Wie groß muss meine Stichprobe sein, damit sie repräsentativ ist? Wie viele Einheiten müssen befragt werden? Was heißt "Repräsentativität"?

von

Peter von der Lippe (Februar 2011)

Stichworte:

Stichprobenumfang, Stichprobenplanung, Repräsentativität, geschichtete Stichprobe, Stichprobenfehler, Zufallsauswahl, Antwortausfälle bei Stichproben

Inhalt

Wie groß muss meine Stichprobe sein, damit sie repräsentativ ist?
Wie viele Einheiten müssen befragt werden?
Was heißt "Repräsentativität?

von

Prof. Dr. Peter von der Lippe

Einführung

Die im Thema genannten Fragen gehören zu den Fragen, die einem als Statistiker wohl am häufigsten gestellt werden. Gleichwohl ist es nicht immer befriedigend, was man dazu sagen kann, oder was man hierzu in Lehrbüchern finden kann. Es ist deshalb sicher nicht unnütz, sich darüber Gedanken zu machen, was man evtl. dem Praktiker[1] raten kann, wenn er eine Stichprobe plant und auf welche Punkte man ihn aufmerksam machen sollte, die besonders zu bedenken sind. Die folgende Darstellung will versuchen, in diesem Sinne mit ein paar praktischen Tipps für die Stichprobenplanung weiter zu helfen.[2]

Wir beginnen im Abschn. 1 mit Formeln, die in vielen Lehrbüchern stehen, und die deshalb in der Praxis sozusagen die erste Wahl sind, insbesondere dann, wenn über die Verteilung der relevanten Merkmale (bzw. des besonders relevanten Merkmals X) in der Grundgesamtheit (insbesondere über die Varianz $\sigma_x^2 = \sigma^2$) wenig bekannt ist. Wir zeigen dann aber (Abschn. 2) einige interessante und geradezu paradox anmutende Implikationen der einfachen Formeln der Lehrbücher, wenn man sie auf eine *geschichtete* Stichprobe anwendet, also mit ihnen den notwendigen Stichprobenumfang getrennt für jede Schicht abschätzen möchte. Es zeigt sich allerdings, dass solche Implikationen wie z.B. die, dass der Auswahlsatz umso größer sein sollte, je kleiner die betreffende Schicht ist,[3] oder dass sich der Stichprobenumfang kaum ändert, wenn die Schicht, aus der man die Stichprobe zieht 2.000 oder 20.000 Einheiten enthält, durchaus konsequent und keineswegs überraschend sind, wenn man eine Formel, die zunächst nur für den Fall einer einfachen (nicht geschichteten) Stichprobe gedacht ist auf den Fall der geschichteten Stichprobe anwendet.

Es wäre dann konsequent, zu fordern, stattdessen die Formeln für die geschichtete Stichprobe anzuwenden, bei deren Anwendung dann auch die "paradoxen" Implikationen nicht auftreten. Die lehrbuchmäßige Darstellung der geschichtete Stichprobe (die hier in Abschn. 3 kurz resümiert wird) ist aber nur bedingt hilfreich; denn es zeigt sich, dass

- diese Formeln eine Planung des Stichprobenumfangs n nur dann verbessern können, wenn man über zusätzliche Kenntnisse, bzw. plausible Annahmen hinsichtlich der Streuung der Variable x verfügt (insbesondere auch was die Varianzen innerhalb der K Schichten σ_1^2, σ_2^2,..., σ_K^2 betrifft),
- die Theorie der geschichteten Stichprobe eine andere Fragestellung behandelt: es geht hier nicht um die Güte der Schätzung der Mittelwerte μ_k *von einzelnen Schichten* (k = 1, ..., K) als solche (für sich genommen) aufgrund der Stichprobenschätzwerte $\bar{x}_k$, son-

[1] Nur wegen der gebotenen Kürze verzichte ich darauf, die Praktikerin zusätzlich zum Praktiker zu nennen. Eine geschlechtsneutrale Bezeichnung, etwa die "Praktizierenden" (analog zu den "Studierenden") schien mir auch nicht sonderlich hilfreich zu sein.

[2] Der Text ist zum großen Teil aus einer längeren internen Notiz für Herrn Dr. Dominik Graf von Stillfried, dem Leiter des Zentralinstituts für die kassenärztliche Versorgung in der Bundesrepublik Deutschland entstanden.

[3] und umgekehrt sollte der Auswahlsatz klein sein bei einer großen Schicht.

dern um die *Schätzung des Gesamtmittelwerts* $\bar{x}$ als Mittel aus allen diesen Schichtmittelwerten $\bar{x}_1, \bar{x}_2, ..., \bar{x}_K$.[4]

In Abschn. 3 werden darüber hinaus aber auch noch einige weitere interessante Zusammenhänge betrachtet, die für geschichtete Stichproben gelten. Andere, z.T. erheblich kompliziertere Stichprobendesigns, wie Klumpenstichprobe oder mehrstufige Stichproben werden in diesem recht elementaren Text allerdings nicht behandelt.

In Abschn. 4 wird auf den Begriff der *Repräsentativität* (ein Ausdruck der Alltagssprache, nicht der Statistik) und im Abschnitt 5 abschließend auf einige Probleme im Zusammenhang mit Fehlern, Antwortausfällen und Hochrechnungen (zur nachträglichen Anpassungen der Struktur der Stichprobe an die Struktur der Grundgesamtheit) eingegangen.

Der Begriff "Repräsentativität" (ein ausgesprochen nicht-statistischer Sprachgebrauch) ist nicht nur unklar, sondern er verführt auch zu offensichtlich unhaltbaren und rein gefühlsmäßigen Schlüssen nach der Art: "nur 100 Befragte kann nicht repräsentativ sein". Statt von "Repräsentativität" zu sprechen ist es in der Fachterminologie üblich (und, wie gezeigt, auch allein sinnvoll), das Konzept des "Stichprobenfehlers" (SF) bei der Beurteilung des Stichprobenumfangs n ins Spiel zu bringen . Wir beginnen deshalb nm Abschn. 1 mit Formeln, in denen n aus SF abgeleitet ist.[5] Dabei soll zunächst noch der problematische Begriff der "Repräsentativität" (bis Abschnitt 4) zurückgestellt werden.

1. Die Lehrbuchformeln für eine einfache Stichprobe

Als Formel für den mindestens erforderlichen Stichprobenumfang n (etwa bei einer geplanten Befragung) findet man meist[6]

$$(1) \qquad n \geq \frac{z^2\,\sigma^2}{e^2} = \left(\frac{z\sigma}{e}\right)^2,$$

worin z Ausdruck der Sicherheit, wie z.B. 90% oder 95% ist, bzw. der Irrtumswahrscheinlichkeit (das Signifikanzniveau) darstellt, was dann entsprechend 10% oder 5% wäre. Der Betrachtung liegt in der Regel die Normalverteilung zugrunde und die z-Werte sind dann (bei der hier üblichen Annahme eines symmetrischen zweiseitig begrenzten Intervalls):

Sicherheit	z
90%	1,6449
95%	1,96 ≈ 2
99%	2,5758

Die Größe σ^2 stellt die Varianz des interessierenden Merkmals X dar und e (error) die absolute Genauigkeit (etwa e = 100, wenn eine Genauigkeit des zu schätzenden Mittelwerts μ der Grundgesamtheit in Höhe von ± 100 € gewünscht wird). Weil es bei einer Stichprobenbefragung in der Regel eine große Zahl von abgefragten Merkmalen gibt (die im unterschiedlichen Maße streuen), und über die Grundgesamtheit aus der die Stichprobe gezogen werden soll meist nichts bekannt ist, stellen sich gleich zwei Fragen

1. Welches meiner Merkmale soll ich bei σ^2 zu Grunde legen?

---

[4] Das war mir bislang nicht so bewusst und erklärt auch dass man sowohl die paradox erscheinenden Ergebnisse von Abschn. 2 als auch ganz andere Ergebnisse mit den Formeln der geschichteten Stichprobe gleichermaßen als logische Konsequenz der betreffenden (unterschiedlichen) Aufgabenstellungen erhält.

[5] Der erforderliche Stichprobenumfang hängt u.a. von der Streuung des interessierenden Merkmals in der Grundgesamtheit ab. Im Extremfall einer Varianz von Null (lauter gleiche Einheiten) reicht eine Stichprobe von n = 1 aus.

[6] Zu Varianten Formel vgl. Anhang 1.

2. Welchen Zahlenwert kann man für σ^2 ansetzen, wo man doch den wahren Wert in der Grundgesamtheit nicht kennt?

Zu 1 gibt es wohl keine brauchbare Antwort. Man könnte argumentieren, es solle ein besonders "wichtiges" Merkmal sein, oder es solle, um sicher zu gehen, die Varianz desjenigen Merkmals sein, das besonders stark streut. Was Frage 2 betrifft, so wird gern Zuflucht genommen zu einer noch zu besprechenden Formel (Gleichungen 2 und 5) mit der man zur Sicherheit einen meist viel zu großen Wert für n erhält. Besser ist es, wenn man in der angenehmen Situation ist, eine zweite oder dritte Befragung der gleichen Art (also eine Panelbefragung) durchführen zu können, dass man dann (versuchsweise) den Wert der ersten oder zweiten (also jeweils der letzten) Befragungs-"Welle" nehmen kann. Verfügt man nicht über solche Daten kann es – wie im Anhang 1 gezeigt wird – evtl. besser sein, mit Annahmen über den relativen statt dem absoluten Fehler (e) und über den Variationskoeffizient statt über σ (wie in Gl. 1) zu operieren.

In jedem Fall wird man statt der wahren Varianz σ^2 eine geschätzte, angenommene oder aufgrund früherer Erhebungen zu vermutende Varianz $\hat{\sigma}^2$ einsetzen. Die Aussage von Gl. 1 ist intuitiv verständlich. Danach hängt die Größe n ab von

1. der quadrierten Genauigkeit, definiert aufgrund des absoluten Fehlers[7] e (große Genauigkeit heißt kleiner Fehler, so dass die Genauigkeit quasi 1/e ist),
2. der Sicherheit, der jeweils ein bestimmter Wert z zugeordnet ist und
3. der Homogenität der Grundgesamtheit σ^2.

Mit dem "Fehler" e ist hier die halbe Breite des Konfidenzintervalls gemeint. Der Begriff ist jedoch nicht eindeutig. Unter dem "Stichprobenfehler" (SF) wird üblicherweise (und so auch im Folgenden) die Standardabweichung $\sigma_{\bar{x}}$, bzw. die (aufgrund der Stichprobe) geschätzte Standardabweichung $\hat{\sigma}_{\bar{x}}$ der Stichprobenverteilung (Verteilung aller möglichen Stichproben vom Umfang n, die man aus einer Grundgesamtheit vom Umfang N ziehen kann) von $\bar{x}$ verstanden und e ist dann das Produkt $z \cdot \sigma_{\bar{x}}$. Das Konfidenzintervall für den mit $\bar{x}$ geschätzten Mittelwert μ der Grundgesamtheit hat bekanntlich bei einer von der Irrtumswahrscheinlichkeit (= Signifikanzniveau) α abhängigen Signifikanzschranke z_α die Grenzen[8]

$\bar{x} \pm z_\alpha \cdot \hat{\sigma}_{\bar{x}} = \bar{x} \pm e$ und der Stichprobenfehler (SF) ist gegeben mit

(1a) $$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad \text{bzw. bei } N \to \infty$$

(1b) $$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$$, was wie Gl. 1 darauf beruht, dass $e = z\hat{\sigma}_{\bar{x}} = z\frac{\hat{\sigma}}{\sqrt{n}}$ nach n aufgelöst wurde. Es ist unmittelbar einsichtig, dass n umso kleiner sein kann, je homogener die Grundgesamtheit ist. Sind dort alle Elemente bezüglich des Merkmals X gleich, dann ist $\sigma_x^2 = \sigma^2 = 0$ und es reicht, wie bereits gesagt, eine Stichprobe von n = 1 zu ziehen, um die Grundgesamtheit komplett und mit Sicherheit (Wahrscheinlichkeit 1) zu kennen. Gl. (1) zeigt, dass n direkt proportional zu z^2 und σ^2 ist. Eine größere Sicherheit und eine weniger homogene Grundge-

[7] e ist die halbe Länge des (symmetrischen zweiseitigen) Schwankungsintervalls (bei Verwendung von σ^2), bzw. Konfidenzintervalls wenn für σ^2 der entsprechende Schätzwert aufgrund der Stichprobe verwendet wird. Die entsprechenden Formeln (1) bis (4) beruhen alle auf einer einfachen Umformung der Gleichungen für e.

[8] Im Folgenden, wie auch in Gl. 1 schreiben wir der Einfachheit halber nur z statt z_α. Der (absolute) Stichprobenfehler SF bestimmt nicht nur die Breite des Konfidenzintervalls sondern auch die Testentscheidung bei Hypothesen über μ.

samtheit verlangen eine größere Stichprobe. Entsprechend ist n proportional zu 1/e (also indirekt proportional zu e), d.h. mehr Genauigkeit (kleineres e, größeres 1/e) bedeutet größerer Stichprobenumfang.

Das Problem, das mit (1) verbunden ist, ist dass man i.d.R. keine Kenntnisse über die Größenordnung von σ^2 hat (außer $\sigma^2 > 0$), wenn X ein Merkmal ist, das "quantitativ" ist (d.h. metrisch skaliert ist und beliebig viele Abstufungen hat).[9] Angenehmer ist in dieser Hinsicht ein "qualitatives" Merkmal mit Merkmalsausprägungen, für die meist nicht eine Ordnung existiert (z.B. beim Familienstand ist "ledig" nicht mehr oder "besser" als "verheiratet") und bei denen man sich für die Wahrscheinlichkeit π des Auftretens einer bestimmten dieser Ausprägung interessiert. In diesem Fall tritt $\pi(1-\pi)$ an die Stelle von σ^2 und weil für die Wahrscheinlichkeit π gilt $0 \le \pi \le 1$ ist die Varianz hier (im Unterschied zu σ^2) nach oben begrenzt $\pi(1-\pi) \le ¼$. Man kann dann n abschätzen weil jetzt σ^2 durch $\pi(1-\pi)$ zu ersetzen ist

(2) $$n \ge \frac{z^2\,\pi(1-\pi)}{e^2} \quad \text{und} \quad \frac{z^2\pi(1-\pi)}{e^2} \le \frac{1}{4}\cdot\frac{z^2}{e^2}.$$

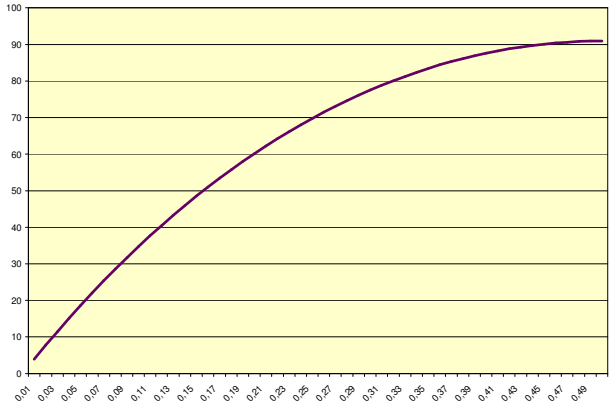
Die meisten Befragungen enthalten neben "quantitativen" Merkmalen auch "qualitative", so dass man sich mit $z^2/4e^2 = (z/2e)^2$ auf der sicheren Seite fühlen kann (vgl. aber auch Anhang 1). Varianten von (1) und (2) sind die Endlichkeit der Grundgesamtheit berücksichtigend[10]

(3) $$n \ge \frac{K}{e^2 + \frac{K}{N}} \quad \text{mit } K = z^2\sigma^2$$

(4) $$n \ge \frac{K^*}{e^2 + \frac{K^*}{N}} \quad \text{mit } K^* = z^2\pi(1-\pi)$$

Der Einfluss von π auf den Stichprobenumfang n nach Gl. 4 ist erheblich.

π (oder 1- π)		π (1- π)	n
0,05	0,95	0,0475	18,65
0,1	0,9	0,09	34,75
0,2	0,8	0,16	60,15
0,3	0,7	0,21	77,49
0,4	0,6	0,25	87,59



In der nebenstehenden Tabelle (und der Graphik darunter) ist mit N = 1000, e = 0,1 und z = 2 gerechnet worden. Wie man sieht, ist der Zusammenhang nicht linear und der Stichprobenumfang kann *erheblich* kleiner sein, wenn sich die relevanten Anteile um 0,1 (0,9) oder 0,2 (0,8) bewegen statt um 0,5. Es ist auch offensichtlich, dass danach der Stichprobenumfang nach Maßgabe der Varianz π (1- π) zunimmt, genauso wie dies auch gem. (1) bei zunehmender Varianz σ^2 der Fall ist. Er dürfte bei dieser Art der Abschätzung von n (nach Gl. 2) überhaupt etwas zu groß sein (vgl. auch Anhang 1).

[9] Man spricht in diesem Fall auch von "heterograd" und bei "qualitativen" (oder "kategorialen") Merkmalen von "homograd".

[10] Es sind die Fälle von Ziehen ohne (statt mit) Zurücklegen (ZoZ statt ZmZ), je nachdem ob eine Endlichkeitskorrektur nicht erforderlich ist (ZmZ) oder erforderlich ist (ZoZ). Mit $N \to \infty$ strebt (3) gegen (1), (4) gegen (2).

Mit $N \to \infty$ erhält man K/e^2 bzw. K^*/e^2 also (1) bzw. (2). Eine beliebte und sehr konservative Abschätzung von n ist mit $\pi(1-\pi) = < ¼$ die Formel

(5) $$n_{max} = \frac{z^2}{4e^2 + z^2/N}$$ bzw. noch einfacher bei einer Sicherheit von etwa 95%, also z = 2

(5a) $$n \geq \frac{N}{1 + e^2 N}.$$

Wir wollen im Folgenden zeigen, was es bedeutet, diese beliebte Formel bei der Planung der Stichprobenumfänge für die einzelnen Schichten einer geschichteten Stichprobe zu benutzen.

2. Planung der Stichprobenumfänge für die Schichten einer geschichteten Stichprobe (mit Formeln für die einfache Stichprobe)

Wendet man für diese Aufgabe Gl. 5a an, so geht man implizit davon aus, dass man über die Varianzen σ_k^2 bzw. $\pi_k(1-\pi_k)$ innerhalb der einzelnen Schichten (z.B. Branchen, Berufsgruppen, Regionen) k = 1, ..., K keine Kenntnisse besitzt und deshalb zur Sicherheit jeweils der maximal mögliche Wert von ¼ (sowie z = 2) angenommen werden könnte bzw. sollte. Es gibt dann jedoch einige sonderbare Konsequenzen.

a) Auswahlsatz indirekt proportional zum Schichtumfang

Dividiert man n in Gl. (5a) durch N (bzw. n_k durch den Schichtumfang N_k)[11] so erhält man

(6) $$\frac{n}{N} \geq \frac{1}{1 + e^2 N}.$$

Tabelle 1 zeigt, was das impliziert, wenn man z = 2 (für eine Sicherheit von ≈ 95%), $\pi_k(1-\pi_k)$ = ¼ und einen absoluten Fehler von e = 0,08 (und damit $e^2 = 0{,}0064$) annimmt, also mit

(6a) $$\frac{n_k}{N_k} = \frac{1}{1 + 0{,}0064 \cdot N_k}$$

rechnet. Man erhält eine hyperbolisch fallende Kurve, d.h. mit wachsendem N_k strebt der Auswahlsatz gegen 0. Bei einer kleinen Schicht ist der Auswahlsatz groß und bei großem N_k ist der Auswahlsatz klein. Der Umfang der Teil-Stichprobe n_k strebt mit wachsendem N_k gem. Gl. 2 wegen z = 2 gegen $1/e^2 = 1/0{,}0064 = 156{,}25$.

Auswahlsatz n_k/N_k in % (und Stichprobenumfang n_k) in Abhängigkeit von N_k

N_k	n_k/N_k in % (bzw. n_k)
10	94
50	76 (n_k = 38)
100	60,9 (n_k = 61)
200	43,8 (n_k = 87,7)
300	34,2 (n_k = 102,7)
400	28,1 (n_k = 112,4)
500	23,8 (n_k = 119)
1000	13,5 (n_k = 135)

N_k	n_k/N_k in % (bzw. n_k)
2000	7,2 (n_k = 145)
3000	4,95 (n_k = 148,5)
5000	3,03 (n_k =151,5)
10.000	1,54 (n_k = 154)
20.000	0,775 (n_k = 155)
30.000	0,518 (n_k = 155,4)
40.000	0,389 (n_k = 155,64)
50.000	0,312 (n_k = 155,76)

[11] Aufteilung der Grundgesamtheit in K Schichten mit den Umfängen N_k (Summation jeweils über k = 1, ..., K) so dass $\Sigma N_k = N_1 + N_2 + ... + N_K = N$. Entsprechend wird der Umfang n der Gesamt-Stichprobe auf K Stichproben wie folgt aufgeteilt $\Sigma n_k = n_1 + n_2 + ... + n_K = n$.

