Andreas Kladroba

Stichprobentheorie

Gliederung

1. Einführung

Stichprobenuntersuchungen gehören für den Menschen der heutigen Zeit fast schon zum Alltag. Fast jeder wurde bereits einmal auf der Straße angesprochen oder zu Hause angerufen und nach seiner Meinung zu einem bestimmten Thema gefragt. Vielen Entscheidungen in der Politik oder in der Wirtschaft gehen Stichprobenuntersuchungen voraus. Deshalb ist es umso wichtiger, dass diese mit der nötigen Sorgfalt und Sachkenntnis vorgenommen werden. Leider ist zu beobachten, dass dies oftmals nicht der Fall ist. Besonders die Vorstellungen bezüglich der Ziehung von Stichproben und der darauf folgenden Anwendung der Methoden der Induktiven Statistik sind oftmals abenteuerlich zu nennen.

Das am häufigsten anzunehmende Missverständnis liegt bereits im Begriff selber. „Stichprobe“ wird häufig mit „Teilgesamtheit“ gleichgesetzt, was aber im Verständnis der statistischen Methodenlehre nicht der Fall ist. Eine Teilgesamtheit ist nur dann eine Stichprobe, wenn sie durch eine Zufallsauswahl gewonnen wurde. Dann und nur dann sind auch die Methoden der Wahrscheinlichkeitsrechnung und der Induktiven Statistik anwendbar. Anders gewonnene Teilgesamtheiten wie z.B. bewusste oder willkürliche Auswahlen ermöglichen keine Schätzungen oder Tests im Sinne der statistischen Schätz- und Testtheorie. Sie sind daher auch nur am Rande Thema dieses Kapitels.

2. Stichprobenpläne

Unter einem Stichprobenplan versteht man die Art der Entnahme von Elementen der Grundgesamtheit. Die einfachste Möglichkeit eines Stichprobenplans besteht in der einfachen Zufallsauswahl, die man sich so vorstellen muss, dass in einer Urne N Elemente der Grundgesamtheit liegen und n Elemente der Stichprobe herausgegriffen werden. Davon ausgehend wird sich dieses Kapitel der Frage zuwenden: Kann man durch andere Gestaltung des Stichprobenplans einen kleineren Stichprobenfehler erhalten als den bei einfacher Zufallsauswahl? Oder wenn die Motivation der Wahl eines zur einfachen Zufallsauswahl alternativen Stichprobenplans eine andere ist als die Verkleinerung des Stichprobenfehlers: Wie verhält sich der Stichprobenfehler bei diesem alternativen Stichprobenplan im Verhältnis zur einfachen Zufallsauswahl? Zur einfacheren Betrachtungsweise soll hier nur der Fall einer

Schätzung des Erwartungswertes μ mit Hilfe des arithmetischen Mittels $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ besprochen werden. Für die Schätzung anderer Parameter möge der Leser speziellere Literatur heranziehen (z.B. Pokropp (1996)).

Im folgenden soll außerdem angenommen werden, dass die Stichprobenvariablen $X_1,\ldots,X_n$ identisch verteilt sind mit $E(X_i) = \mu$ und $Var(X_i) = \sigma^2$.

2.1 Einfache Zufallsauswahl

An dieser Stelle soll noch unterschieden werden zwischen den Fällen des Ziehens mit und ohne Zurücklegen. Im weiteren Vorgehen wird auf diese Unterscheidung aus Gründen der Übersichtlichkeit verzichtet und die Betrachtung auf den Fall des Ziehens mit Zurücklegen beschränkt.

1. Ziehen mit Zurücklegen (ZmZ)

Die Annahme des Ziehens mit Zurücklegen bewirkt, dass die Stichprobenvariablen unabhängig sind. Man erhält somit

$$(1) \qquad V(\bar{X})_{ZmZ} = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left[V(X_1) + V(X_2) + \ldots + V(X_n)\right] = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

2. Ziehen ohne Zurücklegen (ZoZ)

Im Fall des Ziehens ohne Zurücklegen besteht Abhängigkeit zwischen den Stichprobenvariablen:

$$V(\bar{X})_{ZoZ} = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(\sum V(X_i) + \sum_{i\neq j} Cov(X_i, X_j)\right)$$

$$= \frac{1}{n^2}\left(n\sigma^2 - \frac{n(n-1)}{N-1}\sigma^2\right)$$

Der letzte Ausdruck in der Klammer ergibt sich wie folgt: Weil n Stichprobenvariablen mit (n-1) anderen korrelieren, erhält man den Faktor $n(n-1)$. Für die Kovarianz gilt:

$$Cov(X_i, X_j) = -\frac{1}{N-1}\sigma^2$$

Beweis:

$$Cov(X_i; X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

$$E(X_i X_j) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j \qquad \text{für alle } i \neq j \text{ (wegen ZoZ)}$$

Die Summe ergibt sich einfach aus der Definition des Erwartungswertes und der Faktor davor aus der Überlegung, dass alle Elemente gleich wahrscheinlich sein sollen. Also ist bei ZoZ die Wahrscheinlichkeit für zwei Elemente $\frac{1}{N(N-1)}$

Nach einer weiteren Umformung ergibt sich:

$$E(X_i X_j) = \frac{N}{N-1} \left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j - \frac{1}{N^2} \sum_{i=1}^{N} x_i^2 \right)$$

Erläuterung: Die Einschränkung $i \neq j$ in der ersten Summe wird hier durch die zweite Summe ersetzt.

Durch Ausmultiplizieren erkennt man sehr einfach, dass die Doppelsumme in der Klammer μ^2 entspricht. Der zweite Ausdruck in der Klammer ist $\frac{1}{N} E(X^2)$.

$$\Rightarrow E(X_i X_j) = \frac{N}{N-1} \mu^2 - \frac{E(X^2)}{N-1}$$

Also ist:

$$Cov(X_i, X_j) = \frac{N}{N-1} \mu^2 - \frac{E(X^2)}{N-1} - \mu^2$$

$$= \frac{N-(N-1)}{N-1} \mu^2 - \frac{E(X^2)}{N-1} = -\frac{1}{N-1} \left(E(X^2) - \mu^2 \right) = -\frac{\sigma^2}{N-1} \quad \blacktriangledown$$

Damit ist

$$(2) \qquad V(\bar{X})_{ZoZ} = \frac{1}{n} \left(1 - \frac{n-1}{N-1} \right) \sigma^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

Im Vergleich zum Fall des Ziehens mit Zurücklegens stellt man fest, dass Ziehen ohne Zurücklegen zunächst eine kleinere Varianz liefert. Diese Differenz nimmt aber ab je größer der Unterschied zwischen dem Umfang der Grundgesamtheit und dem der Stichprobe ist. Dagegen bietet das Ziehen mit Zurücklegen vor allem für viele theoretischen Betrachtungen den Vorteil der Unabhängigkeit der Stichprobenvari-ablen.

