Stichprobenziehung aus TwixTel*

Ben Jann†

April 2001

Nachfolgend wird beschrieben, wie aus der Telefon-CD-ROM TwixTel eine Zufallsstichprobe gezogen werden kann. Bei dem Verfahren wird ein Anwender simuliert, der nach einem Zufallsverfahren einzelne Einträge in der Datenbank ansteuert und exportiert. Zusätzlich werden einige Vorschläge diskutiert, die sich mit dem Auffinden von Mehrfacheinträgen beschäftigen.

1 Voraussetzungen

Die folgenden Programme werden für das Verfahren benötigt und sollten ordnungsgemäss auf dem Rechner installiert sein:

- TwixTel (Version 23, 11/2000, Einzelplatzversion)[1]
- WinBatch (www.windowware.com)[2]
- ein Tabellenkalkulationsprogramm wie z.B. Microsoft Excel

*Die Grundidee für das Verfahren stammt von Steffen Niemann, Institut für Sozial- und Präventivmedizin, Universität Bern

†Institut für Soziologie, Universität Bern, jann@soz.unibe.ch

[1]Das vorgestellte Verfahren wurde nur an dieser Version von TwixTel getestet. Unter Umständen kann es aber auch – ggf. mit gewissen Modifikationen – für Nachfolgeversionen verwendet werden.

[2]WinBatch kostet US$ 99.95, kann aber von Internet heruntergeladen und 30 Tage kostenlos verwendet werden.

2 Stichprobenziehung

In einem ersten Schritt sollte TwixTel gestartet werden, um die Grundgesamtheit zu bestimmen und einige Export-Einstellungen vorzunehmen.

- Export-Einstellungen: Über *Optionen>Export- und Druckoptionen* kann das Fenster *Exportformate* geöffnet werden. Der nachfolgende Algorithmus verwendet die Zwischenablage (clipboard) für den Datenexport, man wechsle also zu dem Register *Zwischenablage*. Als *Feldtrenner* sollte *Tabulator* gewählt und die Option *Datansatztrenner CR/LF* deaktiviert werden. Man bestimme schliesslich unter *Formate* welche Informationen genau exportiert werden sollen.
- Grundgesamtheit: TwixTel bietet verschiedene Möglichkeiten, die Adresssuche einzuschränken. Für eine Suche unter Deutschschweizer Privatadressen wähle man z.B. alle Kantone ausser *FL* (unter Register *Kantone*, Menu *Optionen>Erweiterte Suchoptionen*; Achtung: Die Auswahl muss dann noch in der Symbolleiste im Hauptfenster von TwixTel aktiviert werden) und aktiviere für *Sprache des Ortes Deutsch* (Register *Sprache/Neu auf CD*, Menu *Optionen>Erweiterte Suchoptionen*). Schliesslich wähle man dann noch die Option *Nur Privateinträge* (in der Symbolleiste im Haputfenster von Twixtel).

Es kann nun eine “Stern”-Suche gestartet werden, um alle Adressen in der Datenbank, die den Suchoptionen entsprechen, in einer Liste anzuzeigen (* in das Feld *Rubrik/Name/Text* eintragen und Suche starten). Insbesondere wird in der Statusleiste unten rechts nach Beendigung der Suche auch die Anzahl gefundener Einträge angezeigt (z.B. 2450192). Man notiere sich diese Nummer. Aus Datenschutz-Gründen kann in TwixTel nicht einfach die ganze Liste der gefundenen Adressen exportiert werden. Wir schlagen hier deshalb vor, ein Programm ausführen zu lassen, das einzelne Einträge aus der Liste nach Zufall auswählt.