Ab einem N_k von etwa von 3000 (dann n etwa 150) nimmt der absolute Stichprobenumfang nur wenig zu. Ist der Schichtumfang N_k 10.000 statt 1000, so steigt n_k nur von $n_k = 135$ auf $n_k = 154$.[12] Mit N bzw. $N_k \to \infty$ erhält man die Formeln (1) und (2) als Grenzfälle von (3) und (4). Mit e = 0,08, z = 2 und $\pi_k(1-\pi_k) = ¼$ erhält man mit (2), wie gesagt den Wert von 156,25.

Sieht man von der 1 im Nenner von (6a) ab – was jedoch nur vertretbar ist bei einem relativ großen N_k,[13] so kann man sagen, dass der Auswahlsatz n_k/N_k umgekehrt (= "indirekt") proportional zur Schichtgröße N_k der Schicht k in der Grundgesamtheit ist; denn es gilt dann

$$(6b) \qquad \frac{n_k}{N_k} \approx \frac{1}{0{,}0064 \cdot N_k}.$$

Im Zusammenhang mit geschichteten Stichproben ist es sehr viel üblicher, gleiche Auswahlsätze vorzusehen (vgl. später Gl. 12), also $\frac{n_k}{N_k} = c$ (c ist eine Konstante) und denkbar wäre auch $\frac{n_k}{N_k} = c \cdot N_k$, also eine *direkte Proportionalität*, d.h. kleiner (großer) Auswahlsatz bei einer kleinen (großen) Schicht. Stichprobenplanung nach (5a) führt genau zum Gegenteil, nämlich *indirekter Proportionalität,* also (6b).

Es spricht im allgemeinen viel für die Vermutung, dass kleine Schichten eher homogen sind und große Schichten eher heterogen sind, so dass indirekte Proportionalität, also ein kleiner (großer) Auswahlsatz bei einer großen (kleinen) Schicht wenig plausibel zu sein scheint.

b) Gesamtstichprobe umso größer je mehr Schichten unterschieden werden

Eine andere Konsequenz der Anwendung von (6a) ist, dass der Umfang der Stichprobe insgesamt davon abhängig ist, wie stark die Grundgesamtheit hinsichtlich der Anzahl der Schichten differenziert ist. Das kann man mit den Zahlen der obigen Tabelle leicht demonstrieren:

Nimmt man eine undifferenzierte Grundgesamtheit von $N_1 + N_2 = 300$ an, so sind nach der Tabelle etwa 103 Einheiten zu befragen, differenziert man sie aber in $N_1 = 100$ und $N_2 = 200$ so sind entsprechend 61 + 88 = 149 Einheiten zu befragen.

Mit zunehmender Differenzierung wird der Unterschied immer krasser, weil ja bei kleinem N ein großer Auswahlsatz zu nehmen ist. Ebenfalls mit den Werten der Tabelle nachzuvollziehen ist: N = 1000 sei als Summe aus 100, 200, 300 und 400 zusammengesetzt. Zu befragen wären dann ohne Differenzierung n = 135 und mit Differenzierung 61 + 88 + 103 + 112 = 364 statt 135 Einheiten (also ein fast 2,7-faches Befragungsvolumen). Das dürfte schwer zu rechtfertigen sein.

c) Gewollte oder ungewollte Konsequenzen?

Man könnte nun sagen, die Abhängigkeit des Umfangs n der Gesamtstichprobe von der Differenzierung nach Schichten (z.B. Berufsgruppen, Branchen etc.) sei ja gerade gewollt, denn man habe ja gerade deshalb auch eine geschichtete Stichprobe gezogen, um die Unterschiede zwischen den einzelnen Gruppen (Schichten) sauber herauszuarbeiten. So gesehen ist ein größeres n bei stärkerer Differenzierung (größere Anzahl K der Schichten), bzw. ein kleineres n bei geringerem K nicht unplausibel. Eine solche Zielsetzung würde es nahe legen, für einzelne Schichten unterschiedlich große Fehler e_k zu postulieren, aber die Stichprobenplanung mit (5a) sieht genau das ja gerade nicht vor.

---

[12] Auch auffallend ist: verdoppelt man N von 5.000 auf 10.000 und dann wieder auf 20.000 und 40.000 dann erhöht sich der Stichprobenumfang gerade mal um 3, 1 Einheiten oder um noch nicht einmal eine Einheit.

[13] Ab $N_k = 1.406$ unterscheidet sich der Auswahlsatz nach (6a) von dem nach (6b) nur noch um maximal 1%.

Entsprechend könnte man argumentieren, eine praktisch umgekehrte (indirekte) Proportionalität zwischen Auswahlsatz und Schichtgröße sei nicht problematisch, denn es könnte ja sein, dass die größeren Schichten in sich homogener seien als die kleineren, so dass $\pi_k(1-\pi_k)$ bei einem großen N_k kleiner ist als bei einer kleineren Schicht (kleines N_k). Ob das in einem konkreten Fall tatsächlich stimmt, also der kleine (große) Auswahlsatz bei einer großen (kleinen) Schicht richtig oder zumindest unschädlich ist, hängt von den Daten ab und kann nach Ziehung der Stichprobe untersucht werden. Fest steht aber, dass die Bestimmung der Stichprobenumfänge für die K Schichten mit einer Formel (5a) erfolgt, bei der gerade davon ausgegangen wurde, dass $\pi_k(1-\pi_k)$ für alle k = 1, …, K Schichten gleich groß ist (und zwar maximal mit dem Wert von ¼), also kein Zusammenhang zwischen Schichtgröße und Homogenität besteht..

Bevor wir zeigen, welche Formeln die Theorie der geschichteten Stichprobe für die Bestimmung der erforderlichen Stichprobenumfänge n_1, n_2, ..., n_K und des Gesamtstichprobenumfangs $n = n_1 + n_2 + ... + n_K$ (der dann evtl. erheblich kleiner sein kann als bei einer einfachen Stichprobe) "anzubieten" hat (Abschn. 3) zeigen wir, dass die Ergebnisse der Abschnitte 2a und 2b zwar sonderbar erscheinen mögen, aber sehr konsequent aus den für eine ungeschichtete Stichprobe geltenden Zusammenhängen folgen

d) Die paradox erscheinenden Ergebnisse sind gleichwohl konsequent

Die Ergebnisse der Abschnitte a (sehr unterschiedliche Auswahlsätze trotz gleicher Streuung) und b (Gesamtstichprobenumfang n hängt von K ab) wirken paradox. Insbesondere das Ergebnis, dass nach Gl. 6, bzw. 6a mit zunehmendem Umfang der Teilgesamtheiten N_k der Auswahlsatz immer kleiner werden soll und gegen null strebt ist überraschend.[14] Das gilt besonders wenn man nicht immer klar zwischen Auswahlsatz n_k/N_k und dem absoluten Stichprobenumfang n_k unterscheidet, denn n_k strebt nicht gegen null nur n_k/N_k.

Normalerweise würde man erwarten, dass ein kleiner Auswahlsatz von etwa $n_1/N_1 = 0{,}03$ (also 3%) sinnvoll ist bei einer homogenen Gruppe (kleines σ_1) und ein großer Auswahlsatz etwa $n_2/N_2 = 0{,}76$ bei einer entsprechend heterogenen Gruppe (großes σ_2).[15] Diese Betrachtung gilt aber nur bei absolut etwa gleich großen Gruppen $N_1 \approx N_2$. oder aber im Fall "mit Zurücklegen" (ZmZ), d.h. wenn keine Endlichkeitskorrektur vorzunehmen ist (N bzw. $N_k \to \infty$), denn dann gilt nach (2) bei gleichem z und e (also $z_1 = z_2 = z$ und $e_1 = e_2 = e$)

(7) $$\frac{n_1}{n_2} = \frac{\sigma_1^2}{\sigma_2^2}$$, bzw. nach (3) bei gleichem N, also $N_1 = N_2 = N$

(8) $$\frac{n_1}{n_2} = \frac{\sigma_1^2}{\sigma_2^2} \cdot \frac{N + z^2\sigma_2^2}{N + z^2\sigma_1^2} = \frac{\sigma_1^2}{\sigma_2^2} \cdot F\,.$$

Zur Illustration nehme man an $\sigma_1 = 20$ und $\sigma_2 = 10$, also Schicht 1 hat eine doppelt so große Standardabweichung σ wie Schicht, bzw. Stichprobe 2. Aus (7) folgt das bekannte Resultat, dass bei doppelter Standardabweichung $n_2 = 4n_1$ ist, also ein vierfacher Stichprobenumfang erforderlich ist. Bei z = 2 erhält man entsprechend $\frac{n_1}{n_2} = 4 \cdot \frac{N+400}{N+1600}$ nach (8), was z.B. bei N = 1000 nur den Wert $n_1/n_2 = 2{,}154$ und bei N = 4000 dagegen aber $n_1 = 3{,}143 \cdot n_2$ ergibt[16] und

[14] Aus (6a) ergibt sich, dass ab $N_k = 99/e^2 = 99/0{,}0064 \approx 15.469$ der Auswahlsatz 1% oder kleiner ist.

[15] Bei einer großen Gruppe N_1 ein kleiner und bei einer kleinen Gruppe N_2 ein großer Auswahlsatz wirkt sehr sonderbar, zumal für beide Gruppen eine gleich große Streuung $\pi_k(1-\pi_k) = ¼$ angenommen wurde.

[16] Der Faktor F strebt mit wachsendem N gegen 1.

folglich (bei angenommenem gleichen N) auch ein 2,15- bzw. 3,14-fachen Auswahlsatz bedeutet.

So gesehen erscheint es reichlich paradox, wenn man gem. Abschn. 2a je nach Schichtgröße N_k sehr unterschiedliche Auswahlsätze (die noch dazu gegen null streben) erhält. Das sehr sonderbar erscheinende Ergebnis nur 3% bei $N_1 = 5000$ aber 76% bei $N_2 = 50$ trotz gleicher (maximaler) Streuung ist aber gleichwohl konsequent aus der Sicht der Theorie der einfachen (nicht geschichteten) Stichprobe.

Betrachtet man nämlich die Gl. 3 zugrundeliegende Formel für den absoluten Fehler e des Mittelwerts $\bar{x}$ so erhält man durch Umformung von (3) die Gleichung

$$(9) \qquad e = \sqrt{\frac{K}{n}\left(1-\frac{n}{N}\right)} = z\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N}} = z\hat{\sigma}_{\bar{x}}\,.$$

Setzt man $z_1 = z_2 = z$ so erhält man hieraus als Bedingung für $e_1 = e_2 = e$

$$(9a) \qquad \hat{\sigma}_{\bar{x}_1} = \frac{\hat{\sigma}_1}{\sqrt{n_1}}\sqrt{1-\frac{n_1}{N_1}} = \hat{\sigma}_{\bar{x}_2} = \frac{\hat{\sigma}_2}{\sqrt{n_2}}\sqrt{1-\frac{n_2}{N_2}}\,.$$

Gleichsetzen von $\hat{\sigma}^2_{\bar{x}_1}$ und $\hat{\sigma}^2_{\bar{x}_2}$ bei Geltung von $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\pi}(1-\hat{\pi})$ liefert mit $f_k = n_k/N_k$ für den Auswahlsatz

$$(9b) \qquad \frac{1-f_1}{f_1} = \frac{N_1}{N_2}\cdot\frac{1-f_2}{f_2},$$

was sich leicht verifizieren lässt mit den wiederholt genannten Zahlen $N_1 = 50$, $N_2 = 5000$ sowie $f_1 = 0{,}757575$ und $f_2 = 0{,}030303$.[17] Man erhält dann auf der linken und rechten Seite von (9b) den Wert 0,32. Damit folgt der Umstand, dass der Auswahlsatz f_k praktisch indirekt proportional zur Schichtgröße N_k ist einfach daraus, dass die Stichprobenfehler $\hat{\sigma}^2_{\bar{x}_k}$ gleich gesetzt worden sind.

Dass diese Planung des Stichprobenumfangs mit Gl. (5) und (6) darauf hinausläuft, dass man für verschiedene Stichproben einen gleich großen Fehler postuliert, kann man auch wie folgt zeigen: Es ist leicht zu sehen, dass man für $\Phi_k = \left(1-\frac{n_k}{N_k}\right)\Big/ n_k$ in Gl. (9) den Wert e_k^2 erhält, wenn man die Größen n_k/N_k und n_k gem. Gl. (6) und (5a) bestimmt, denn dann ist

$\Phi_k = \left(1-\frac{n_k}{N_k}\right)\Big/ n_k = \left(1-\frac{1}{1+e_k^2N_k}\right)\Big/\left(\frac{N_k}{1+e_k^2N_k}\right) = e_k^2$. Es gilt also

$\frac{\sigma_1}{\sqrt{n_1}}\sqrt{1-\frac{n_1}{N_1}} = \sigma_1\sqrt{\Phi_1} = \sigma_1 e_1$ und entsprechend $\frac{\sigma_2}{\sqrt{n_2}}\sqrt{1-\frac{n_2}{N_2}} = \sigma_2 e_2$.

Die Gleichheit von e_1 und e_2 nach (9) führt also zudem genannten Zusammenhang für die Auswahlsätze $f_1 = n_1/N_1$ und $f_2 = n_2/N_2$.[18]

Das paradox erscheinende Ergebnis ist also nichts anderes als eine Konsequenz der Annahme gleicher Varianzen innerhalb der Schichten $\sigma_1^2 = \sigma_2^2 = ... = \sigma_K^2$, die hier in Unkenntnis besserer

[17] Man beachte, dass oben in der Tabelle mit 76% und 3,03% nur gerundete Werte angegeben wurden. Die oben angegebenen genaueren Zahlen erhält man aus Gl. 6a, indem man für N 50 bzw. 5000 einsetzt.

[18] Man erhält in diesem konkreten Fall die Gleichung $e_1 = z_1\sigma_1 e_1 = e_2 = z_2\sigma_2 e_2$, die bei Herleitung von (5) und (6) davon ausgegangen wurde, dass $z_1 = z_2 = 2$ und $\sigma_1 = \sigma_2 = (1/4)^{1/2} = 0{,}5$ ist, so dass gilt $z_1\sigma_1 = z_2\sigma_1 = 1$.

Erkenntnisse und aus Vorsichtsgründen mit dem maximalem Wert ¼ angesetzt wurden. Dies führt dazu, dass dann auch die Stichprobenfehler (und damit die Breite der Konfidenzintervalle) der Schätzwerte[19] $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_N$ (heterograd) bzw.[20] $\hat{\pi}_1$, $\hat{\pi}_2$, ... (homograd[21]) absolut gleich groß sind. Das spricht an sich dafür, dass eine Stichprobenplanung nach Art dieses Abschnitts ganz vernünftig zu sein scheint, angesichts

a) der Ausgangssituation (keine Kenntnisse über unterschiedliche Varianzen und über die Größenordnung der Varianzen überhaupt)[22] und

b) des Ergebnisses, dass die K Stichproben, jeweils (jede für sich) einen gleich großen absoluten Fehler (und damit ein gleich breites Konfidenzintervall) für π haben.[23]

Gleichwohl ist eine Stichprobenplanung, die davon ausgeht, dass man für jede Teilstichprobe vom Umfang n_k aus einer Schicht N_k einfach die Formeln für den Abschätzung des Stichprobenumfangs n einer "einfachen" (d.h. ungeschichteten) Stichprobe übernimmt nicht voll befriedigend. Die Implikationen dieser Vorgehensweise, wie sie in Abschn. 2a und 2b dargestellt sind, erscheinen geradezu kontraintuitiv, obgleich sie - wie gerade gezeigt - in gewisser Weise durchaus "richtig" sind.

Im Folgenden wird gezeigt, dass man zu ganz anderen Ergebnissen gelangt, wenn man Formeln der Theorie der geschichteten (statt der "einfachen", also ungeschichteten) Stichprobe anwendet. Insbesondere vermeidet man damit gerade die kontraintuitiven Ergebnisse (trotz gleicher Varianz unterschiedliche Stichprobenumfänge und Auswahlsätze) der Vorgehensweise dieses Abschnitts.

Mit den speziellen Lehrbuch-Formeln für geschichtete Stichprobe, die im folgenden Abschnitt behandelt werden, erhält man insbesondere das Ergebnis, dass man, wenn man keine Kenntnisse über die Varianzen relevanter Merkmale in den Schichten hat mit gleichen Auswahlsätzen operieren sollte, statt bei kleinen Schichten große Auswahlsätze und bei großen Schichten kleine Auswahlsätze vorzusehen.

3. Planung der Stichprobenumfänge für die Schichten mit den Formeln für die geschichtete Stichprobe

Dieser Abschnitt stellt in den Teilen a bis d weitgehend bekanntes Lehrbuchwissen dar (entnommen aus meinem Lehrbuch im Oldenbourg Verlag). In Abschn. e vergleichen wir die Bestimmung der Stichprobenumfänge $n_1, n_2, \ldots, n_K$ und auch $n = n_1 + n_2 + \ldots + n_K$ nach den Formeln für geschichtete Stichproben mit denen einer Stichprobenplanung nach Art von Abschnitt 1. Zwei Zahlenbeispiele mögen helfen, unsere Aussagen zu veranschaulichen.

---

[19] Für die Mittelwerte $\mu_1, \mu_2, ..., \mu_N$ der Schichten (in der Grundgesamtheit).

[20] Für die Anteile (Wahrscheinlichkeiten) $\pi_1, \pi_2, ..., \pi_N$.

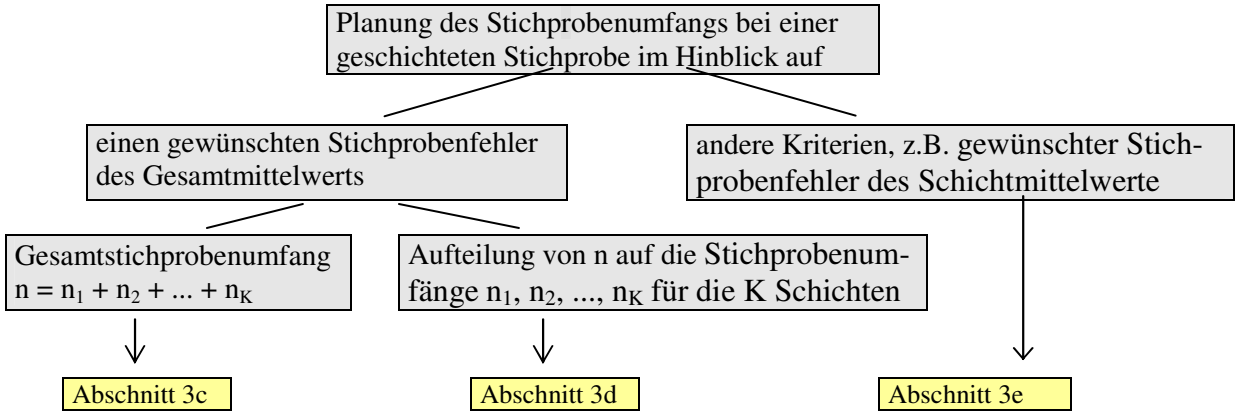
[21] wie in der Vorgehensweise nach Gl. (6).

[22] Die Bestimmung der Stichprobenumfänge bzw. Auswahlsätze gem. (5) und (6) geht also in der Tat von der Annahme gleicher und (für den homograden Fall) maximaler, weil unbekannter Varianzen aus. Dem steht natürlich nicht entgegen, dass man bei Kenntnis der unterschiedlichen Größe von Varianzen innerhalb der K Schichten zu besseren Ergebnissen gelangen kann.

