Bootstrap mit SAS

Paul Eberhard Rudolph
Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere
FB Genetik und Biometrie
Wilhelm-Stahl-Allee 2
18198 Dummerstorf
rudolph@fbn-dummerstorf.de

Armin Tuchscherer
Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere
FB Genetik und Biometrie
Wilhelm-Stahl-Allee 2
18198 Dummerstorf
atuchsch@fbn-dummerstorf.de

Bernd Jäger
Ernst-Moritz-Arndt-Universität Greifswald
Institut für Biometrie und Medizinische Informatik
Wilhelm-Stahl-Allee 2
17487 Greifswald
bjaeger@biometrie.uni-greifswald.de

Margret Tuchscherer
Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere
FB Verhaltensphysiologie
Wilhelm-Stahl-Allee 2
18198 Dummerstorf
mtuchsch@fbn-dummerstorf.de

Zusammenfassung

Die Verfügbarkeit leistungsfähiger Rechentechnik hat auch der Anwendung von Bootstrap-Methoden z. B. bei der Bestimmung des Standardfehlers eines Schätzers oder der Prüfung einer Statistik auf Abweichung von der Erwartung unter einer gewissen Hypothese erheblichen Auftrieb gegeben. Voraussetzung für die Anwendung von Bootstrap-Methoden unter Nutzung der in SAS verfügbaren Prozeduren ist ein Werkzeug zur Erzeugung von Stichproben aus einer gegebenen Stichprobe (Resampling, Resimulation).

Zwei derartige Werkzeuge werden in diesem Beitrag vorgestellt. Beim ersten handelt es sich um ein von SAS bereitgestelltes SAS-Programm, das Macros zur Durchführung von Bootstrap-Analysen enthält. Das zweite Werkzeug ist ein eigenes SAS-Macro unter Verwendung von SAS IML. Die Macros werden kurz beschrieben und ihre Verwendbarkeit am Beispiel der Bootstrap-Schätzung für den Standardfehler des Mittelwertes als Schätzer für den Erwartungswert einer Verteilung demonstriert.

Keywords: Bootstrap, SAS-Macro, Schätzer, Mittelwert, Standardfehler

1 Einleitung

Unter Bootstrap versteht man eine Methode zur Schätzung der Verteilung einer Statistik (z. B. eines Schätzers oder einer Teststatistik) einer Stichprobe, indem

aus der ursprünglichen Stichprobe neue Stichproben (Bootstrap-Stichproben) gezogen werden (resampling).
Es seien $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ eine (realisierte) Stichprobe vom Umfang n einer unbekannten Verteilung F und $s(\mathbf{x})$ eine Funktion (Statistik) dieser Stichprobe, z. B. ein Schätzer $\hat{\vartheta} = s(\mathbf{x})$ für einen unbekannten Parameter ϑ der Verteilung F. Ist die Genauigkeit dieses Schätzers auf analytischem Wege nicht bestimmbar (z. B. bei unbekannter Verteilung F), dann kann mit Hilfe der hierzu von Efron(1979) eingeführten Bootstrap-Methode diese Genauigkeit rechner-experimentell bestimmt werden.
Die Bootstrap-Methode z. B. zur numerischen Bestimmung des Standardfehlers für den Schätzer $\hat{\vartheta} = s(\mathbf{x})$ lässt sich in die folgenden Schritte untergliedern:

1. Aus der Ausgangsstichprobe $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ werden unabhängig voneinander m Bootstrap-Stichproben

$$\mathbf{x}_j^* = \{x_{j1}^*, x_{j2}^*, ..., x_{jn}^*\}, \; j = 1, ..., m \tag{1.1}$$

gezogen, wobei die Elemente x_{ji}^*, $i = 1, ..., n$ der Bootstrap-Stichproben durch zufällige Auswahl von Elementen der Ausgangsstichprobe mit der Wahrscheinlichkeit 1/n mit Zurücklegen gebildet werden.

2. Für die m Bootstrap-Stichproben werden die entsprechenden Stichprobenfunktionen

$$\hat{\vartheta}_j^* = s(\mathbf{x}_j^*), \quad j = 1, ..., m \tag{1.2}$$

und damit eine empirische Verteilung der Stichprobenfunktion s berechnet.