Generierung von Zufallszahlen

Die Bestimmung der auszuwählenden Einträge erfolgt am einfachsten mit Hilfe der Generierung von Zufallszahlen, die die Positionen der Einträge in der Liste bezeichnen. Man bilde also mit einem Programm wie z.B. Excel eine Liste von *n* Zufallszahlen im Bereich zwischen 1 und der Anzahl gefundener Einträge (z.B. 2450192), wobei *n* der gewünschten Stichprobengrösse entspricht.[3]

Umrechnung der Distanzen in Tastaturanschläge

Die Distanzen zwischen den Zufallszahlen müssen jetzt in “pagedowns” und “downs” umgerechnet werden (“Wie oft muss ich die Taste “pagedown” und die Taste “down” bzw. “↓”

[3]Mit der Funktion “Zufallszahlengenerierung” des Add-Ins “Analyse-Funktionen” von Excel, Verteilung: Gleichverteilt. Hinweis: die Zufallszahlen sollten auf ganzzahlige Werte gerundet und doppelte Zahlen eliminiert werden.

drücken, um zu dem gewünschten Eintrag zu gelangen?").[4] Es wird empfohlen, die Angaben zusätzlich noch mit einer Laufnummer zu versehen. Die ganzen Zahlen sollten dann in eine Textdatei exportiert werden, wobei die einzelnen Spalten z.B. mit Tabulatoren getrennt werden (*Speichern unter...*, Dateityp: *Text (Tabs getrennt)*). Die Datei sollte etwa so aussehen:

```
224     7    16   1
3664    132  8    2
6879    123  17   3
7403    20   4    4
9347    74   20   5
...     ...  ...  ...
2447051 46   0    1599
2449220 83   11   1600
```

In der ersten Spalte stehen hier die Zufallszahlen, es folgt die Anzahl "PageDowns", die Anzahl "Downs" und die Laufnummer (Stichprobengrösse ist hier 1600).

Man könnte die gewünschten Adressen natürlich von Hand ansteuern, nur würde man dafür Tage brauchen. Einfacher ist es, mit WinBatch ein Programm zu erstellen, das einem die Arbeit abnimmt. Wir schlagen hier den folgenden Code vor:

```
indat = FileOpen("zufallszahlen.txt","READ")
outdat = FileOpen("stichprobe.txt","APPEND")
while @TRUE
      x = FileRead(indat)
      if x == "*EOF*" then break
      pagedown = ItemExtract(2,x,@TAB)
      down = ItemExtract(3,x,@TAB)
      id = ItemExtract(4,x,@TAB)
      WinActivate("TwixTel")
      if id == "1" then SendKey("{TAB 7}")
                  else MouseClick(@LCLICK,0)
      SendKey("{PGDN %pagedown%}{DOWN %down%}")
      MouseClick(@LCLICK,0)
      SendKey("!b~")
      adresse = ClipGet( )
      filewrite(outdat,"%id%%@TAB%%adresse%")
endwhile
FileClose(indat)
FileClose(outdat)
```

[4] Dazu muss erst gezählt werden, wie viele Einträge bei einem Pagedown in TwixTel übersprungen werden (z.B. 26; hängt von der Grösse des TwixTel-Fensters ab). Achtung: beim allerersten Pagedown überspringt TwixTel die doppelte Anzahl Einträge (also z.B. 52), weil sich der Cursor noch zuoberst im angezeigten Teil der Liste befindet. Dies muss bei den Berechnungen für die erste Zufallszahl berücksichtigt werden. Bei den nachfolgenden Bewegungen bleibt der Cursor immer auf der untersten vollständig angezeigten Zeile der Liste stehen.

Zuerst werden zwei Variablen für den Aufruf von externen Dateien definiert. Es handelt sich dabei um die Datei mit den Zufallszahlen, aus der Informationen gelesen werden sollen, und eine Datei, in die die ausgewählten Adressen exportiert werden.[5] Es folgt dann eine Schleife, in der die eigentliche Ziehung abläuft. Bei jedem Durchgang wird eine neue Zeile der Input-Datei gelesen und in die Variablen "pagedown", "down" und "id" aufgeteilt (falls das Ende der Datei erreicht ist (`*EOF*`), wird die Schlaufe beendet). Es wird dann TwixTel aktiviert und die entsprechende Anzahl "pagedowns" und "downs" werden ausgeführt. Wenn das Programm bei dem gewünschten Eintrag angelangt ist, wird ein Mausklick ausgeführt um den Eintrag zu markieren. Der Eintrag wird in die Zwischenablage kopiert (`!b~`) und die Zwischenablage inkl. Identifikationsnummer in die Output-Datei geschrieben. Der if-Befehl in dieser Sequenz hat zur Folge, dass beim ersten Fall vor der Ausführung der "pagedowns" noch von den TwixTel-Eingabefeldern in den Listenbereich navigiert (`SendKey("{TAB 7}")`)[6] und bei allen nachfolgenden Fällen jeweils die vorherige Auswahl mit einem Mausklick deaktiviert wird.

