Einige Bemerkungen zur maschinellen Ziehung von Zufallsstichproben

Kirschner, Hans-Peter

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

EINIGE BEMERKUNGEN ZUR MASCHINELLEN ZIEHUNG VON ZUFALLSSTICHPROBEN

1. Einleitung

Bei der ZUMA-Mitarbeit an Projekten der verschiedensten Art ist immer wieder festzustellen, daß ganz im Gegensatz zum hohen Bekanntheitsgrad der Möglichkeiten maschineller Datenanalyse die Techniken der maschinellen Stichprobenziehung bisher nicht die wünschenswerte Verbreitung gefunden haben. Es sind bei kooperierenden Wissenschaftlern häufig Vorstellungen anzutreffen, die sich ausschließlich an den verschiedenen Verfahrensweisen des "manuellen Ziehens" orientieren, etwa an derjenigen des systematischen Ziehens mit Zufallsstart. Hinzu kommt, daß von den gebräuchlichsten Programmpaketen in den Sozialwissenschaften (SPSS, BMDP, OSIRIS) keines bisher die Entwicklungen in der Theorie des maschinellen Stichprobenziehens verifiziert hat. Anzutreffen ist lediglich die sehr einfache Methode des sog. "binomial sampling", deren theoretischer Hintergrund einer der Gegenstände des vierten Abschnitts ist.

Selbstverständlich ist es einerseits in vielen Fällen unangemessen oder einfach unmöglich, eine Stichprobe anders als "von Hand" zu erstellen. Unbestreitbar gibt es aber auch andererseits eine Vielzahl von Fällen, in denen sorgfältig abgewogen werden muß, ob man die Vorteile einer maschinellen Ziehung z.B. durch hohe Kosten für Loch- bzw. Schreibarbeiten "erkaufen" will. Dies gilt insbesondere z.B. für Grundmaterial zu Textstichproben oder Adressenmaterial in sich teilweise überdeckenden Listen, die möglicherweise noch unterschiedlichen Informationsgehalt haben.

Man kann also sagen, daß die Entscheidung, Hilfsmittel der elektronischen Datenverarbeitung für Stichprobenziehungen zu verwenden, häufig durchaus eingehender Begründungen bedarf. Diesen Gesichtspunkt weiter zu vertiefen, soll allerdings einer der späteren Ausgaben der ZUMANACHRICHTEN vorbehalten bleiben. Im folgenden wird deshalb stets angenommen, daß die Auswahlgrundlage für die Stichprobe auf Band, Platte usw. in geeigneter Form vorliegt.

Verwendet man zur Erstellung einer Stichprobe Zufallszahlen (vgl. zweiter Abschnitt), so lassen sich die wesentlichen Techniken zur Ziehung in zwei Klassen einteilen: zum einen wird die Stichprobe vermöge einer Sequenz von Zufallszahlen bestimmt, wobei diese jeweils den einzelnen Elementen der Stichprobe zugeordnet sind. Nur in degenerierten Fällen sind diese Zufallszahlen alle gleich. Zum anderen wird die gesamte Stichprobe durch die Angabe einer (oder mehrerer) Zufallszahl(en) und ein daran anschließendes deterministisches Verfahren vollständig festgelegt.

Zur ersten Klasse gehören z.B. einfache Zufallsstichproben (simple random samples) "ohne Zurücklegen" oder "mit Zurücklegen", wohingegen alle systematischen Verfahren mit einem oder mehreren Zufallsstarts der zweiten Klasse zugerechnet werden müssen. Für beide Klassen werden im dritten und vierten Abschnitt entsprechende Verfahren erläutert und im Rahmen des Programmpakets OSIRIS jeweils Beispiele angegeben.