3. Aus der empirischen Verteilung der Stichprobenfunktion können die interessierenden Statistiken berechnet werden. Eine Bootstrap-Schätzung für den Standardfehler $se_F(\hat{\vartheta})$ des Schätzers $\hat{\vartheta} = s(\mathbf{x})$ ergibt sich damit aus der Standardabweichung von

$$\hat{\vartheta}^* = \left\{\hat{\vartheta}_1^* = s(\mathbf{x}_1^*), \hat{\vartheta}_2^* = s(\mathbf{x}_2^*), ..., \hat{\vartheta}_m^* = s(\mathbf{x}_m^*)\right\} \tag{1.3}$$

als

$$\hat{se}_{boot} = \left[\frac{1}{m-1}\sum_{j=1}^{m}(\hat{\vartheta}_j^* - \bar{\hat{\vartheta}}^*)^2\right]^{\frac{1}{2}} \quad \text{mit} \quad \bar{\hat{\vartheta}}^* = \frac{1}{m}\sum_{j=1}^{m}\hat{\vartheta}_j^* \, . \tag{1.4}$$

Die rechentechnische Durchführung der oben beschriebenen Bootstrap-Methode ist in SAS einfach realisierbar. Dazu kann einmal auf das auf einer Internetseite von SAS Institute bereitgestellte SAS-Programm `jackboot.sas` zurückgegriffen werden, das eine Reihe von SAS-Macros für Jackknife- und Bootstrap-Analysen enthält. Die Bootstrap-Macros dieses Programms werden in Kapitel 2 kurz erläutert. Das Standardvorgehen zur Beschreibung der Genauigkeit eines Schätzers wird am Beispiel des Mittelwertes als Schätzer für den Erwartungswert einer Verteilung im Abschnitt 2.1 vorgestellt. Die Bootstrap-Schätzung für die Standardabweichung des Mittelwertes wird im Abschnitt 2.2 in einem Beispiel mit den Macros des SAS-Pogramms `jackboot.sas` demonstriert. Im Kapitel 3 wird ein eigenes Bootstrap-Macro unter Verwendung von SAS-IML vorgestellt sowie dessen Handhabung am gleichen Beispiel wie im Abschnitt 2.2 beschrieben.

2 Bootstrap-Macros im SAS-Programm JACKBOOT.SAS

Auf der SAS-Internetseite http://ftp.sas.com/techsup/download/stat/ werden Beispielprogramme zur Datenanalyse zur Verfügung gestellt. Darunter ist auch das SAS-Programm jackboot.sas, das eine Reihe von SAS-Macros für Jackknife- und Bootstrap-Analysen enthält. Umfangreiche Kommentare erläutern die Handhabung der einzelnen Macros. Der Anwender sollte jedoch über ausreichende Kenntnisse der SAS-Macrosprache verfügen. Auf den mathematischen Hintergrund wird in den Kommentaren nicht eingegangen. Hinweise auf die den Verfahren zugrunde liegende Literatur sind jedoch vorhanden.

Zur Durchführung von Bootstrap-Analysen sind im SAS-Programm `jackboot.sas` die beiden Macros `%BOOT` und `%BOOTCI` verfügbar.
Zunächst ist allerdings ein Macro `%ANALYZE` zu schreiben, das die gewünschten Statistiken berechnet. Dies geschieht in der Regel unter Verwendung geeigneter SAS-Prozeduren.
Mit dem Macro `%BOOT` können elementare nichtparametrische Bootstrap-Analysen für einfache Zufallsstichproben realisiert werden. Dazu gehören die Berechnungen approximativer Standardfehler, biaskorrigierter Schätzwerte und Konfidenzintervalle unter der Voraussetzung, dass die entsprechenden Bootstrap-Stichprobenfunktionen näherungsweise normalverteilt sind. Dieses Macro kann außerdem bei Regressionsmodellen zur Erzeugung von Bootstrap-Beobachtungen bzw. –residuen verwendet werden.
Das Macro `%BOOT` verwendet ein Macro `%ANALYZE` zur Berechnung der entsprechenden Statistiken sowohl für die Ausgangsstichprobe als auch für die Bootstrap-Stichproben. Dieses Macro ist vom Anwender problemadäquat zu erstellen.
Das Macro %BOOTCI wird zur Bestimmung approximativer Konfidenzintervalle nach unterschiedlichen Methoden in den Fällen verwendet, wenn die

Bootstrap-Stichprobenfunktionen nicht näherungsweise normalverteilt sind. Dabei geht einem Aufruf des %BOOTCI-Macros stets ein Aufruf des %BOOT-Macros voraus, in dem die Bootstrap-Stichproben erzeugt werden.