Wichtig: Weil in dem Programm Mausklicks verwendet werden (geht leider nicht anders), muss der Mauszeiger vor Ausführung des Programms an eine geeignete Stelle gesetzt werden. Der Zeiger muss auf die unterste vollständig sichtbare Zeile im TwixTel-Fenster weisen.

Hinweis: Falls die erste Zufallszahl sehr klein ist (kleiner als die Anzahl Einträge, die beim ersten Pagedown übersprungen wird; 52 in unserem Beispiel) führt der Algorithmus zu einem fehlerhaften Ergebnis und muss entsprechend modifiziert werden.

Die Ziehung der Einträge erfolgt relativ schnell (ca. 1000 pro Stunde; hängt vom Verhältnis von Stichprobengrösse zu Grundgesamtheit ab). Leseverzögerungen beim CD-ROM sollten kein Problem darstellen, da das Programm jeweils erst bei Bereitschaft des CD-ROMs weiterarbeitet. Die gezogene Stichprobe (`stichprobe.txt`) kann problemlos z.B. in Excel eingelesen und weiterverarbeitet werden.

3 Mehrfacheinträge

An eine einfache Wahrscheinlichkeitsauswahl wird die Anforderung gestellt, dass alle Elemente der Grundgesamtheit die gleiche Chance besitzen, ausgewählt zu werden. Bei der Ziehung aus der Telefon-CD-ROM ergeben sich diesbezüglich einige Probleme. Einerseits sind natürlich nicht alle Haushalte in dem Verzeichnis aufgeführt (z.B. weil seit einigen Jahren die Aufnahme ins Verzeichnis verhindert werden kann und es auch Haushalte ohne Telefonanschluss gibt). Andererseits können einzelne Haushalte mit mehreren Einträgen im Verzeichnis auftreten (z.B. mehrere Anschlüsse oder mehrere Einträge unter dem selben Anschluss). Das erste Problem kann bei der Ziehung aus dem Telefonverzeichnis nicht gelöst werden, mit dem

[5] Falls diese Datei schon besteht, wird ihr Inhalt nicht gelöscht, sondern sie wird einfach mit weiterer Information aufgefüllt. Da dies normalerweise nicht gewünscht ist, müssen Daten aus früheren Versuchen ggf. vorher manuell entfernt werden.

[6] Die Anzahl notwendiger Tabs kann je nach Einstellung der Suchoptionen variieren.

zweiten sollte man sich aber etwas näher beschäftigen. Die Existenz von Mehrfacheinträgen verzerrt die Auswahlwahrscheinlichkeiten und es gibt prinzipiell zwei Möglichkeiten damit umzugehen. Erstens könnte man die Auswahlwahrscheinlichkeiten bei der Ziehung korrigieren, indem man jeweils einen Teil der Elemente in der Stichprobe, die mehrere Einträge im Verzeichnis aufweisen, nach Zufall aus der Stichprobe eliminiert (z.B. haben Haushalte mit 2 Einträgen doppelte Auswahlwahrscheinlichkeit, die Hälfte dieser Adressen sollte also nach Zufall wieder aus der Stichprobe gelöscht werden). Zweitens kann man die Stichprobe so belassen wie sie ist und bei der Datenauswertung eine Gewichtung einführen, die die Auswahlwahrscheinlichkeiten korrigiert. In beiden Fällen muss man aber die Anzahl Einträge der einzelnen Haushalte in der Stichprobe kennen.