2.2 Geschichtete Stichproben

Wie bereits erwähnt wird in diesem Kapitel die Frage gestellt, ob man durch einen zur einfachen Zufallsauswahl alternativen Stichprobenplan eine Verringerung des Stichprobenfehlers erreichen kann.

Die Idee geschichteter Stichproben besteht darin, die Grundgesamtheit in K Teilgesamtheiten (Schichten) in der Form aufzuteilen, dass die Schichten in sich sehr homogen, untereinander dagegen aber heterogen sind. Aus jeder Schicht wird dann nach einem bestimmten Stichprobenplan eine bestimmte Anzahl an Elementen gezogen.

Im folgenden wird unterstellt, dass die Schätzfunktion so zerlegbar ist, dass sie durch eine einfache Linearkombination wieder aggregiert werden kann:

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

woraus sich aufgrund der Annahme eines Ziehens mit Zurücklegen ergibt

(4a) $$E\left(\hat{\theta}\right) = \sum_{k=1}^{K} \frac{N_k}{N} E\left(\hat{\theta}_k\right)$$ (4b) $$Var\left(\hat{\theta}\right) = \sum_{k=1}^{K} \left(\frac{N_k}{N}\right)^2 Var\left(\hat{\theta}_k\right)$$

Diese Forderung wird z.B. durch das arithmetische Mittel erfüllt. Da in jeder Schicht die Stichprobenziehung durch eine einfache Zufallsauswahl erfolgt, erhält man für den Fall des Ziehens mit Zurücklegen:

(5) $$Var\left(\overline{X}\right) = \sum_{k=1}^{K} \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k}$$

1. Proportionale Aufteilung

Ein naheliegender Stichprobenplan ergibt sich aus der Überlegung, dass der Anteil der einzelnen Schichten in der Stichprobe der gleiche sein soll wie in der Grundgesamtheit:

(6) $$\frac{n_k}{n} = \frac{N_k}{N} \Rightarrow n_k = n \frac{N_k}{N}$$

Setzt man dies in den Varianzausdruck (5) ein, erhält man:

$$(7) \qquad \mathrm{Var}\left(\overline{X}\right)_{prop} = \sum \frac{N_k^2}{N} \frac{\sigma_k^2}{n\frac{N_k}{N}} = \frac{1}{n} \sum \frac{N_k}{N} \sigma_k^2$$

Für einen Vergleich mit dem Stichprobenfehler bei einfacher Zufallsauswahl muss bezüglich der Grundgesamtheitsvarianz folgende Überlegung angestellt werden:

Die Streuung der Variable X in der Grundgesamtheit erfolgt auf zwei Weisen (Varianzzerlegung):

1. Es erfolgt eine Streuung unter den Schichten (externe Varianz). Dabei wird jede Schicht durch ihren Mittelwert μ_k repräsentiert und man erhält:

   $$(8) \qquad V_{ext} = \sum \frac{N_k}{N} (\mu_k - \mu)^2$$

2. Die Variable streut innerhalb der Schichten. Die daraus resultierende interne Varianz erhält man als arithmetisches Mittel der Schichtenvarianzen σ_k^2:

   $$(9) \qquad V_{int} = \sum \frac{N_k}{N} \sigma_k^2$$

Die Gesamtvarianz ergibt sich als Summe von externer und interner Varianz:

$$(10) \quad \sigma^2 = V_{ext} + V_{int}$$

Wie man aus Gleichung (7) sieht, ist

$$(11) \quad nV\left(\overline{X}\right)_{prop} = \sum \frac{N_k}{N} \sigma_k^2 = V_{int} \ ,$$

was selbstverständlich höchstens so groß werden kann wie

$$(12) \quad nV\left(\overline{X}\right)_{e\,inf} = \sigma^2 = V_{int} + V_{ext} \ .$$

Der Grenzfall, dass beide Varianzen gleich sind, tritt auf, wenn die externe Varianz Null ist, also $\mu_1 = \mu_2 = ... = \mu_n = \mu$ ist, alle Schichten also identisch sind. Im Regelfall (unterschiedliche Schichten) ist aber der Stichprobenfehler bei einer geschichteten Stichprobe und proportionaler Aufteilung kleiner als bei einer einfachen Zufallsauswahl (Schichtungseffekt).

Beispiel 1:

Eine Wetterstation misst jeden Tag um 12.00 Uhr die aktuelle Tagestemperatur. Um die Durchschnittstemperatur über 10 Jahre (N = 3600 Tage) zu ermitteln soll eine

Stichprobe im Umfang n = 100 gezogen werden. Dabei lässt sich die Grundgesamtheit gemäß den Jahreszeiten Frühjahr (F), Sommer (S), Herbst (H) und Winter (W) in vier Schichten mit $n_k = 900$ aufteilen. Es ist bekannt, dass folgende Varianzen gelten: $\sigma_F^2 = 25$, $\sigma_S^2 = 9$, $\sigma_H^2 = 16$, $\sigma_W^2 = 36$. Die Gesamtvarianz betrage $\sigma^2 = 100$.

Bei einer einfachen Zufallsauswahl ergibt sich für die Varianz von $\bar{x}$:

ZmZ: $$V(\bar{X})_{ZmZ} = \frac{\sigma^2}{n} = \frac{100}{100} = 1$$

ZoZ: $$V(\bar{X})_{ZoZ} = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{100}{100}\frac{3600-100}{3600-1} = 0{,}9725$$

Für eine geschichtete Stichprobe mit proportionaler Aufteilung gilt für jede Schicht $n_k = 25$, so dass sich die Varianz ergibt:

$$V(\bar{x})_{prop} = \frac{1}{n}\sum\frac{N_k}{N}\sigma_k^2 = \frac{1}{100}\cdot\frac{1}{4}(25+9+16+36) = 0{,}215$$

2. Nicht-proportionale Aufteilung

Als nicht-proportionale Aufteilung wollen wir hier verschiedene Formen „optimaler“ Aufteilungen betrachten. Dabei sollen zwei Konstruktionsprinzipien der „Optimierung“ im Vordergrund stehen:

1. *Minimierung des Stichprobenfehlers:* Da ein absolut gesehen minimaler Stichprobenfehler (nämlich $V(\bar{X}) = 0$) nur bei einer Totalerhebung erreicht werden kann, kann eine Minimierung nur unter einer Nebenbedingung erfolgen. Dabei sind denkbar

   a) Einhaltung eines bestimmten Stichprobenumfangs n

   b) Einhaltung bestimmter Kosten

2. *Minimierung der Erhebungskosten:* Auch diese Forderung kann nur unter einer Nebenbedingung erfolgen, da ansonsten „minimale Kosten = keine Erhebung“ bedeuten würde.