[23] Man kann auch als eine Art, den Stichprobenumfänge $n_1, n_2, ...$ zu bestimmen (alternativ zur Auflösung nach n der Formel für den Stichprobenfehler e der einen Variable x), von folgender Aufgabenstellung ausgehen: Stichprobenumfänge so planen, dass eine vorgegebene (meist gleich große) Genauigkeit in bestimmten Teilgesamtheiten (z.B. Bundesländern) eingehalten werden kann (das ist v.a. in der amtlichen Statistik ein wichtiges Kriterium zur Bestimmung von Stichprobenumfängen). Wie man sieht, bedeutet das obige Ergebnis, dass (quasi als Nebenprodukt) auch dieses Kriterium erfüllt wird. Auch bei solchen amtlichen Erhebungen wird ja ein größerer Auswahlsatz für ein Bundesland, wie Bremen und ein kleinerer Auswahlsatz für Nordrhein Westfalen geplant.

a) Übersicht

Die unterschiedlichen Aussagen über erforderliche Stichprobenumfänge in den Abschnitten 2 und 3 ergeben sich daraus, dass den beiden Abschnitten unterschiedliche Fragestellungen zugrunde liegen. Mit der folgenden Übersicht wird versucht, die Aufgabenstellungen (und damit auch die Gliederung) des Abschnitts 3 übersichtlich darzustellen:



Eine der im Abschnitt 2 verfolgten Aufgabenstellung entsprechende Frage wird erst im Abschn. 3e behandelt. Darüber hinaus werden einige allgemeine Beziehungen zwischen einer einfachen und einer geschichteten Stichprobe (z.B. die Frage, unter welchen Voraussetzungen man mit letzterer einen "Schichtungsgewinn" erzielen kann) im Anhang behandeln.

b) Grundlegende Konzepte und Zusammenhänge bei geschichteten Stichproben

Mit den Definitionen der Größen N_k und n_k (vgl. Fußnote 11) erhält man für den Gesamtmittelwert der Grundgesamtheit [24]

$$(10) \qquad \mu = \frac{N_1}{N}\cdot\mu_1 + \ldots + \frac{N_K}{N}\cdot\mu_K = \sum_k \frac{N_k}{N}\cdot\mu_k$$

und den Gesamtmittelwert $\bar{x}$ der Stichprobe (er dient der Schätzung von μ, man könnte also auch $\hat{\mu}$ statt $\bar{x}$ schreiben)

$$(11) \qquad \bar{x} = \frac{N_1}{N}\cdot\bar{x}_1 + \ldots + \frac{N_K}{N}\cdot\bar{x}_K = \sum_k \frac{N_k}{N}\cdot\bar{x}_k \,.$$

Dieser Schätzer ist *vor* Ziehung einer Stichprobe eine Zufallsvariable (angedeutet durch große Buchstaben $\bar{X}$), oder Stichprobenfunktion,[25] danach eine Realisation dieser Zufallsvariable, bzw. ein konkreter Funktionswert (kleine Buchstaben $\bar{x}$) und erwartungstreu, denn

$$E(\bar{X}) = \sum \frac{N_k}{N} E(\bar{X}_k) = \mu \,,$$

d.h. ein konkretes $\bar{x}$ ist nicht (oder allenfalls rein zufällig) gleich dem wahren Wert μ, wohl aber ist $\bar{x}$ "im Mittel" gleich μ, und das schon bei kleinen Stichproben, also nicht erst bei $n \to \infty$ wie bei asymptotischer Erwartungstreue. Die Funktion (11) ist zu unterscheiden von

$$(11a) \qquad \bar{x}^* = \frac{n_1}{n}\cdot\bar{x}_1 + \ldots + \frac{n_K}{n}\cdot\bar{x}_K = \sum_k \frac{n_k}{n}\cdot\bar{x}_k \,.$$

---

[24] Die Summation erfolgt im Folgenden, sofern nichts anderes angegeben ist, stets über die K Schichten.

[25] Synonyme: Schätzfunktion oder Schätzer.

Bei gleichen Auswahlsätzen $\frac{n_1}{N_1} = \frac{n_2}{N_2} = ... = \frac{n_k}{N_k} = \frac{\sum_k n_k}{\sum_k N_k} = \frac{n}{N}$ oder (gleichbedeutend)

$$(12) \qquad \frac{n_k}{n} = \frac{N_k}{N} = \frac{n}{N} = c$$

für alle k = 1, 2, ..., K (was man *proportionale Aufteilung* [von n in n_1, n_2, ...,n_K] nennt, denn für die Stichprobe gelten die gleichen Proportionen $n_1:n_2:\ldots,:n_K$ wie für die Grundgesamtheit $N_1:N_2:\ldots:N_K$), und nur dann gilt $\bar{x}^* = \bar{x}$. In der Regel sind die beiden Schätzer für μ aber unterschiedlich und $\bar{x}^*$ ist im Unterschiede zu $\bar{x}$ nicht erwartungstreu.

Interessanter für die Beurteilung der Güte (quasi "Repräsentativität") einer Stichprobe als der Erwartungswert ist die Varianz eines Schätzers wie $\bar{X}$. Man erhält für die Varianz $\sigma^2_{\bar{x}}$, also der im Folgenden primär interessierenden Größe

$$(13) \qquad \sigma^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k} \text{ und}$$

$$(14) \qquad \sigma^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k} \frac{N_k - n_k}{N_k - 1},$$

je nachdem, ob Ziehen ohne (14) oder Ziehen mit (13) Zurücklegen vorliegt,[26] also die Endlichkeitskorrekturen[27] $(N_k - n_k)/(N_k - 1) \approx 1 - n_k/N_k$ zu berücksichtigen sind oder nicht.[28] Aus (13) folgt auch, dass der Standardfehler $\sigma_{\bar{x}}$ um so kleiner ist und damit die Schätzung um so besser ist, je homogener die K Schichten (je kleiner die K Varianzen σ_k^2) sind.

Da man die wahren Varianzen σ_k^2 innerhalb der Schichten in der Grundgesamtheit nicht kennt sind die geschätzten Varianzen $\hat{\sigma}_k^2$ einzusetzen und man erhält für die geschätzte Varianz $\hat{\sigma}^2_{\bar{x}}$ auch unter Berücksichtigung der Definition der geschätzten Stichprobenfehler $\hat{\sigma}_{\bar{x}_k} = \frac{\hat{\sigma}_k}{\sqrt{n_k}}$ für die einzelnen Schichtmittelwerte

$$(15) \qquad \hat{\sigma}^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \frac{\hat{\sigma}_k^2}{n_k} = \sum \omega_k^2 \hat{\sigma}^2_{\bar{x}_k}$$

mit den Schichtanteilen $0 < \omega_k = N_k/N < 1$ und $\hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2}{n_k - 1}$ anstelle von (13) und

$$(15a) \qquad \hat{\sigma}^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \frac{\hat{\sigma}_k^2}{n_k} \frac{N_k - n_k}{N_k} = \sum \omega_k^2 \hat{\sigma}^2_{\bar{x}_k}$$

anstelle von (14). Mit $\hat{\sigma}_k^2$ anstelle von σ_k^2 in Gl. (14) lauten die K finite multipliers $(N_k - n_k)/N_k$ statt $(N_k - n_k)/(N_k - 1)$.

Man beachte, dass der quadrierte Stichprobenfehler (SF) des Gesamtmittels $\hat{\sigma}^2_{\bar{x}}$ nicht einfach ein gewogenes Mittel der quadrierten SF der Schichtmittelwerte ist, weil sich die Summe der Gewichte zu weniger als 100% addiert, denn

---

[26] Für den Erwartungswert (im Unterschied zur Varianz) des arithmetischen Mittels spielt das keine Rolle.

[27] finite multipliers

[28] Wie man leicht sieht streben die K finite multipliers mit $N_k \to \infty$ gegen 1.

$1 = \left(\sum_k \omega_k\right)^2 = \sum_k \omega_k^2 + \sum_k \sum_{j \neq k} \omega_k \omega_j$, so dass $\sum_k \omega_k^2 < 1$.

Entsprechend ist auch e^2 als Maß für die Genauigkeit (eigentlich, wie gesagt, aber der quadrierte Fehler) bei einem Signifikanzniveau von z nicht einfach ein Mittel aus den e^2-Werten der Schichtstichproben $e_k^2 = z^2 \cdot \hat{\sigma}^2_{\bar{x}_k}$. Man kann (15) bzw. (15a) auch schreiben als

(16) $\quad e^2 = z \cdot \hat{\sigma}^2_{\bar{x}} = \sum \omega_k^2 e_k^2$,

was eine Gleichung ist, auf die Aussagen zum Gesamtstichprobenumfang n gestützt werden können (vgl. hierzu den folgenden Abschnitt e).

Im homograden Fall erhält man die analogen Formeln mit $\sigma_k^2 = \pi_k(1-\pi_k)$ bzw. $\hat{\sigma}_k^2 = \hat{\pi}_k(1-\hat{\pi}_k)$ Es gilt dann

(13a) $$\sigma^2_{\hat{\pi}} = \sum \frac{N_k^2}{N^2} \frac{\pi_k(1-\pi_k)}{n_k}$$

(14a) $$\sigma^2_{\hat{\pi}} = \sum \frac{N_k^2}{N^2} \frac{\pi_k(1-\pi_k)}{n_k} \frac{N_k - n_k}{N_k - 1}$$, bzw.

(14b) $$\hat{\sigma}^2_{\hat{\pi}} = \sum \frac{N_k^2}{N^2} \frac{\hat{\pi}_k(1-\hat{\pi}_k)}{n_k - 1} \frac{N_k - n_k}{N_k}$$.[29]

Für die Varianz von $\hat{\pi}$ gilt das für $\overline{X}$ Gesagte analog, also $E(\hat{\pi}) = \pi$ (eigentlich müsste man um konsequent zu sein auch $\hat{\Pi}$ statt $\hat{\pi}$ schreiben).

c) Gesamtstichprobenumfang bei gewünschter Größe des Stichprobenfehlers des Gesamtmittelwerts $\bar{x}$ bzw. $\hat{\pi}$

Durch Auflösen der Gleichung für $e = z\hat{\sigma}_{\bar{x}}$ bzw. $e = z\hat{\sigma}_{\hat{\pi}}$ erhält man auch hier Formeln für den mindestens notwendigen Stichprobenumfang um μ bzw. π zu schätzen. Wir betrachten zunächst den heterograden Fall. Mit (16) erhält man in Verbindung mit (13)

(17) $$e^2 = z^2 \cdot \hat{\sigma}^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \cdot \hat{\sigma}^2_{\bar{x}_k} = \sum \frac{N_k^2}{N^2} \frac{\hat{\sigma}_k^2}{n_k}.$$

Hierin ist $\hat{\sigma}_k^2$ die geschätzte Varianz der Variable x *innerhalb* der k-ten Schicht. Ein Mittelwert dieser Varianzen ist bekannt als "*interne*" Varianz V_{int}. Für die geschätzte V_{int} gilt also

(17a) $$\hat{V}_{int} = \sum \frac{N_k}{N} \cdot \hat{\sigma}_k^2.$$

Für die weitere Betrachtung sind zwei Dinge zu beachten.

- Da in dieser Gleichung kein n vorkommt müssen Annahmen über die Aufteilung des Stichprobenumfangs getroffen werden und
- für $\hat{\sigma}^2_{\bar{x}_k}$ sind die entsprechenden Werte einzusetzen, je nachdem ob man mit oder ohne der Endlichkeitskorrektur rechnet.

Neben der mit (12) bereits eingeführten proportionalen Aufteilung ist als wichtigste nichtproportionale Aufteilung die optimale Aufteilung zu nennen. Für sie gilt

[29] Man beachte, dass hier die finite multiplier etwas anders aussehen.

(18) $$\frac{n_k}{n} = \frac{N_k\,\sigma_k}{\sum N_k\,\sigma_k} = \frac{N_k\,\sigma_k}{N\overline{\sigma}}\,.$$

Auf die Herleitung dieser Formel und die ihr zugrundeliegenden Überlegungen soll erst im nächsten Abschnitt d) eingegangen werden.

c1) Heterograd, proportionale Aufteilung

Wir betrachten zunächst die proportionale Aufteilung für alle zu unterscheidenden Fälle (13) und (14) im heterograden und (13a) und (14a) im homograden Fall und erst danach die entsprechenden Formeln für die optimale Aufteilung.

Setzt man (12) in (17) ein, so erhält man unter Beachtung von (13) bzw. (15)

$$\frac{e^2}{z^2} = \hat{\sigma}^2_{\bar{x}} = \sum \frac{N_k^2}{N^2}\hat{\sigma}^2_{\bar{x}_k} = \sum \frac{N_k^2}{N^2}\frac{\hat{\sigma}_k^2}{n_k} = \frac{1}{n}\sum \frac{N_k}{N}\cdot\hat{\sigma}_k^2 = \frac{1}{n}\hat{V}_{int}\,,$$ was nach n aufgelöst

(19) $$n_{prop} \geq \frac{z^2\hat{V}_{int}}{e^2}$$

ergibt. Das ist ein Stichprobenumfang, der i.d.R. nicht unerheblich kleiner ist als der entsprechende Stichprobenumfang bei einer einfachen Stichprobe, in der die Gesamtvarianz in der Grundgesamtheit (als Summe der internen und externen Varianz) erscheint; denn nach (1) gilt

(20) $$n_{einf} \geq \frac{z^2\hat{\sigma}^2}{e^2} = \frac{z^2\left(\hat{V}_{ext} + \hat{V}_{int}\right)}{e^2},$$

wobei hier wie in (19) für die Gesamtvarianz bzw. die interne Varianz Schätzwerte, aus früheren Erhebungen gewonnene Werte oder vermutete Werte einzusetzen sind.

Für die Varianzzerlegung gilt folgender definitorischer Zusammenhang

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

was analog für die entsprechenden Schätzwerte $\hat{\sigma}^2$ usw. gilt. Es gilt also

(19a) $$n_{prop} \geq \frac{z^2}{e^2}\sum \frac{N_k}{N}\sigma_k^2$$ im Vergleich zu (20); BZW:

(20a) $$n_{einf} \geq \frac{z^2}{e^2}\cdot\left(V_{int} + V_{ext}\right) = \frac{z^2}{e^2}\cdot\left(\sum \frac{N_k}{N}\sigma_k^2 + \sum \frac{N_k}{N}\left(\mu_k - \mu\right)^2\right)$$

Man beachte, dass die Mittelwerte $\mu_1, \ldots, \mu_K$ keine Rolle spielen für n_{prop}, wohl aber die Standardabweichungen $\sigma_1, \ldots, \sigma_K$. Das Problem der Abschätzung von n_{prop} und n_{einf} ist in der Praxis allerdings meist, dass man die Varianzen nicht kennt. Es dürfte in jedem Fall nützlich sein, die entsprechenden Größen mit Stichprobenwerten zu schätzen also mit $\hat{\sigma}_k^2$, $\hat{\mu}_k = \bar{x}_k$ und $\hat{\mu} = \bar{x} = \sum \frac{N_k}{N}\cdot\bar{x}_k$. Man kann $n_{prop} < n_{einf}$ (oder auch, damit zusammenhängend und äquivalent $\sigma^2_{\bar{x}(prop)} < \sigma^2_{\bar{x}(einf)}$)[30] als Schichtungsgewinn bezeichnen. Bevor wir versuchen mit einem Zahlenbeispiel die Größenrelation $n_{prop} < n_{einf}$ zu demonstrieren soll noch auf (14) bzw. (15a) eingegangen werden. Man erhält jetzt (auch wegen $n_k/n = N_k/N$) einen wegen der Größe S kleineren notwendigen Stichprobenumfang als in (19), nämlich

[30] vgl. Anhang 3.

(19b) $$n^*_{prop} \geq \frac{z^2}{e^2}\sum\frac{N_k}{N}\hat{\sigma}_k^2\left(1-\frac{n_k}{N_k}\right) = \frac{z^2}{e^2}\hat{V}_{int} - \frac{z^2}{e^2}\sum\frac{n_k}{N}\hat{\sigma}_k^2 = n_{prop} - S\,.$$

Hierzu ein kleines Zahlenbeispiel (***Beispiel 1***)

$N_1 = 300$	$\mu_1 = 80$	$\sigma_1 = 20$
$N_2 = 700$	$\mu_2 = 200$	$\sigma_2 = 30$

(10) $\mu = 0{,}3\cdot 80 + 0{,}7\cdot 200 = 164$

(17a) $V_{int} = 0{,}3\cdot 400 + 0{,}7\cdot 900 = 750$

(21) $V_{ext} = 0{,}3(80-164)^2 + 0{,}7(200-164)^2 = 3024$

Da die externe Varianz erheblich größer ist als die interne Varianz kann man durch eine geschichtete Stichprobe erheblich gewinnen. Bei z = 2 und e = 10 (entspricht einem relativen Fehler von etwa 6%, weil $\bar{x}$ im Bereich von 160 zu erwarten ist) erhält man mit

(19) $n_{prop} = 30$ (bei prop Aufteilung würde das bedeuten $n_1 = 9$ und $n_2 = 21$) und

(20) $n_{einf} = 150{,}96 \approx 151$ (Aufteilung 45 : 106)

weil die Gesamtvarianz mit 3024 + 750 = 3774 gut fünfmal so groß ist wie die für n_{prop} maßgebliche interne Varianz.[31] Eine Berücksichtigung der Endlichkeitskorrektur nach (19b) ändert wegen des geringen Stichprobenumfangs von etwa 2% wegen $n_k/N_k = n/N = 0{,}021$ fast nichts. Man erhält für die Größe S in (19b) nur den Wert 0,9, also noch nicht einmal eine Einheit, die weniger zu befragen wäre.

c2) Homograd, proportionale Aufteilung

Im homograden Fall erhält man die entsprechenden Formeln wenn man (14b) in (17) einsetzt und mit n_k statt n_k-1 rechnet

(22) $$n^*_{prop} \geq \frac{z^2}{e^2}\sum\frac{N_k}{N}\left(\hat{\pi}_k(1-\hat{\pi}_k)\right)\left(1-\frac{n_k}{N_k}\right),$$

so dass eigentlich nur $\hat{\pi}_k(1-\hat{\pi}_k)$ an die Stelle von $\hat{\sigma}_k^2$ tritt. Bei proportionaler Aufteilung entfällt im Fall ZmZ (ohne Endlichkeitskorrektur) der Faktor $\Theta = \left(1-\frac{n_k}{N_k}\right)$ der konstant 1- n/N ist. Setzt man für jede Schicht $\pi_k(1-\pi_k) = \pi\,(1-\pi)$ an (also gleiche Varianzen innerhalb der Schichten) so gewinnt man nichts gegenüber Gl. (2) bzw. mit dem Faktor $\Theta = (N-n)/N$ nichts gegenüber (4). Das ist intuitiv verständlich, denn bei lauter gleichen Schichten, kann man mit einer Schichtung nichts gewinnen. Ist $\pi_k = \pi$ für alle k, dann gibt es auch keine externe Varianz. Auch hier ergibt sich, wie im heterograden Fall, dass es entscheidend darauf ankommt, dass die Varianzen innerhalb der Schichten unterschiedlich sind.

Wegen $\pi = \sum(N_k/N)\pi_k$ gilt

Mindeststichprobenumfang n für die Gesamtstichprobe

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

*) vgl. (23a) **) vgl. (23)

[31] wie man sieht ist auch der Stichprobenumfang etwa fünfmal so groß, was sich auch nach einem Vergleich von (20) mit (19) so ergibt.

Worauf es ankommt ist also, inwieweit

$\sigma^2_{\hat{\pi}(prop)} = V_{prop} = \sum_k \frac{N_k}{N} \cdot \pi_k (1-\pi_k)$ abweicht von (genauer: kleiner ist als)

$\sigma^2_{\hat{\pi}(einf)} = V_{einf} = \pi(1-\pi)$, wobei $\pi = \sum_k \frac{N_k}{N} \cdot \pi_k$.

In der Tabelle auf S. 15 unten findet sich die Formel für die proportionale Aufteilung mit

(23) $$n \geq \frac{z^2 \sum \omega_k \hat{\pi}_k (1-\hat{\pi}_k)}{e^2 + \frac{1}{N} z^2 \sum \omega_k \hat{\pi}_k (1-\hat{\pi}_k)}$$ bei ZoZ und als Spezialfall

(23a) $$n \geq \frac{z^2 \sum \omega_k \hat{\pi}_k (1-\hat{\pi}_k)}{e^2}$$ bei ZmZ.

Da im Zähler jeweils mit $\sum \omega_k \hat{\pi}_k (1-\hat{\pi}_k)$ eine mittlere Varianz innerhalb der Schichten (also eine interne Varianz) steht und diese nicht größer als die größtmögliche Varianz $\pi(1-\pi) = ¼$ sein kann, ist der Stichprobenumfang nach (23) und (23a) kleiner oder gleich dem in Abschn. 1 mit den Gl. 3ff bestimmte Umfang (was dann ja auch in Abschn. 2 problematisiert wurde). Außerdem gilt für die interne Varianz und die Gesamtvarianz

(23b) $$\sum \omega_k \hat{\pi}_k (1-\hat{\pi}_k) \leq \sum \omega_k \hat{\pi}_k \left(1 - \sum \omega_k \hat{\pi}_k\right) = \pi(1-\pi),$$

so dass man schlimmstenfalls mit Schichtung und proportionaler Aufteilung einen genauso großen Stichprobenumfang erhält wie im Falle einer einfachen Stichprobe, i. d. R. aber einen kleineren. Der Zusammenhang ist analog zur Zerlegung der Varianz σ^2 der Grundgesamtheit nach (21) zu sehen, denn es gilt

(23c) $$\sum_{k=1}^{K} \frac{N_k}{N} (\pi_k - \pi)^2 + \sum_{k=1}^{K} \frac{N_k}{N} \pi_k (1-\pi_k) = A + B = \pi(1-\pi)$$

Wir wollen hierzu wieder ein Beispiel betrachten und rechnen der Einfachheit halber mit dem Fall "Ziehen mit Zurücklegen" (ZmZ).

