Schätzung der Varianz bei Stichprobenerhebungen mit einem Jackknife Verfahren

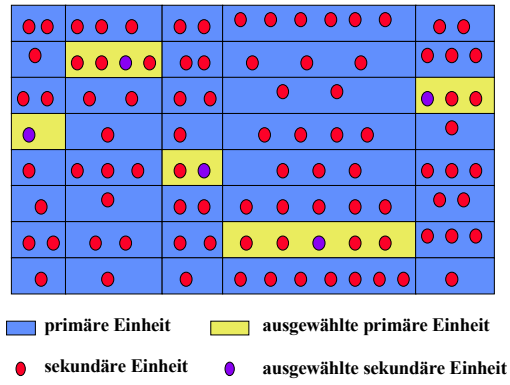
Christian Vonlanthen, Markus Eichenberger
Bundesamt für Statistik, CH-2010 Neuchatel, Bundesamt für Informatik, CH-3003 Bern

Kurzfassung

Bei statistischen Auswertungen von Stichprobenerhebungen muss oft die Varianz geschätzt werden, etwa zur Berechnung von Konfidenzintervallen. Wenn es sich dabei nicht um eine einfache Zufallsstichprobe handelt, sollten die normalen Prozeduren der verbreiteten statistischen Software-Pakete wie SAS nicht verwendet werden, da die Varianz eher unterschätzt wird. Ferner gibt es in diesen Fällen oft auch keine geschlossene Formel, um letztere exakt zu bestimmen. Eine Möglichkeit ist der Einsatz von sogenannten Resampling-Verfahren. Unter diesen ist das Jackknife-Verfahren eine Variante. Der Beitrag zeigt eine Implementation dieses Algorithmus am Beispiel der schweizerischen Gesundheitsbefragung.

Stichprobenplan der Schweizerischen Gesundheitsbefragung 1992

Die Stichprobe der Schweizerischen Gesundheitsbefragung setzt sich aus 4 Unterstichproben für je eine Jahreszeit zusammen. Jede Unterstichprobe stammt von einer zweistufigen geschichteten Stichprobe. Die 12 Schichten (Variable Ort) repräsentieren geographische Einheiten, und zwar Gruppen von Schweizer Kantonen. In jeder Schicht wird eine zweistufige Zufallsstichprobe gezogen (vgl. nebenstehende Graphik). Die Privathaushalte stellen die Auswahleinheiten erster Stufe (primäre Einheiten) dar; die 15-jährigen oder älteren in diesen Haushalten lebenden Personen sind die Auswahleinheiten zweiter Stufe (sekundäre Einheiten).



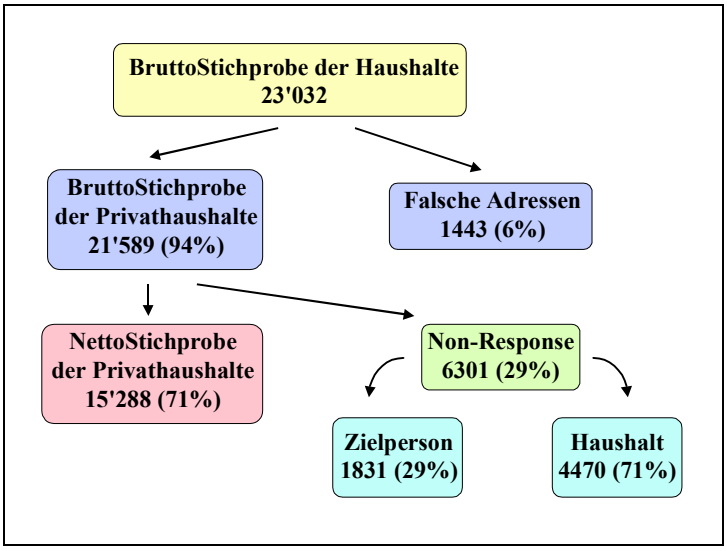
Die landesweite Bruttostichprobe umfasst 23'032 Adressen (Telefonnummer), darunter 1443 Stichprobenfehler, die auf folgende Quellen zurückzuführen sind:

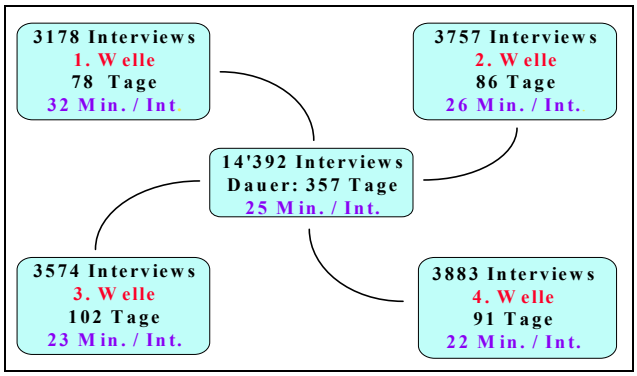
- Qualität des Stichprobenrahmens (41% der Fälle),
- Adressen von Kollektivhaushalten (31,9% der Fälle),
- Adressen von Ferienhäusern (20,5% der Fälle),
- andere (6,6% der Fälle).