Zuvor seien jedoch einige grundsätzliche Bemerkungen vorangeschickt, um möglichen Mißverständnissen vorzubeugen.
Die gesamte folgende Diskussion bezieht sich auf die probabilistischen Modelle, die den Ziehungen zugrundeliegen. Insbesondere Vergleiche zwischen verschiedenen Verfahren dürfen nur unter diesen Gesichtspunkten bewertet werden. Wenn sich also zeigen wird, daß etwa die Varianz beim "binomial sampling" größer ist als diejenige für vergleichbare einfache Zufallsstichproben, wird man für die reale Situation nicht ohne weiteres sagen können, die zweite dieser Prozeduren sei "besser". Vielmehr ist stets zunächst zu prüfen, inwieweit das angenommene Modell als Näherung an die wirklich gegebenen Umstände betrachtet werden kann. Hat man z.B. Grund zu der Annahme, daß sich die nicht-neutralen Ausfälle nicht "zufällig" verteilen werden, kann man nicht erwarten, daß Güte- oder gar Signifikanzaussagen besonders tragfähig sind. In einer solchen Situation wird man die Theoreme der Theorie lediglich als Indiz für die Verhälnisse unter guter Modellanpassung sehen dürfen.

Natürlich ist dies ein ernstzunehmender Einwand gegen die Verwendung von Zufallsstichproben an sich. Man sollte aber bei der Verwendung alternativer Verfahren (Quoten, bewußte Auswahl usw.) stets bedenken, daß allein Zufallsstichproben sich dadurch auszeichnen, daß man zumindest im "Idealfall" (100% Ausschöpfung, keine sonstigen Verzerrungen usw.) mathematisch fundierte Resultate über die Nähe der üblicherweise berechneten Schätzwerte zu den "wahren" Werten kennt.

2. Zufallszahlen

Ganz allgemein wird eine Zahl U zwischen Null und Eins als Zufallszahl bezeichnet, wenn sie als Ausprägung einer auf dem Einheitsintervall definierten Gleichverteilung aufgefaßt werden kann. Für alle Zahlen Z mit $0 \leq Z < 1$ ist dann die Wahrscheinlichkeit dafür, daß U kleiner oder gleich Z ist, gleich Z. Hat man eine Folge von Zufallszahlen, wird unterstellt, daß diese voneinander (stochastisch) unabhängig sind.

In die Definition einer Zufallszahl wird bewußt keine "Konstruktionsvorschrift" aufgenommen, sie ist lediglich eine mathematische Setzung, die sich auf ein abstraktes Axiomensystem stützt. Alle physikalischen oder arithmetischen Verfahren, die "Zufallszahlen" erzeugen sollen, sind daher mehr oder minder gute Verifizierungen dessen, was modellhaft gefordert wird.

An dieser Stelle von besonderem Interesse sind die sogenannten Zufallszahlengeneratoren, die auf elektronischen Rechenanlagen in der Regel als Unterprogramme zur Verfügung stehen. Die Arbeitsweise dieser Programme ist ganz überwiegend deterministisch und läßt sich stark vereinfachend wie folgt darstellen. Ist etwa N_o eine positive ganze Startzahl und ist M eine fest vorgegebene (möglichst große) positive ganze Zahl, so ist die folgende "Zufallszahl" der (ganzzahlige) Divisionsrest, der sich ergibt, wenn die auf ebenfalls fest vorgegebene Weise arithmetisch manipulierte Zahl N_o durch M dividiert wird, usw. Die Zahl M hat zumeist die Größenordnung der auf der jeweiligen Maschine größten darstellbaren ganzen Zahl (auf der SIEMENS 4004 z.B. $2^{31}-1$),und die resultierenden "Zufallszahlen" sind ebenfalls ganze Zahlen zwischen 0 und M. Die Reduktion auf das Einheitsinter-

vall geschieht einfach vermöge Division durch M, also U = N/M.[+]
Auf diese Weise erzeugte Zufallszahlen heißen auch "Pseudo-Zufallszahlen".