Beispiel 2:

Variante 1

N = 1000, N_1 = 300, N_2 = 700 π_1 = 0,2 und π_2 = 0,6

Man erhält dann[32] $V_{einf} = \pi(1-\pi) = 0{,}48 \cdot 0{,}52 = 0{,}2496$ weil $\pi = 0{,}3 \cdot 0{,}2 + 0{,}7 \cdot 0{,}6 = 0{,}48$ und V_{prop} = $B = 0{,}3 \cdot 0{,}2 \cdot 0{,}8 + 0{,}7 \cdot 0{,}6 \cdot 0{,}4 = 0{,}216 < 0{,}2496$ was die Ungleichung (23b) verifiziert. Betrachtet man den Ausdruck A in diesem Fall, so erhält man $A = 0{,}3 \cdot (0{,}2 - 0{,}48)^2 + 0{,}7 \cdot (0{,}6 - 0{,}48)^2 = 0{,}0336$ und die Summe A + B = 0,0336 + 0,216 ist in der Tat 0,2496, was (23c) verifiziert.

Mit $z^2 = 4$ und $e^2 = (0{,}1)^2 = 0{,}01$ also $z^2/e^2 = 400$ erhält man dann für die einfache Stichprobe einen Mindeststichprobenumfang von $400 \cdot 0{,}2496 = 99{,}84$ und bei der geschichteten Stichprobe $400 \cdot 0{,}216 = 86{,}4$.[33] Der Unterschied wäre erheblich größer (als 86 zu 100), wenn sich die Varianzen innerhalb der beiden Schichten stärker unterschieden, wie das in der folgenden Variante der Fall ist.

---

[32] wir betrachten hier stets die n-fachen Varianzen des Mittelwerts (x-quer).

[33] Mit der maximalen Streuung von $\pi(1-\pi) = ¼$ erhält man n = 100.

Variante 2

N = 1000, N_1 = 300, N_2 = 700 π_1 = 0,1 und π_2 = 0,7 sowie $z^2/e^2 = 2^2/0{,}1^2 = 400$. Es ist jetzt π = 0,52 und somit V_{einf} wie bisher 0,2496 so dass auch der Stichprobenumfang bei einer einfachen Stichprobe wieder 99,84 also praktisch 100 ist (wie bei $\pi(1-\pi) = ¼$). Für die geschichtete Stichprobe erhält man jetzt B = V_{prop} = 0,3·0,1·0,9 + 0,7·0,7·0,3 = 0.174. Auch hier ist wieder (23b) verifiziert und für A in (23c) erhält man 0,0756 und die Summe von A und B ist wieder 0,2496.

Die interne Varianz B ist jetzt nur noch 0,174, also 57,8% von 0,2496. Entsprechend kleiner ist der Stichprobenumfang. Man erhält jetzt bei Multiplikation von B mit $z^2/e^2 = 400$ den Wert n_{prop} = 69,6 also etwa 70 Einheiten statt 100, die man in beiden Varianten erhalten würde, wenn man bei einer einfache Stichprobe mit einer Varianz von $\pi(1-\pi) = ¼$ rechnen würde. Die proportionale Aufteilung im Verhältnis N_1:N_2 von n = 70 führt zu n_1 = 0,3·70 = 21 und n_2 = 0,7·70 = 49. Die Auswahlsätze sind n_1/N_1 = 21/300 = 0,07 und n_2/N_2 = 49/700 = 0,07, also einheitlich 7%.

Offensichtlich gewinnt man durch Schichtung mehr, wenn die Varianzen innerhalb der Schichten recht unterschiedlich sind. Das zeigt der Vergleich von 70 zu 86 bei den beiden Varianten.

c3) Heterograd, optimale Aufteilung

Betrachtet man nun den Fall der optimalen Aufteilung, so ist eine Umformung von (18) nützlich. Man erhält nämlich

(18a) $$\frac{n_k}{n} = \frac{N_k\,\sigma_k}{\sum N_k\,\sigma_k} = \frac{(N_k/N)\sigma_k}{\sum N_k\,\sigma_k/N} = \frac{N_k}{N}\cdot\frac{\sigma_k}{\overline{\sigma}}$$, so dass wegen $$\frac{n_k}{N_k} = \frac{n}{N}\cdot\frac{\sigma_k}{\overline{\sigma}}$$

der Auswahlsatz n_k/N_k größer/kleiner als bei proportionaler Aufteilung ist (dort ist er für alle Schichten gleich und n/N), je nach dem, ob die Standardabweichung (nicht die Varianz, wie oben bei den Betrachtungen mit V_{int}) σ_k größer oder kleiner ist als die mittlere Standardabweichung $\overline{\sigma}$. Setzt man $\frac{N_k}{N} = \frac{n_k}{n}\cdot\frac{\overline{\sigma}}{\sigma_k}$ in (17) ein, so erhält man im Fall ZmZ

$$\frac{e^2}{z^2} = \frac{1}{n}\overline{\sigma}\sum\frac{N_k}{N}\cdot\sigma_k = \frac{\overline{\sigma}^2}{n}$$ und damit

(24) $$n_{opt} \geq \frac{z^2\overline{\sigma}^2}{e^2},$$

was einen kleineren Wert ergibt als (19) bzw. (19a), denn es ist ja bekannt, dass bei der Varianz für die ersten beiden Momente m_1 und m_2 gilt

$$\sigma^2 = \sum\frac{N_k}{N}\cdot x_k^2 - \left(\sum\frac{N_k}{N}\cdot x_k\right)^2 = \sum h_k x_k^2 - \overline{x}^2 = m_2 - m_1^2$$ (mit $m_r = \sum h_k m_k^r$ als r-tes Moment).

Der entsprechende Zusammenhang bei ZoZ geht aus von

(24a) $$\frac{e^2}{z^2} = \frac{1}{n}\overline{\sigma}\sum\frac{N_k}{N}\cdot\sigma_k\left(1-\frac{n_k}{N_k}\right) \leq \frac{\overline{\sigma}^2}{n}$$

weil die finite multiplier zwischen 0 und 1 liegen. Das Problem hierbei ist, dass man n_k nicht kennt, so lange man auch n nicht kennt und (18) anwenden kann. Allerdings ist die entscheidende Größe n_k/N_k, also der Auswahlsatz für die kte Schicht und müsste hierfür versuchsweise Werte über oder unter dem Durchschnitt n/N, den man vielleicht im Auge haben mag, ansetzen. Die finite multipliers $(1 - n_1/N_1)$, $(1 - n_2/N_2)$, …, $(1 - n_K/N_K)$, sind in jedem Fall kleiner als 1 (sobald $n_k > 0$ ist), so dass man mit

(24b) $$n = \frac{z^2}{e^2}\overline{\sigma}\sum\frac{N_k}{N}\cdot\sigma_k\left(1-\frac{n_k}{N_k}\right) = \frac{z^2}{e^2}\overline{\sigma}^2 - \frac{z^2}{e^2}\overline{\sigma}\sum\frac{n_k}{N}\cdot\sigma_k \le \frac{z^2}{e^2}\overline{\sigma}^2$$

einen kleineren Stichprobenumfang erhält als mit (22). Es sind insbesondere die Ausdrücke n_k/N sehr kein und sie addieren sich nicht zu 1.

Die zu (19) bzw. (19a) also ZmZ analoge Formeln für ZoZ ist somit

(25) $$n^*_{prop} \ge \frac{z^2}{e^2}\sum\frac{N_k}{N}\hat{\sigma}_k^2\left(1-\frac{n_k}{N_k}\right) = \frac{z^2}{e^2}\hat{V}_{int} - \frac{z^2}{e^2}\sum\frac{n_k}{N}\hat{\sigma}_k^2 = n_{prop} - S$$

c4) Homograd, optimale Aufteilung

Die Formel für die optimale Aufteilung gewinnt man entsprechend indem man in (17) für $\hat{\sigma}^2_{\bar{x}}$ den Ausdruck $\hat{\sigma}^2_{\hat{\pi}}$ gem. (14a) einsetzt und versucht, bei Größen wie N_k/N die Bedingung für die optimale Aufteilung (also Gl. (18)) zu benutzen. So kann man $N_k^2\pi_k(1-\pi_k)$ ersetzten durch $(N_k\,\sigma_k)^2 = N_k^2\pi(1-\pi) = \left(\frac{n_k N\overline{\sigma}}{n}\right)^2$ und erhält so

(26) $$n^*_{opt} = \frac{z^2\overline{\sigma}^2}{e^2} - \frac{z^2\overline{\sigma}^2}{e^2}\sum\frac{n_k^2}{nN_k} = \frac{z^2\overline{\sigma}^2}{e^2} - S^* \text{ mit } \overline{\sigma}^2 = \left(\sum\frac{N_k}{N}\sqrt{\pi_k(1-\pi_k)}\right)^2,$$

wobei der Subtrahend S* entfällt bei ZmZ. Es ist übrigens verschieden von S in (25).

Erst im folgenden Unterabschnitt sollen die beiden Zahlbeispiele für den Fall der optimalen Aufteilung weitergeführt werden.

d) Aufteilung einer Stichprobe vom gegebenen Gesamtumfang n auf die K Schichten

Die Formel (18) $\frac{n_k}{n} = \frac{N_k\,\sigma_k}{\sum N_k\,\sigma_k}$ für die optimale Aufteilung wurde von Jerzy Neyman entwickelt. Sie beruht auf der Minimierung der Varianz $V(\overline{X}) = \sigma^2_{\bar{x}}$ nach Gl. (1b) unter der Nebenbedingung $n = \sum n_k$. Die gleiche Überlegung liegt dem Anhang 2 dieses Papiers zugrunde.

Es ist leicht zu sehen, dass die optimale Aufteilung nach (18) dann auf die proportionale (12) hinausläuft, wenn alle $\sigma_k = \overline{\sigma}$ sind, und dass man dann nichts durch eine optimale Aufteilung gegenüber einer proportionalen Aufteilung gewinnt. Es ist auch klar, dass sich unter solchen Voraussetzungen (19a) zu (24) reduziert. Die optimale Aufteilung ist also besser als die proportionale wenn und insofern die Standardabweichungen innerhalb der Schichten streuen, also die Varianz $V(\sigma_k) = \sum\frac{N_k}{N}\cdot\sigma_k^2 - \overline{\sigma}^2 > 0$ ist. Es gilt also

- Sind die Mittelwerte μ_k alle gleich (d.h. ist die externe Varianz V_{ext} von x null), so folgt aus (20a), (21) und (21a), dass man durch eine geschichtete Stichprobe bei proportionaler Aufteilung gegenüber einer einfachen Stichprobe nichts gewinnt.[34]
- Sind alle Varianzen (und damit auch alle Standardabweichungen) innerhalb der Schichten gleich $\sigma_1 = \sigma_2 = \ldots = \sigma_K$ (also die Varianz der Standardabweichungen $V(\sigma_k) = 0$),

[34] Man könnte auch auf den Gedanken kommen, dass man sich bei $V_{ext} > 0$ mit einer geschichteten Stichprobe stets besser stellt als mir einer einfachen Stichprobe. Das gilt jedoch nur bei proportionaler Aufteilung. Zu weiteren Einzelheiten vgl. Anhang 3 dieses Papiers.

dann gewinnt man nichts durch eine optimale Aufteilung gegenüber einer proportionaler Aufteilung.

Zur Berechnung der Auswahlsätze n_k/N_k für jede Schicht an nach der "optimalen Aufteilung" gem. (18) sind also die Standardabweichungen (bzw. geschätzten Standardabweichungen $\hat{\sigma}_k$) entscheidend. Man kann jetzt zwei interessante Interpretationen von (18) geben.

1. Umformung von (18) liefert

(18a) $$\frac{n_k}{N_k} = n \cdot \frac{\sigma_k}{\sum N_k\,\sigma_k} = \frac{n}{N} \cdot \frac{\sigma_k}{\overline{\sigma}},$$

so dass der Auswahlsatz über- oder unterdurchschnittlich ist, je nachdem ob σ_k größer oder kleiner als der Durchschnitt $\overline{\sigma}$ ist.

2. Aus (18a) folgt auch, dass sich die Auswahlsätze zueinander verhalten wie die (absoluten) Standardabweichungen, denn

(18b) $$\frac{n_j/N_j}{n_k/N_k} = \frac{\sigma_j}{\sigma_k}$$

und eine große Standardabweichung σ_k (große Heterogenität der Schicht k) bedeutet ein großer Auswahlsatz n_k/N_k, ein kleines σ_k ein kleiner Auswahlsatz; alles ganz im Gegensatz zu den Feststellungen im Abschnitt 2a, wo der Auswahlsatz mit zunehmender Schichtgröße abnahm.

Wir betrachten nun an den beiden Zahlenbeispielen, wie sich die optimale Aufteilung nach (18) im konkreten Fall berechnen lässt, und wie sich das Ergebnis von der proportionalen Aufteilung unterscheidet.

Um n für ***Beispiel 1*** nach (24) zu berechnen muss zunächst die durchschnittliche Standardabweichung berechnet werden. Man erhält $\overline{\sigma} = 0{,}3\cdot 20 + 0{,}7\cdot 30 = 27$ so dass man (wieder bei z = 2 und e = 10) mit (22) für $n_{opt} = 29{,}16$ erhält, was nur unwesentlich weniger ist als $n_{prop} = 30$. Das liegt daran, dass $\overline{\sigma}^2 = 21^2 = 729$ kaum kleiner ist als die interne Varianz mit 750. Die Stichprobe n = 30 wurde proportional mit $n_1 = 9$ und $n_2 = 21$ aufgeteilt. Jetzt erhält man ebenfalls bei n = 30 allerdings eine andere Aufteilung, nämlich in $n_1 = 6{,}67 \approx 7$ und $n_2 = 23{,}33 \approx 23$ statt 9:21, denn die Anteile n_1/n und n_2/n sind nach (18) 2/9 = 0,2222 und 7/9 = 0,7778. Wenn man mit dem Übergang von proportional zu optimal hinsichtlich n nicht gewann, so wird dies auch für die Stichprobenfehler gelten. Man erhält wenn man mit $n_{opt} = 29{,}16 \approx n_{prop} = 30$ rechnet

$$\sigma^2_{\bar{x}(opt)} = V\left(\overline{X}\right)_{opt} = \frac{1}{n}\left(\sum \frac{N_k}{N}\sigma_k\right)^2 = \frac{\overline{\sigma}^2}{n} = \frac{27^2}{30} = 24{,}3 \text{ im Vergleich zu}$$

$$\sigma^2_{\bar{x}(prop)} = V\left(\overline{X}\right)_{prop} = \frac{1}{n}\sum \frac{N_k}{N}\sigma_k^2 = \frac{V_{int}}{n} = \frac{750}{30} = 25.$$

Wir führen die entsprechende Betrachtung im ***Beispiel 2*** durch:

Variante 1:

Man erhält $\sigma_1 = (0{,}2\cdot 0{,}8)^{1/2} = 0{,}16^{1/2} = 0{,}4$ und $\sigma_2 = 0{,}24^{1/2} = 0{,}4898$. Die mittlere Standardabweichung nach (26) ist dann $0.3\cdot 0{,}4 + 0{,}7\cdot 0{,}4898 = 0{,}4629$. Der quadrierte Wert ist 0,2143 und das multipliziert mit $z^2/e^2 = 400$ ergibt 85,72 im Unterschied zu 86,4 bei proportionaler Aufteilung (der Unterschied bezüglich n ist also auch hier wieder gering).

Anwendung von (18) liefert die Anteile $n_1/n = 0{,}253$ und $n_2/n = 0{,}7408$ und damit bei n = 86 die Werte $n_1 = 22{,}3 \approx 22$ und $n_2 = 63{,}7 \approx 64$, also 22:64 statt 26:60.[35] Das Verhältnis 64 zu 22 ist etwa 2,9 was

[35] Bei proportionaler Aufteilung erhält man $n_1 = 0{,}3\cdot 86 = 25{,}8 \approx 26$ und $n_2 = 0{,}7\cdot 86 = 60{,}2 \approx 60$.

auch dem Quotient $N_1\sigma_1/N_2\sigma_2$ entspricht. Die Auswahlsätze sind jetzt ungleich und zwar n_1/N_1 = 22/300 = 0,0733 also 7,3% und n_2/N_2 = 64/700 = 0,0914, also etwa 9,1% statt einheitlich 8,6% bei proportionaler Aufteilung. Der Auswahlsatz bei Schicht 2 ist größer, weil auch die Standardabweichung größer ist.[36]

Variante 2

Man erhält $\sigma_1 = (0{,}1\cdot 0{,}9)^{1/2} = 0{,}3$ und $\sigma_2 = 0{,}21^{1/2} = 0{,}45826$. Die mittlere Standardabweichung nach (26) ist dann $0.3\cdot 0{,}4 + 0{,}7\cdot 0{,}4898 = 0{,}4108$. Der quadrierte Wert ist 0,1687 und das multipliziert mit $z^2/e^2 = 400$ ergibt 67,5 ≈ 68 (wenig verschieden von 70 bei der proportionalen Aufteilung, aber durchaus ein Gewinn gegenüber der Variante 1).[37]

Anwendung von (18) liefert die Anteile n_1/n = 0,2191 und damit n_1 = 14,89 ≈ 15 und n_2/n = 0,7809 und damit n_2 = 53,1 ≈ 53 (also 15:53 gegenüber 21:49 bei proportionaler Aufteilung). Die Auswahlsätze sind jetzt n_1/N_1 = 0,050 und n_2/N_2 = 0,076, während sie bei proportionaler Aufteilung einheitlich 7% waren. Man kann die Ergebnisse hinsichtlich der Stichprobenumfänge wie folgt zusammenfassen:

Schicht	einfache Stichprobe*	Variante 1		Variante 2	
		prop.	optim.	prop.	optim.
n_1	da $\pi(1-\pi)$ =	26	22	21	15
n_2	0,2496 ≈ ¼	60	64	49	53
n	n ≈ 100 (maximaler Wert)	86,4 ≈ 86	85,7 ≈ 86	69,6 ≈ 70	67,5 ≈ 68

* gleicher Wert bei Variante 1 und 2.

Von weiteren Konzepten der optimalen Aufteilung bei der Planung einer geschichteten Stichprobe wäre vor allem die kostenoptimale Aufteilung zu nennen. Auch hier wird $\sigma_{\bar{x}}$ minimiert, aber unter der Nebenbedingung[38] gegebener Gesamtkosten $C = c_0 + \sum_{k=1}^{K} c_k n_k$, wobei c_0 fixe Kosten und c_k variable Kosten (der Erhebung) pro Einheit in der k-ten Schicht darstellen. Die Aufteilung ist jetzt in Analogie zu (18) gegeben mit

$$\text{(18c)} \qquad \frac{n_k}{n} = \frac{\frac{1}{\sqrt{c_k}} N_k \sigma_k}{\sum_{k=1}^{K} \frac{1}{\sqrt{c_k}} N_k \sigma_k}.$$

e) Vergleich mit der Stichprobenplanung nach Abschnitt 1 und 2

Nach der Übersicht auf Seite 11 oben war das Gemeinsame der Betrachtung in den Abschnitten a bis d, dass es um den Stichprobenfehler des Gesamtmittels (der Gesamtstichprobe, über alle Schichten gerechnet), also um $\sigma_{\bar{x}}$ (bzw. $\sigma_{\hat{\pi}}$) ging, und nicht um die Stichprobenfehler bei den einzelnen Schichten (also $\sigma_{\bar{x}_k}$, k = 1, …, K und entsprechend $\sigma_{\hat{\pi}_k}$).

Die Bestimmung des Stichprobenumfangs nach Abschn.3 erfolgte im Hinblick auf eine gewünschte, bzw. minimale Varianz[39]

$$\text{(13)} \qquad \sigma_{\bar{x}}^2 = V(\bar{X}) = \sum \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k} \text{ und}$$

[36] Zum Vergleich der Auswahlsätze siehe (18b).

[37] Wie man sieht, ist die optimale Aufteilung umso günstiger je unterschiedlicher die Varianzen innerhalb der Schichten sind.

[38] Es ist wieder, wie auch Anhang 2 und (18) eine Anwendung von Lagrange Multiplikatoren.

[39] Der homograde Fall ist analog zu behandeln.