Die landesweite Bruttostichprobe der Privathaushalte setzt sich somit effektiv aus 21'589 Einheiten zusammen. Die 4470 Antwortverweigerungen von primären Einheiten (Privathaushalte) sowie die 1831 von sekundären Einheiten (Zielpersonen) reduzieren die landesweite Nettostichprobe auf 15'288 Personen, dies bei einer Beteiligungsquote von knapp über 70%. Im folgenden sind die wichtigsten genannten Gründe für eine Nichtteilnahme an der Schweizerischen Gesundheitsbefragung aufgeführt:

	Welle 1		Welle 2		Welle 3		Welle 4		Total	
	Person	Haushalt	Person	Haushalt	Person	Haushalt	Person	Haushalt	Person	Haushalt
Zeitmangel	22%	14%	12%	9%	9%	11%	8%	11%	12%	12%
Kein Interesse	16%	28%	25%	43%	35%	39%	36%	40%	29%	36%
Einstellung gegenüber Befragungen	15%	18%	12%	16%	17%	13%	13%	11%	14%	15%
Zu persönliches Thema	5%	5%	5%	3%	4%	3%	2%	2%	4%	3%
Alter	12%	9%	12%	10%	11%	14%	9%	10%	11%	10%
Andere Gründe	30%	26%	34%	19%	24%	20%	32%	26%	30%	24%

Das nebenstehende Diagramm fasst den Weg zur Ermittlung der landesweiten Nettostichprobe zusammen.





Die nebenstehende Graphik zeigt die Dauer der 574 Proxy- und der 13'818 telefonischen Interviews gegliedert nach Wellen.

Aus der mittleren Interviewdauer geht hervor, dass der Fragebogen für die telefonischen Interviews trotz seiner insgesamt rund 400 Fragen nicht zu lang ist. Aufgrund der eingebauten Filter muss keine der Zielpersonen sämtliche Fragen beantworten.

Mangels einer genügenden Infrastruktur zur Durchführung von CATI- (computer-assisted telephone interviews) und Face-to-Face-Interviews sowie mangels praktischer Erfahrungen in der Führung eines Befragungsdienstes hat das Bundesamt für Statistik ein Erhebungsinstitut mit der Realisierung der Interviews bei den Zielpersonen beauftragt.

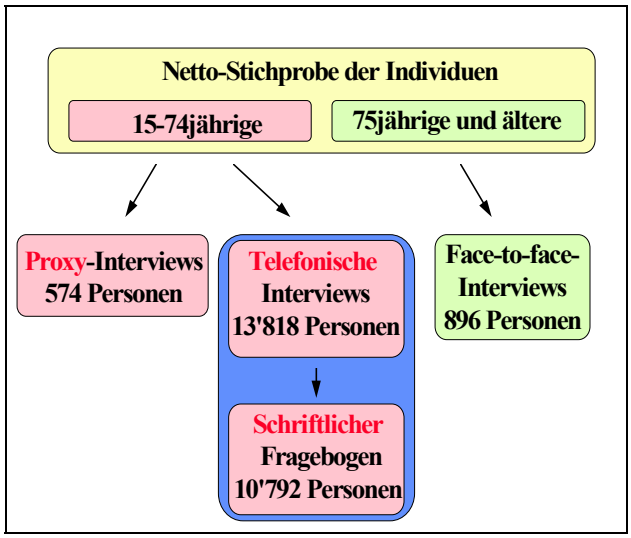
Interviewmethoden

Wie nachfolgend dargestellt, wurden bei der Realisierung der 15‘288 Interviews 4 Methoden angewandt.

1. Die 896 an 75-jährige oder ältere Personen gerichteten Befragungen erfolgten in Form von "Face-to-Face"-Interviews.

2. Die Personen unter 75 Jahren wurden per telefonischen Interviews befragt.
   574 Auswahleinheiten dieser Gruppe erhielten einen Proxy-Fragebogen, bei dem eine Drittperson die Fragen für die Zielperson beantworten muss.

3. Die 13'818 Personen, die telefonisch Auskunft gegeben hatten, erhielten nach dem Befragungsgespräch noch einen schriftlichen Fragebogen. Die Beteiligungsquote an dieser Untererhebung betrug etwas mehr als 78% (10'792 Interviews bei 13'818 maximal möglichen Befragungen).