zu 1a)

Minimierung des Stichprobenfehlers unter Einhaltung eines bestimmten Gesamtstichprobenumfang bedeutet

(13) $$\min_{n_k} \sum \left(\frac{N_k}{N}\right)^2 \frac{\sigma_k^2}{n_k} \text{ u.d.R. } \sum n_k = n$$

Die Minimierung erfolgt mit Hilfe des Lagrange-Ansatzes. Wir bekommen damit folgende Lagrange-Funktion:

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

Partielles Ableiten nach n_k ergibt:

(15) $$\frac{\partial L}{\partial n_1} = -\left(\frac{N_1}{N}\right)^2 \frac{\sigma_1^2}{n_1^2} + \lambda = 0 \Rightarrow n_1\sqrt{\lambda} = \frac{N_1}{N}\sigma_1$$

$$\frac{\partial L}{\partial n_2} = -\left(\frac{N_2}{N}\right)^2 \frac{\sigma_2^2}{n_2^2} + \lambda = 0 \Rightarrow n_2\sqrt{\lambda} = \frac{N_2}{N}\sigma_2$$

$$\vdots$$

$$\frac{\partial L}{\partial n_K} = -\left(\frac{N_K}{N}\right)^2 \frac{\sigma_K^2}{n_K^2} + \lambda = 0 \Rightarrow n_K\sqrt{\lambda} = \frac{N_K}{N}\sigma_K$$

Als Summe der K Gleichungen erhält man:

(16) $$\sum n_k\sqrt{\lambda} = n\sqrt{\lambda} = \frac{1}{N}\sum_{kj=1}^{K} N_k\sigma_k$$

Dividierte man die k-te Gleichung durch (16), erhält man:

(17) $$\frac{n_k}{n} = \frac{N_k\sigma_k}{\sum N_k\sigma_k}$$

und damit ergibt sich für die k-te Schicht der folgende Stichprobenumfang:

(18) $$n_k = n\frac{N_k\sigma_k}{\sum N_k\sigma_k}$$

Gleichung (17) drückt aus, dass der Anteil einer Schicht an der Stichprobe umso größer sein muss

- je größer der Umfang dieser Schicht in der Grundgesamtheit ist und
- je größer die Varianz dieser Schicht ist (beides relativ zu den anderen Schichten).

Zur Ermittlung des Stichprobenfehlers wird dies in den Varianzausdruck (5) eingesetzt und man erhält nach einigen einfachen Umformungen:

$$(19) \quad V\left(\overline{X}\right)_{opt} = \frac{1}{n}\left(\sum \frac{N_k}{N}\sigma_k\right)^2 = \frac{1}{n}\left[\sum \frac{n_k}{n}\left(\frac{\sum N_k\sigma_k}{N}\right)\right]^2$$

Anmerkung: Die letzte Umformung ist leichter rückwärts nachzuvollziehen, wenn man für n_k den Ausdruck $n_k = n\frac{N_k\sigma_k}{\sum N_k\sigma_k}$ einsetzt.

Zum Vergleich mit dem Stichprobenfehler bei proportionaler Aufteilung wird die Differenz zwischen den beiden Varianzen (7) und (19) gebildet:

$$V\left(\overline{X}\right)_{prop} - V\left(\overline{X}\right)_{opt}$$

$$= \frac{1}{n}\sum \frac{n_k}{n}\sigma_k^2 - \frac{1}{n}\left[\sum \frac{n_k}{n}\frac{\sum N_k\sigma_k}{N}\right]^2$$

Der letzte Bruch in diesem Ausdruck ist nichts anderes als das arithmetische Mittel der k Standardabweichungen der Schichten in der Grundgesamtheit:

$$= \frac{1}{n}\sum \frac{n_k}{n}\sigma_k^2 - \frac{1}{n}\left[\sum \frac{n_k}{n}\overline{\sigma}\right]^2$$

$$= \frac{1}{n}\left[\sum \frac{n_k}{n}\sigma_k^2 - \left(\sum \frac{n_k}{n}\overline{\sigma}\right)^2\right]$$

Für die letzte Klammer gilt, dass $\frac{1}{n}\sum n_k = \frac{n}{n} = 1$ ist

$$= \frac{1}{n}\left(\frac{1}{n}\sum n_k\sigma_k^2 - \overline{\sigma}^2\right)$$

Jetzt ist der erste Ausdruck in der Klammer das arithmetische Mittel der σ_k^2:

$$= \frac{1}{n}\left(\overline{\sigma^2} - \overline{\sigma}^2\right)$$

Der Klammerausdruck entspricht aber dem Verschiebungssatz zur Berechnung von Varianzen (quasi die Varianz einer Varianz). Diese wird aber niemals negativ. Also ist $V\left(\overline{X}\right)_{prop} \geq V\left(\overline{X}\right)_{opt}$. Identisch sind die beiden Varianzen, wenn gilt

$\sigma_1^2 = \sigma_2^2 = ... = \sigma_K^2 = \sigma^2$, da in diesem Fall die optimale Aufteilung $\frac{n_k}{n} = \frac{N_k \sigma}{\sigma \sum N_k} = \frac{N_k}{N}$

gleich der proportionalen Aufteilung ist.

zu 1b)

An dieser Stelle soll eine Aufteilung mit minimaler Varianz unter der Nebenbedingung der Einhaltung eines bestimmten Kostenbudgets erfolgen. Es wird angenommen, dass die Kosten in jeder Schicht „pro Stück“, also pro erhobene Einheit und damit insgesamt proportional zur Größe der Stichprobe anfallen.

(20) $$c = \sum_{k=1}^{K} c_k n_k$$

Der Lagrange-Ansatz lautet hier:

(21) $$L = \sum \frac{N_k^2}{N^2} \frac{\sigma_k^2}{n_k} + \lambda\left(\sum c_k n_k - c\right)$$

Der weitere Ablauf erfolgt ähnlich wie unter 1a).