(14) $$\sigma_{\bar{x}}^2 = V(\bar{X}) = \sum \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k} \frac{N_k - n_k}{N_k - 1},$$

des *Mittelwerts der Gesamtstichprobe* (bzw. jeweils die geschätzte oder angenommene Varianz). Die halbe Breite des Konfidenzintervalls e, die bei der Bestimmung von n im Vordergrund steht ist abgesehen von z allein von $\sigma_{\bar{x}}^2$ abhängig und beträgt $e = z\sigma_{\bar{x}}$. Bei der Stichprobenplanung, wie sie in den Abschnitten 1 und 2 behandelt wurde, d.h. bei Anwendung der Formeln für die einfache Stichprobe auf jede Schicht einzeln, ging es dagegen um die Varianz der *Mittelwerte jeder einzelnen Schicht*, also $\sigma_{\bar{x}_k}^2 = V(\bar{X}_k) = \frac{\sigma_k^2}{n_k} = \frac{e_k^2}{z_k^2}$ woraus folgt, dass n_k nach (1) mindestens den Wert $n_k = \left(\frac{z_k \sigma_k}{e_k}\right)^2$ haben sollte.

Der Fall ZoZ mit $\sigma_{\bar{x}_k}^2 = V(\bar{X}_k) = \frac{\sigma_k^2}{n_k} \frac{N_k - n_k}{N_k - 1}$, was dann (3) liefert, sei zunächst zurückgestellt.

Ausgehend von (1) ist der Gesamtstichprobenumfang wenn in jeder Schicht nach Art von (1) berechnet wird

(27) $$n = \sum n_k = \sum \left(\frac{z_k \sigma_k}{e_k}\right)^2$$ und bei gleichem z und e für alle K Schichten erhält man

(27a) $$n = \frac{z^2}{e^2}\left(\sigma_1^2 + ... + \sigma_K^2\right) = \frac{z^2 \sum \sigma_k^2}{e^2}$$

im Unterschied zu Gl. 19, bzw. 19a (geschichtete Stichprobe bei proportionaler Aufteilung)

(19) $$n_{prop} = \frac{z^2}{e^2} V_{int} = \frac{z^2}{e^2} \sum \frac{N_k}{N} \sigma_k^2$$ und (24) bei optimaler Aufteilung.

Man beachte: $\sum \sigma_k^2$ in (27a) ist das N-fache *ungewogene* Mittel während V_{int} ein gewogenes Mittel darstellt. Wir vergleichen das auch mit der einfachen Stichprobe, für die gilt

(20) $$n_{e\,inf} = \frac{z^2}{e^2}\left(V_{int} + V_{ext}\right).$$

Wären alle Schichtumfänge N_k gleich groß, nämlich $N_k = N/K \;\forall\, k$, dann wäre wegen der Stichprobenumfang bei der Vorgehensweise nach (27a) immerhin K mal so groß wie gem. (19) also $n = K \cdot n_{prop}$.[40] Um die Beziehungen zwischen den Gleichungen deutlicher zu machen greifen wir wieder auf unser Zahlenbeispiel zurück.

Wir zeigen am ***Beispiel 1***, wie unterschiedlich die Stichprobenumfänge sind. Mit den Zahlen des Beispiels erhält man mit $z^2/e^2 = 0{,}04$ und $V_{int} = 750$ sowie $V_{ext} = 3024$ so dass $\sigma^2 = 3774$. Daraus ergibt sich nach (20) und (19)

$n_{einf} = 3774 \cdot 0{,}04 = 150{,}96$ und $n_{prop} = 750 \cdot 0{,}04 = 30$ mit der Aufteilung $n_1 = 9$ und $n_2 = 21$.[41]

---

[40] Nach dem bekannten Zusammenhang zwischen Stichprobenfehler und n (bei ZmZ) bewirkt eine K-fache Vergrößerung des Stichprobenumfangs nur eine Ver-$K^{1/2}$-fachung der "Genauigkeit"

[41] Ferner ist, wie bereits erwähnt, die mittlere Standardabweichung $0{,}3 \cdot 20 + 0{,}7 \cdot 30 = 27$ so dass $n_{opt} = 0{,}04 \cdot (27)^2 = 29{,}16$ nur geringfügig kleiner ist als n_{prop}.

Mit einer Stichprobenplanung nach Art der Abschnitte 1 und 2, indem man also bei jeder Schicht nach Gl. 1 vorgeht ergibt sich nach (26a)

$$n = 0{,}04\cdot(20^2 + 30^2) = 0{,}04\cdot 1300 = 52.$$[42]

Die 52 setzen sich wie folgt zusammen $n_1 = 0{,}04\cdot 20^2 = 16$ und $n_2 = 0{,}04\cdot 30^2 = 36$. Die Auswahlsätze sind dann 16/300 = 0,053 und 36/700 = 051 (insgesamt 52/1000 =0,052), während sie bei einer geschichteten Stichprobe mit proportionaler Aufteilung einheitlich 0,03 sind.

Wir haben bereits festgestellt, dass die Berücksichtigung der Endlichkeit von N_1 und N_2 bei n_{prop} praktisch keine Wirkung hat und n_{prop} nur um 0,9 verringert. Sie wirkt sich auch auf n = 52 kaum aus, denn es gilt nach (3):

$K_1 = 2^2\cdot 20^2 = 1600$, $N_1 = 300$, $e^2 = 100 \rightarrow n_1 = 15{,}18 \approx 15$ und

$K_2 = 2^2\cdot 30^2 = 3600$, $N_2 = 700$, $e^2 = 100 \rightarrow n_1 = 34{,}24 \approx 34$

so dass der Gesamtstichprobenumfang 49 statt 52 wäre.

Interessant ist nun ein Vergleich der Stichprobenfehler für die Teilgesamtheiten (Schichten) und die gesamte Grundgesamtheit je nachdem, wie n geplant wird.

Stichproben-umfang nach	Schicht 1 $\sigma_{\bar{x}_1} = \sigma_1 \big/ \sqrt{n_1}$	Schicht 2 $\sigma_{\bar{x}_2} = \sigma_2 \big/ \sqrt{n_2}$	insgesamt $\sigma_{\bar{x}}$ nach (13)*
geschicht. Stich-probe mit prop. Aufteilung	$\frac{20}{\sqrt{9}} = 6{,}667$	$\frac{30}{\sqrt{21}} = 6{,}546$	$\sqrt{0{,}3^2\frac{20^2}{9} + 0{,}7^2\frac{30^2}{21}} = \sqrt{25} = 5$
Gl. (27) vgl. Ab-schnitte 1 und 2	$\frac{20}{\sqrt{15}} = 5{,}164$ **	$\frac{30}{\sqrt{34}} = 5{,}145$ **	$\sqrt{0{,}3^2\frac{20^2}{15} + 0{,}7^2\frac{30^2}{34}} = \sqrt{15{,}37} = 3{,}92$

* bei der Herleitung der Formel ist zu berücksichtigen, dass (11) für den Zusammenhang zwischen den Mittelwerten gilt und man beachte, dass der Stichprobenfehler kein gewogenes Mittel der einzelnen Stichprobenfehler ist. Selbst dann, wenn die einzelnen Stichprobenfehler gleich sind ist der des Gesamtmittels nicht gleich, sondern kleiner wegen der *quadrierten* Gewichte 0,3 und 0,7. Auch die Gesamtvarianz ist nach (21), wegen der externen Varianz nicht einfach ein Mittelwert der Varianzen innerhalb der Schichten. Entsprechend ist auch e^2 gem. (16) nicht einfach ein Mittel der $(e_k)^2$.

** die beiden Werte sind nicht aufgrund von Rundungsfehleren ungleich, sondern tatsächlich (auch mit nicht gerundeten Werten für n_1 und n_2 gerechnet) geringfügig verschieden.

Die paradoxen Ergebnisse von Abschnitt 2 kommen v.a. dadurch zustande, dass für jede Schicht mangels Kenntnisse über σ_1^2, …, σ_K^2 mit gleichen Varianzen gerechnet wurde, so dass sich die nach (3), bzw. (4) berechneten Stichprobenumfänge n_k und n_j nur durch die Verschiedenheit von N_k und N_j unterscheiden konnten. Allerdings gilt *nicht* $\frac{n_k}{n_j} = \frac{N_k}{N_j}$, wie dies bei einer geschichtete Stichprobe bei proportionaler Aufteilung gilt, und zwar sowohl im Fall ZmZ als auch (wie gleich gezeigt wird) bei ZoZ. Hinsichtlich des Verhältnisses zweier Teilstichproben n_k/n_j gilt beim Vorgehen nach Art von Abschn. 1 vielmehr

$$\frac{n_k}{n_j} = 1 \text{ bei ZmZ weil } n_k = n_j = \frac{z^2\sigma^2}{e^2} \text{ und bei ZoZ}$$

$$\frac{n_k}{n_j} = \frac{N_k}{N_j}\cdot\frac{N_j e^2 + z^2\sigma^2}{N_k e^2 + z^2\sigma^2} \text{ weil } n_k = \frac{z^2\sigma^2}{e^2 + \frac{z^2\sigma^2}{N_k}} = \frac{N_k z^2\sigma^2}{N_k e^2 + z^2\sigma^2}.$$

[42] Dass n kleiner ist als n_{einf} liegt vor allem an der sehr großen externen Varianz in diesem Beispiel. Rechnet man mit den Mittelwerten 80 und 90 statt 80 und 200, so erhält man $V_{ext} = 9$, $\sigma^2 = 759$ und damit $n_{einf} = 0{,}04.759 = 30{,}36$. Jetzt ist n = 52 zwar größer als n_{einf}, aber man hat auch praktisch keinen Schichtungsgewinn, weil ja weiterhin gilt $n_{prop} = 30$.

Bei den Betrachtungen in Abschn. 2 war e = 0,08, z = 2 und $\sigma_2 = \pi(1-\pi) = ¼$, so dass man z.B. bei N_1 = 10.000 und N_2 = 200 erhält man $\frac{n_1}{n_2} = \frac{10000}{200} \cdot \frac{2,28}{65} = 1,754$. Ist z.B. n_2 = 88, dann ist $n_1 = 1,754 \cdot 88 = 154,4$ was den Relationen der Tabelle auf S. 6 entspricht. Bei weniger extremen Situationen, etwa N_1 = 500 und N_2 = 200 ist die Relation nur $\frac{5}{2} \cdot \frac{2,28}{4,2} = 1,357$ oder 2,7 : 2 statt 5 : 2. Bei proportionaler Aufteilung berechnet sich der Stichprobenumfang für die k-te Schicht mit

$$n_{k(prop)} = \frac{N_k}{N} n_{prop} = \frac{N_k}{N} \cdot \frac{z^2\sigma^2}{e^2} \cdot \sum \frac{N_k}{N}\left(1 - \frac{n_k}{N}\right).$$

Der dritte Faktor ist 1 im Fall ZmZ. Der zweite und dritte Faktor ist für all k gleich, so dass sowohl bei ZmZ als auch bei ZoZ gilt $\frac{n_{k(prop)}}{n_{j(prop)}} = \frac{N_k}{N_j}$. Die folgende Tabelle mag dies noch einmal verdeutlichen:

Stichprobenplanung* nach	N_1 = 10.000 und N_2 = 200	N_1 = 500 und N_2 = 200
$n_1 : n_2$ bei geschichteter Stichprobe (prop. Aufteilung) ZoZ und Auswahlsätze n_1/N_1 und n_2/N_2	$\frac{n_1}{n_2} = \frac{N_1}{N_2} = \frac{10000}{200} = \frac{50}{1}$ Auswahlsätze gleich groß	$\frac{n_1}{n_2} = \frac{N_1}{N_2} = \frac{500}{200} = \frac{2,5}{1}$ gleich, also $n_1/N_1 = n_2/N_2$
Formel für einfache Stichprobe (Gl. (4))	$\frac{n_1}{n_2} = \frac{1,754}{1} \neq \frac{N_1}{N_2}$	$\frac{n_1}{n_2} = \frac{1,357}{1} \neq \frac{N_1}{N_2}$
Auswahlsätze n_k/N_k	$\frac{n_1}{N_1} = \frac{1,754 \cdot n_2}{50 \cdot N_2} = 0,035 \cdot \frac{n_2}{N_2}$	$\frac{n_1}{N_1} = \frac{1,357 \cdot n_2}{2,5 \cdot N_2} = 0,543 \cdot \frac{n_2}{N_2}$

* **wichtig**: bei jeweils gleicher Varianz, bzw. Standardabweichung innerhalb der Schichten (Teilgesamtheiten) also $\sigma_1 = \sigma_2$.

Wie man sieht treten bei einer geschichteten Stichprobe mit proportionaler Aufteilung die in den Abschnitten 2a und 2b als paradox bezeichneten Konsequenzen hinsichtlich Stichprobenumfang und Auswahlsatz in Abhängigkeit von N nicht auf. Allerdings ist die Fragestellung bei der geschichteten Stichprobe anders als bei Verwendung der Formeln (1) bzw. (4) für die einfache Stichprobe. Hebt man besonders auf die Schätzung von $\mu_1,\ldots, \mu_K$ und deren Unterschiede ab und nicht auf den Gesamtmittelwert, mag das Vorgehen nach Abschn. 2 gerechtfertigt sein. Man erhält dann jedoch stets sehr viel größere Stichprobenumfänge für eine nicht im gleichen Maße zunehmende Schätzqualität. Hinzu kommt die Paradoxie eines geringen Auswahlsatzes bei einer großen und damit i.d.R heterogenen Schicht und einem kleinen Auswahlsatz bei einer kleinen und damit i.d.R homogeneren Schicht.

Aus der obigen Übersicht und der folgenden Tabelle folgt auch: will man die Paradoxien von Abschn. 2a und 2b vermeiden und hat man keine Kenntnis von den evtl. sehr verschiedenen Varianzen σ_k^2 innerhalb der Schichten, empfiehlt es sich nach

(19) $\quad n \geq \left(\frac{z\overline{\sigma}}{e}\right)^2$, statt nach (27) den Gesamtstichprobenumfang n zu bestimmen und nach (18) eine proportionale Aufteilung von n vorzunehmen.

Abschließend noch die Auswirkung der Planung mit (2) bzw. (4) für n im ***Beispiel 2***. Man kann die folgenden Umfänge für die Stichproben berechnen ($e^2 = 0{,}01$).[43]

	Schicht 1	Schicht 2	insgesamt
Variante 1 n nach (2) ZmZ	64	96	160
n nach (4) ZoZ	53	84	137
prop. Aufteil. Seite 16	26	60	86
Variante 1 n nach (2) ZmZ	36	84	120
n nach (4) ZoZ	32	75	107
prop. Aufteil. Seite 17	21	49	70
mit $\pi_k(1-\pi_k) = ¼$ in beiden Schichten nach (5a) (beide Varianten)	75	88	163

Jeweils gerundete Werte.

Wie man sieht erhält man recht große Zahlen für die Teilstichproben wenn man nach (2) bzw. (4) oder gar nach (5a) vorgeht. Weil der Stichprobenfehler $\hat{\sigma}^2_{\hat{\pi}_k}$ für die Teilstichproben nicht von n_k, sondern von $\sqrt{n_k}$ abhängt und $\hat{\sigma}^2_{\hat{\pi}}$ auch nicht einfach ein gewogenes Mittel von $\hat{\sigma}^2_{\hat{\pi}_1}$ und $\hat{\sigma}^2_{\hat{\pi}_2}$ ist, kann man auch nicht sagen, dass sich durch die Vergrößerung von n_1, n_2 und $n = n_1 + n_2$ der Stichprobenfehler entsprechend verringert.

4. Der Begriff "Repräsentativität"

In diesem Abschnitt soll versucht werden, systematisch zusammenzustellen, wie das schillernde, unexakte und wohl besser zu vermeidende Wort "Repräsentativität" benutzt wird. In diesem Sinne sollen sechs Repräsentativitätsbegriffe, wie RS, RV,... unterschieden werden. Eine wichtige Quelle, auf die ich hier stark bezug nehme ist neben einem eigenen Aufsatz[44] vor allem ein bereits etwas älterer dreiteiliger Aufsatz von Kruskal und Mosteller (zitiert als KM).[45] Der Abschnitt dürfte klar machen, dass für das, was in der Regel mit "Repräsentativität" gemeint ist, nämlich ein Qualitätsmerkmal einer Stichprobe allein der Stichprobenfehler ein brauchbares Konzept darstellt. Alle anderen Versuche, die Idee der "Repräsentativität" zu operationalisieren sind bei genauerer Betrachtung unbrauchbar.

a) Strukturkonzept der Repräsentativität (RS) und der Stichprobenfehler (SF)

Es ist bemerkenswert, dass der Begriff der "Repräsentativität", der im Alltagssprachgebrauch im Zusammenhang mit Stichproben geradezu als Synonym für Vertrauenswürdigkeit und Seriosität benutzt wird, in Statistik-Lehrbüchern so gut wie nicht vorkommt. Dafür gibt es einige Gründe. Einen besonders naheliegenden sieht man schnell: Es gibt kein Maß R für die Repräsentativität, wonach etwa bei Stichprobe 1 die Repräsentativität $R_1 = 4{,}7$ oder 19,8 beträgt, oder R_1 85% (von was?) ist, oder auch nur einfach $R_1 > R_2$ ist.

Bevor es ein Maß geben kann, sollte es zumindest eine halbwegs exakte Definition dieses Begriffs geben. Aber es gibt noch nicht einmal das, sondern nur eine Reihe unterschiedlicher und meist sehr vager Vorstellungen, die man mit dem Konzept verbindet, und die auch unterschiedlich sind, je nachdem, in welcher Art Text der Begriff gebraucht wird.

---

[43] Mit $e^2 = 0{,}0064$ würde man sehr große, auch N überschreitende Werte für n erhalten.

[44] P. v. d. Lippe u. A. Kladroba, Repräsentativität von Stichproben, in Marketing 24 (2002), S. 227 - 238.

[45] William Kruskal and Frederick Mosteller, Representative Sampling, I: Non-scientific Literature, part II: Scientific Literature Excluding Statistics, part III: The Current Statistical Literature, in: International Statistical Review, 47 (1979), 13 - 24, 111 - 127, 245 -265. Die Arbeit wird zitiert als KM, Seitenzahl.

Die häufigste Vorstellung ist wohl die, dass eine konkrete Stichprobe dann "repräsentativ" ist, wenn die **Struktur** der Stichprobe[46] ähnlich der der Grundgesamtheit ist. Herrscht z.B. in der Grundgesamtheit eine Aufteilung bezüglich des Geschlechts: männlich $\pi = 0{,}5$ und weiblich $1\text{-}\pi = 0{,}5$ und erhält man mit einer Stichprobe $\hat{\pi} = 0{,}5$ oder $\approx 0{,}5$, so gilt diese Stichprobe als repräsentativ und man hält sie ist auch für repräsentativer als eine Stichprobe, die zu einer Aufteilung 40:60 (also $\hat{\pi} = 0{,}4$) oder gar 30:70 führt. Dass an diesem RS-Konzept etwas nicht zu stimmen scheint, wird schnell erkennbar, wenn man sieht, dass danach eine Stichprobe von 3 Männer und 3 Frauen genauso gut wie eine Stichprobe von 30 Männer und 30 Frauen, aber erheblich besser sein müsste als eine von 305 Männer und 295 Frauen sein müsste.[47] Zweifel können einem auch kommen, wenn man sich fragt, welche von zwei Stichprobe bei einem trichotomen statt dichotomen Merkmal besser ist:

Familienstand	Grundge-samtheit	Stich-probe 1	Stich-probe 2
ledig (L)	36%	34%	38%
verheiratet (V)	52%	54%	51%
sonstige (S)	12%	12%	11%

Der Gedanke an ein "getreues Abbild" kann auch eine Mindestgröße der Stichprobe verlangen. Die Grundgesamtheit in der Tabelle kann nicht abgebildet werden mit einer Stichprobe von $n < 25$ weil es 9 L-Einheiten, 13 V- und 3 S-Einheiten sein müssen bzw. ein Vielfaches dieser Zahlen, damit die Proportionen stimmen.[48] Auf weitere Probleme des RS-Konzepts wird später noch hingewiesen. Wie man leicht sieht, steht der RS-Gedanke auch Pate bei der sog. Quotenauswahl oder auch beim Versuch, durch Hochrechnung bei "wichtigen" Merkmalen die Randverteilungen in der Stichprobe zu korrigieren, um sie den Randverteilungen in der Grundgesamtheit ähnlicher zu machen.[49]

Die Rolle, die dem Begriff "Repräsentativität" (der - wie gesagt - in der statistischen Fachliteratur praktisch nicht vorkommt) als Qualitätsnachweis zugedacht ist, spielt in der Statistik der Begriff "Stichprobenfehler" $\sigma_{\bar{x}}$ einer (die Stichprobe) beschreibenden Kennzahl wie etwa ein Mittelwert $\bar{x}$ oder eine Varianz σ^2, bzw. eine Standardabweichung σ. Zwischen RS und $\sigma_{\bar{x}}$ bestehen jedoch ganz erhebliche Unterschiede. Die Standardabweichung $\sigma_{\bar{x}}$ der Stichprobenverteilung von $\bar{x}$ (d.h. der Verteilung von $\bar{x}$ wenn man alle $\binom{N}{n}$ Stichproben vom Umfang n zöge) bezieht sich nicht wie das Konzept RS auf eine einzelne konkrete Stichprobe, sondern auf alle Stichproben vom Umfang n, die man überhaupt aus einer Grundgesamtheit vom Umfang N ziehen kann und ist somit eine Wahrscheinlichkeitsaussage.[50]

---

[46] Wir sprechen deshalb vom Strukturkonzept der Repräsentativität (oder kurz RS-Begriff).