Analyse der Antwortausfälle

Bereits vor der Realisierung der Erhebung war uns bewusst, dass die gewählte Stichprobenmethode sowohl die Jungen (aus Mobilitätsgründen) als auch die Betagten (da sie in Kollektivhaushalten, d.h. Altersheimen, leben) benachteiligen würde. Die folgende Tabelle illustriert diese Unter- und Übervertretungen:

	Stichprobe		Population		Ziffer[1]
Umfang	15'288		5'683'260		
Männer	6855	44.84%	2'746'590	48.33%	0.79
Frauen	8433	55.16%	2'936'670	51.67%	1.07
15-19 Jahre	687	4.49%	404'773	7.12%	0.63
20-24 Jahre	1115	7.29%	492'458	8.67%	0.84
25-29 Jahre	1810	11.84%	583'948	10.27%	1.15
30-34 Jahre	1731	11.32%	567'516	9.99%	1.13
35-39 Jahre	1538	10.06%	522'274	9.19%	1.09
40-44 Jahre	1334	8.73%	508'207	8.94%	0.98
45-49 Jahre	1261	8.25%	495'304	8.72%	0.95
50-54 Jahre	1137	7.44%	409'134	7.20%	1.03
55-59 Jahre	1031	6.74%	367'355	6.46%	1.04
60-64 Jahre	972	6.36%	330'997	5.82%	1.09
65-69 Jahre	918	6.00%	296'771	5.22%	1.15
70-74 Jahre	792	5.18%	245'794	4.32%	1.20
75-79 Jahre	418	2.73%	197'188	3.47%	0.79
80 Jahre und älter	544	3.56%	261'541	4.60%	0.77
Schweizer	13'203	86.36%	4'698'152	82.67%	1.04
Ausländer	2085	13.64%	985'108	17.33%	0.79

Wie zu erwarten war, sind auch die Ausländer in unserer Stichprobe untervertreten.

Gewichtung

Ursprüngliche Gewichtung

Sei

N	=	Anzahl Privathaushalte in der Bevölkerung
n	=	Anzahl Privathaushalte in der Stichprobe
M	=	Grösse des Haushalts

dann ist die Einschlusswahrscheinlichkeit einer Einzelperson gegeben durch

$$\pi_i = \frac{n}{N}\frac{1}{M}$$

Daraus folgt das ursprüngliche Gewicht

$$Schicht_i = \frac{1}{\pi_i}$$

Die Mengen

$$\hat{\boldsymbol{Y}} = \sum_{k \in S} \frac{\boldsymbol{y}_k}{\boldsymbol{\pi}_k}$$

$$\hat{\boldsymbol{P}} = \frac{\mathbf{1}}{\sum_{k \in S} \boldsymbol{\pi}_k} \sum_{k \in S} \frac{\boldsymbol{y}_k}{\boldsymbol{\pi}_k}$$

definieren somit einerseits den Horvitz-Thomson-Schätzer eines Totals und den Hajek-Schätzer eines Prozentwerts.

Schlussendliche Gewichtung

Zur Reduktion des Bias im Zusammenhang mit Antwortverweigerungen und Stichprobenfehlern wird eine Kalibrierung vorgenommen. Die Ränder sind wie folgt gegeben:

1. Geschlecht * Ort
2. Alter * Ort (15≤Alter<35; 35≤Alter<50; 50≤Alter<65; 65≤Alter)
3. Nationalität. (schweizerische/ausländische)

Auf Grund dieser Gewichtungsmethode ist keine Varianzformel wirklich geeignet. Die Jackknife-Methode liefert jedoch eine gute Schätzung der Varianz.

Jackknife-Varianzschätzer

S beschreibt eine Stichprobe, die zufällig in g Teilstichproben von ganz oder fast identischer Grösse unterteilt wird. Diese g Teilstichproben, ausgedrückt S(α), werden *Abbilder der Population* gennannt.

Ist S eine einfache geschichtete Zufallstichprobe, so lässt sich jede Schicht h zufällig in g Gruppen von identischer Grösse aufteilen; diese Gruppen definieren die *Abbilder der Schichten.*

In diesem Fall gelten die *Abbilder der Population* als Summe der Abbilder des Schichten.

Für jedes der g Abbilder der Schichten werden die g Pseudowerte konstruiert

$$Y_\alpha = g\hat{Y} - (g-1)Y_{(\alpha)} \qquad \alpha = 1,2,...,g.$$

wobei $\hat{Y}$ der Schätzer einer Untersuchungsvariablen Y ist;
$Y_{(\alpha)}$ ist der Schätzer von Y nach Entfernung des α-ten Abbildes der Population.

Zum Beispiel im Fall des Schätzer des Totals

$$\hat{Y} = \sum_{k \in S} \omega_k y_k$$

wobei ω_i das Gewicht des Individuums k ist, können die $Y_{(\alpha)}$ geschätzt werden durch

1. $\frac{g}{g-1} \sum_{k \in S - S_{(\alpha)}} \omega_k y_k$
2. $\frac{\sum_{k \in S} \omega_k}{\sum_{k \in S - S_{(\alpha)}} \omega_k} \sum_{k \in S - S_{(\alpha)}} \omega_k y_k$

Der "Jackknife"-Schätzer wird dann als Mittel der Y_α definiert, d.h.

$$Y_J = \frac{1}{g} \sum_{\alpha=1}^{g} Y_\alpha$$

Und schliesslich wird der Schätzer der Varianz von Y_J

$$\text{var}(Y_J) = \frac{1}{g(g-1)} \sum_{\alpha=1,2,...,g} \left(Y_\alpha - Y_J\right)^2$$

als Schätzer der Varianz Y verwendet.

Hinweis: Im Grunde genommen müsste für jeden Wert $Y_{(\alpha)}$ eine Neugewichtung vorgenommen werden. Aus Einfachheitsgründen wurde hier jedoch auf dieses Verfahren verzichtet.