(22) $$\frac{\partial L}{\partial n_1} = -\frac{N_1^2}{N^2} \frac{\sigma_1^2}{n_1^2} + \lambda c_1 = 0 \Rightarrow \lambda c_1 n_1^2 = \frac{N_1^2}{N^2} \sigma_1^2 \Rightarrow n_1 = \frac{N_1 \sigma_1}{N\sqrt{\lambda c_1}}$$

$$\frac{\partial L}{\partial n_2} = -\frac{N_2^2}{N^2} \frac{\sigma_2^2}{n_2^2} + \lambda c_2 = 0 \Rightarrow \lambda c_2 n_2^2 = \frac{N_2^2}{N^2} \sigma_2^2 \Rightarrow n_2 = \frac{N_2 \sigma_2}{N\sqrt{\lambda c_2}}$$

$$\vdots$$

$$\frac{\partial L}{\partial n_K} = -\frac{N_K^2}{N^2} \frac{\sigma_K^2}{n_K^2} + \lambda c_K = 0 \Rightarrow \lambda c_K n_K^2 = \frac{N_K^2}{N^2} \sigma_K^2 \Rightarrow n_K = \frac{N_K \sigma_K}{N\sqrt{\lambda c_K}}$$

Summenbildung ergibt:

(23) $$n = \frac{1}{N\sqrt{\lambda}} \sum \frac{N_k \sigma_k}{\sqrt{c_k}}$$

Dividieren der k-ten Gleichung durch (23) ergibt schließlich:

(24) $$\frac{n_k}{n} = = \frac{N_k \sigma_k}{\sqrt{c_k}} \frac{1}{\sum \frac{N_k \sigma_k}{\sqrt{c_k}}}$$

Im Vergleich zu Gleichung (15) sagt Gleichung (24) zusätzlich aus, dass der Anteil der k-ten Schicht an der Grundgesamtheit um so größer ausfallen muss, je kleiner die Stückkosten der Schicht sind.

zu 2

Im Vergleich zu 1b) liegt der Gedanke nahe eine Erhebung mit möglichst minimalen Kosten durchzuführen, wobei allerdings der Stichprobenfehler Beschränkungen aufgelegt bekommt um zu vermeiden einen unakzeptabel hohen Stichprobenfehler durch eine zu kleine Stichprobe zu bekommen. Das Optimierungsproblem lautet also hier:

$$(25) \quad \min_{n_k} \sum_{k=1}^{K} c_k n_k \text{ u.d.R. } V^* = \sum \left(\frac{N_k}{N} \right)^2 \frac{\sigma_k^2}{n_k},$$

wobei V^* ein vorgegebenes Niveau der Varianz ist.

Damit ergibt sich der Lagrange-Ansatz:

$$(26) \quad L = \sum_{k=1}^{K} c_k n_k + \lambda \left(\sum \left(\frac{N_k}{N} \right)^2 \frac{\sigma_k^2}{n_k} - V^* \right)$$

Partielles Ableiten führt zu:

$$(27) \quad \frac{\partial L}{\partial n_1} = c_1 - \lambda \left(\frac{N_1^2}{N^2} \frac{\sigma_1^2}{n_1^2} \right) = 0 \Rightarrow n_1 = \sqrt{\lambda} \frac{N_1}{N} \frac{\sigma_1}{\sqrt{c_1}}$$

$$\frac{\partial L}{\partial n_2} = c_2 - \lambda \left(\frac{N_2^2}{N^2} \frac{\sigma_2^2}{n_2^2} \right) = 0 \Rightarrow n_2 = \sqrt{\lambda} \frac{N_2}{N} \frac{\sigma_2}{\sqrt{c_2}}$$

$$\vdots$$

$$\frac{\partial L}{\partial n_1} = c_K - \lambda \left(\frac{N_K^2}{N^2} \frac{\sigma_K^2}{n_K^2} \right) = 0 \Rightarrow n_K = \sqrt{\lambda} \frac{N_K}{N} \frac{\sigma_K}{\sqrt{c_K}}$$

Summenbildung ergibt:

$$(28) \quad n = \frac{\sqrt{\lambda}}{N} \sum \frac{N_k \sigma_k}{\sqrt{c_k}}$$

Division der k-ten Gleichung durch (28) führt zu

(29) $$\frac{n_k}{n} = \frac{\frac{N_k \sigma_k}{\sqrt{c_k}}}{\sum \frac{N_k \sigma_k}{\sqrt{c_k}}}$$

und damit zur gleichen Aufteilung wie unter 1b).

Beispiel 2:

Führt man das oben begonnene Beispiel (Temperaturmessung) fort, bekommt man folgende Stichprobenaufteilung bei minimaler Varianz (optimale Aufteilung):

$$n_F = n\frac{N_F \sigma_F}{\sum N_k \sigma_k} = 100 \cdot \frac{900 \cdot 5}{900(5+3+4+6)} = 100\frac{5}{18} = 27{,}\overline{7} \approx 28$$

$$n_S = 100\frac{3}{18} = 16{,}\overline{6} \approx 17$$

$$n_H = 100\frac{4}{18} = 22{,}\overline{2} \approx 22$$

$$n_W = 100\frac{6}{18} = 33{,}\overline{3} \approx 33$$

Die Stichprobenvarianz lautet damit:

$$V(\bar{x})_{opt} = \frac{1}{n}\left(\sum \frac{N_k}{N}\sigma_k\right)^2 = \frac{1}{100} \cdot \frac{1}{4^2}(5+3+4+6)^2 = 0{,}2025$$

Angenommen die Datenerfassung sei (aus welchen Gründen auch immer) im Herbst und im Winter teurer als im Frühjahr und im Sommer. Ein Ablesevorgang koste im Herbst und im Winter 1,- DM, sonst nur 0,50 DM. Dann gilt folgende Aufteilung:

$$n_F = 100\frac{\frac{900 \cdot 5}{\sqrt{0{,}5}}}{\frac{900 \cdot 5}{\sqrt{0{,}5}} + \frac{900 \cdot 3}{\sqrt{0{,}5}} + \frac{900 \cdot 4}{\sqrt{1}} + \frac{900 \cdot 6}{\sqrt{1}}} = 33{,}18 \approx 33$$

$$n_S = 100\frac{\frac{900 \cdot 3}{\sqrt{0{,}5}}}{\frac{900 \cdot 5}{\sqrt{0{,}5}} + \frac{900 \cdot 3}{\sqrt{0{,}5}} + \frac{900 \cdot 4}{\sqrt{1}} + \frac{900 \cdot 6}{\sqrt{1}}} = 19{,}91 \approx 20$$

$$n_H = 100\frac{\frac{900 \cdot 4}{\sqrt{1}}}{\frac{900 \cdot 5}{\sqrt{0{,}5}} + \frac{900 \cdot 3}{\sqrt{0{,}5}} + \frac{900 \cdot 4}{\sqrt{1}} + \frac{900 \cdot 6}{\sqrt{1}}} = 18{,}76 \approx 19$$

$$n_W = 100\frac{\frac{900\cdot 6}{\sqrt{1}}}{\frac{900\cdot 5}{\sqrt{0{,}5}}+\frac{900\cdot 3}{\sqrt{0{,}5}}+\frac{900\cdot 4}{\sqrt{1}}+\frac{900\cdot 6}{\sqrt{1}}} = 28{,}15 \approx 28$$

Problematisch ist hier, dass das zur Verfügung stehende Budget c in Gleichung (24) nicht mehr auftaucht. Bei der Beispielrechnung wurde weiterhin ein Stichprobenumfang von $n = 100$ angenommen, was Gesamtkosten in Höhe von 73,50 DM ergibt. Sollte das Budget niedriger (z.B. 50 DM) sein, muss ein entsprechend geringer Stichprobenumfang angesetzt werden. Ein optimales Ergebnis ist nur durch Ausprobieren zu erhalten.