[47] Befragt man 20 Männer und 22 Frauen könnte man nach der RS-Logik zwei (egal welche!) Frauen aus der Stichprobe entfernen und hätte dann bessere Daten. Man kann auch leicht Beispiele konstruieren, bei denen eine Zufallsauswahl (also eine echte Stichprobe, die nach statistischem Verständnis stets "repräsentativ ist) vorgenommen wird und überhaupt keine der Stichproben "repräsentativ im Sinne von RS ist. Im Beispiel mit dem Geschlecht würde jedes ungerade n (etwa n = 193) nur Stichproben liefern die alle nicht "repräsentativ" sind, denn es kann ja nicht 96,5 Männer und 96,5 Frauen geben.

[48] Repräsentativ können allenfalls Stichproben vom Umfang 25, 50, 75 usw. sein.

[49] Es gibt auch Versuche, "Repräsentativität" im Sinne von RS quantifizierbar und überprüfbar zu machen, indem man mit dem χ^2 Test überprüft, ob die Häufigkeitsverteilung der Stichprobe von der (dann als bekannt vorauszusetzenden oder hypothetisch anzunehmenden) Verteilung der Grundgesamtheit signifikant abweicht.

[50] Auch die üblichen Gütekriterien wie Erwartungstreue, Effizienz und Konsistenz sind alle Wahrscheinlichkeitsaussagen, die sich auf eine Stichprobenverteilung beziehen (also alle möglichen Stichproben), nicht auf eine einzelne konkrete Stichprobe, wie RS.

	Stichprobenfehler (SF-Konzept)	"Repräsentativität" (RS-Konzept)
Gütekriterium	$\sigma_{\bar{x}}$ klein (schmales Konfidenzintervall)	ähnliche "Struktur" wie Grundges.
Die Aussage bezieht sich auf ...	eine Wahrscheinlichkeitsverteilung nämlich die Stichprobenverteilung aller mögichen Stichproben, nicht auf eine konkrete (Wahrscheinlichkeitsaussage)	auf eine einzelne konkrete Stichprobe (*nach* Ziehung); keine Wahrscheinlichkeitsaussage über alle Stichproben unter sonst gleichen Umständen
Bedeutung des Zufalls	bei gleicher Grundgesamtheit und Zufallsauswahl sind alle "repräsentativ", auch wenn einzelne Stichproben sehr unterschiedlich sein können	wie eine konkrete Stichprobe ausfällt ist vom Zufall bestimmt; eine Stichprobe kann unter gleichen Umständen repräsentativ sein, eine andere nicht
Grundgesamtheit	Kenntnis der Struktur der Grundgesamth. nicht vorausgesetzt	vergleicht eine einzelne Stichprobe mit Grundges., die aber nicht bekannt ist
Ist das Konzept operational? Gibt es Messwerte?	ein metrisch skaliertes Maß ($\sigma_{\bar{x}}$ ist eine reelle Zahl), so dass quantitative Vergleiche unbegrenzt möglich sind; bezieht sich auf ein Merkmal, näml. x	ein Maß für mehr oder weniger Strukturähnlichkeit zweier Verteilungen ist denkbar aber kontrovers und unüblich; nicht klar auf welche(s) Merkmal(e) sich RS beziehen soll
Ist die Formel für das Konzept plausibel? Was kann man aus ihr ableiten?	$\sigma_{\bar{x}}$ hängt vom Stichprobenumfang n und der Varianz von x [1)] ab und ist auch entscheidend für die Breite des Konfidenzintervalls (es ist somit auch ein anschauliches Konzept)	kein Zusammenhang mit dem Stichprobenumfang; auch aus einem Maß (wenn es dies gäbe) für RS folgt keine Formel für eine Punkt- bzw. Intervallschätzung[3)]

1) Es ist unmittelbar einleuchtend, dass eine kleine oder große Streuung des Merkmals x in der Grundgesamtheit bei gleichem Stichprobenumfang bedeutet, dass die Stichprobe mehr oder weniger "repräsentativ" ist.
2) Es gibt keine Formel, aus der hervorgeht, wie viel "repräsentativer" eine Stichprobe von 50 Frauen und 50 Männern ist als eine Stichprobe von 51 Frauen und 49 Männern.
3) Selbst wenn man wüsste, dass RS für eine konkrete Stichprobe z.B. 4 beträgt, so hätte die Zahl 4 keine Bedeutung für irgendwelche andere Größe (z.B. einen Mittelwert), die aufgrund der Stichprobenwerte zu schätzen ist.

Dem Konzept RS liegt die Vermutung zugrunde, dass eine Auswahl, die in der Lage ist, die Proportionen bezüglich eines Merkmals (also eine Randverteilung) "richtig" wiederzugeben auch geeignet sein dürfte eine "gute" (woran gemessen?) Schätzung eines Mittelwerts $\bar{x}$ zu liefern. Es bleibt offen, worauf sich diese Erwartung stützen soll,[51] insbesondere dann, wenn man sich wegen nichtzufälliger (trotzdem "repräsentativer" im Sinne von RS) Auswahl nicht der Wahrscheinlichkeitsrechnung bedienen kann. Man sieht hier nicht nur, dass das SF-Konzept direkt ansetzt, an dem, worauf es ankommt, nämlich auf die Aufgabe der Schätzung ($\sigma^2_{\bar{x}}$ ist eben ein Maß dafür, ob eine Schätzung von $\bar{x}$ besser oder schlechter ist), es ist auch ein Unterschied ob man gem. RS gleiche Proportionen ("Strukturen") verlangt oder sich damit begnügt, dass gem. SF ähnliche Strukturen *wahrscheinlicher* als ganz unähnliche Strukturen sind. Ähnliche Randverteilungen müssen auch nicht ähnliche Strukturen der gemeinsamen Verteilung bedeuten. Im folgenden Zahlenbeispiel findet man zwei sehr verschiedene gemeinsame (hier bivariate) Verteilung von Familienstand und Alter trotz gleicher Randverteilungen (was eine Frage der Korrelation zwischen den beiden Merkmalen ist):[52]

	jung	alt	Σ
ledig (L)	0,08	0,28	0,36
verheiratet (V)	0,17	0,35	0,52
sonstige (S)	0,07	0,05	0,12
Σ	0,32	0,68	1

	jung	alt	Σ
ledig (L)	0,13	0,23	0,36
verheiratet (V)	0,16	0,36	0,52
sonstige (S)	0,03	0,09	0,12
Σ	0,32	0,68	1

[51] Trotz RS kann eine Stichprobe nicht unerheblich "biased" sein (KM, 249).

[52] Die Tabelle hat bei 3 Zeilen und 2 Spalten nur (3-1)(2-1)=2 Freiheitsgrade. Legt man zwei Zahlen, etwa im linken Fall die Zahlen 0,08 und 0,17 fest, so sind damit alle anderen Zahlen auch gegeben.

Die Zahlen der Randverteilungen stehen in den schattierten Feldern. Das RS Konzept ist so beliebt, weil es offenbar intuitiv verständlich ist. Es ist dagegen sehr viel schwieriger, plausibel zu machen, dass bei gleicher Grundgesamtheit jede (Zufalls-)Stichprobe vom gleichen Umfang n gleich "repräsentativ" ist, obgleich die konkrete Stichprobe (gerade wegen der Zufälligkeit der Ziehung) sehr unterschiedlich ausfallen kann.

Zusätzlich zu den in der Tabelle zusammengestellten Unterschieden zwischen den beiden Konzepten wäre noch zu erwähnen,[53] dass es durchaus Sinn machen kann, eine Auswahl so vorzunehmen, dass die Proportionen ganz bewusst andere sind als in der Grundgesamtheit, z.B. kleine Bundesländer stärker repräsentiert werden sollen als große oder generell eine geschichtete Stichprobe mit nichtproportionaler Aufteilung gezogen wird. Ein anderer Fall bewusster Nichtrepräsentativität" (im Sinne von RS) ist das Konzentrationsprinzip (cutt off principle), wenn z.B. nur Betriebe ab 20 Beschäftigte (und dies zu 100%) befragt werden.[54]

Zusammenfassend könnte man sagen: nach dem SF-Konzept heißt eine Stichprobe "repräsentativ", wenn gilt "that it ordinarily can be used for responsible inference" (KM, 113) während beim RS Konzept überhaupt keinen Bezug zur Wahrscheinlichkeitsrechnung und zu einer darauf aufbauenden statistischen Inferenz hat. Ohne ein aus dem Schätzproblem abgeleiteten Maß ist die Aussage, eine Untersuchung sei "repräsentativ" nur ein verbales Urteil: "the concept of representativeness is used primarily as an assertive talisman, or as means of sounding more scientific" (KM, 16).

b) Das Miniaturkonzept der Repräsentativität (RM)

Nach diesem besonders vagen und daher ziemlich unbrauchbaren Konzept sollte die Stichprobe eine "getreue" Verkleinerung der Grundgesamtheit sein. Wenn mit "getreu" die gleichen Strukturen (Verteilungen) bezüglich der Merkmale gemeint sind, ist das Konzept natürlich mit RS verwandt. Bei RS geht es um die Verteilung von Merkmalen, bei RM (und ähnlichen Konzepten, die noch dargestellt werden) auch um die Anwesenheit oder Abwesenheit bestimmter Einheiten in der Auswahl damit diese "repräsentativ" ist.[55]

Das Konzept RM ist - wie gesagt - im besonderen Maße vage und unklar. Es lässt offen, wie beurteilt werden soll, ab wann man von einer "Miniatur" sprechen kann, es legt sich auch nicht fest (wie die nächsten beiden Konzepte, RV und RA), welche Einheiten in die Auswahl gelangen sollten.

Des Weiteren ist das Konzept RM auch nicht hilfreich, um daraus Aussagen über den notwendigen Stichprobenumfang n abzuleiten. Kann bei N = 1000 eine Auswahl von n = 50 eine akzeptable Miniatur sein, oder reicht vielleicht schon n = 10? Wenn 10 ausreicht, wie erkennt man, dass die Miniatur nicht schlechter ist als bei n = 50?[56]

---

[53] Vgl. hierzu insbes. von der Lippe u. Kladroba.

[54] Man könnte von einem Grenzfall der geschichteten Stichprobe sprechen. Aber bei "Auswahlsätzen" von 100% und 0% kann man nicht mehr von einer "Auswahl" sprechen, weshalb das Konzentrationsprinzip mit einer Abschneidegrenze (z.B. 20 Beschäftigtre) auch nicht zu den Verfahren der Zufallsauswahl (das kennzeichnend ist für echte Stichproben) gerechnet wird.

[55] Der Gedanke einer Miniaturausgabe der Grundgesamtheit spielt eine Rolle bei der Klumpenauswahl, weil dort Klumpen zu 100% ausgezählt werden, also eine Auswahl nur auf der ersten Stufe (Auswahl der Klumpen) stattfindet, nicht auf der zweiten (Einheiten innerhalb des Klumpens). Von einem Klumpen (z.B. einer ausgewählten Gemeinde) wird gefordert, dass er in sich inhomogen sein soll, d.h. die Vielfalt der Grundgesamtheit auch in ihren richtigen Proportionen wiedergeben soll. Im Gegensatz dazu sollte eine Schicht k in sich homogen sein (ein Schichtungsgewinn ist zu erwarten wenn die σ_k klein sind), also gerade keine Miniatur sein.

[56] Der Gedanke, dass es bei jeder Miniatur vielleicht eine noch kleinere Miniatur geben mag, kann bei physischen Objekten zutreffend sein (das immer noch kleinere Spielzeugauto) ist aber wohl nicht direkt übertragbar auf das Stichprobenproblem. Man findet in der Literatur in diesem Zusammenhang aber auch oft den Homunkulus Gedanken zitiert, d.h. die früher einmal für möglich gehaltene Vorstellung, wonach der Mann in seinen

Auf Einheiten Bezug nehmende Konzepte haben Schwierigkeiten in Fällen, in denen es eine Auswahl von unterscheidbaren "Einheiten" kaum gibt, wie z.B. die zufällige Entnahme einer bestimmten Menge Flüssigkeit aus einem Tank. Dabei dürfte man mit der Vorstellung einer Miniaturausgabe der Grundgesamtheit gerade im Fall einer Stichprobe aus einer gut durchmischten Flüssigkeit noch relativ wenig Probleme haben, weil hier wohl in einem höheren Maße von Homogenität gesprochen werden kann als z.B. bei einer Personengesamtheit.

Die kritische Frage ist beim Konzept RM wie bei RS, wie genau man es mit der Ähnlichkeit zwischen der Grundgesamtheit und dem Miniaturexemplar der Grundgesamtheit meint. Betrachtet man nur ein Merkmal mit einer groben Klassifizierung der Merkmalsausprägungen (z.B. Geschlecht mit den Ausprägungen männlich und weiblich) mögen bereits zwei Menschen eine Miniatur darstellen. Verlangt man dagegen von einem "getreuen Abbild", auch dass es den Umstand "getreu" wiederspiegelt, dass in der Grundgesamtheit viele ältere berufstätige und alleinerziehende Frauen gibt, dagegen wenige jüngere blonde Männer, die einen Sportwagen fahren, dann wird man mit $n = 2$ nicht mehr auskommen. Je mehr Aspekten Rechnung zu tragen ist damit die Stichprobe im Sinne von RM "repräsentativ" ist, desto größer muss die Miniatur werden, bis sie schließlich nicht kleiner ist als das Original.

Beim im Folgenden behandelten Konzept RV ist es dagegen genau umgekehrt. Aus ihm folgt, dass eine im Sinne von RV "repräsentative" Stichprobe eher relativ klein sein müsste. Zuvor jedoch noch eine Bemerkung zum Unterschied zwischen RS und RM. Er besteht - wie gesagt - darin, dass RS mehr Anhaltspunkte zur Unterscheidung zwischen mehr oder weniger "Repräsentativität" liefern dürfte als RM. Man kann versuchen, quantitative Konzepte für mehr oder weniger gleiche "Strukturen" zu finden (z.B. ähnliche relative Häufigkeiten für bestimmte Merkmalsausprägungen), aber es dürfte schwer sein, zu entscheiden ob etwas mit mehr oder weniger Berechtigung als "Miniatur" akzeptiert werden kann.

Beide Konzepte, RM und RS (sowie auch alle anderen noch darzustellenden Konzepte) haben gemeinsam, dass sie keinen Zusammenhang zum Schätzproblem erkennen lassen. Selbst wenn man sich sicher sein könnte, dass die Auswahl eine "Miniatur" darstellt, weiß man deshalb noch nicht wie groß $\sigma_{\bar{x}}$ und damit die Breite des Konfidenzintervalls von μ ist oder ob $\bar{x}$ signifikant verschieden von einem bestimmten $\mu = \mu_0$ ist.

Es ist schließlich auch zu bedenken *wie* die Miniatur gebildet wird: "...the idea of a sample as a mirror or miniature of the population is rarely appropriate… a miniature is usually constructed purposefully rather than through a process of probability sampling." (KM, 120).

c) Das Stellvertreter (Vize) Konzept der Repräsentativität (RV)

Man kann unter "repräsentativ" auch - ausgehend von der wörtlichen Übersetzung - verstehen, dass die ausgewählten Einheiten die nichtausgewählten vertreten ("repräsentieren") können. Um das zu können, müssten sie den Nichtausgewählten weitgehend gleich oder zumindest "ähnlich" sein. Das dürfte jedoch schwer zu beurteilen sein, weil man die Nichtausgewählten nicht kennt, und selbst wenn man sie kennen würde, wäre es schwer zu sagen, welches Maß (!) an Ähnlichkeit ausreicht um jemand gleichwertig zu vertreten.

Spermien alle seine Nachkommen in Miniaturform (Homunkulus) in sich trägt. Wenn das richtig wäre, müsste er genau genommen auch noch die zweite, dritte,... Generation in sich tragen, jede jeweils in entsprechend noch kleinerer Miniaturform als die vorangegangene Generation. Eine vergleichbare Situation hat man, wenn man bedenkt, dass es in der Regel legitim ist, aus einer Stichprobe wiederum eine Stichprobe zu ziehen, und daraus auch wieder eine Stichprobe (mehrphasige Auswahl). Ist die n-te (Sub-) Stichprobe ebenfalls "repräsentativ" im Sinne des RM-Konzepts, kann man einwenden, dass die jeweils vorletzte (also (n-1)-te) Stichprobe noch nicht die gewünschte Miniatur ist, sondern noch "eine Nummer zu groß ist". Nach der Logik des RM Konzepts müsste jeweils die Miniatur der Miniatur als noch repräsentativer gelten.

Von einer "repräsentativen" Einheit kann auch verlangt werden, dass sie stellvertretend für alle (nicht nur für nichtausgewählte Einheiten) fungieren kann, weil sie "typisch"[57] für alle Einheiten der Grundgesamtheit ist. Das Problem dieses Konzepts ist, dass es auf den Durchschnitt bezüglich aller relevanter Merkmale abstellt (jemand ist "typisch" für alle, wenn er in jeder Hinsicht dem Durchschnitt entspricht), aber

- die Kenntnis der Verteilung (und damit auch der Durchschnitte) der interessierenden Merkmale in der Grundgesamtheit ist unvollständig und man nimmt gerade deshalb eine Stichprobenuntersuchung vor,[58] und
- gäbe es real eine Einheit[59] (z.B. eine Person), die in jeder Hinsicht genau dem Durchschnitt entspricht, so würde eine Stichprobe mit n = 1, also dieser einen Einheit, ausreichen, um "repräsentativ" zu sein, solange es bei der Stichprobe um die Feststellung von Mittelwerten bezüglich der Merkmale geht
- geht es bei der Stichprobe hingegen auch um die Feststellung der Streuung der Merkmale, dann wäre eine Stichprobe von lauter gleichermaßen "typischen" Einheiten (wie sie im Konzept RV gefordert wird) ohnehin völlig ungeeignet, weil das ja eine Varianz von null für die Stichprobe "impliziert". Es wären ja gerade die von den Mittelwerten abweichenden Einheiten vielleicht in der Grundgesamtheit, nicht aber in der (nach dem Konzept RV idealen) Stichprobe vorhanden. Eine Stichprobenvarianz von $\hat{\sigma}^2 = 0$ ist kein geeigneter Schätzer für eine Varianz $\sigma^2 > 0$ in der Grundgesamtheit.

Stattdessen könnte man an Repräsentativität im Sinne einer korrekten (vollständigen) Wiedergabe der in der Grundgesamtheit vorhandenen *Vielfalt* denken. Das führt zum folgenden Konzept der "Repräsentativität"

d) "Coverage" oder Arche-Noah Konzept der Repräsentativität (RA)

Repräsentativität im Sinne von "coverage" eines samples "might be defined to mean inclusion in the sample of at least one member of each class" (KM, 14), so wie in der Arche von jeder Tiergattung wenigstens ein Exemplar vorhanden war.[60] Das Prinzip ist hier "Selektivität" (bewusste Auswahl) und scheint insofern quasi das Gegenteil einer Zufallsauswahl zu sein. Tatsächlich schließt aber das eine das andere nicht aus. Eine Kombination von Zufallsauswahl mit selektiven Elementen liegt z.B. bei einer geschichteten Stichprobe vor. Mit der geschichteten Stichprobe kann nämlich verhindert werden, dass eine bestimmte Schicht der Grundgesamtheit überhaupt nicht in der Stichprobe vertreten ist. Insofern ist Selektivität und Zufallsauswahl nicht ein Widerspruch.