Viele statistische Tests zeigen nun, daß diese "(linear) kongruente" Methode der Zahlengeneratoren in der Tat Zahlenfolgen produziert, die als "zufällig" betrachtet werden können und somit eine brauchbare Basis für die maschinelle Ziehung einer Stichprobe abgeben.

Wie nähere Untersuchungen der bei ZUMA vorhandenen Generatoren in den Paketen SPSS und OSIRIS allerdings ergeben haben, sind diese von der Güte her keineswegs äquivalent. Nach Anwendung verschiedener Testverfahren ergab sich ein uneinheitliches Bild; dies war Anlaß zur Installation einer (weitgehend maschinenunabhängigen) Spielart des von KNUTH (1969: 30f) empfohlenen Algorithm M.
Die vergleichenden Betrachtungen zu diesem Algorithmus sind jedoch noch nicht abgeschlossen.
Ausführliche Darstellungen zur maschinellen Erzeugung von Zufallszahlen finden sich in KNUTH (1969:Kap. 3.2) und VADUVA (1976).

3. Systematisches Ziehen

Das systematische Ziehen einer Stichprobe mit einem (mehreren) Zufallsstart(s) und konstanter (konstanten) Schrittweite(n) ist außerordentlich weit verbreitet. Dies liegt in erster Linie wohl an der sehr einfachen manuellen technischen Durchführbarkeit, jedenfalls im Vergleich zum Hantieren mit Tabellen von Zufallszahlen. Dieser operationale Gesichtspunkt tritt ohne Zweifel beim maschinellen Ziehen stark in den Hintergrund; folglich wird in der gängigen Software für die Sozialwissenschaften (SPSS, BMDP, OSIRIS usw.) systematisches Ziehen nicht explizit unterstützt. Dabei spielt sicher auch der operationale und wohlbekannte Nachteil eine Rolle, daß bei dieser Art zu ziehen die Stichprobengröße häufig von dem eigentlich gewünschten Wert erheblich abweicht (wenn also die Anzahl der Fälle in der

[+] Man beachte, daß z.B. die Zahl $\pi/4$ sicher nicht als Quotient N/M darstellbar ist. D.h. in Strenge überdecken diese Zufallszahlen entgegen der Forderung nicht das gesamte Einheitsintervall.

Datei nicht durch die Stichprobengröße teilbar ist). Dies läßt sich zwar durch Kunstgriffe vermeiden (vgl. KONIJN 1973: 360ff), jedoch erscheint deren Implementierung wenig reizvoll, weil sich - wie der folgende Abschnitt zeigen wird - einfache Zufallsstichproben mit vorgegebener Stichprobengröße sehr einfach erzeugen lassen und zudem die Voraussetzungen, unter denen systematisches Ziehen theoretische Vorteile gegenüber anderen Techniken besitzt, häufig bei sozialwissenschaftlichen Untersuchungen nicht zu verifizieren sind.

Außer den Kunstgriffen, die üblicherweise in der Literatur zu finden sind, wird gelegentlich ein Verfahren benutzt, dessen Verwendung keinesfalls empfohlen werden kann. Ein Beispiel möge die abstrakte Behandlung hier ersetzen:

Angenommen, es liegt eine Datei mit N = 140 Fällen vor und es soll eine Stichprobe der Größe n = 50 gezogen werden. Man errechnet N/n zu 2,8 und argumentiert, daß für die ganzen Zahlen j von 0 bis 49 gilt: $[k + j * 2{,}8] \leq 140$, wenn k = 1,2 oder 3 ist und - allgemein - [m] die größte ganze Zahl kleiner gleich m ist. Man rundet also stets ab. Ferner gilt für k = 1,2 oder 3, daß $[k + 50 * 2{,}8] = k + 50 * 2{,}8 > 140$. Man wird also alle Einheiten mit den Nummern $[k + j * 2{,}8]$ in die Stichprobe nehmen, wenn man zuvor zufällig eine Zahl k zwischen 1 und 3 ausgewählt hat und anschließend j von 1 bis 49 laufen läßt. Auf diese Weise wird sichergestellt, daß die Stichprobe genau 50 Fälle enthält. Eine einfache Überlegung zeigt nun bereits an dieser Stelle, daß ein solches Design höchst unerwünschte Implikationen hat. Zunächst ist offensichtlich, daß drei Startzahlen zugelassen werden müssen. Würde man sich etwa auf 1 und 2 beschränken, könnten die beiden möglichen Stichproben der Größe 50 nicht die Grundgesamtheit von 140 Fällen überdecken, d.h. die Fälle 5, 8, 11 usw. hätten keine Chance gezogen zu werden. Läßt man jedoch die Startzahlen 1, 2 und 3 zu, tritt natürlich eine "Über-Deckung" der Grundgesamtheit ein, d.h. die drei möglichen Stichproben müssen gemeinsame Elemente enthalten. So werden die Stichproben zur Startzahl 1 und 3 u.a. die Einheiten mit den Nummern 17 und 31 gemeinsam haben; folglich ist die Wahrscheinlichkeit dafür, daß Einheit 17 sich in der Stichprobe befindet gleich 2/3, wohingegen die Wahrscheinlichkeit dafür, daß Einheit 4 in der Stichprobe vertreten ist, 1/3 beträgt. Selbstverständlich muß gemäß dieser unterschiedlichen Auswahlwahrscheinlichkeiten gewichtet werden.

Insgesamt produziert dieses Design also Stichproben mit ungleichen Auswahlwahrscheinlichkeiten, wobei die Ermittlung der Gewichte rechnerisch unangenehm ist. Die gewünschte Stichprobengröße wird demnach zu einem hohen Preis erzwungen.

Das Problem der Stichprobengrößen beim systematischen Ziehen mit Zufallsstart kann natürlich nachrangig werden, wenn bekannt ist, daß die Einheiten in der gegebenen Auswahlgrundlage gerade Voraussetzungen erfüllen, unter denen die systematischen den einfachen Zufallsstichproben überlegen sind, z.B. wenn ein linearer Trend vorliegt.

Man wird dann mit ganzzahligem Zufallsstart und ganzzahliger Schrittweite arbeiten und diese so kalkulieren, daß sich eine akzeptable Stichprobengröße ergibt.

Die maschinelle Durchführung einer solchen Aufgabe ist nun mit den Hilfsmitteln des Programmpakets OSIRIS leicht möglich. Leider ist dies weitgehend unbekannt. Daher enthält der Anhang A dieses Artikels als Beispiel ein OSIRIS-SETUP, mit dem systematisch "jeder fünfte" gezogen werden kann. (Ein analoges SETUP scheint mit SPSS nicht möglich zu sein.)

Es sei noch darauf hingewiesen, daß sich dieses Beispiel auf das systematische Ziehen von Stichproben in verschiedenen Schichten verallgemeinern läßt - was übrigens "von Hand" häufig schon Probleme aufwirft.

4. Einfache Zufallsstichproben

Wie in dem vorangegangenen Abschnitt schon erwähnt, sind systematisch gezogene Stichproben nur unter bestimmten Bedingungen gleich gut oder sogar besser als einfache u.U. auch schichtweise gezogene Zufallsstichproben. Da aber das Vorliegen solcher Bedingungen oft nur sehr schwer oder überhaupt nicht überprüft werden kann, befindet sich der Wissenschaftler offensichtlich auf der "sicheren Seite", wenn er seine Stichproben nicht systematisch zieht, sondern ein Verfahren der einleitend angesprochenen ersten Klasse verwendet.