2.1 Bootstrap-Schätzung für den Standardfehler des Mittelwertes

Es seien $\mathbf{x}$ eine reelle Zufallsgröße mit der Verteilungsfunktion F und $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ eine mathematische Stichprobe von $\mathbf{x}$. Werden Erwartungswert bzw. Varianz von $\mathbf{x}$ mit $E_F(\mathbf{x}) = \mu_F$ bzw. $V_F(\mathbf{x}) = \sigma_F^2$ bezeichnet, schreibt man häufig auch

$$\mathbf{x} \sim (\mu_F, \sigma_F^2) \ . \qquad (2.1)$$

Für das arithmetische Mittel $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ gilt dann $E_F(\bar{\mathbf{x}}) = \mu_F$ bzw. $V_F(\bar{\mathbf{x}}) = \frac{\sigma_F^2}{n}$ und damit

$$\bar{\mathbf{x}} \sim (\mu_F, \frac{\sigma_F^2}{n}) \ . \qquad (2.2)$$

Der Standardfehler $se_F(\bar{\mathbf{x}})$ des Mittelwertes als Quadratwurzel aus der Varianz, d. h.,

$$se_F(\bar{\mathbf{x}}) = \left[V_F(\bar{\mathbf{x}})\right]^{\frac{1}{2}} = \left[\frac{\sigma_F^2}{n}\right]^{\frac{1}{2}} = \frac{\sigma}{\sqrt{n}} \qquad (2.3)$$

ist dann ein Maß für die Genauigkeit des Mittelwertes als Schätzer für den Erwartungswert $E_F(\mathbf{x}) = \mu_F$. Theoretischer Hintergrund ist die Tatsache, dass der Mittelwert $\bar{\mathbf{x}}$ nach dem zentralen Grenzwertsatz approximativ normalverteilt ist, d. h., es gilt:

$$\bar{\mathbf{x}} \sim N\left(\mu_F, \frac{\sigma_F^2}{n}\right) . \qquad (2.4)$$

Das bedeutet insbesondere, dass für große n annähernd die folgende Gleichung gilt:

$$P\left\{\left|\bar{\mathbf{x}} - \mu_F\right| < \frac{2\sigma_F}{\sqrt{n}}\right\} \approx 0{,}954\,. \tag{2.5}$$

Um eine Bootstrap-Schätzung für den Standardfehler des Mittelwertes $\bar{\mathbf{x}}$ zu erhalten, geht man wie bereits in Kapitel 1 beschrieben vor mit $\hat{\vartheta} = s(\mathbf{x}) = \bar{\mathbf{x}}$.

1. Aus einer vorliegenden konkreten Stichprobe $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ werden zunächst m Bootstrap-Stichproben

$$\mathbf{x}_j^* = \left\{x_{j1}^*, x_{j2}^*, ..., x_{jn}^*\right\}, \;\; j = 1, ..., m\,, \tag{2.6}$$

gezogen.

2. Aus den Bootstrap-Stichproben (2.6) werden die Mittelwerte

$$\bar{\mathbf{x}}_j^* = \frac{1}{n}\sum_{i=1}^{n} x_{ji}^*\,, \;\; j = 1, ..., m \tag{2.7}$$

berechnet.

3. Eine Bootstrap-Schätzung $\widehat{se}_{boot}(\bar{\mathbf{x}})$ für den Standardfehler $se_F(\bar{\mathbf{x}})$ des Schätzers $\bar{\mathbf{x}}$ ergibt sich damit aus der Standardabweichung von

$$\bar{\mathbf{x}}^* = \left\{\bar{\mathbf{x}}_1^*, \bar{\mathbf{x}}_2^*, ..., \bar{\mathbf{x}}_m^*\right\} \tag{2.8}$$

als

$$\widehat{se}_{boot}(\bar{\mathbf{x}}) = \left[\frac{1}{m-1}\sum_{j=1}^{m}(\bar{\mathbf{x}}_j^* - \bar{\bar{\mathbf{x}}}^*)^2\right]^{\frac{1}{2}} \tag{2.9}$$

mit

$$\bar{\bar{\mathbf{x}}}^* = \frac{1}{m}\sum_{j=1}^{m}\bar{\mathbf{x}}_j^*\,. \tag{2.10}$$

2.2 Beispiel

Es liege die Stichprobe $\mathbf{x} = \{x_1, x_2, ..., x_{10}\}$ aus einer Normalverteilung mit Erwartungswert $\mu = 0$ und Varinz $\sigma^2 = 1$ in der temporären SAS-Datei `stich1` vor. Zur Erzeugung einer derartigen Stichprobe kann z. B. das Macro PNORMAL (s. Tuchscherer u.a. 1999) verwendet werden.