Im Unterschied zu RV wird man mit diesem Konzept (also RA) eher für einen großen Stichprobenumfang plädieren. Fordert man, sich nur auf typische Einheiten zu beschränken (im Sinne von RV) wird man ein kleines n bevorzugen.[61] Das dürfte intuitiv nicht sehr überzeugend sein. Das Plädoyer für ein großes n, wie auf der Basis des RA Konzepts steht demge-

---

[57] "Typicalness is an ambiguous idea: it might mean a sample each of whose members was itself typical, or it might mean a sample typical as a whole. The latter sense is close to the miniature idea; the former sense requires explication of typicalness for a single unit…" und endet dann meist beim Fall n = 1, also dem "case study approach" (KM, 253), dem man aus guten Gründen nicht uneingeschränkt vertraut.

[58] Dieser Punkt gilt bei allen Konzepten die an der Auswahl statt am Auswahlprozess ansetzen (also bei RS, RM RV und RA). Sie alle setzen Kenntnisse über die Grundgesamtheit voraus, wie sie in der Regel nicht, oder nicht im geforderten Maße vorhanden sind.

[59] Eine entsprechend *konstruierte* Einheit kann es ohnehin nur einmal geben.

[60] Dieser Gedanke steht auch Pate bei der "repräsentativen" Demokratie, bei der im Parlament alle Schichten der Bevölkerung vertreten sein sollten. Man würde es als nicht im Einklang mit diesem Prinzip beklagen, wenn im Parlamente nur Beamte und Gewerkschaftsfunktionäre vertreten wären.

[61] "The idea of typicalness naturally leads to samples of one." KM, 121.

genüber mehr im Einklang mit dem "gesunden Menschenverstand", nach dem grundsätzlich bessere Ergebnisse wenn man den Stichprobenumfang vergrößert.

Dem RA Konzept würde es entsprechen wenn bei einer geschichteten Stichprobe jede der K Schichten mit einer und nur einer ausgewählten Einheit repräsentiert wird (vorausgesetzt, dass die Schichtbildung so ist, dass Heterogenität in genau der Differenziertheit interessiert, wie sie der Unterscheidung in K Schichten zugrunde liegt). Dieser Gedanke ist aber der geschichteten Stichprobe völlig fremd. Es ist im Gegenteil die Regel, dass die Stichprobenumfänge n_1, n_2 ,..., n_K nicht alle gleich groß sind oder gar $n_1 = n_2 =... = n_K = 1$ betragen.

e) Das Nichtselektivitätskonzept der Repräsentativität (RN)

Nach diesem Konzept liegt Repräsentativität vor bei "absence of selective forces" (KM, 14), also von Einwirkungen, um die Auswahl einer bestimmten Einheit zu fördern oder zu behindern, bis hin zur Sicherstellung oder Verhinderung der Auswahl einer ganz bestimmten Einheit. Das Problem von RN ist aber, dass man nie sicher sein kann, dass wirklich jede Selektivität ausgeschlossen ist.

Aus dieser Überlegung heraus, und um nicht bei jeder Einheit die Unparteilichkeit der Auswahl "beweisen" zu müssen, erhält das Prinzip der Zufallsauswahl seine überragende Bedeutung. Entscheidend ist die "Blindheit" gegenüber der konkreten Einheit. Es kommt bei einer Zufallsauswahl nicht darauf an, welche konkrete Einheit ausgewählt wird, sondern welche Auswahlwahrscheinlichkeit die Einheiten gleichen Typs haben. Entscheidend ist der Auswahlmechanismus, der eine a priori bestimmte Auswahlwahrscheinlichkeit garantieren muss, nicht die konkret erfolgte Auswahl. Der Weg ist das Ziel. Repräsentativität ist die Qualität eines Prozesses, nicht eines Ergebnisses.

Aber, wie bereits beim RA-Konzept angedeutet, ist Nichtselektivität nicht gleichbedeutend mit Zufallsauswahl; denn es ist durchaus möglich, eine Zufallsauswahl mit Elementen der bewussten Auswahl (purposive oder judgmental selection) zu verbinden (und umgekehrt kann auch eine ausgesprochene Selektivität, wie Auswahl "typischer" Elemente "repräsentativ" im Sinne von RA sein, und auch generell zu brauchbaren Ergebnissen führen oder sogar aus praktischen Gründen geboten sein[62]).

Eine Kombination von bewusster Auswahl und Sicherstellung der Nichtselektivität durch *Zufalls*-auswahl ist die geschichtete Stichprobe. Deswegen können bei einer geschichteten Stichprobe auch die typischen (nicht mit dem Stichprobenfehler $\sigma_{\bar{x}}$ gemessenen) Fehler einer bewussten Auswahl auftreten, nämlich

- nicht ausreichende oder nicht aktuelle *Kenntnisse* über die Grundgesamtheit
- unzutreffende, subjektive *Annahmen* über die Verteilung von Merkmalen in der Grundgesamtheit (z.B. Wahl eines "falschen", weil mit dem Untersuchungsmerkmal wenig korrelierten Schichtungsmerkmals).

Das bewusste Verhindern einer Nichtziehung eines Elements aus der k-ten Schicht (bei einer geschichteten Stichprobe wenn $n_k > 0$) kann Repräsentativität im Sinne von einem geringen Stichprobenfehler $\sigma_{\bar{x}}$ verbessern im Vergleich zum Fall einer "einfachen" (nicht geschichteten) Stichprobe, bei der quasi alle Elemente in einen Topf geworfen werden. Das gilt insbesondere wenn die Schichten in sich homogen sind. Man kann also nicht einfach Nichtselektivität und Repräsentativität gleichsetzen, wie es das RN Konzept tut.

---

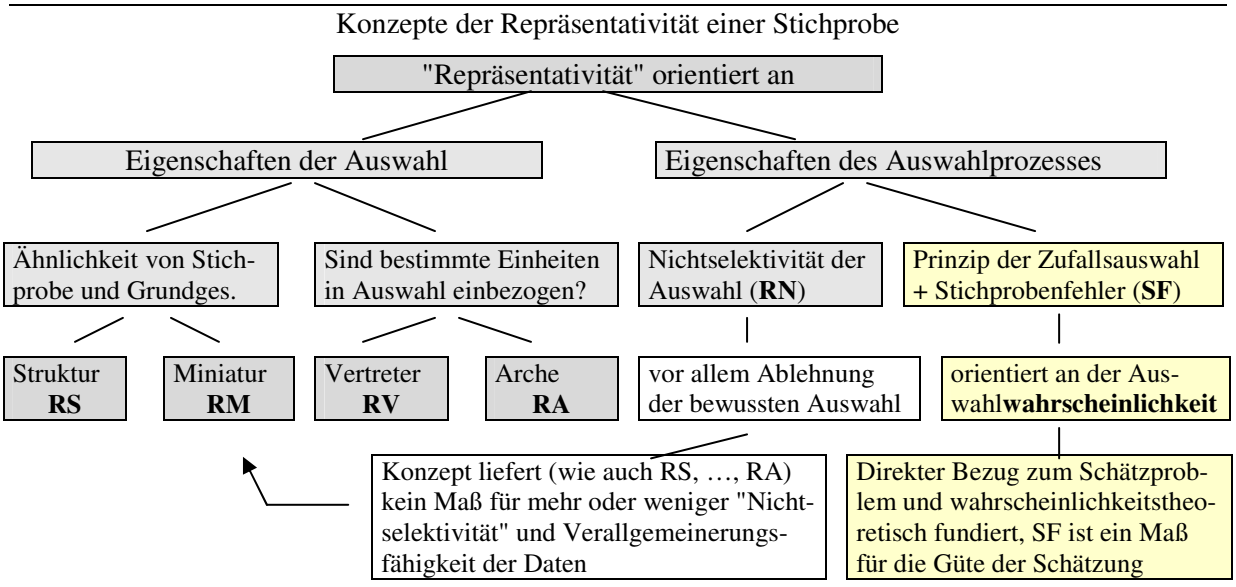
[62] Ein Beispiel für eine üblicherweise bewusste Auswahl ist die Auswahl von Waren und Geschäften bei Verbraucherpreisindizes.

Eine im Sinne von RN (Vermeidung von Einseitigkeit, wie auch bei RA) "repräsentative" Auswahl kann auch gleichwohl selektiv sein wenn die Auswahlgesamtheit einer Selbstselektion unterliegt (z.B. Teilnehmer bei Diskussionsrunden im Internet, Anrufer beim Rundfunk im Fall von Meinungsfragen). Selbstselektion tritt auch auf in Form von Nichtbeantwortung (alle nichtzufällige Antwortausfälle) auf. Zu damit zusammenhängenden Fragen vgl. Abschn. 5 dieses Papiers.

f) Zufallsauswahl und Stichprobenfehler (SF)

Zufall wird gerne mit Willkür verwechselt und es nicht einfach, den Unterschied zwischen einer zufälligen und einer willkürlichen Auswahl (man befragt z.B. irgendwelche Passanten, der gerade vorbeikommen) deutlich zu machen. Kennzeichnend für eine Zufallsauswahl (random sample) ist, dass für jede Einheit a priori (vor Ziehung der Stichprobe) die Wahrscheinlichkeit, in die Stichprobe zu gelangen (also die Auswahlwahrscheinlichkeit) bekannt ist[63] (bei einer einfachen, ungeschichteten Stichprobe ist sie für alle Einheiten gleich). Nur dann kann man die Wahrscheinlichkeitsrechnung anwenden, nur auf diese Art der Auswahl bezieht sich die "Stichprobentheorie" und auch nur bei dieser "echten" Stichprobe sollte man deshalb von einer "Stichprobe" sprechen.

Nur bei einem random sample kann man auch sinnvoll den absoluten Stichprobenfehler (SF), also $\sigma_{\bar{x}}$ oder den relativen Fehler RSF, also $\sigma_{\bar{x}}/\bar{x}$ berechnen und einen geringen RSF als Ausweis von "Repräsentativität" betrachten. Alle anderen Konzepte der "Repräsentativität" sind demgegenüber, wie gezeigt wurde, problematisch, wenn nicht schlicht unbrauchbar. Es mag nützlich sein, an dieser Stelle die Überlegungen in einer Übersicht zusammenzufassen:



g) Bedeutung des Auswahlrahmens für die Repräsentativität

Bei allen Versuchen, ein operationales Konzept zu finden für das, was unter Repräsentativität im Sinne eines Gütekriteriums verstanden werden könnte sollte man nicht vergessen, dass die Qualität des Auswahlrahmens für eine Stichprobe entscheidend ist. Geht es beispielsweise um eine Untersuchung über die Verbreitung einer bestimmten Krankheit, macht es einen großen Unterschied, ob die Auswahl aus einer Patientendatei (in die man ganz ohne eine Krankheit

[63] Es ist nicht bekannt, was die Wahrscheinlichkeit bestimmt, als Passant bei einer willkürlichen Auswahl gerade im rechten Zeitpunkt vorbei zu kommen.

kaum gelangt), dem Mitgliederverzeichnis eines Vereins (ob Sportverein oder Heimatverein dürfte auch durchaus relevant sein) oder aus dem Telefonbuch erfolgt. Es kann bei einem ungeeigneten Auswahlrahmen auch eine Zufallsauswahl nicht "repräsentativ" sein.

Im Prinzip liegt hier ein nicht-stichprobenbedingter Fehler in Gestalt einer fehlerhaften Feststellung der Grundgesamtheit vor (mehr zu dieser Art Fehler im folgenden Abschnitt 5).

Man kann auch generell von "Repräsentativität" (egal nach welchem Konzept definiert und wie die Auswahl vorgenommen wurde) kaum sprechen und streng genommen auch die üblichen Berechnungen von Konfidenzintervallen und Testgrößen nicht durchführen, wenn nicht klar gesagt werden kann, was die Grundgesamtheit sein könnte, die hier "repräsentiert" werden soll. Gegen diese Forderung wird besonders gerne verstoßen. Bei Zeitreihen in der Ökonometrie ist es z. B. üblich, die Werte für aufeinanderfolgende Jahre (eben einer Zeitreihe), etwa 1995, 1996, 1997 ... als Stichprobe zu betrachten, obgleich nicht klar ist, wer sie, wie, und aus welcher Grundgesamtheit gezogen hat und warum er nicht auch zufällig 1995, 1743, 1832, ... gezogen haben könnte. Die Lehrbuchweisheit lautet: die Grundgesamtheit sollte zeitlich, räumlich und sachlich genau abgegrenzt sein. Was ist die zeitliche Abgrenzung wenn man eine Zeitreihe als Stichprobe auffasst?

5. Nichtbeantwortung, Fehler und Hochrechnung

In diesem Abschnitt können einige Probleme nur kurz angerissen werden. Insbesondere der Umgang mit Antwortausfällen und die Hochrechnung (oft gedacht als Verfahren, um "Repräsentativität" [wohl im Sinne von RS] herzustellen) kann hier aus Platzgründen nicht vertieft werden.

a) Fehlerarten

Betrachtet man "Repräsentativität" als geringer absoluter oder relativer Stichprobenfehler SF (bzw. RSF) so ist zu beachten, dass

- es bei Stichproben auch andere Fehler als "nur" stichprobenbedingte Fehler (die mit $\sigma_{\bar{x}}$ gemessen werden) gibt, und
- einer dieser nichtstichprobenbedingter Fehler die Nichtbeantwortung (gemessen als geringe Antwortquote oder "Ausschöpfungsquote") ist, die auch bei einer korrekten Zufallsauswahl auftreten kann und die Aussagefähigkeit einer (vor allem freiwilligen) Befragung erheblich beeinträchtigen kann.

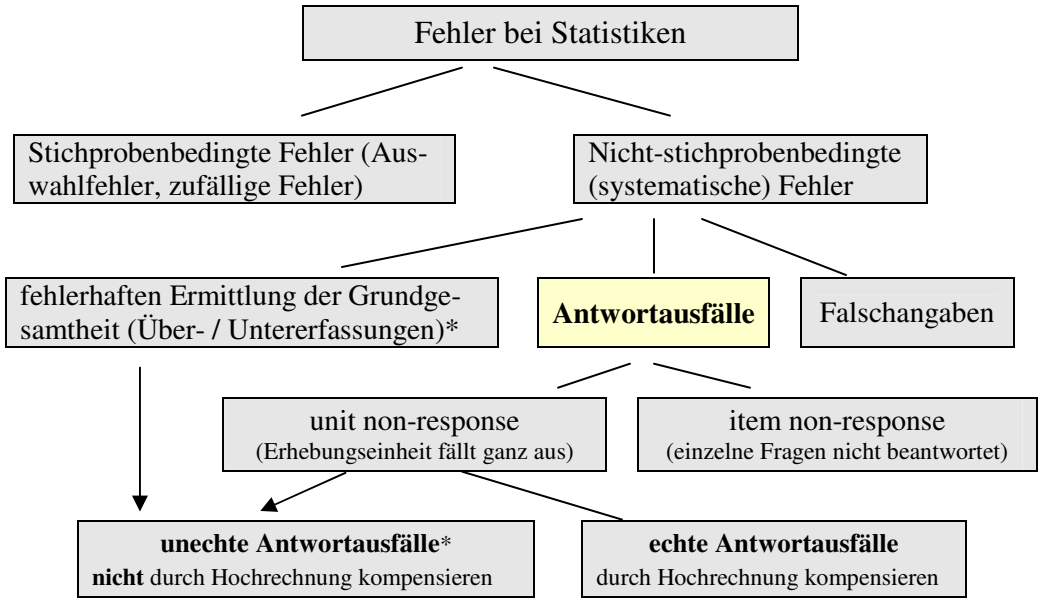
Man mag sich fragen, ab welcher Non-response-Quote man vielleicht nicht mehr von einer Zufallsauswahl sprechen kann. Denn das Argument wäre ja, dass man eigentlich nur eine Zufallsauswahl aus den Antwortwilligen und nicht eine Zufallsauswahl aus allen Befragten gezogen hat. Auch das ist eine Frage, auf die man in Statistik Lehrbüchern meist keine Antwort finden wird. Wäre die Verteilung der interessierenden Merkmale in der Grundgesamtheit der Nichtantwortenden exakt gleich der Verteilung bei den Antwortenden könnte man argumentieren, dass hier nur eine ungewollte Verringerung des Stichprobenumfangs vorliegt.[64]

Stichprobenbedingte Fehler (gemessen als Stichprobenfehler $\sigma_{\bar{x}}$) bestehen darin, dass bei einer Stichprobe nicht alle Einheiten in die Stichprobe einbezogen sind. Man nennt sie deshalb auch "Auswahlfehler". Die Besonderheit einer (echten) Stichprobe ist dass der Auswahl-

[64] Um nicht in die Fallstricke des RS oder RM Konzepts der Repräsentativität zu geraten könnte man vielleicht besser formulieren: Wäre es nur vom Zufall bestimmt, ob man der einen oder der anderen Gruppe angehört. Im Unterschied zum RS liegt auch bei einer "bloßen" ungewollten Verringerung des Stichprobenumfangs n eine Verschlechterung im Sinne eines größeren Stichprobenfehlers vor.

fehler ein Zufallsfehler ist, also mit den Gesetzen der Wahrscheinlichkeitsrechnung zu quantifizieren ist.[65]

Nicht-stichprobenbedingte Fehler sind demgegenüber neben echten Antwortausfällen z. B. auch Falschangaben (vgl. die folgende Übersicht). Sie sind i.d.R. im Unterschied zu stichprobenbedingten Fehlern nicht-zufällig (sondern "systematisch"[66]) und auch schwerer zu quantifizieren. Durch Einsatz von Plausibilitätskontrollen und Rückfragen können Falschangaben weitgehend erkannt und korrigiert werden.



* Fehlerhafte oder nicht aktualisierte Erhebungsregister, z.B. durch nicht erfasste Neugründungen, Weiterführen von erloschenen Einheiten ("Karteileichen", übrigens ein von Statistikern "erfundener" Begriff), Fehlzuordnungen (zu Wirtschaftszweigen) und nicht berücksichtigte Umwidmungen. Untererfassungen bzw. Übererfassungen bedeuten, dass eine Einheit zur Zielgesamtheit gehört (nicht gehört) aber im Auswahlrahmen nicht ist (trotzdem vorhanden ist)

Bei einer Untererfassung spricht man auch von "unechten Antwortausfällen". Sie bleiben – anders als echte Antwortausfälle – in der Hochrechnung unberücksichtigt.[67]

b) Echte Antwortausfälle (non-response) und Hochrechnung

Echte Antwortausfälle[68] können – *wenn sie zufällig auftreten* (verteilt sind) – im Rahmen der "Hochrechnung" durch Korrektur (Erhöhung) des Hochrechnungsfaktors der Erhebungseinheiten der gleichen Ziehungsschicht "eingeschätzt" werden (Zuschätzung echter Antwortausfälle).[69]

---

[65] Bei anderen Arten der Teilerhebung gilt das nicht. Der Vorteil, den Auswahlfehler quantifizieren zu können besteht nur wenn eine Zufallsauswahl vorliegt, also bei einer (echten) Stichprobe.

[66] Das heißt u.a. auch, dass sie sich "in der großen Zahl" nicht ausgleichen.

[67] Daraus resultiert, dass bei Vorliegen unechter Antwortausfälle die hochgerechneten Ergebnisse der Erhebung in der Fallzahl (Anzahl der Einheiten) immer niedriger als die der Zielgesamtheit sind. Das Auftreten von unechten Antwortausfällen in den Schichten bewirkt eine Verstärkung der Merkmalsstreuung und damit ein Anwachsen von durch die zufällige Auswahl der Stichprobeneinheiten bewirkten Schätzfehlern.

[68] Wenn keine unechten Antwortausfälle vorliegen erhält man sie als Differenz zwischen Fragebogenversand und Fragebogenrücklauf. Neben echten Antwortausfällen auf der Ebene der Einheiten (unit non-response) gibt es auch Antwortausfälle auf der Ebene einzelner Fragen (Merkmale), die jedoch weniger bedeutsam sind wenn sie nicht zur Unbrauchbarkeit der Angaben insgesamt und damit zu einem unit non-response Fall führen.

[69] Diese Vorgehensweise ist immer dann verzerrungsfrei, wenn das Auftreten der echten Antwortausfälle innerhalb der Schicht als Zufallsereignis angesehen werden darf.

Dadurch ändert sich faktisch das n (absolute Häufigkeit) eines Tabellenfelds[70] zu n* = n·f um den Faktor f (es ist somit größer oder kleiner als die Anzahl n der Einheiten in der Stichprobe), was sich auf den Stichprobenfehler $\sigma_{\bar{x}}$ auswirkt. Will man, zur Beurteilung der Güte einer Stichprobe den Stichprobenfehler ausweisen, so ist es ein Unterschied, ob dies vor oder nach Hochrechnung geschieht.

Üblicherweise stellen aber echte Antwortausfälle *systematische* Fehler dar, die nicht durch Hochrechnung[71] ausgeglichen werden können, weil sie in verschiedenen Schichten auch unterschiedlich häufig sind, d.h. auch innerhalb einer Schicht häufiger auftreten als in einer anderen Schicht.