Zum Schluss noch einige Anmerkungen zu geschichteten Stichproben:

1. Schichtenwahl

   Wir sind bisher davon ausgegangen, dass die einzelnen Schichten und deren Varianzen σ_k^2 gegeben waren. In der Praxis sind sie natürlich erst zu bilden, was auf einige Schwierigkeiten stößt. Relativ unproblematisch ist die Schichtenbildung, wenn sie quasi institutionell bereits vorliegt. Zu denken ist hier die Verwendung von Bundesländern, Gemeinden u.ä. als Schichten. Ist dies nicht der Fall, sind die Schichten so zu konstruieren, dass die Varianzen σ_k^2 möglichst klein sind. Dies führt allerdings zu einem Optimierungsproblem, dessen Lösung relativ schwierig ist. In der Praxis versucht man dieses Problem daher in der Regel durch die Verwendung von Schichtungsmerkmalen mit bereits bekannter Verteilung zu lösen. Die Idee ist dabei, dass die Grundgesamtheit bezüglich des Schichtungsmerkmals so in Schichten eingeteilt wird, dass die Varianzen möglichst klein sind und dass, wenn das Erhebungsmerkmal und das Schichtungsmerkmal hinreichend hoch korreliert sind, dann auch die Varianzen des Erhebungsmerkmals entsprechend klein sind.

2. Nachträgliche Schichtung

   Es ist denkbar, dass sich erst nach der Ziehung einer einfachen Zufallsauswahl eine „natürliche Schichtung“ herausstellt, z.B. dass sich zeigt, dass die Variabilität innerhalb der Geschlechter relativ gering ist und damit eine geschichtete Stichprobe angezeigt gewesen wäre. Das Konzept der nachträglichen Schichtung versucht nun den Schichtungseffekt nachträglich dadurch zu nutzen, dass das arith-

metische Mittel der Gesamtstichprobe als gewichtetes arithmetisches Mittel der einzelnen Schichten berechnet wird. Als Gewicht wird dabei der Anteil, den die Schicht in der Grundgesamtheit hat, verwendet. Pokropp (1996) weist nach, dass damit ein asymptotisch erwartungstreuer Schätzer erreicht wird und zumindest ein Teil des Schichtungsgewinns erzielt werden kann.

2.3 Klumpenstichproben

Methodisch deutlich schwieriger als die geschichtete Stichprobe stellt sich die Klumpenstichprobe dar. Die Idee hierbei ist, die Grundgesamtheit in in sich heterogene, untereinander aber möglichst homogene Teilgesamtheiten (Klumpen) aufzuteilen. Die Zufallsziehung erfolgt dann im Ziehen einer bestimmten Anzahl von Klumpen, aus denen

- alle Elemente ausgewertet werden (einstufig)
- ein Teil der Elemente wieder zufällig ausgewählt werden (zweistufig).

Die Vorteile der Klumpenstichprobe gegenüber der geschichteten Stichprobe bestehen vor allem im praktischen Bereich:

- Verringerung der Reisekosten
- Es ist keine vollständige Auswahlgrundlage (z.B. Einwohnermeldekartei) erforderlich. Entsprechende Unterlagen über die ausgewählten Klumpen reichen aus.
- Es sind kaum Kenntnisse über die Grundgesamtheit erforderlich. Oftmals ist es nicht einmal notwendig den Umfang der Grundgesamtheit oder auch nur der einzelnen Klumpen zu kennen.

Als Hauptnachteil muss, neben der bereits erwähnten methodischen Schwierigkeit, der Umstand genannt werden, dass eine Klumpenstichprobe (anders als eine geschichtete Stichprobe) nicht zwangsweise zu einem kleineren Stichprobenfehler als bei uneingeschränkter Zufallsauswahl führt.

Es gelte folgende Notation:

M	Anzahl der Klumpen in der Grundgesamtheit
N_i	Anzahl der Elemente im i-ten Klumpen
x_{ij}	j-te Element im i-ten Klumpen

$$X_i = \sum_{j=1}^{N_i} x_{ij}$$ Merkmalssumme im i-ten Klumpen

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$$ Durchschnitt im i-ten Klumpen

$$\overline{X}_K = \frac{1}{M}\sum_{i=1}^{M} X_i$$ durchschnittliche Merkmalssumme/Klumpen

Betrachten wir im folgenden eine einstufige Auswahl mit dem Spezialfall, dass alle Klumpen gleich groß sind, also dass $N_i = \overline{N}$ gilt. Ziel ist es weiterhin einen Schätzer für den Erwartungswert der Grundgesamtheit μ anzugeben.

Für eine einstufige Auswahl werden m der M Klumpen durch eine einfache Zufallsauswahl gezogen. Dann kann leicht gezeigt werden, dass

(30) $$\hat{\mu} = \frac{1}{\overline{N}m}\sum_{i=1}^{m} X_i = \frac{1}{m}\sum_{j=1}^{m}\mu_j$$

ein erwartungstreuer Schätzer für μ ist.

Die Varianz von $\hat{\mu}$ ist:[1]

(31) $$V(\hat{\mu}) = \frac{1}{\overline{N}^2 m}\left(1-\frac{m}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(X_i - \overline{X}_K\right)^2$$

Im folgenden soll gezeigt werden, dass die Frage, ob die Durchführung einer einstufige Klumpenstichprobe ähnlich wie eine geschichtete Stichprobe zu einer kleineren Varianz als bei einer einfachen Zufallsauswahl führt, vom Zusammenhang der Merkmalswerte innerhalb der Klumpen abhängt. Die Summe in Gleichung (31) lässt sich wie folgt auflösen:

(32) $$\sum_{i=1}^{M}\left(X_i - \overline{X}_K\right)^2 = \sum_{i=1}^{M}\left(\sum_{j=1}^{\overline{N}} x_{ij} - \overline{N}\mu\right)^2 = \sum_{i=1}^{M}\left[\sum_{j=1}^{\overline{N}}\left(x_{ij}-\mu\right)\right]^2$$

$$= \sum_{i=1}^{M}\left[\sum_{j=1}^{\overline{N}}\left(x_{ij}-\mu\right)^2 + \sum_{j=1}^{M}\sum_{\substack{\varphi=1\\ j\neq\varphi}}^{M}\left(x_{ij}-\mu\right)\left(x_{i\varphi}-\mu\right)\right]$$

[1] Die genaue Herleitung dieser Varianz kann bei Cochran (1972), S. 286ff. nachgelesen werden.

$$= (N-1)\sigma^2 + \sum_{i=1}^{K}\sum_{j=1}^{M}\sum_{\substack{\varphi=1 \\ j\neq\varphi}}^{M}(x_{ij}-\mu)(x_{i\varphi}-\mu)$$

Der zweite Summand kann als Maßzahl für den Zusammenhang innerhalb der Klassen verwendet werden. Man nennt

(33) $$\rho = \frac{1}{(\overline{N}-1)(N-1)\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{M}\sum_{\substack{\varphi=1 \\ j\neq\varphi}}^{M}(x_{ij}-\mu)(x_{i\varphi}-\mu)$$

den Intraklasskorrelationskoeffizienten. Er kann Werte im Intervall $-\frac{1}{\overline{N}-1} \le \rho \le 1$ annehmen.