Wiederum kann man dies von Hand ermittelt, in dem man z.B. alle Einträge unter einer bestimmten Telefonnummer in TwixTel sucht, was aber wohl wenig effizient ist. Es folgen nun zwei weitere WinBatch-Codes die diesen Vorgang zumindest teilweise automatisieren.

Suche nach mehreren Einträgen unter der gleichen Telefonnummer:

```
indat = FileOpen("stichprobe.txt","READ")
outdat = FileOpen("eintraege_tel.txt","APPEND")
while @TRUE
      x = FileRead(indat)
      if x == "*EOF*" then break
      id = ItemExtract(1,x,@TAB)
      tel = ItemExtract(8,x,@TAB)
      ClipPut("")
      WinActivate("TwixTel")
      SendKey("{TAB 7}%tel%~")
      SendKey("!b~")
      clipboard = ClipGet( )
      if clipboard==""
            anzahl = AskLine("","","1")
            filewrite(outdat,"%id%%@TAB%%anzahl%")
      else
            filewrite(outdat,"%id%%@TAB%1")
            SendKey("{TAB}")
      endif
endwhile
FileClose(indat)
FileClose(outdat)
```

Das Programm führt für jede Telefonnummer in der Stichprobe eine Suche durch.[7] Immer wenn unter einer Telefonnummer mehrere Einträge gefunden werden, meldet sich das Programm mit einem Dialogfeld, in das die Anzahl Einträge manuell eingetragen werden muss

[7] Vielen Telefonnummern wird beim Export in die Datei `stichprobe.txt` ein * vorangestellt. Diese Zeichen sollten in `stichprobe.txt` vor Ausführung des Programms gelöscht werden (z.B. mit der Ersetzen-Funktion eines Texteditors).

(die genaue Anzahl lässt sich m.W. nicht automatisch ermitteln). Liegt nur ein Eintrag vor, so fährt das Programm automatisch weiter. Das Resultat ist eine Datei `eintraege_tel.txt`, in der für jede Adresse der Stichprobe die Anzahl Einträge aufgeführt wird.

Suche nach mehreren Anschlüssen unter gleichem Namen und gleicher Adresse:

```
indat = FileOpen("stichprobe.txt","READ")
outdat = FileOpen("eintraege_adr.txt","APPEND")
while @TRUE
      x = FileRead(indat)
      if x == "*EOF*" then break
      id = ItemExtract(1,x,@TAB)
      vorn = ItemExtract(2,x,@TAB)
      name = ItemExtract(3,x,@TAB)
      adr = ItemExtract(4,x,@TAB)
      plz = ItemExtract(6,x,@TAB)
      ort = ItemExtract(7,x,@TAB)
      ClipPut("")
      WinActivate("TwixTel")
      SendKey("%name% {TAB}%vorn% {TAB}%adr% {TAB}%plz% %ort%~")
      SendKey("!b~")
      clipboard = ClipGet( )
      if clipboard==""
            anzahl = AskLine("","","1")
            filewrite(outdat,"%id%%@TAB%%anzahl%")
      else
            filewrite(outdat,"%id%%@TAB%1")
            SendKey("{TAB}")
      endif
endwhile
FileClose(indat)
FileClose(outdat)
```

Dieses Programm funktioniert ähnlich wie das Programm zur Identifizierung von Mehrfacheinträgen unter der gleichen Telefonnummer, ausser dass hier nach Einträgen mit gleichem Namen und gleicher Adresse gesucht wird (Haushalte mit mehreren Anschlüssen). Zu beachten ist hier, dass sich Überschneidungen zum ersten Programm ergeben können (Einträge unter gleichem Namen, gleicher Adresse *und* gleicher Telefonnummer). Als Anzahl sollte daher hier jeweils nur die Anzahl Einträge mit *unterschiedlichen* Telefonnummern erfasst werden. Zu bemerken ist zudem, dass mit dem Verfahren unter Umständen nicht alle Anschlüsse eines Haushaltes erfasst werden können (wenn sich der Name unterscheidet).