Das Nonresponse-Problem (oder auch das der "selektiven" panel attrition) besteht darin, dass über die nichtteilnehmenden Einheiten (Subskript m = missing) keine Informationen außer denen des sampling frames vorliegen und die Gruppe anders strukturiert sein kann als die Gruppe der teilnehmenden Einheiten (Subskript r = responding). Die Umfänge in Grundgesamtheit und Stichprobe setzen sich wie folgt zusammen[72]

n = r + m in der Stichprobe und

N = R + M in der Grundgesamtheit

entsprechend gilt für die Mittelwerte in der Stichprobe

$$\bar{y}_n = \frac{r}{n}\bar{y}_r + \frac{m}{n}\bar{y}_m , \tag{28}$$

und es gilt $\bar{y}_n$ zu schätzen obgleich nur $\bar{y}_r$ bekannt ist. Nichtbeachtung von $\bar{y}_m$, also Inferenz allein mit $\bar{y}_r$ ist nur zu rechtfertigen wenn gilt $\bar{y}_m = \bar{y}_r$. Das Problem ist aber, dass man bei nennenswerter Nichtbeantwortung – wie gesagt – eigentlich keine Stichprobe aus der Grundgesamtheit, sondern nur aus der Gesamtheit der Antwortbereiten gezogen hat. Außerdem ist auch zu beachten, dass wegen

$$\hat{\sigma}^2_{\bar{y}_n} = \left(\frac{r}{n}\right)^2 \hat{\sigma}^2_{\bar{y}_r} + \left(\frac{m}{n}\right)^2 \hat{\sigma}^2_{\bar{y}_m} + 2 \cdot \frac{r}{n} \cdot \frac{m}{n} \cdot \left(\hat{\rho}_{\bar{y}_r \bar{y}_m} \hat{\sigma}_{\bar{y}_r} \hat{\sigma}_{\bar{y}_{rn}}\right) \tag{29}$$

die Varianz (und damit der Stichprobenfehler) von $\bar{y}_n$ falsch eingeschätzt wird, wenn man nur mit der Varianz von $\bar{y}_r$ rechnet. Für die Korrelation ρ der Schätzungen von $\bar{y}_r$ und $\bar{y}_m$ gilt natürlich $-1 \le \rho \le +1$. Setzt man für $\hat{\sigma}^2_{\bar{y}}$ jeweils den Quotient aus Varianz und Stichprobenumfang (Ziehen mit Zurücklegen) ein, so erhält man

$$\hat{\sigma}^2 = \frac{r\hat{\sigma}^2_r + m\hat{\sigma}^2_m}{n} + \frac{2\sqrt{rm}}{n} \cdot \hat{\rho}_{\bar{y}_r \bar{y}_m} \hat{\sigma}_r \hat{\sigma}_m , \tag{30}$$

so dass bei non-response nicht nur der Stichprobenfehler von $\bar{y}$, sondern auch die Varianz von y falsch eingeschätzt werden dürfte, und zwar selbst dann, wenn die Antwortausfälle zu-

---

[70] Es sei denn, alle Merkmalskombinationen (Tabellenfelder) werden mit dem gleichen Hochrechnungsfaktor "hochgerechnet".

[71] "Hochrechnung" ist ein Begriff, der praktisch nur in der deutschen statistischen Literatur vorkommt. Es gibt im Englischen nicht so etwas wie eine "high calculation", sondern nur eine Punktschätzung (point estimation). Gemeint ist mit "Hochrechnung" die Schätzung einer Merkmalssumme (in der Grundgesamtheit) aufgrund eines Mittelwerts in der Stichprobe oder einer Anzahl N_s von Einheiten aufgrund eines Anteils n_s/n in der Stichprobe.

[72] Die entsprechende Betrachtung wird ungleich komplizierter wenn man auch zwischen non-response und einem withdrawal of a panelist (also dem Problem der Panelmortalität, dass eine bekannte Einheit, die früher teilgenommen hat, aber bei dieser Welle nicht mehr antwortet) unterscheiden will.

fällig sind und sogar $\bar{y}_m = \bar{y}_r$ gilt. Man beachte auch, dass sich zwar r/n und m/n zu eins addieren, nicht aber $(r/n)^2$ und $(m/n)^2$.

Für die Punktschätzung (von μ [oder μ_y] aufgrund von $\bar{y}$) und für die "Hochrechnung" von Merkmalssummen (von Σy_j [j = 1, …, N] in der Grundgesamtheit aufgrund von Σy_i [i = 1, …, r] in der Stichprobe) sind drei Parameter relevant, nämlich die

- Nonresponse-Quote[73] m/n (zusammen mit dem Auswahlsatz n/N), die
- Unterschiedlichkeit der Mittelwerte $\bar{y}_r$ und $\bar{y}_m$ und die
- Kovarianz zwischen dem Response-Indikator ρ_i und der zur Diskussion stehenden Variable y_i wobei gilt $\rho_i = \begin{cases} 1 \text{ wenn Teilnahme} \\ 0 \text{ keine Teilnahme} \end{cases}$

die für den "non-response bias" verantwortlich sind.

Die üblichen –aber nicht unproblematischen – Methoden im Umgang mit Nonresponse sind

- *ignorieren*
- *redressment* (mit geeigneten Gewichtungsverfahren) und
- *imputation* (Schätzung eines oder mehrerer Werte für den fehlenden Wert
  [eine naheliegende Option ist z.B. den Mittelwert der Teilgesamtheit zu der die missing unit gehört als Schätzung der missing observation einzusetzen]).

Welche Methode angemessen ist hängt vor allem von der Nonresponse-Quote[74] und der Zufälligkeit/Nichtzufälligkeit (oder Korreliertheit mit der Response-Gruppe mit der Non-Response-Gruppe) des Nichtantwortens ab. Aus Platzgründen können wir auf weiterte Einzelheiten in diesem Papier nicht eingehen.

Anhang

1. Relative Fehler und Variationskoeffizient

Eine Variante von (1) ist

$$(A1) \qquad n \geq \frac{z^2 \sigma^2}{e^2} = \frac{z^2 V^2}{e^{*2}}$$

mit dem Variationskoeffizient $V = \hat{\sigma}/\bar{x}$ und dem *relativen* Fehler $e^* = e/\bar{x}$. Es mag evtl. einfacher sein, den Fehler als relativer Fehler (in Prozent des Mittelwerts auszurücken) und entsprechend auch die Standardabweichung in Form des Variationskoeffizienten zu relativieren. Hat man z.B. eine mehr oder weniger vage Vorstellung, dass vielleicht ca 68% der Werte zwischen x_{1u} und x_{1o} liegen könnten (etwa zwischen 70 und 90) und dass x normalverteilt sein könnte,[75] dann ist V = 10/80 = 0,125 (also 12,5%) und ein relativer Fehlere von 5% (e* = 0,05

[73] leider oft auch Nonresponse-Rate genannt, wie ja überhaupt heutzutage "Rate" und "Quote" ganz nach Belieben benutzt wird, während das früher einmal deutlich verschiedene Begriffe waren (eine Unterscheidung, die zudem eigentlich sehr leicht zu begreifen ist). Die "Nichtantworterquote" q (etwa q = 02 also 20%) ist das Gegenstück zur "Ausschöpfungsquote" 1 – q (das Verhältnis von auswertbaren Fragebögen zu den insgesamt versendeten Fragebögen).

[74] In diesem Sinne wird unterschieden zwischen missing completely at random (MCAR), missing at random (MAR) usw.

[75] Mit anderen stetigen (auch nichtsymmetrischen) Verteilungen f(x) kann man natürlich auch rechnen, wenn man die Wahrscheinlichkeit F(90) - F(80) durch Integration über f(x) bestimmt. Die Wahrscheinlichkeit wird

und somit e = 0,05·80 = 4), dann würde man mit der Stichprobe den Mittelwert mit einer Genauigkeit von ± 4 schätzen können und der erforderliche Stichprobenumfang wäre bei z = 1 (Sicherheit nur 68%) nur 7 wegen n ≥ (0,125/0,05)2 = 6,25 und bei z = 2 eben 4.6,25 = 25.

Ein anderes Beispiel: man kann annehmen, dass x normalverteilt ist und ca. 95,45% der x-Werte zwischen x_{2u} = 800 und x_{2o} = 2000 liegen (so dass μ etwa 1400 sein dürfte). Dann ist V etwa 300/1400 = 0,2143 denn z = 2 = (2000 - 1400)/σ so dass σ = 300. Für z = 1 erhält man einen Stichprobenumfang von $(0,21/0,05)^2$ = 17,64 wenn e* wieder 5% sein soll (was jetzt nur eine Genauigkeit von 0,05.1400 = ± 70 bedeuten würde).[76] Entsprechend wäre bei z = 2 mit dem vierfachen Stichprobenumfang also etwa n = 71 zu rechnen.

Wie man sieht, führt die pessimistische Variante, bei völliger Unkenntnis der Varianz mit π(1-π) = ¼ zu sehr viel größeren Stichprobenumfängen. Die (A1) entsprechende Formel ist , nämlich (bei z = 1 und e =)

(A1a) $$n \geq \frac{z^2\,(1-\pi)}{e^{*2}\pi}$$ (wobei i.d.R. für π ein Wert *angenommen* werden muss)

denn der Variationskoeffizient ist im homograden Fall $V = \sqrt{\pi(1-\pi)}/\pi = \sqrt{(1-\pi)/\pi}$ und nimmt man den maximalen Wert *¼ für π(1-π)* an, dann ist er 1 (also *ein Variationskoeffizient von 100%*!!) wegen π = 1-π = ½ und so gerechnet erhält man bei e* = 0,05 einen Stichprobenumfang von n = (1/0,05)2 = 400 statt 17,64 und ähnlich wie oben im heterograden Fall bei e* = 0,01 und z = 1 n = $(1/0,01)^2$ = 10.000 (!) statt 441.

Die beliebte Abschätzungsformel für den Stichprobenumfang bei Annahme von π(1-π) = ¼ dürfte also zu viel zu hohen Werten für n führen, was auch verständlich ist, wenn man bedenkt, dass der Formel ein Variationskoeffizient von 100% zugrunde liegt. Man kann also eigentlich die so beliebte Formel (2) mit π = ½ bzw. (5) nicht wirklich empfehlen.

Es mag etwas verwirrend sein, dass im homograden Fall nicht nur der relative Fehler e*, sondern auch der absolute Fehler (genauso, wie auch π) eine relative Größe (Prozentangabe) ist. Bei π = 0,4 (also 40%) und e = 0,1 ist e* = 0,25. Der relative Fehler ist 25%, der absolute dagegen nur 10%. Ein absoluter Fehler von 0,08 wie oben auf S. 6 gerechnet bedeutetet bei π = ½ immerhin ein e* = e/π = 0,16, also 16%.

2. Gleiche Auswahlsätze (proportionale Aufteilung des gesamten Stichprobe n) sind optimal wenn man gleiche Varianzen in allen Schichten annimmt

Im Folgenden wird gezeigt, dass man immer dann, wenn man nichts über die (evtl. unterschiedlich großen) Varianzen innerhalb der Schichten weiß[77] die Varianz der Schätzfunktion (z.B. des Mittelwerts $\overline{X}$ als Schätzung für μ) minimiert, wenn man *gleiche* Auswahlsätze bei jeder Schicht vorsieht, (also den gesamten Stichprobenumfang n proportional aufteilt).

Für die Varianz der Stichprobenverteilung von $\hat{\pi}$ gilt bei gleichen maximalen geschätzten Varianzen $\hat{\pi}_k(1-\hat{\pi}_k) = 1/4$ in allen K Schichten gilt nach (14b)

$$\hat{\sigma}^2_{\hat{\pi}} = \frac{1}{4N^2}\sum \frac{N_k^2}{n_k - 1}\cdot\frac{N_k - n_k}{N_k} \approx \frac{1}{4N^2}\sum \frac{N_k^2}{n_k}\cdot\frac{N_k - n_k}{N_k} = \frac{1}{4N^2}\sum N_k\left(\frac{N_k}{n_k} - 1\right), \text{ also}$$

dann natürlich nicht etwa 68,27% betragen, wie bei der Normalverteilung für den 1σ-Bereich, sondern einen anderen Wert annehmen.

[76] Wie man leicht sieht vergrößert sich der Stichprobenumfang erheblich, wenn man z.B. mit e* = 0,01 (also ± 14) rechnen würde: man erhält dann n = 441 statt nur 17,64.

[77] und deshalb der Einfachheit halber nur davon ausgehen kann, sie seien alle gleich groß.

(A2) $$\hat{\sigma}^2_{\hat{\pi}} = \frac{1}{4N^2}\sum\frac{N_k^2}{n_k} - \frac{1}{4N}\,.$$ [78]

Man kann zeigen, dass diese Größe $\hat{\sigma}^2_{\hat{\pi}}$ genau dann minimal ist, wenn man eine proportionale Aufteilung annimmt, also $N_k/n_k = N/n$ für alle k gilt. Man erhält dann nämlich

$$\hat{\sigma}^2_{\hat{\pi}} = \frac{0{,}25}{N^2}\left(\frac{N}{n}-1\right)\sum_k N_k = \frac{0{,}25}{N}\left(\frac{N}{n}-1\right) \approx \frac{0{,}25}{n} \quad \text{weil } \sum_k N_k = N \text{ und dieser Wert}$$

$$\hat{\sigma}^2_{\hat{\pi}} = (\hat{\sigma}^2_{\hat{\pi}})_{min} = 1/4n$$

ist auch der kleinstmögliche Wert, wenn für die K Varianzen einheitlich ¼ angenommen wird. Um das zu zeigen ist $\hat{\sigma}^2_{\hat{\pi}}$ zu minimieren unter der Nebenbedingung $\sum_k n_k = n$ (die Minimierung von $\sigma^2_{\bar{x}}$ im heterograden Fall ist ganz analog). Hierzu ist die Lagrangefunktion

$$L = \sum_k\left(\frac{N_k}{N}\right)^2\frac{1}{n_k} + \lambda\left(\sum_k n_k - n\right)$$

für jedes k nach n_k abzuleiten und die Ableitung Null zu setzen. Man erhält dann K Gleichungen der folgenden Art $\frac{\partial L}{\partial n_k} = -\left(\frac{N_k}{N}\right)^2\frac{1}{n_k^2} + \lambda$ (mit k = 1, 2, …, K). Nullsetzen $\frac{\partial L}{\partial n_k} = 0$ liefert dann K Gleichungen $n_k\sqrt{\lambda} = \frac{N_k}{N}$. Die Summe aller dieser K Gleichungen ergibt $n\sqrt{\lambda} = 1$, also $\sqrt{\lambda} = \frac{1}{n}$ weil ja gilt $\Sigma n_k = n$. Dividiert man jede Gleichung durch die Summengleichung, so fällt die Wurzel aus dem Lagrange Multiplikator (also $\sqrt{\lambda}$) weg und man erhält $\frac{n_k}{n} = \frac{N_k}{N}$, was ja die Bedingung für die proportionale Aufteilung ist.

Eine nichtproportionale Aufteilung kann somit dann, wenn sich bei den K Schichten die Varianzen innerhalb der Schichten nicht wesentlich unterscheiden, zu schlechteren Ergebnissen führen als eine proportionale Aufteilung oder sogar eine ungeschichtete Stichprobe. Die hier soeben angestellte Beweisführung liegt auch der Herleitung der optimalen Aufteilung (Gleichung 18) zugrunde.

3. Herleitung des Schichtungseffekts

Bei einfacher Stichprobe (ZmZ) gilt für die Varianz der Stichprobenverteilung von $\bar{x}$

(A3) $$\sigma^2_{\bar{X}} = V(\bar{X})_{einf} = \frac{\sigma^2}{n},$$

wobei sich die Varianz σ^2 von x in der Grundgesamtheit wie folgt in externe und interne Varianz zerlegen lässt

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

[78] Bei ZmZ fällt der ohnehin sehr kleine Subtrahend 1/4N weg (er spielt auch keine Rolle bei der im Folgenden durchgeführten Minimierung unter der Nebenbedingung $\Sigma n_k = n$).

Für die Varianz von $\bar{x}$ gilt bei einer geschichteten Stichprobe im Falle von ZmZ (also ohne Endlichkeitskorrektur) nach Gl. 13 und 14

(13) $$\hat{\sigma}^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \frac{\hat{\sigma}_k^2}{n_k}$$ und (14) $$\hat{\sigma}^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \frac{\hat{\sigma}_k^2}{n_k} \frac{N_k - n_k}{N_k}.$$

Rechnet man mit der proportionaler Aufteilung, also $N_k/N = n_k/n$ so ergibt sich bei (13)

(A4) $$n\hat{\sigma}^2_{\bar{x}} = n\hat{V}(\bar{X})_{prop} = \sum \frac{n_k}{n} \cdot \hat{\sigma}_k^2 = \sum \frac{N_k}{N} \cdot \hat{\sigma}_k^2$$

wobei der ganz rechts stehende Ausdruck die geschätzte interne Varianz ist. Entsprechend mit Berücksichtigung der Endlichkeitskorrektur (also ZoZ) den noch kleineren Ausdruck (da ja stets $0 \le (1 - n_k/N_k) \le 1$ gilt)

(A4a) $$n\hat{\sigma}^2_{\bar{x}} = n\hat{V}(\bar{X})_{prop} = \sum \frac{N_k}{N} \cdot \hat{\sigma}_k^2 \left(1 - \frac{n_k}{N_k}\right).$$

Im Ergebnis erhält man also für den Schichtungseffekt (Größenvergleich von $n\hat{\sigma}^2_{\bar{x}}$ zwischen einfacher und geschichteter [bei proportionaler Aufteilung] Stichprobe)

(A5) $$nV(\bar{X})_{prop} = \sum \frac{N_k}{N} \sigma_k^2 = V_{int} \le \sigma^2 = V_{ext} + V_{int}$$ und damit

(A6) $$V(\bar{X})_{prop} = \frac{V_{int}}{n} \le V(\bar{X})_{e\,inf} = \frac{V_{int}}{n} + \frac{V_{ext}}{n}.$$

Man kann festhalten:

> Sobald eine externe Varianz auftritt (zwischen den Schichten große Unterschiede sind) ist der Standardfehler der Schätzung bei Schichtung kleiner als bei einfacher Stichprobe. Der Schichtungseffekt ist umso größer je homogener (je kleiner V_{int} ist) die Schichten sind.

Das heißt nicht, dass man grundsätzlich etwas gewinnt, sobald es eine externe Varianz $V_{ext} > 0$ gibt. Wenn es keine externe Varianz gibt kann man sogar etwas verlieren. Ein *Schichtungsverlust* statt Schichtungsgewinn ist möglich wenn man statt einer proportionalen Aufteilung eine sehr ungünstige Aufteilung wählt (z.B. genau umgekehrt vorgeht wie bei der optimalen Aufteilung). Schichtung ist also nicht automatisch ein Gewinn. Auch das spricht dafür, dass man dann, wenn man keine Daten hat, um eine optimale Aufteilung zu bestimmen, besser keine ungleichen Auswahlsätze vorsieht, also eine proportionale Aufteilung wählt.

Die Betrachtung im Falle einer optimalen Aufteilung gestaltet sich etwas komplizierter. Mit $\sum N_k\sigma_k = N\bar{\sigma}$ und (bei optimaler Aufteilung) $N_k\sigma_k = \frac{n_k}{n} \cdot N\bar{\sigma}$ erhält man, wenn man dies einsetzt in

(13) $$\sigma^2_{\bar{x}} = \sum \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k}$$ die folgende Feststellung

(A7) $$nV(\bar{X})_{opt} = \sum \frac{n_k}{n} \cdot \bar{\sigma}^2 = \bar{\sigma}^2 \le nV(\bar{X})_{prop} = \sum \frac{n_k}{n} \sigma_k^2,$$

wonach man mit optimaler Aufteilung noch einmal einen Schichtungsgewinn (gegenüber der proportionalen Aufteilung) erzielt. Dabei ist auch zu beachten, dass n bei optimaler Auftei-

lung kleiner sein kann als bei proportionaler Aufteilung. Die Gleichung wird etwas komplizierter, wenn man statt (13) den Ausdruck $\sigma_{\bar{x}}^2 = \sum \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k}\left(1 - \frac{n_k}{N_k}\right)$ betrachtet. Es gilt dann

(A7a) $$nV(\overline{X})_{opt} = \overline{\sigma}^2 \sum \frac{n_k}{n} \cdot \left(1 - \frac{n_k}{N_k}\right)$$

Die Gleichungen A6 und A7 können auch als Ausgangspunkt für die Herleitung des notwendigen (Gesamt-) Stichprobenumfangs n dienen (vgl. Abschn. 3c).

Die Behandlung des homograden Falls verläuft ganz analog und kann deshalb hier entfallen.