Unter Berücksichtigung von ρ wird die Varianz in Gleichung (31) somit zu:

(34) $$V(\hat{\mu}) = \frac{1}{\overline{N}^2 m}\left(1-\frac{m}{M}\right)\frac{1}{m-1}\left[(N-1)\sigma^2 + (\overline{N}-1)(N-1)\sigma^2\rho\right]$$

$$= \frac{1}{\underbrace{\overline{N}m}_{=\frac{1}{n}}}\left(1-\frac{m}{M}\right)\underbrace{\frac{\overline{N}M-1}{\overline{N}(M-1)}}_{\approx 1}\sigma^2\left[1+(\overline{N}-1)\rho\right]$$

$$\approx \frac{\sigma^2}{n}\left(1-\frac{m}{M}\right)\left[1+(\overline{N}-1)\rho\right] = V(\overline{x})_{einf}\left[1+(\overline{N}-1)\rho\right]$$

Der Ausdruck in der eckigen Klammer wird oftmals auch Varianzaufblähungsfaktor genannt, da er die Varianz gegenüber der einfachen Zufallsauswahl (Ziehen ohne Zurücklegen) entweder vergrößert $(\rho > 0)$ oder verkleinert $(\rho < 0)$. Bei einer negativen Korrelation ist also das Klumpenverfahren genauer als die einfache Zufallsauswahl, bei einer positiven Korrelation ist dagegen die einfache Zufallsauswahl genauer.

Bei einer zweistufigen Klumpenstichprobe lässt sich die Varianz aufteilen in eine Varianz zwischen (between) den Klumpen (σ_b^2) und einer Varianz innerhalb (within) des i-ten Klumpen $(\sigma_{w_i}^2)$. Sie lautet (ohne Beweis):

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

mit: $$\hat{\sigma}_b^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(\mu_i - \overline{X}_K\right)^2$$

$$\hat{\sigma}_{w_i}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(x_{ij} - \mu_i\right)^2$$

Beispiel 3:

Das folgende Beispiel soll vor allem den Einfluss des Intraklasskorrelationskoeffizienten verdeutlichen. Dazu werden zwei extreme Grundgesamtheiten verwendet.

1. Grundgesamtheit: (1, 1); (2, 2); (3, 3); (4, 4); (5, 5)

In dieser Grundgesamtheit sind die Klumpen in sich homogen, untereinander aber sehr heterogen.

2. Grundgesamtheit: (1, 5); (2, 4); (3, 3); (4, 2); (5, 1)

Hier ist das genaue Gegenteil der Fall. Die Klumpen sind in sich sehr heterogen, untereinander aber homogen (alle $\mu = 3$)

Für die 1. Grundgesamtheit ergibt sich somit einen Erwartungswert von $\mu_1 = 3$ und einer Grundgesamtheitsvarianz $\sigma_1^2 = 2{,}\overline{2}$. Die Dreifachsumme in Gleichung (33) ergibt 20 und damit eine Intraklasskorrelation von $\rho_1 = \frac{1}{(2-1)(10-1)2{,}\overline{2}}20 = 1$. Damit ergibt sich ein Varianzaufblähungsfaktor von 2. Die Varianz ist also wegen der Homogenität innerhalb der Klumpen und gleichzeitigen Heterogenität zwischen den Klumpen doppelt so hoch wie bei einer einfachen Zufallsauswahl ohne Zurücklegen.

In der 2. Grundgesamtheit gelten der gleiche Erwartungswert und die gleiche Grundgesamtheitsvarianz wie für die 1. GG. Allerdings beträgt hier die Intraklasskorrelation $\rho_2 = -1$, so dass der Varianzaufblähungsfaktor und somit die Schätzervarianz Null beträgt. Dies verwundert auch nicht weiter, da bereits die Ziehung eines einzigen Klumpen genügt um den exakten Mittelwert μ zu bestimmen.

2.4 P-Stichproben

In der bisherigen Betrachtung haben wir vorausgesetzt, dass jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit besitzt in die Stichprobe zu kommen. Durch Aufheben dieser Annahme kann aber ebenfalls eine Verringerung der

Schätzervarianz im Vergleich mit einer einfachen Zufallsauswahl erreicht werden. Allerdings ist die Durchführung formal relativ anspruchsvoll, so dass wir hier auf eine ausführliche Beschreibung verzichten und uns nur auf einige Andeutungen beschränken wollen.

Zunächst ist eine Wahrscheinlichkeitsfunktion zu definieren, mit deren Hilfe den einzelnen Elementen der Grundgesamtheit eine Wahrscheinlichkeit p_j in die Stichprobe gezogen zu werden zugeordnet werden kann. Dann bildet im Fall des Ziehens mit Zurücklegen der Hasen-Hurwitz-Schätzer

(36) $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{Np_i}$$

unabhängig von der zugrundegelegten Verteilung einen erwartungstreuen Schätzer für μ. Die Varianz von $\hat{\mu}$ beträgt:

(37) $$V(\hat{\mu}) = \frac{1}{n}\sum_{j=1}^{N} p_j\left(\frac{x_j}{Np_j} - \mu\right)^2 = \frac{1}{n}\left[\sum_{j=1}^{N}\frac{1}{p_j}\left(\frac{x_j}{N}\right)^2 - \mu^2\right]$$

Es stellt sich jetzt natürlich die Frage, welche Verteilung zugrundegelegt werden sollte um eine möglichst kleine Schätzervarianz zu erhalten. Beachtet man, dass sich Gleichung (37) umformen lässt in

(38) $$V(\hat{\mu}) = \frac{1}{nN^2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\left(\frac{x_i}{p_i} - \frac{x_j}{p_j}\right)^2 p_i p_j \ ,$$

dann lässt sich leicht zeigen, dass für eine Verteilung, bei der die Wahrscheinlichkeiten proportional zur Größe der Elemente (pps = probability proportional to size) angenommen werden, die Varianz minimal (hier sogar null) wird. Der Beweis ist zu führen, indem man in Gleichung (38) $p_j = \left(\frac{y_j}{N\mu}\right)$ für alle j setzt. Das Problem ist dabei natürlich, dass die Grundgesamtheit unbekannt ist und eine entsprechende Modellierung der Wahrscheinlichkeiten nicht möglich ist. Denkbar wäre die Möglichkeit eine andere mit der zu untersuchenden Größe stark korrelierende Variable zu verwenden.