Ein wichtiger Vertreter dieser Klasse ist das sogenannte "binomial sampling". Mit dessen Hilfe läßt sich aus einer Datei (mit bekannter oder unbekannter Fallzahl) ein vorgegebener Prozentsatz von Fällen zufällig entnehmen. Konkret stellt sich der Algorithmus wie folgt dar:

Ist $p * 100$ der gewünschte Prozentsatz, wird Fall für Fall eine (jeweils neue) Zufallszahl erzeugt und geprüft, ob diese kleiner oder gleich p ist. Trifft das zu, wird der entsprechende Fall zur Stichprobe genommen. Andernfalls wird der Fall verworfen.

Da die praktische Durchführung dieses Algorithmus' in den großen Paketen SPSS, BMDP und OSIRIS überaus einfach ist und in SPSS der SAMPLE-Befehl allein schon diese Methode beinhaltet, soll hier kein Beispiel gegeben, sondern mehr - vor allem im Anhang - der theoretische Hintergrund beleuchtet werden.

Beim "binomial sampling" ist ganz offensichtlich die Größe $\tilde{n}$ der resultierenden Stichprobe selber zufallsabhängig und diese daher in Strenge keine einfache Zufallsstichprobe. Man kann leicht zeigen, daß $\tilde{n}$ einer Binomialverteilung ("binomial sampling"!) mit der Erfolgswahrscheinlichkeit p genügt und daher den Erwartungswert $N * p$ besitzt, wenn N die (evtl. unbekannte) Fallzahl der Datei ist.

Dieser Sachverhalt der nur ungefähr vorhersagbaren Stichprobengröße $\tilde{n}$ führt natürlich zu der Frage, wieso die Prozedur überhaupt im Abschnitt über einfache Zufallsstichproben besprochen wird. Die Antwort besteht darin, daß sich bereits bei relativ geringen Fallzahlen in der Auswahlgrundlage das "binomial sampling" ähnlich verhält wie die Ziehung einer einfachen Zufallsstichprobe der Größe $N * p$. Diese Aussage muß selbstverständlich präzisiert werden. Da das bisher in der Literatur nicht ausreichend geschehen ist, enthält der Anhang B für den mehr methodisch interessierten Leser eine eingehendere Diskussion über diejenigen Bereiche, in denen von der Gleichwertigkeit beider Verfahren gesprochen werden kann.

Eine Datei wie beim "binomial sampling" zur Stichprobenziehung einmal sequentiell zu durchlaufen, hat den zuweilen sehr erwünschten Vorteil, daß die Eingabedatei nach der Ziehung dieselbe Anordnung wie die Ausgabedatei besitzt.

Spielt dieser Gesichtspunkt keine Rolle, kann man aus einer Datei (nicht notwendig bekannter Größe) eine einfache Zufallsstichprobe fester Länge n maschinell wie folgt erstellen:

Jedem Fall wird fortlaufend eine Zufallszahl zugeordnet. Die ersten n Einheiten der nach den Zufallszahlen geordneten Datei stellen dann eine einfache Zufallsstichprobe der Größe n dar.

Dieses sehr einfache Verfahren soll hier nicht weiter vertieft werden; es sei nur darauf hingewiesen, daß es insofern von der Qualität des jeweiligen Zufallszahlengenerators abhängt, daß dieser aus theoretischen Gründen bei der sukzessiven Erzeugung der Zufallszahlen keine Doubletten produzieren darf.