Beschreibung der Makros

Allgemeines

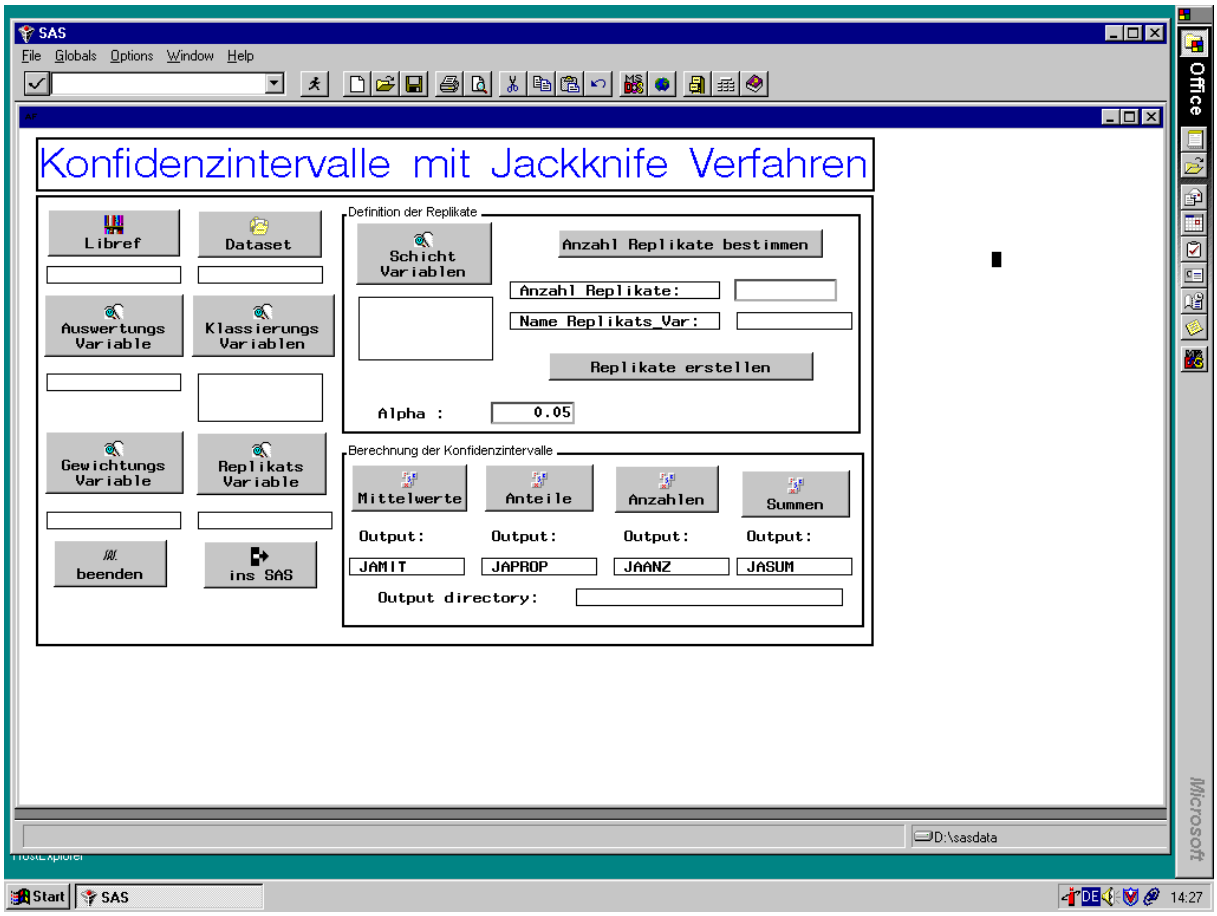
Für die Berechnung von Konfidenzintervallen bei geschichteten Stichprobenerhebungen wurden 4 Makros auf IML Basis entwickelt, und zwar :

%jackmit : für Mittelwerte
%jackprop: für prozentuale Anteile
%jacksum: für Summen
%jackanz : für Anzahlen

Folgende Parameter werden den Makros übergeben :

data=SAS_DATASET Auswertungsdatei, daher Angabe zwingend.
var=BERECHNUNGS_VARIABLE . Für diese Variable wird die Varianz und das Konfidenzintervall berechnet, daher zwingend. Bei Anteilen und Anzahlen ist die Variable qualitativer, bei Mittelwerten und Summen quantitativer Art
gewvar=GEWICHTUNGS_VARIABLE. Optionaler Parameter, fehlt diese Angabe, hat jede Observation das Gewicht 1.
Class=GRUPPIERUNGS_VARIABLE. Optional. Hier kann eine Unterteilung festgelegt werden, etwa nach Geschlecht, Altersklassen etc. Fehlt die Variable, erfolgt die Auswertung über die ganze Stichprobe.
Repl=REPLIKAT_VARIABLE. Optional. Dann wird eine Zufallsstichprobe angenommen und im Verfahren jeweils nur eine Beobachtung entfernt.