Bezüglich der weiteren Vorgehensweise sei auf die weiterführende Literatur (besonders Pokropp (1996)) verwiesen. Dort werden auch die besonderen Probleme, die beim Ziehen ohne Zurücklegen auftreten, besprochen.

3. Einige spezielle Fragestellungen

In diesem Kapitel sollen einige Problemstellungen behandelt werden, die bei Stichprobenuntersuchungen immer oder zumindest sehr häufig auftreten.

3.1 Notwendiger Stichprobenumfang

Bei den Stichprobenverfahren des vergangenen Kapitels sind wir davon ausgegangen, dass die Größe der Stichprobe bereits feststeht. Es stellt sich allerdings die Frage, wie diese Größe bei einer empirischen Untersuchung zu bestimmen ist. Dass diese von extremer Bedeutung ist, lässt sich an einem einfachen Beispiel leicht zeigen: Angenommen für eine potenzielle Käuferschaft im Umfang von 100.000 Personen wurde ein Produkt entwickelt, dass einer Stichprobe zur Begutachtung vorgelegt wird. Dabei wird den Mitgliedern der Stichprobe der Einfachheit halber nur die Frage gestellt, ob ihnen das Produkt gefällt. Es antworten 60% mit „ja“. Würde diese Stichprobe nur aus $n_1 = 10$ Personen bestehen, ergäbe sich ein 95%-Konfidenzintervall mit den Grenzen $[0{,}28; 0{,}92]$. Das heißt, wir hätten eine Spannbreite von großer Ablehnung des Produkts (72%) bis zu fast einstimmiger Zustimmung. Es ist leicht einsehbar, dass diese Aussage für den Hersteller absolut wertlos ist. Dagegen weist eine Stichprobe im Umfang $n_2 = 100$ beim gleiche Ergebnis bereits eine mehrheitliche Zustimmung von ca. 50 – 70% aus. Eine Stichprobe im Umfang $n_3 = 1000$ ergibt das Konfidenzintervall $[0{,}57; 0{,}63]$ und damit schon eine recht genaue Vorstellung, in welcher Größenordnung sich die Zustimmung zum Produkt bewegt. Daher wird der Stichprobenumfang in der Regel so bestimmt, dass ein bestimmter Stichprobenfehler nicht überschritten wird. Andererseits soll die Stichprobe auch nicht überdimensioniert werden, da eine empirische Untersuchung in der Regel auch mit großen Kosten verbunden ist.

Aus den Gleichungen () – () ergeben sich somit folgende notwendigen Stichprobenumfänge

a) bei Mittelwertschätzungen

(39) $$n^* = \frac{z^2\sigma^2}{e^2}$$ ZmZ

(40) $$n^* \geq \frac{z^2\sigma^2 N}{e^2(N-1) + z^2\sigma^2}$$ ZoZ

wobei e der absolute Fehler (= halbe Breite des Konfidenzintervalls) ist.

b) bei Anteilswertschätzungen

(41) $$n^* \geq \frac{z^2 \pi^* (1-\pi^*)}{e^2}$$ ZmZ

(42) $$n^* \geq \frac{z^2 \pi^* (1-\pi^*) N}{e^2 (N-1) + z^2 \pi^* (1-\pi^*)}$$ ZoZ

Dabei bezeichnet π^* eine vorsichtige Abschätzung von π. Hat man z.B. bei einer Wahlumfrage eine ungefähr Vorstellung, in welchem Bereich sich der Wähleranteil einer bestimmten Partei bewegt, sollte man ein π^* wählen, das ein wenig darüber (bei $\pi < 0{,}5$) oder darunter (bei $\pi > 0{,}5$) liegt. Hat man überhaupt keine Anhaltspunkte, dann ist der ungünstigste Fall mit $\pi^* = 0{,}5$ (größtmögliche Varianz) anzunehmen.

Um eine möglichst kleine Stichprobe bei konstantem Stichprobenfehler zu erhalten lassen sich auch Schichtungseffekte bei der Berechnung des notwendigen Stichprobenumfangs verwenden. Man erhält:

a) bei proportionaler Aufteilung

(43) $$n^* \geq \frac{z^2}{e^2} \sum_{k=1}^{K} \sigma_k^2 \frac{N_k}{N}$$

b) bei optimaler Aufteilung

(44) $$n^* \geq \frac{z^2}{e^2} \left(\sum_{k=1}^{K} \sigma_k \frac{N_k}{N} \right)^2$$

3.2 Antwortverweigerung und Antwortvariabilität

Die bisher angegebenen Schätzfunktionen sind davon ausgegangen, dass alle Befragten wahrheitsgemäß geantwortet haben. Dies trifft vor allem in der Umfragepraxis sicherlich nicht zu. Im Gegenteil wird man hier mehr oder weniger stark auf das Problem treffen, dass

1. Befragte die Antwort verweigern bzw. dass Antworten nicht eingeholt werden können. Gründe dafür können sein:

   - falsche Adressen oder der zu befragende wird auch bei mehrmaligen Besuchen nicht angetroffen

- Interesselosigkeit des Befragten, was vor allem bei Fragebogenaktionen oft auftritt.
- generelles Misstrauen gegenüber Befragungen

2. Befragte wissentlich oder unwissentlich falsche Angaben machen. Gründe hierfür können sein:
   - Dem Befragten ist die Antwort nicht bekannt. Er verlässt sich auf eine Schätzung, die bei mehrfachen Befragungen unterschiedliche Werte annehmen kann.
   - Der Befragte denkt, die korrekte Antwort sei z.B. gesellschaftlich nicht konform (z.B. eine Bejahung bei Fragen nach regelmäßigem Drogengenuss).