Tabelle 1: Stichprobenelemente

x
-0.32659
1.54244
0.25903
-0.55583
0.77872
-0.65520
-0.02210
-0.77265
0.72423
1.26331

Ein erforderliches Macro `%ANALYZE` zur Berechnung des arithmetischen Mittels der Ausgangs- und der Bootstrap-Stichproben hat bei Verwendung der SAS-Prozedur `MEANS` z. B. die folgende Gestalt:

```
%macro analyze(data=,out=);
   proc means noprint data=&data  vardef=DF ;
     output out=&out (drop=_type_ _freq_
                      rename=(_STAT_=STAT));
     var X ;
     %bystmt;
   run;
   %if &syserr=0 %then %do;
      data &out;
         set &out;
         where STAT='MEAN' ;
      run;
   %end;
%mend;
```

Mit dem Aufruf des Macros `%BOOT`

```
title1 'Beispiel mit JACKBOOT';
title2 'Bootstrap Analysis uniform resampling';

%boot(data= STICH1, samples=200, balanced=0,
      id=stat, random=123);
```

werden 200 Bootstrap-Stichproben vom Umfang 10 der Ausgangsstichprobe gleichverteilt aus der Ausgangsstichprobe gezogen und eine Bootstrap-Analyse der Statistik `'MEAN'` aus `PROC MEANS` durchgeführt. Mit `random=123` wird zur Erzeugung gleichverteilter Zufallszahlen der feste Startwert 123 für den Parameter `seed` in der SAS-Funktion `UNIFORM` gesetzt, um die Reproduzierbarkeit der Ergebnisse zu garantieren. Für praktische Anwendungen sollte mit `random=0` die Systemzeit als zufälliger Startwert gewählt werden.
Durch balanced = 0 wird jedes Element der Ausgangsstichprobe mit der gleichen Wahrscheinlichkeit (1/n) als Element einer Bootstrap-Stichprobe ausgewählt (uniform resampling). Das bedeutet jedoch nicht, dass in m Bootstrapstichproben jedes Element der Ausgangsstichprobe genau m-mal vorkommt. Um eine derartige Balanciertheit zu erreichen (balanced resampling), ist balanced = 1 zu setzen (Gleason, 1988).

Als Ergebnisse der Bootstrap-Analyse erhält man das Histogramm für das arithmetische Mittel (s. Abb. 1) und statistische Maßzahlen für den Schätzer `'MEAN'`(s. Abb.2).

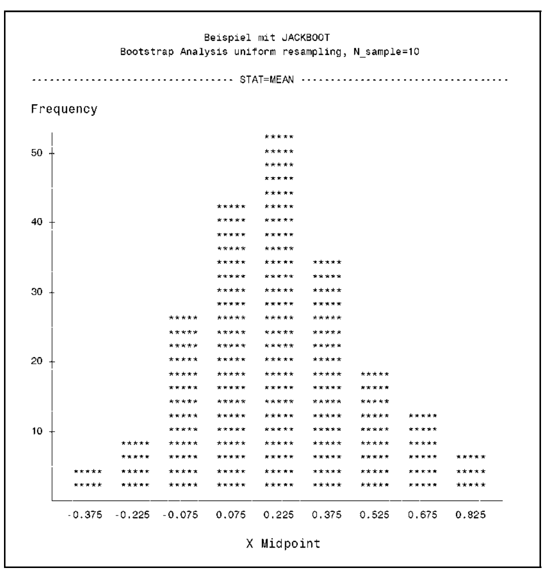


Abbildung 1: Histogramm für das arithmetische Mittel

Beispiel mit JACKBOOT
Bootstrap Analysis uniform resampling, N_sample=10

stat	Name	Observed Statistic	Bootstrap Mean	Approximate Bias	Approximate Standard Error	Approximate Lower Confidence Limit	Bias-Corrected Statistic
MEAN	X	0.22354	0.22722	.003688310	0.25813	-0.28608	0.21985

stat	Name	Approximate Upper Confidence Limit	Confidence Level (%)	Method for Confidence Interval	Minimum Resampled Estimate	Maximum Resampled Estimate	Number of Resamples
MEAN	X	0.72578	95	Bootstrap Normal	-0.42527	0.88835	200

Abbildung 2: Statistische Maßzahlen

In der Abb. 2 ist unter der Bezeichnung `Observed Statistic` das arithmetische Mittel der Ausgangsstichprobe und unter `Approximate Standard Error` die Bootstrap-Schätzung nach Formel (2.9) zu finden.