Einbindung der Makros in eine AF-Oberfläche



Mit der AF-Oberfläche kann der Benutzer dieselben Parameter den Makros übergeben, wie beim Aufruf im Programm Editor, ohne sich um die Makro-Syntax kümmern zu müssen. Darüber hinaus bietet diese Variante noch die Möglichkeit, die Replikatsvariable zu generieren, falls diese nicht schon im SAS-Data Set vorhanden ist. Ist letzteres der Fall, so kann diese direkt mit ihrem Nanem eingegeben werden. Ferner kann man sich bezüglich der Anzahl Replikate einen Vorschlag berechnen lassen oder die Anzahl explizit eingeben. Die vorgeschlagene Anzahl beruht auf der Faustregel, die Anzahl Einheiten der kleinsten Schicht zu nehmen. Beim Aufruf im Programm Editor muss die Replikatsvariable schon vorhanden sein.
Beispiel :

```
title1 "Aufruf im SAS Programm Editor";
%jackmit(data=ges92ch,var=gewicht,gewvar=wght,class=sex gesund,repl=rep)
```

Ein Vergleich mit den Jackknife-Makros von SAS Institute

Auf dem WWW-Server von SAS Institute stehen Makros für Jackknife- und Bootstrap-Verfahren zur Analyse von einfachen Zufallsstichproben („simple random samples“) zur Verfügung. Die Anwendung derselben ist aber für einen Endbenutzer eher ungeeignet, da Kenntnisse in Makro-Programmierung nötig sind. Beim Jackknife wird jeweils eine Beobachtung aus dem Datensatz entfernt, die Programme sind bei grösseren Files sehr rechenintensiv. (Für Download: www.sas.com/techsup/download/stat/jackboot.sas)

Konfidenzintervall : Mittelwert Gewicht

Geschlecht	Gesundheitszustand	Name	Observed Statistic	Jackknife Mean	Estimated Bias	**Estimated Standard Error**
1	1	MITTEL	75.8025	75.8025	0.000145	0.31708
1	2	MITTEL	75.3851	75.3851	0.000075	0.21856
1	3	MITTEL	75.8623	75.8623	0.000912	0.48854
1	4	MITTEL	75.9788	75.9788	0.002974	1.44479
1	5	MITTEL	75.9440	75.9440	0.021344	2.02982
2	1	MITTEL	60.1189	60.1189	-0.000041	0.23674

Geschlecht	Gesundheitszustand	Name	Estimated Lower Confidence Limit	Bias-Corrected Statistic	Estimated Upper Confidence Limit
1	1	MITTEL	75.1809	75.8024	76.4238
1	2	MITTEL	74.9566	75.3850	75.8134
1	3	MITTEL	74.9039	75.8614	76.8190
1	4	MITTEL	73.1441	75.9758	78.8076
1	5	MITTEL	71.9443	75.9226	79.9010
2	1	MITTEL	59.6549	60.1189	60.5829

Geschlecht	Gesundheitszustand	Name	Confidence Level (%)	Method for Confidence Interval	Minimum Resampled Estimate
1	1	MITTEL	95	Jackknife	75.6679
1	2	MITTEL	95	Jackknife	75.3411
1	3	MITTEL	95	Jackknife	75.7680
1	4	MITTEL	95	Jackknife	75.3379
1	5	MITTEL	95	Jackknife	75.1662
2	1	MITTEL	95	Jackknife	60.0771

Maximum

Geschlecht	Gesundheitszustand	Name	Resampled Estimate	Number of Resamples
1	1	MITTEL	75.8418	15288
1	2	MITTEL	75.4032	15288
1	3	MITTEL	75.9372	15288
1	4	MITTEL	76.4798	15288
1	5	MITTEL	76.7253	15288
2	1	MITTEL	60.1503	15288

Konfidenzintervall : Mittelwert Groesse

Geschlecht	Gesundheitszustand	Name	Observed Statistic	Jackknife Mean	Estimated Bias	**Estimated Standard Error**
2	2	MITTEL	61.1825	61.1825	0.000003	0.17890
2	3	MITTEL	61.9865	61.9865	0.000642	0.38629
2	4	MITTEL	61.3473	61.3473	0.000573	0.72764
2	5	MITTEL	63.0496	63.0496	0.007528	1.90929

Geschlecht	Gesundheitszustand	Name	Estimated Lower Confidence Limit	Bias-Corrected Statistic	Estimated Upper Confidence Limit
2	2	MITTEL	60.8319	61.1825	61.5331
2	3	MITTEL	61.2287	61.9858	62.7429
2	4	MITTEL	59.9206	61.3467	62.7729
2	5	MITTEL	59.2999	63.0420	66.7842

Geschlecht	Gesundheitszustand	Name	Confidence Level (%)	Method for Confidence Interval	Minimum Resampled Estimate
2	2	MITTEL	95	Jackknife	61.1519
2	3	MITTEL	95	Jackknife	61.9006
2	4	MITTEL	95	Jackknife	61.1675
2	5	MITTEL	95	Jackknife	62.1553

Geschlecht	Gesundheitszustand	Name	Maximum Resampled Estimate	Number of Resamples
2	2	MITTEL	61.1952	15288
2	3	MITTEL	62.0355	15288
2	4	MITTEL	61.5320	15288
2	5	MITTEL	63.7555	15288

Die IML-Makros ergeben das gleiche Resultat, ohne Angabe einer Replikatsvariable wird ebenfalls eine Beobachtung entfernt. Die Rechenzeit relativ zu den Makros von SAS ist dagegen bescheiden.