Diesen Problemen sollte in erster Linie durch einen entsprechenden Aufbau der Befragungsaktion und durch eine entsprechende Schulung der Interviewer begegnet werden. Bei Interesselosigkeit der Befragten könnte das Interesse in einem persönlichen Gespräch geweckt werden. Ebenso könnte durch entsprechende Aufklärung bestehendes Misstrauen ausgeräumt werden. Bei Befragungen zu sensiblen Themen (wie z.B. Drogengenuss) könnte versucht werden durch zusätzliche Anonymisierung die Angst vor Entdeckung beim Befragten zu nehmen. Dies erfolgt in vielen Fällen durch die Formulierung einer Alternativfrage. Betrachten wir folgendes Beispiel:

Von Interesse sei bei einer Befragung der Drogengenuss der befragten Personen. Daher wird die Frage formuliert (Frage 1): „Nehmen Sie regelmäßig Drogen?“ Als Alternativfrage wird Frage 2 formuliert: „Ist Ihre Sozialversicherungsnummer eine gerade Zahl?“. Dem Befragten wird eine der beiden Fragen gestellt, ohne dass der Interviewer weiß, welche Frage das ist. Dies kann z.B. dadurch erreicht werden, dass die Frage in einem geschlossenen Umschlag überreicht wird. Der Befragte muss die Frage nur noch mit „ja“ oder „nein“ beantworten und kann die Frage dann sogar selber vernichten. Ist bei der Auswertung jetzt bekannt, wie oft welche der beiden Fragen gestellt wurde und ist auch der Anteil der geraden Sozialversicherungsnummern bekannt, kann aus dem Anteil der „ja“-Antworten mit Hilfe des Satzes der Totalen Wahrscheinlichkeit relativ einfach auf den Anteil der Drogenkonsumenten geschlossen werden. Angenommen die Sozialversicherungsfrage sei in 75% aller Fälle gestellt worden $(P(S) = 0{,}75)$, es seien 50% der Sozialversicherungsnummern gerade

$(P(ja|S) = 0{,}5)$ und es hätten 20% aller Befragten mit „ja“ geantwortet. Dann haben unter denjenigen, die nach regelmäßigem Drogenkonsum befragt wurden $\frac{P(ja) - P(ja|S)}{P(D)} = \frac{0{,}4 - 0{,}5 \cdot 0{,}75}{0{,}25} = 0{,}1$, also 10%, diese Frage mit „ja“ beantwortet.

Hat man z.B. bei einer Fragebogenaktion eine Rücklaufquote von unter 100%, so kann man, wenn ein Anteilswert geschätzt werden soll, zumindest ein Intervall für diesen Anteilswert angeben. Lautet die Frage z.B. „Stehen Sie der Partei ABC nahe?“ und hat man bei einer Rücklaufquote von 75% einen Anteil von „ja“-Antworten von 40%, so kann man sagen, dass der Anteil der Sympathisanten der Partei auf jeden Fall zwischen 30% und 55% liegen muss, da der Anteil in der Nichtbeantwortergruppe ja auf jeden Fall zwischen 0 und 100% liegt. Kann man darüber hinaus noch weitere Vorinformationen verwenden, lässt sich dieses Intervall sogar noch weiter eingrenzen. Weiß man z.B. dass gerade Anhänger dieser Partei Befragungen gegenüber sehr skeptisch sind und dass der Anteil der Parteianhänger unter den Nichtantwortern deutlich höher sein muss als unter den Antwortern (z.B. über 50%), lässt sich das Intervall einengen auf den Bereich zwischen 42,5% und 55%.

Neben diesen einfachen Überlegungen lassen sich Antwortausfälle und Antwortvariabilitäten auch modellmäßig erfassen. Dazu sei aber auf die spezielle Literatur verwiesen.

3.3 Repräsentativität

Ein zumindest in der öffentlichen Darstellung wichtiger Aspekt einer Stichprobe ist der ihrer „Repräsentativität“. In der allgemeinen Vorstellung hat eine Stichprobe repräsentativ zu sein, bzw. haftet ihr ein Makel an, wenn behauptet wird, sie sei es nicht. Repräsentativität ist also gleichsam ein Qualitätskriterium für eine Stichprobe und damit für die komplette mit ihr durchgeführte Untersuchung. Dabei versteht man im allgemeinen Sprachgebrauch unter Repräsentativität Strukturähnlichkeit zwischen Stichprobe und Grundgesamtheit. Die Vorstellung, auf der sich die große Verbreitung des Konzepts gründet, ist die, dass die Schätzung einer Grundgesamtheitseigenschaft umso besser ist, je mehr sich Stichprobe und Grundgesamtheit in ihrer Struktur ähneln. Die Stichprobe soll also eine verkleinerte Abbildung der Grundgesamtheit sein.

Zunächst ist dazu zu sagen, dass der Ausdruck „repräsentative Stichprobe“ sprachlich falsch ist, da – wie wir bereits gesehen haben - in der Stichprobentheorie unter Stichproben nur Zufallsauswahlen verstanden werden. Eine repräsentative Auswahl ist aber eine bewusste Auswahl und somit nicht mehr zufällig. „Repräsentativität“ ist auch kein statistischer Fachausdruck. Man spricht hier lieber von der „Quotenauswahl“.

Zur Bewertung des Konzepts der Repräsentativität ist folgendes zu sagen:

1. Da es sich hier wie bereits erwähnt nicht um eine Zufallsauswahl handelt, ist auch die Wahrscheinlichkeitsrechnung nicht anwendbar. Damit ist auch auf das Instrumentarium der Induktiven Statistik (Schätzen, Testen) zu verzichten.
2. Der Grad von Repräsentativität ist nicht messbar. Damit kann nicht entschieden werden, ob eine Stichprobe mehr oder weniger repräsentativ ist. Darüber hinaus gibt es keinerlei Theorie, die die Annahme einer „besseren“ Schätzung bei hoher Repräsentativität im Vergleich zu einer „schlechteren“ Schätzung bei niedriger Repräsentativität stützt. Somit ist auch keinerlei z.B. dem Stichprobenfehler vergleichbare Formel bekannt, in der „Repräsentativität“ als explizite Größe auftaucht.
3. Wie wir am Beispiel der optimalen Aufteilung bei geschichteten Stichproben gesehen haben, wird in der Stichprobentheorie oftmals sogar bewusst gegen das Prinzip der Repräsentativität verstoßen um einen kleineren Stichprobenfehler zu erhalten. Repräsentativität steht somit oftmals in einem Gegensatz zu für die Güte von Schätzungen wichtigen Größen wie dem Stichprobenumfang oder der an der Varianz gemessenen Homogenität der Grundgesamtheit.

Zusammenfassend lässt sich sagen, dass Repräsentativität in seiner Bedeutung oftmals stark überschätzt wird. Es ist im Gegenteil sogar zu bezweifeln, ob Repräsentativität im Vergleich zum Konzept des Stichprobenfehlers überhaupt ein sinnvolles Gütekriterium für Stichproben ist.

Weiterführende Literatur

[1] Cochran, W.G. (1972), Stichprobenverfahren, Berlin/New York

[2] Kreienbrock, L. (1993), Einführung in die Stichprobenverfahren, München/Wien

[3] Pokropp, F. (1996), Stichproben: Theorie und Verfahren, München/Wien

[4] Stenger, H. (1986), Stichproben, Heidelberg/Wien

[5] Strecker, H./R. Wiegert (1994), Stichproben, Erhebungsfehler, Datenqualität, Göttingen