Neben der hier dargestellten Anzahl von 200 Bootstrap-Stichproben wurden analoge Analysen auch für jeweils 100, 200, 500 und 1000 Bootstrap-Stichproben durchgeführt für Ausgangsstichproben vom Umfang $N \in \{10, 50, 100\}$. Die Ergebnisse sind in den Tabellen 2 und 3 zusammengefasst.

Tabelle 2: Bootstrap_Analysen mit N=10, 50, 100 (Uniform Resampling)

N	N_BOOT	MEAN	STDERR	BOOT_MEAN
10	100	0.21666	0.22982	0.26610
10	200	0.21666	0.23864	0.23350
10	500	0.21666	0.24380	0.20419
10	1000	0.21666	0.24838	0.21787
50	100	0.05463	0.15435	0.04996
50	200	0.05463	0.13692	0.03975
50	500	0.05463	0.13467	0.05147
50	1000	0.05463	0.14116	0.04939
100	100	0.05109	0.10179	0.04106
100	200	0.05109	0.10705	0.04319
100	500	0.05109	0.10274	0.05340
100	1000	0.05109	0.10446	0.05115

Tabelle 3: Bootstrap_Analysen mit N=10, 50, 100 (Balanced Resampling)

N	N_BOOT	MEAN	STDERR	BOOT_MEAN
10	100	0.21666	0.25810	0.21666
10	200	0.21666	0.23276	0.21666
10	500	0.21666	0.24302	0.21666
10	1000	0.21666	0.25305	0.21666
50	100	0.05463	0.14696	0.05463
50	200	0.05463	0.14069	0.05463
50	500	0.05463	0.14488	0.05463
50	1000	0.05463	0.13730	0.05463
100	100	0.05109	0.10475	0.05109
100	200	0.05109	0.09226	0.05109
100	500	0.05109	0.10048	0.05109
100	1000	0.05109	0.09937	0.05109

Der in Abb. 2 angegebene Bias (`Approximate Bias`) ist die Differenz zwischen dem arithmetischem Mittel der Ausgangsstichprobe und dem arithmetischen Mittel der arithmetischen Mittel der Bootstrap-Stichproben. Erfolgt das Ziehen der Bootstrap-Stichproben balanciert, ist dieser Bias erwartungsgemäß gleich Null, wie man auch in Tab. 3 sieht.
Die Bootstrapschätzung für den Standardfehler des Mittelwertes ändert sich kaum mit wachsender Zahl der Bootstrap-Stichproben. Dies deckt sich mit Empfehlungen zur Wahl der Anzahl der Bootstrapstichproben von ca.200 bei Verteilungen, die näherungsweise normal sind, und von ca. 1000 bei Verteilungen, die stärker von der Normalverteilung abweichen.

Tabelle 4: Mittelwerte von 1000 Wiederholungen der Bootstrap-Analysen von Tabelle2

N	N_BOOT	MEAN	STDERR	BOOT_MEAN
10	100	-.001141790	0.28899	0.000727154
10	200	-.001141790	0.29101	-.002384338
10	500	-.001141790	0.29034	-.001536216
10	1000	-.001141790	0.29043	-.001186102
50	100	0.001287094	0.13930	0.001021826
50	200	0.001287094	0.13929	0.000942768
50	500	0.001287094	0.13919	0.001626480
50	1000	0.001287094	0.13922	0.001094297
100	100	0.002188066	0.09874	0.002086736
100	200	0.002188066	0.09940	0.002257538
100	500	0.002188066	0.09946	0.002399593
100	1000	0.002188066	0.09922	0.002128158

Die Ausgangsstichproben wurden jeweils als unabhängige Stichproben einer N(0,1)-Verteilung erzeugt. Werden viele Stichproben mit gleichem Umfang erzeugt, ist im Mittel zu erwarten, dass mit wachsendem N auch die Schätzung (MEAN) für den Erwartungswert der N(0,1)-Verteilung näher an 0 liegt. In Tabelle 4 sind die Mittelwerte für 1000 Wiederholungen zusammengestellt

3 Ein Bootstrap-Macro mit SAS-IML

Die richtige Anwendung von `JACKBOOT.SAS` erfordert vom Nutzer gute Macroprogrammierkentnisse, um das Macro `%ANALYZE` zu schreiben und eine Portion Mühe zum Verständnis der in `JACKBOOT.SAS` ablaufenden Rechenprozesse.