Konfidenzintervall : Mittelwert Gewicht

Dataset : GES92CH / Gewichtungs_Var : WGHT

SEX	GESUND	Anzahl	M_GEWICHT	**Std_err**	U_Konf_grz	O_Konf_grz	Var_koeff
1	1	2006	75.80	0.317	75.181	76.424	0.42

1	2	3819	75.39	0.219	74.957	75.813	0.29
1	3	720	75.86	0.489	74.905	76.820	0.64
1	4	192	75.98	1.445	73.147	78.811	1.90
1	5	43	75.94	2.030	71.966	79.922	2.67
2	1	2182	60.12	0.237	59.655	60.583	0.39
2	2	4617	61.18	0.179	60.832	61.533	0.29
2	3	1142	61.99	0.386	61.229	62.744	0.62
2	4	307	61.35	0.728	59.921	62.773	1.19
2	5	42	63.05	1.909	59.307	66.792	3.03

Berechnung von Konfidenzintervallen ohne Einbezug der Schichtung (als SRS), mit den IML-Makros und der Software Wesvar unter Berücksichtigung der Schichtung

Die Grösse der Stichprobe der Gesundheitsbefragung 1992 beträgt 15288 Einheiten. In diesem Abschnitt soll ein Vergleich gemacht werden, wie sich die Vertrauensintervalle verhalten, wenn man die Daten als einfache Zufallsstichprobe (SRS), oder unter Berücksichtigung der Schichtung betrachtet. Im Falle der Schichtung wurde noch ein anderes Softwareprodukt herangezogen, nämlich Wesvar von der US Firma WESTAT. Diese Software ist bis Version 2.1 frei erhältlich und kann im Internet heruntergeladen werden.

Simple Random Sample (mit Procedure Means)

Konfidenzintervalle nach Simple Random Sample

SEX	GESUND	MITTEL	STDM	DF	U_GRENZE	O_GRENZE
1	1	75.8025	0.25939	2005	75.2938	76.3112
1	2	75.3851	0.18683	3818	75.0188	75.7514
1	3	75.8623	0.44022	719	74.9981	76.7266
1	4	75.9788	1.16467	191	73.6815	78.2761
1	5	75.9440	1.77582	42	72.3602	79.5277
2	1	60.1189	0.20436	2181	59.7181	60.5196
2	2	61.1825	0.15063	4616	60.8872	61.4778
2	3	61.9865	0.34108	1141	61.3172	62.6557
2	4	61.3473	0.62637	306	60.1148	62.5798
2	5	63.0496	1.85025	41	59.3129	66.7862

SAS-Code für obigen Output (generiert mit SAS/ASSIST) :

```
options linesize=76 pagesize=58 nodate number pageno=1;
title "Konfidenzintervalle nach Simple Random Sample";
footnote;
proc means noprint nway  data=me.ges92ch vardef=wgt;
   var gewicht ;
   class sex gesund;
   weight  wght;
   output out =sasast1
          n   =n1
          mean=mean1
          std =std1;
run;
data sasast2;
   set sasast1;
   namelist = "GEWICHT";
   array num   { 1 } n1;
   array avg   { 1 } mean1;
```

```
   array stdv  { 1 } std1;
   halfalf = 1 - ( (1 - .95) / 2 );

   do i = 1 to 1;
      level = 95;
      df = num{ i } - 1;
      if ( df < 0 ) then df = . ;
      mittel = avg{ i };
      stdm = stdv{ i } / sqrt( num{ i } );
      est = tinv( halfalf , df ) * stdv{ i } / sqrt( num{ i } );
      u_grenze = avg{ i } - est;
      o_grenze = avg{ i } + est;
      output;
      end;
run;
proc print data=sasast2 noobs;
   var sex gesund mittel stdm df u_grenze o_grenze;
run;
```

IML-Makros

The SAS System

Dataset : GES92CH / Gewichtungs_Var : WGHT

SEX	GESUND	Anzahl	M_GEWICHT	Std_err	U_Konf_grz	O_Konf_grz	Var_koeff
1	1	2006	75.80	0.328	75.160	76.445	0.43
1	2	3819	75.39	0.233	74.928	75.843	0.31
1	3	720	75.86	0.485	74.911	76.813	0.64
1	4	192	75.98	1.409	73.216	78.741	1.86
1	5	43	75.94	2.067	71.893	79.995	2.72
2	1	2182	60.12	0.222	59.685	60.553	0.37
2	2	4617	61.18	0.177	60.835	61.530	0.29
2	3	1142	61.99	0.397	61.209	62.764	0.64
2	4	307	61.35	0.779	59.821	62.874	1.27
2	5	42	63.05	1.922	59.283	66.816	3.05

