Merkmal  $Y$  (etwa  $Y$ : Fläche bei der Schätzung von  $X$ : Ernteertrag) geschätzt als

$$(10.29) \quad w_i^* = \frac{y_i}{\sum y_i} \text{ und } w_i^* \text{ ist (anders als } w_i) \text{ mit dem Anteil } p_i \text{ nicht identisch.}$$

$$\text{Es gilt bei Stichproben ZmZ } E(\bar{X}) = \frac{1}{n} \sum_{j=1}^n E\left(\frac{X_j}{Nw_j}\right) = \frac{1}{nN} \sum_{j=1}^n \sum_{i=1}^N \frac{X_i}{w_i} p_i \text{ und}$$

$$(10.30) \quad \sigma_{\bar{x}}^2 = V(\bar{X}) = V\left(\frac{1}{n} \sum_{j=1}^n \frac{X_j}{Nw_j}\right) = \left(\frac{1}{n}\right)^2 \sum_{j=1}^n \sum_{i=1}^N p_i \left(\frac{X_i}{Nw_i} - \mu\right)^2 = \frac{1}{n} \left[ \frac{1}{N^2} \sum_{i=1}^N \frac{X_i^2}{w_i^2} p_i - \mu^2 \right].$$

Für  $w_i = 1/N$  erhält man die bekannten Ergebnisse für die einfache Stichprobe  $E(\bar{X}) = \mu$  und  $V(\bar{X}) = \sigma^2 / n$ .  $V(\bar{X})$  nach Gleichung 10.30 ist erwartungstreu zu schätzen mit:

$$\hat{\sigma}_{\bar{x}}^2 = \frac{1}{n} \sum_{j=1}^n \left( \frac{X_j}{Nw_j} - \bar{x} \right)^2 w_j [1 - (n-1)w_j].$$

Die Differenz zwischen der Varianz der Stichprobenverteilung von  $\bar{X}$  bei einfacher Stich-

probe und bei PPS ist somit  $\Delta V = \frac{1}{n} \left[ \sum_{i=1}^N \left( x_i^2 - \frac{x_i^2}{N^2 w_i^2} \right) p_i \right]$

Mit  $w_i = 1/N$  gibt es keinen Genauigkeitsgewinn ( $\Delta V=0$ ) und die Varianz  $V(\bar{X})$  bei PPS ist dann kleiner ( $\Delta V > 0$ ) als bei einfacher Stichprobe, wenn der Ausdruck in den eckigen Klammern positiv ist. Das ist z.B. der Fall, wenn für  $w_i$  Gl. 10.22 gilt, denn dann ist dieser Ausdruck  $\bar{x}^2 - \sum \mu^2 p_i = \bar{x}^2 - \mu^2 = \sigma^2 > 0$ , wobei  $\bar{x}^2$  das zweite und  $\mu$  das erste Anfangsmoment der endlichen GG ist.

## b) Mehrphasige Stichproben

Eine zweiphasige Stichprobe (auch double sampling) genannt liegt vor, wenn aus einer Stichprobe von  $n$  aus  $N$  Elementen der GG (z.B. eine einfache Stichprobe) erneut eine Stichprobe von  $m$  aus (diesen)  $n$  Elementen gezogen wird.<sup>12</sup> Das Verfahren ist i.d.R. nur sinnvoll, wenn der Vorteil einer Schichtung (nach dem Merkmal  $Y$ ) oder einer gebundenen Hochrechnung (bei Benutzung der Information über  $Y$ ) den Nachteil der Reduktion des Umfangs von  $n$  auf  $m$  überkompensiert und z.B. die Variable  $Y$  (anders als das Untersuchungsmerkmal  $X$ ) leicht aus einer Kartei zu entnehmen ist. Die erste Stichprobe dient der Beschaffung von Informationen über  $Y$ , mit der eine zweite, kleinere ( $m < n$ ) Unterstichprobe gezogen und untersucht werden kann.

<sup>12</sup> Der Begriff zweiphasig wird auch im Sinne von zweistufig benutzt. So z.B. im Buch von E. P. Billeter-Frey und V. Vlach, Grundlagen der statistischen Methodenlehre, UTB Bd. 1163, S. 192 ff.