Einfacher, leichter durchschaubar und ohne Voraussetzung von Macroprogrammierkenntnissen ist die folgende Vorgehensweise zur Realisierung von Bootstrap-Analysen:

1. Erzeugung einer vorgegebenen Anzahl von Bootstrap-Stichproben aus der Ausgangsstichprobe mit dem Macro `Bootstrap`

2. Berechnung der interessierenden Statistik für die Ausgangsstichprobe und alle Bootstrap-Stichproben unter Verwendung einer geeigneten SAS-Prozedur

3. Beschreibung der Verteilung der verwendeten Statistik in der Regel mit den SAS-Prozeduren `MEANS` bzw. `UNIVARIATE` und gegebenenfalls ergänzenden `data steps`

Im nächsten Abschnitt wird das unter Verwendung von SAS IML erstellte und im Anhang vollständig wiedergegebene Macro `%BOOTSTRAP` kurz beschrieben.

3.1 Das SAS-Macro %BOOTSTRAP

Analog zu den Möglichkeiten der Macros in JACKBOOT.SAS lassen sich mit dem Macro %BOOTSTRAP gleichverteilte oder balancierte Bootstrap-Stichproben aus einer uni- oder multivariaten Ausgangsstichprobe erzeugen.

In

```
%macro BOOTSTRAP(M, N_RESAMPLE, SAMPLE, START,
                 BALANCE, BOOT, NAMES)
```

haben die Macroparameter folgende Bedeutung:

`M`	Anzahl der Bootstrap-Stichproben
`N_RESAMPLE`	Umfang der Bootstrap-Stichproben Kann nur bei `BALANCE=0` verschieden vom Umfang der Ausgangsstichprobe sein!
`START`	Startwert für den Zufallszahlengenerator bei der Resampling-Prozedur: `START=0`: zufälliger Startwert durch Systemzeit `START=G`: fester Startwert mit der positiven ganzen Zahl G
`SAMPLE`	SAS-Datei der Ausgangsstichprobe mit der/den Variablen der Ausgangsstichprobe (z.B.: `X`, `Y`)
`BALANCE`	Parameter zum Einstellen der Resampling-Prozedur: `BALANCE=1`: balanciertes Resampling `BALANCE=0`: gleichverteiltes Resampling
`BOOT`	SAS-Datei der M Bootstrap-Stichproben vom Umfang `N_RESAMPLE`: erste Spalte: Nummer der Bootstrap-Stichprobe (`N_BOOT`) zweite Spalte: Nummer des Stichprobenelements der Ausgangstichprobe `SAMPLE` (`NR_Obs`) restliche Spalten: Variable/n der Ausgangsstichprobe (z.B.: `X`, `Y`)
`NAMES`	Variablenbezeichnung in BOOT z.B.: `{'N_BOOT' 'NR_Obs' 'X' 'Y'}`

Vorausetzung für die Anwendung des Macros ist das Vorliegen der SAS-Datei `SAMPLE`, die die Ausgangsstichprobe enthält.
Die Anwendung dieses Macros erfordert keine Kenntnisse der Macro-Programmierung. Man muß nur die Parameter beim Aufruf des Macros setzen.

Mit dem Aufruf

```
%Bootstrap(200,10,123,STICH1,0,BOOTSTICH,{'N_BOOT'
           'NR_Obs' 'X'});
```

werden z.B. `200` Bootstrap-Stichproben vom Umfang `10` aus der Ausgangsstichprobe mit der Variablen `X` gezogen, die in der temporären SAS-Datei `STICH1` vorliegt. Wegen `BALANCE=0` erfolgt gleichverteiltes Resampling. In diesem Fall könnte der Umfang der Ausgangsstichprobe verschieden von `10` sein. Der feste Startwert `START=123` liefert reproduzierbare Ergebnisse, die in der temporären SAS-Datei `BOOTSTICH` mit den 3 Spalten `'N_BOOT'`, `'NR_Obs'` und `'X'` abgelegt werden.