Wesvar 2.11 for Windows 95

TABLE REQUEST : SEX * GESUND

SEX	GESUND	STATISTIC	EST_TYPE	ESTIMATE	STDERROR	LOWER	UPPER
1	1	m_gewicht	VALUE	75.80	0.328	75.16	76.44
1	2	m_gewicht	VALUE	75.39	0.233	74.93	75.84
1	3	m_gewicht	VALUE	75.86	0.485	74.91	76.81
1	4	m_gewicht	VALUE	75.98	1.409	73.22	78.74
1	5	m_gewicht	VALUE	75.94	2.067	71.89	80.00
1	MARGINAL	m_gewicht	VALUE	75.57	0.180	75.22	75.93
2	1	m_gewicht	VALUE	60.12	0.222	59.68	60.55
2	2	m_gewicht	VALUE	61.18	0.177	60.83	61.53
2	3	m_gewicht	VALUE	61.99	0.397	61.21	62.76
2	4	m_gewicht	VALUE	61.35	0.779	59.82	62.87
2	5	m_gewicht	VALUE	63.05	1.922	59.28	66.82
2	MARGINAL	m_gewicht	VALUE	61.02	0.126	60.78	61.27

Warning: 218 observations were excluded from the preceding table.
These observations were excluded because they contained

```
        one or more requested variables with missing values.
```

Ein Vergleich der Werte zeigt, dass der Standardfehler bei der Zufallstichprobe unterschätzt wird und somit bei Stichprobenerhebungen die „klassischen Prozeduren“ – bei SAS etwa MEANS, UNIVARIATE - mit Vorsicht bzw. nicht anzuwenden sind.