## Anhang

Terminologische Festlegungen aus Übersicht 8.1 und 8.2 meines Buchs im Oldenbourg Verlag

	direkter Schluß	indirekter Schluß
Name des Intervalls	Schwankungs- <sup>a)</sup> , Prognose- oder Toleranzintervall <sup>b)</sup>	Konfidenz-, Vertrauens- oder Mutungsintervall
$1 - \alpha$	Sicherheitswahrscheinlichkeit oder Prognosewahrscheinlichkeit	Sicherheitswahrscheinl.t oder Vertrauens- oder Konfidenzniveau (oder -grad)
$\alpha$	Irrtumswahrscheinlichkeit oder beim Testen: Signifikanzniveau	

<sup>a)</sup> Der Begriff wird auch benutzt für ein Intervall auf der  $x$ -Achse (meist symmetrisch um  $\mu$ ) der  $N(\mu, \sigma^2)$  - Verteilung statt auf der  $\bar{x}$  - Achse bei der Stichprobenverteilung von  $\bar{x}$ , also der Verteilung  $N(\mu, \sigma^2/n)$ .

<sup>b)</sup> Dieser Begriff wird auch anders gebraucht.

	Grundgesamtheit <sup>a)</sup>	Stichprobe(n) <sup>b)</sup>
allgemeine Terminologie	Parameter $\theta$	Kennzahl (Schätzer) $\hat{\theta}$
Beispiele	Mittelwert $\mu$ (endl. GG) oder Erwartungswert $\mu = E(X)$	$\bar{x}$ (Stichprobenmittelwert)
1. Mittelwert (heterograd)		
2. Anteilswert (homograd)	$\pi$ Zweipunktverteilung $Z(\pi)$	$p$ (Anteil in der Stichprobe)
3. Varianz	$\sigma^2$ (bzw. homogr. $\sigma^2 = \pi(1-\pi)$ )	$\hat{\sigma}^2$ oder $s^2$
4. Mittelwertdifferenz (heterograd)	$\mu_1 - \mu_2$ (Mittel- oder Erwartungswertdifferenz der GG)	$\bar{x}_1 - \bar{x}_2$ (Mittelwerte der Stichpr., Umfänge $n_1$ und $n_2$ )

a) endliche (Umfang  $N$ ) oder unendliche Grundgesamtheit

b) **vor** Ziehung einer Stichprobe einer Zufallsvariable (Schätzfunktion vgl. Def. 7.4)  $\hat{\theta} = f(X_1, X_2, \dots, X_n)$ ; **nach** Ziehung eine Realisation dieser Zufallsvariable (ein Funktionswert der Stichprobenfunktion)  $\hat{\theta} = f(x_1, x_2, \dots, x_n)$  (entsprechend  $\bar{X}$  und  $\bar{x}$  usw.).

Übersicht 8.8:

*Punktschätzung von Mittel- bzw. Anteilswert und Varianz*

a) Mittel- bzw. Anteilswert , Stichprobe Ziehen **mit Zurücklegen (ZmZ)**

	heterograd	homograd
Schätzfunktion für $\mu$ bzw. $\pi$	$\hat{\mu} = \bar{X} = \frac{1}{n} \sum X_i$	$\hat{\pi} = P = \frac{X}{n} \quad (X = \sum X_i)$
Eigenschaften der Schätzfunktion	$E(\bar{X}) = \mu, V(\bar{X}) = \frac{\sigma^2}{n}$ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$	$E(P) = \pi, V(P) = \frac{\pi(1-\pi)}{n}$ $P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

b) Varianz

Schätzfunktion für die Varianz	$\hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$	$\frac{n}{n-1} P Q$ mit $Q = 1 - P$
Eigenschaften der Schätzfunktion	$E(\hat{\sigma}^2) = \sigma^2, \text{plim } \hat{\sigma}^2 = \sigma^2$	Erwartungstreue, Konsistenz wie im heterograden Fall *

c) Stichproben aus endlicher Grundgesamtheit **ohne Zurücklegen (ZoZ)**; Schätzfunktionen  $\bar{X}$  und P wie unter a)

Varianz der Schätzfunktion $\bar{X}$ bzw. P für $\mu$ bzw. $\pi$	$V(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} < \frac{\sigma^2}{n}$	$V(P) = \frac{\pi(1-\pi)}{n} \frac{N-n}{N-1}$
--	--	---

\*  $E(PQ) = \frac{n-1}{n} \pi(1-\pi)$  , so dass die Schätzfunktion (Stichprobenfunktion)  $\frac{n}{n-1} \cdot PQ$  (statt PQ) erwartungstreu ist.