Die Methode des "binomial sampling" ist offenbar sowohl für die Situation bekannter als auch für die unbekannter Fallzahlen anwendbar. Sie ist nun für bekannte Fallzahlen von FAN, MULLER und REZUCHA (1962) sowie unabhängig davon durch T. G. JONES (1962; CACM, S. 343), so verfeinert worden, daß die Ziehung von einfachen Zufallsstichproben fester Länge möglich wurde. Nähere Ausführungen dazu finden sich bei KNUTH (1969: 122 ff). Dieser Algorithmus läßt sich mit den Hilfsmitteln des Programmpakets OSIRIS darstellen. Da dies ähnlich wie beim systematischen Ziehen weitgehend unbekannt ist, wird in Anhang C als Beispiel ein OSIRIS-SETUP angegeben, mit dem sich aus einer Datei eine einfache Zufallsstichprobe fester Länge ziehen läßt. (Ein analoges SETUP scheint in SPSS nicht möglich zu sein.) Es gilt auch hier, daß sich dieses Beispiel auf das Ziehen einfacher Zufallsstichproben in verschiedenen Schichten während eines Durchlaufs durch die Datei verallgemeinern läßt. Da dieser Artikel jedoch mehr grundsätzlichen Aspekten gewidmet ist, soll hierauf nicht näher eingegangen werden. Es sei lediglich angemerkt, daß eine Vielzahl möglicherweise recht kompliziert definierter Schichten deren vorherige Auszählung u. U. ineffizient erscheinen läßt. Es liegt dann näher, eine andere Ziehungsmethode zu verwenden. Dieses sogenannte "RESERVOIR SAMPLING" findet sich ebenfalls bei FAN, MULLER und REZUCHA (1962), sowie bei KNUTH (1969: 123). Der Algorithmus hängt nicht von der Kenntnis der

Fallzahlen ab, benötigt dafür aber mehr als einen Durchlauf durch die Datei. Er wurde bei ZUMA noch nicht praktisch erprobt, so daß an dieser Stelle kein Beispiel gegeben werden kann.

5. Schlußbemerkung

Da zu erwarten ist, daß in Zukunft immer mehr für den Sozialwissenschaftler relevante Daten auf elektronischen Speichermedien vorhanden sein werden, ist es recht plausibel zu vermuten, daß die maschinelle Ziehung von Stichproben immer häufiger angewendet werden wird. Es ist also sicher sinnvoll, geeignete Ziehungsalgorithmen sowie die zugehörige Software allgemein zu propagieren. Deshalb sei an dieser Stelle die Bitte an alle Leser erlaubt, eigene Erfahrungen, eigene Entwicklungen usw. ZUMA mitzuteilen, um so eine weiterführende Diskussion zu ermöglichen. ZUMA selber wird in absehbarer Zeit ein Programm implementieren, daß u.a. auch die oben besprochenen Algorithmen enthält und OSIRIS-kompatibel ist.

Anhang A

Wie im dritten Abschnitt angekündigt, wird im folgenden ein OSIRIS-SETUP angegeben, mit dessen Hilfe aus einer Datei "jeder fünfte" systematisch mit Zufallsstart gezogen werden kann. Pro Fall enthält die Datei 12 Variablen. Die Stichprobe wird durch eine Null-Eins-Variable R13 - die "Stichprobenvariable" - charakterisiert ("1": Fall gehört zur Stichprobe, "0" sonst); mit Hilfe des Programms "TRANS" wird R13 an die bereits vorhandenen Variablen angefügt.

```
ZWEI FILE-KOMMANDOS
FUER DIE EINGABEFILES
                                  } Systemabhängig !
ZWEI FILE-KOMMANDOS
FUER DIE AUSGABEFILES

$RUN TRANS
$RECODE
        CARRY (R1, R2)
        IF R2 GE 1 THEN GO TO FLAG
        R1 = RAND(0,5)
 FLAG   R2 = R2 + 1
        R3 = (R2 - R1)/5
        R4 = TRUNC(R3)
        IF R4 EQ R3 THEN R13 = 1
       CELSE R13 = 0
$SETUP
TITLE
TRAN,WIDTH = 1 *
V1 - V12 *
R13 *
```

Anhang B

Ohne auf mathematische Einzelheiten einzugehen, soll im folgenden dargestellt werden, daß das Ziehen einer Stichprobe mit Hilfe des "binomial sampling" in weiten Bereichen dem Ziehen einfacher Zufallsstichproben äquivalent ist.
Als erstes seien einige spezielle Notationen eingeführt.
Die Datei möge N Fälle enthalten, wobei die Ausprägungen einer später zu erhebenden interessierenden Variablen mit $y_1, \ldots, y_N$ bezeichnet seien. Es sei weiter δ eine Null-Eins-wertige Funktion der Stichprobe $\tilde{\zeta}$ und Fallnummern i mit

$$\delta(\tilde{\zeta}, i) = \begin{cases} 1, & \text{der i-te Fall gehört zu } \tilde{\zeta} \\ 0, & \text{der i-te Fall gehört nicht nicht zu } \tilde{\zeta}\,. \end{cases}$$