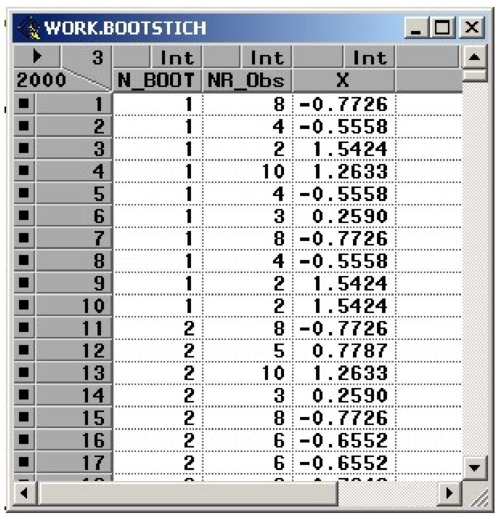
3.2 Beispiel

Es werden die gleichen Beispieldaten wie im Abschnitt 2.2 (SAS-Datei `STICH1`) verwendet. Zunächst werden 200 Bootstrap-Stichproben mit dem Aufruf des Macros `%BOOTSTRAP` erzeugt

```
title1 'Beispiel mit Macro %BOOTSTRAP';
title2 'Bootstrap Analysis uniform resampling,
N_sample=10';

%Bootstrap(200,10,123,STICH1,0,BOOTSTICH,{'N_BOOT'
'NR_Obs' 'X'});
```

und in der SAS-Datei `BOOTSTICH` (Abb. 3) abgelegt:

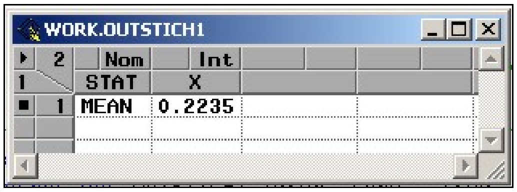

WORK.BOOTSTICH

3 / 2000	Int N_BOOT	Int NR_Obs	Int X
1	1	8	-0.7726
2	1	4	-0.5558
3	1	2	1.5424
4	1	10	1.2633
5	1	4	-0.5558
6	1	3	0.2590
7	1	8	-0.7726
8	1	4	-0.5558
9	1	2	1.5424
10	1	2	1.5424
11	2	8	-0.7726
12	2	5	0.7787
13	2	10	1.2633
14	2	3	0.2590
15	2	8	-0.7726
16	2	6	-0.6552
17	2	6	-0.6552

Abbildung 3: Erzeugte Bootstrap-Stichproben

Dann erfolgt die Berechnung des arithmetischen Mittels (Abb. 4) für die Ausgangsstichprobe mit dem folgenden SAS-Programm:

```
proc means noprint data=STICH1  vardef=DF ;
     output out=OUTSTICH1 (drop=_type_ _freq_
                           rename=(_STAT_=STAT));
     var X ;
run;
data OUTSTICH1;
    set OUTSTICH1;
    where STAT='MEAN' ;
run;
```



WORK.OUTSTICH1

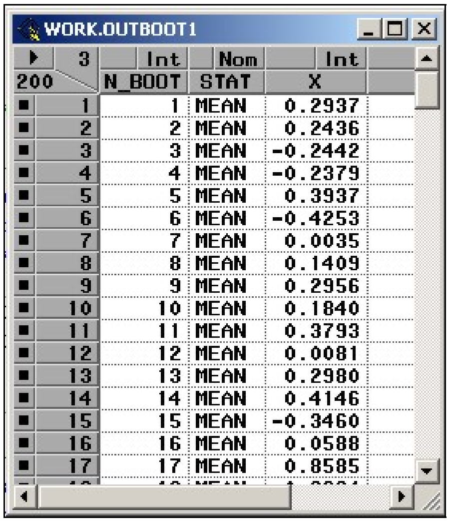
	Nom	Int
	STAT	X
1	MEAN	0.2235

Abbildung 4: Mittelwert der Ausgangsstichprobe

Die Berechnung des arithmetischen Mittels für die Bootstrap-Stichproben realisiert das folgende SAS-Programm:

```
proc means noprint data=BOOTSTICH  vardef=DF ;
     output out=OUTBOOT1 (drop=_type_ _freq_
                          rename=(_STAT_=STAT));
     var X ;
     by N_BOOT;
run;
data OUTBOOT1;
    set OUTBOOT1;
    where STAT='MEAN' ;
run;
```

Einen Ausschnitt der in der Datei dann enthaltenen Daten zeigt Abb. 5.