**Klumpenstichprobe**

**Interpretation der externen Varianz  $\sigma_b^2$  bei der Herleitung der Varianz von  $\bar{X}$**

In der Varianz  $V(\bar{X}) = \sigma_x^2 = \frac{\sigma_b^2}{m} \frac{M-m}{M-1}$  erscheint mit  $\sigma_b^2$  die externe Varianz (zwischen den Klumpen, die als Einheiten betrachtet werden)

$$\sigma_b^2 = \frac{1}{N} \sum_{i=1}^M N_i (\mu_i - \mu)^2 = \frac{1}{N} \sum \bar{N} (\mu_i - \mu)^2 = \frac{1}{M} \sum (\mu_i - \mu)^2$$

Wir betrachten nun die Berechnung von  $\mu_i$  aufgrund der Einzelwerte innerhalb des i-ten Klumpens

$$\sigma_b^2 = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{\bar{N}} \sum_{k=1}^{\bar{N}} x_{ik} - \mu \right)^2$$

Multiplikation des Klammerausdrucks mit  $\bar{N}$  und Division durch

$$\bar{N}^2 \text{ liefert } \sigma_b^2 = \frac{1}{M \bar{N}^2} \sum_{i=1}^M \left( \sum_{k=1}^{\bar{N}} x_{ik} - \bar{N} \mu \right)^2$$

da  $\sum_{k=1}^{\bar{N}} \mu = \bar{N} \cdot \mu$  kann man dies auch schreiben als

$$\sigma_b^2 = \frac{1}{MN^2} \sum_{i=1}^M \left( \sum_{k=1}^{\bar{N}} (x_{ik} - \mu) \right)^2$$

Zur Analyse dieses Ausdrucks  $\sum[\sum(\dots)]^2$  betrachten wir das Beispiel  $\bar{N}=3, M=5$ . Man erhält dann

$$\frac{1}{5 \cdot 3} \left\{ \underbrace{[(x_{11} - \mu) + (x_{12} - \mu) + (x_{13} - \mu)]^2}_{\text{erster Klumpen (i = 1)}} + \dots + \underbrace{[(x_{51} - \mu) + (x_{52} - \mu) + (x_{53} - \mu)]^2}_{\text{fünfter Klumpen (i = M = 5)}} \right\}$$

die erste eckige Klammer ergibt mit  $z_{ik} = x_{ik} - \mu$

$$\left[ \sum_{i=1}^{\bar{N}} \right]^2 = z_{11}^2 + z_{12}^2 + z_{13}^2 + 2z_{11}z_{12} + 2z_{11}z_{13} + 2z_{12}z_{13}$$

und ist eine Summe von  $M\bar{N} = 5 \cdot 3$  Gliedern des Typs  $(x_{ik} - \mu)^2 = z_{ik}^2$  und entsprechendes gilt für die übrigen Klumpen. Insgesamt erhält man also

$$\sigma_b^2 = \frac{1}{MN^2} \left[ (z_{11}^2 + z_{12}^2 + z_{13}^2 + \dots + z_{51}^2 + z_{52}^2 + z_{53}^2) \right. \\ \left. + \frac{1}{MN^2} \left[ \underbrace{2z_{11}z_{12} + \dots + 2z_{12}z_{13}}_{\binom{3}{2} = \binom{\bar{N}}{2} = \frac{\bar{N}(\bar{N}-1)}{2} \text{ Glieder}} + \dots + \underbrace{2z_{51}z_{52} + 2z_{51}z_{53} + 2z_{52}z_{53}}_{\binom{3}{2} = 3 \text{ Glieder des 5ten Klumpens}} \right] \right]$$

oder einzeln ausgeschrieben (statt  $2z_{11}z_{12}$  geschrieben  $z_{11}z_{12} + z_{12}z_{11}$  usw.)

Der Gesamtausdruck  $(z_{11}^2 + \dots + z_{53}^2)$  der *ersten Zeile* ist die N-fache Gesamtvarianz  $\sigma^2$ , so dass gilt  $\frac{1}{MN^2} (\dots) = \frac{\sigma^2}{N}$  weil  $N = M \cdot \bar{N}$ .

Die *zweite Zeile* besteht aus  $M \cdot \bar{N}(\bar{N} - 1)$  Glieder der Art  $z_{ik}z_{il} = (x_{ik} - \mu)(x_{il} - \mu)$  und das Mittel über diese Produkte  $z_{ik}z_{il}$  ist  $\frac{1}{MN} \cdot \frac{1}{(\bar{N} - 1)} \sum_i \left[ \sum_k \sum_l (x_{ik} - \mu)(x_{il} - \mu) \right]$  die durchschnittliche Kovarianz der Beobachtungen innerhalb der Klumpen

$\frac{1}{M} \left[ \frac{1}{\bar{N}(\bar{N} - 1)} \sum_k \sum_l (x_{ik} - \mu)(x_{il} - \mu) \right] = \frac{1}{M} (\sigma_{kl,1} + \dots + \sigma_{kl,M}) = \sigma_{kl}$  und ein Maß für die Unterschiedlichkeit der Werte innerhalb der Klumpen.