Liegt konkret die Stichprobe $\tilde{\zeta}$ vor mit der Größe $\tilde{n}$ (ζ), notiert sich der übliche Schätzer für $\bar{Y} = (\sum_{i=1}^{N} y_i)/N$ zu

$$\tilde{\bar{Y}} = \begin{cases} \sum_{i=1}^{N} y_i * \delta(\tilde{\zeta}, i)/\tilde{n}, & \text{falls } \tilde{n} \neq 0 \\ c, & \text{falls } \tilde{n} = 0 \end{cases}$$ (c sei eine passend gewählte Konstante).

Diese Schätzfunktion, die auf einer Stichprobe variabler Größe basiert, ist in naheliegender Weise zu vergleichen mit der Schätzfunktion

$$\hat{\bar{Y}} = (\sum_{i=1}^{N} y_i * \delta(\zeta, i))/n,$$

die sich auf eine einfache Zufallsstichprobe ζ ohne Zurücklegen der festen Größe $n = N * p$ stützt - es sei der Einfachheit halber angenommen, daß $N * p$ ganzzahlig ist. Zum Vergleich herangezogen wird die mittlere quadratische Abweichung von $\tilde{\bar{Y}}$ bzw. $\hat{\bar{Y}}$ vom "wahren" Wert $\bar{Y}$, also $E(\tilde{\bar{Y}} - \bar{Y})^2$ und $E(\hat{\bar{Y}} - \bar{Y})^2$, wobei letzteres gerade die Varianz von $\hat{\bar{Y}}$ ist.

Es gilt nun, daß $\tilde{\bar{Y}}$ nicht unverzerrt ist, d.h.

$$E\tilde{\bar{Y}} = c * (1-p)^N + \bar{Y} * (1-(1-p)^N), \text{ und daß}$$

$$E(\tilde{\bar{Y}} - \bar{Y})^2 = S_Y^2 * K(p,N) + (1-p)^N * (c-\bar{Y})^2 \quad \text{sowie}$$

$$E(\hat{\bar{Y}} - \bar{Y})^2 = S_Y^2 * (\frac{1}{N * p} - \frac{1}{N}),$$

wenn S_Y^2 die Varianz in der Gesamtdatei ist, also

$S_Y^2 = (\sum_{i=1}^{N} (y_i - \bar{Y})^2)/N-1$; $K(p,N)$ ist eine von p und N abhängige Zahl.

Man stellt zunächst fest, daß $\tilde{\bar{Y}}$ unverzerrt wäre, wenn man für c von vorneherein den in der Regel unbekannten Wert $\bar{Y}$ einsetzen würde. Man wird also c in der Nähe des vermuteten $\bar{Y}$ wählen. Allerdings kann die Verzerrung von $\tilde{\bar{Y}}$ praktisch als unbedeutend bezeichnet werden, da z.B. für $p = 0,1$ und $N = 100$ die Zahl $(1-p)^N$ nur von der Größenordnung 10^{-5} ist. Für größere Werte p und N werden die Verzerrungen noch weitaus geringer.