SAS Code für die Makro JACKMIT

```
%macro jackmit(data=,var=,gewvar=,class=,repl=);
  /* Jackknife zur Berechnung von Konfidenzintervallen
     von Mittelwerten mit Elimination von Replikaten
 data   : Auswertungsdatei     (required)
 var    : Variable für Konfidenzintervalles (required)
 gewvar : Gewichtungsvariable (optional)
 class  : Variablen für die Bildung von Untergruppen (optional)
 repl   : Identifikation des Replikates
  Die Variablen werden alle als numerisch vorausgesetzt
  */
%let alpha=0.05;
options  linesize=96 pagesize=40  nodate pageno=1;
title3  " Dataset : &data  /  Gewichtungs_Var : &gewvar";
 %if &repl ne  %then %do;
  data sub0;
    keep &var &gewvar &class &repl;
    set &data;
  run;
 %end;
 %else %do;
  %let repl=repl;
  data sub0;
    keep &var &gewvar &class &repl;
    set &data;
    repl=_n_;
  run;
 %end;
proc sort data=sub0;
     by &repl;
run;
 %let jd1=;
 %let jd2=;
 %let jd3=;
%if &class ne  %then  %do;
 %let anzcla=1;
 %let jd1=jadat1;
 %let cl1=%scan(&class,1);
 %let cl2=%scan(&class,2);
 %let cl3=%scan(&class,3);
 %let clst=cl=compress(left(trim(&cl1)),' ');
 %if  &cl2 ne  %then %do;
    %let jd2=jadat2;
    %let anzcla=2;
    %let clst=&clst||' '||compress(left(trim(&cl2)),' ');
 %end;
 %if  &cl3 ne  %then %do;
    %let jd3=jadat3;
```

```
    %let anzcla=3;
    %let clst=&clst||' '||compress(left(trim(&cl3)),' ');
 %end;
 proc  sort  data=sub0    out=tot;
       by &class  &repl;
 run;
 proc  means  data=tot noprint  nway;
       class &class;
       output  out=str  n=n;
 run;
 data _null_;
  call symput('anzstr',trim(left(put(nobs,8.))));
   if  0 then set str nobs=nobs;
   stop;
 run;

 data tot;
   length cl $ 20;
   set tot;
   &clst;
 run;

 data str2;
  length  cl $ 20;
   set str;
   &clst;
 run;

 data _null_;
    set str2;
     call symput('wert'!!left(_n_),left(trim(cl)));
 run;
 %do i=1 %to &anzstr;
    data sub&i;
    keep &var &gewvar &class &repl;
      set tot;
      if cl = "&&wert&i"  then  output;
    run;
 %end;
%end;
%else  %do;
  %let anzstr=1;
  %let wert1=.;
  data sub1;
    set sub0;
  run;
%end;

proc  iml  workspace=4096;
  flagr=1;
  reset noname;
 use sub0;
 read all var {&repl} into r where(&var ^= .);
 read all var {&var}  into x where(&var ^= .);
 ru=unique(r)`;
```

```
 nruanz=nrow(ru);
 n=nrow(x);
 if n = nruanz  then flagr = 0;
 free x r;
%do  ks=1  %to  &anzstr;
 use   sub&ks;
 %if  &gewvar ne  %then  %do;
 read all var  {&var &gewvar  &repl} into gw where(&var ^= .) ;
 %end;
 %else %if &gewvar =  %then  %do;
 read all var{&var &repl} into  gw   where(&var ^= .) ;
 %end;
 %if &gewvar ne  %then %do;
 x=gw[,1];
 w=gw[,2];
 r=gw[,3];
 n=nrow(x);
 if flagr = 0 then do;
   do i=1 to n;
     r[i]=i;
   end;
 end;
 %end;
 %else  %do;
 x=gw[,1];
 r=gw[,2];
 n=nrow(x);
 w=j(n,1,1);
 if flagr = 0 then do;
   do i=1 to n;
     r[i]=i;
   end;
 end;
 %end;
 if min(r) = 0  then r=r+1;
 ru=unique(r)`;
 nru=nrow(ru);
 if flagr=1 & (min(ru) ^= 1 | nru ^= nruanz | max(ru) > nruanz)
 then do;
 do  i=1  to  n;
     wrp=r[i];
    do  j=1  to  nru;
      if wrp=ru[j]  then    do;
         r[i]=j;
         goto next;
       end;
    end;
   next:
  end;
 end;

 sumwx=sum(x#w);
 sumw=sum(w);
 xmit=sumwx/sumw;
```

```
 swxrep=j(nruanz,1,0);
 swrep=j(nruanz,1,0);
 do i=1 to n;
    jj=r[i];
    swxrep[jj]=swxrep[jj] + x[i]#w[i];
    swrep[jj]=swrep[jj] + w[i];
 end;
 xmjack=j(nruanz,1,0);
 pse=j(nruanz,1,0);
   do re=1 to  nruanz;
    sumwr=sumw;
    sumwxr=sumwx;
        sumwr=sumwr-swrep[re];
        sumwxr=sumwxr-swxrep[re];
        xmjack[re]=sumwxr/sumwr;
        pse[re]=nruanz#xmit-(nruanz-1)#xmjack[re];
   end;
  xmtm= sum(pse)/nruanz;
  xme = sqrt(ssq(pse-xmit)/(nruanz#(nruanz-1)));
    zt=probit(1-&alpha/2);
    lcl=xmit-zt#xme;
    ucl=xmit+zt#xme;
    varcoef=(xme/xmit)#100;
    %if &jd1 ne  %then %do;
      %let bw1=%scan(&&wert&ks,1);
      bl1={&bw1};
      bd1=bd1//bl1;
    %end;
    %if &jd2 ne  %then %do;
      %let bw2=%scan(&&wert&ks,2);
      bl2={&bw2};
      bd2=bd2//bl2;
    %end;
    %if &jd3 ne  %then %do;
      %let bw3=%scan(&&wert&ks,3);
      bl3={&bw3};
      bd3=bd3//bl3;
    %end;
    zn=n||xmit||xme||lcl||ucl||varcoef;
    jd=jd//zn;
%end;
    %if &jd1 ne  %then %do;
      create  &jd1 var{&cl1} ;
      append from  bd1;
    %end;
    %if &jd2 ne  %then %do;
      create  &jd2 var{&cl2} ;
      append from  bd2;
    %end;
    %if &jd3 ne  %then %do;
      create  &jd3 var{&cl3} ;
      append from  bd3;
    %end;

create  jasta var{n mittel stderr ungr obgr varcoe} ;
```

```
append from  jd;
options missing=' ';
%if &jd1 ne  %then %do;
 data jadat;
   merge  &jd1 &jd2 &jd3  jasta;
   label  n="Anzahl"  mittel="M_&var"
        stderr="Std_err"  ungr="U_Konf_grz"
        obgr="O_Konf_grz"  varcoe="Var_koeff";
   format  n 6.  mittel 9.2  stderr 8.3  ungr 9.3 obgr 9.3 varcoe 8.2;
 run;
%end;
%if &jd1 eq  %then %do;
 data jadat;
   set jasta;
   label  n="Anzahl"  mittel="M_&var"
        stderr="Std_err"  ungr="U_Konf_grz"
        obgr="O_Konf_grz"  varcoe="Var_koeff";
   format  n 6.  mittel 9.2  stderr 8.3  ungr 9.3 obgr 9.3 varcoe 8.2;
 run;
%end;
proc print data=jadat  label  uniform  noobs;  run;
quit;
%mend  jackmit;
```

Literatur

Statistische Methoden der Schweizerischen Gesundheitsbefragung 1992/93, BFS Bern

Rudi Peters et Beat Hulliger, La technique de pondération des données: application à l'enquête suisse sur la santé 1994, BFS Bern

K.M. Wolter, Introduction to Variance Estimation, Springer Verlag, New York 1985

W.G Cochran , Sampling techniques, John Wiley 1977

L. Kish, Survey Sampling, John Wiley 1965

J-C Deville, C-E Särndal, O Sautory, Generalized raking procedures in survey sampling, Journal of the American Statistical Association, vol 88, n°423 pp 1013-1020, 1993.

O Sautory, Redressements d'échantillons d'enquête aurpès des ménages par calge sur marges, Actes des journées de méthodologie statistique, INSEE-Méthodes n°29-30-31, 13 et 14 mars 1992