Für die mittleren quadratischen Abweichungen läßt sich zeigen, daß bei wachsendem N der Quotient

$$K(p,N)/\left(\frac{1}{N*p} - \frac{1}{N}\right) \text{ gegen 1 strebt.}$$

(Mathematisch interessierte Leser mögen dies als anregende Denkaufgabe verstehen.)
Folgerung: Für große N sind Zähler und Nenner praktisch gleich. Die Verhältnisse bei mittelgroßem N lassen sich aus der folgenden Tabelle ablesen. Die obere Hälfte der Tabellenfelder weist den Wert für $(1/N * p)-(1/N)$ aus, in der unteren Hälfte stehen jeweils die Werte für $K(p,N)$:

N \ p	0,5	0,4	0,3	0,2	0,1
100	0,010000 0,010206	0,015000 0,015391	0,023333 0,024159	0,040000 0,042210	0,090000 0,101524
200	0,005000 0,005051	0,007500 0,007596	0,011667 0,011867	0,020000 0,020524	0,045000 0,047502
300	0,003333 0,003356	0,005000 0,005043	0,007778 0,007866	0,013333 0,013563	0,030000 0,031071
400	0,002500 0,002513	0,003750 0,003774	0,005833 0,005883	0,010000 0,010128	0,022500 0,023091
500	0,002000 0,002008	0,003000 0,003015	0,004667 0,004698	0,008000 0,008081	0,018000 0,018375

Man beachte, daß mit wachsendem N die Abweichungen zwischen $(1/N * p) - (1/N)$ und $K(p,N)$ - allerdings in starker Abhängigkeit von p - sehr klein werden.

Insgesamt kann also gesagt werden, daß bei Mittelwertschätzern $\tilde{\bar{Y}}$, die mit Hilfe des "binomial sampling" gewonnen werden, schon bei relativ geringen Fallzahlen in der Auswahlgrundlage die mittleren quadratischen Abweichungen vom wahren Wert $\bar{Y}$ denjenigen gut entsprechen, die sich bei vergleichbaren einfachen Zufallsstichproben (ohne Zurücklegen, mit fester Größe) ergeben.

Anhang C

Wie im vierten Abschnitt angekündigt, wird im folgenden ein OSIRIS-SETUP angegeben, mit dessen Hilfe aus einer Datei mit 912 Fällen eine einfache Zufallsstichprobe der Größe 110 gezogen werden kann. Pro Fall enthält die Datei 12 Variablen. Die wie im Anhang A anzufügende "Stichprobenvariable" R13 hat genau 110 mal den Wert Eins und genau 802 mal den Wert Null:

```
ZWEI FILE-KOMMANDOS
FUER DIE EINGABEFILES
                        } Systemabhängig !
ZWEI FILE-KOMMANDOS
FUER DIE AUSGABEFILES

$RUN TRANS
$RECODE
        CARRY(R1,R2)
        R3 =RAND (Ø,2147483647)
        R4 =R3/2147483647
        R5 =(912 - R1) * R4
        R6 =11Ø - R2
        R1 =R1 + 1
        IF R5 GE R6 THEN R13 = Ø
      CELSE R13 = 1 AND R2 = R2 + 1
$SETUP
TITLE
TRAN, WIDTH = 1 *
V1 - V12 *
R13 *
```

Die Zahl $2147483647 = 2^{31}-1$ wird aus technischen Gründen als die größte ganze auf den Maschinen SIEMENS 4004 und IBM 360/370 darstellbare Zahl gewählt.

Für die Behandlung stichprobentheoretischer Problemstellungen ist bei ZUMA Hans-Peter Kirschner verantwortlich, der auch diesen Artikel verfaßt hat.

Literatur:

FAN, C.T., M.E. MULLER & I.REZUCHA. Development of sampling plans by using sequential (item by item) selection techniques and digital computers. J. American Statist. Ass. 1962, 57, 387-402.

KNUTH, D.E. The art of computer programming. Seminumerical algorithms. Mass.: Addison-Wesley Publishing Company, 1969.

KONIJN, H.S. Statistical theory of sample survey design and analysis. Amsterdam: North Holland Publishing Company, 1973.

VADUVA, J. Revage. A subroutine library for computer generation of random numbers, random variables, random vectors and stochastic processes. GMD-Mitteilung, 1976, 39.