WORK.OUTBOOT1

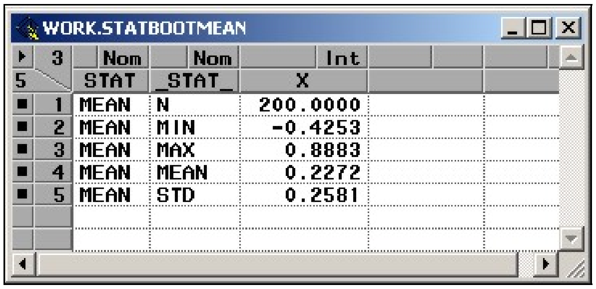
	Int	Nom	Int
	N_BOOT	STAT	X
1	1	MEAN	0.2937
2	2	MEAN	0.2436
3	3	MEAN	-0.2442
4	4	MEAN	-0.2379
5	5	MEAN	0.3937
6	6	MEAN	-0.4253
7	7	MEAN	0.0035
8	8	MEAN	0.1409
9	9	MEAN	0.2956
10	10	MEAN	0.1840
11	11	MEAN	0.3793
12	12	MEAN	0.0081
13	13	MEAN	0.2980
14	14	MEAN	0.4146
15	15	MEAN	-0.3460
16	16	MEAN	0.0588
17	17	MEAN	0.8585

Abbildung 5: 200 Mittelwerte der 200 Bootstrap-Stichproben

Anschließend erfolgt die Berechnung statistischer Masszahlen (`N MIN MAX MEAN STD`) für die Statistik `STAT=MEAN` aus `WORK.OUTBOOT1` mit der Prozedur `MEANS`.

```
proc sort data=OUTBOOT1; by STAT; run;
proc means data=WORK.OUTBOOT1  noprint   vardef=DF ;
  var X ;
  id STAT ;
  output out=WORK.STATBOOTMEAN (drop=_type_ _freq_);
  by STAT ;
run;
```

Die Ergebnisse dieses SAS-Programms werden in der Datei `statbootmean` abgelegt (s. Abb. 6).



WORK.STATBOOTMEAN

3	Nom STAT	Nom _STAT_	Int X
1	MEAN	N	200.0000
2	MEAN	MIN	-0.4253
3	MEAN	MAX	0.8883
4	MEAN	MEAN	0.2272
5	MEAN	STD	0.2581

Abbildung 6: Statistische Maßzahlen der 200 Mittelwerte der Bootstrap-Stichproben

Nach Zusammenführen der Ergebnisse in die Datei `BOOTMASSZAHL` erfolgt die Berechnung des approximativen Bias, des Bias-korrigierten Mittelwerts und approximativer Konfidenzgrenzen unter Voraussetzung von Normalverteilung mittels folgender `data steps` und anschließendem Druck der Ergebnisse, um die gleichen Ausgabedaten wie bei der Verwendung des Macros `%BOOT` aus dem Programm `JACKBOOT` zu erhalten (vgl. Abschnitt 2.2).

```
proc transpose data=STATBOOTMEAN out=STATBOOTMEANT
                                  prefix=BOOT_;
   id _STAT_;
   by STAT;
run;
proc sort data=OUTSTICH1; by STAT ; run;
```

```
proc transpose data=OUTSTICH1 out=OUTSTICH1T
                              prefix=WERT_;
   by STAT ;
run;

Data BOOTMASSZAHL (rename=(_name_=NAME
                           WERT_1=WERT));
  merge OUTSTICH1T STATBOOTMEANT;
  by STAT;
  BIAS=BOOT_MEAN-WERT_1;
  MEAN_CORR=WERT_1-BIAS;
  ALPHA=0.05;
  APP_NORMAL_CI_LOW=MEAN_CORR-probit(1-ALPHA/2)
                    *BOOT_STD;
  APP_NORMAL_CI_UPP=MEAN_CORR+probit(1-ALPHA/2)
                    *BOOT_STD;
  label STAT = 'Statistic'
   _NAME_  ='Name Variable'
   WERT_1 ='Observed Statistic'
   BOOT_MEAN='Bootstrap Mean'
   BIAS  ='Approximate Bias'
   MEAN_CORR='Bias-Corrected Statistic'
   BOOT_STD='Approximate Standard Error'
   APP_NORMAL_CI_LOW  ='Approximate Lower Confidence
                        Limit'
   APP_NORMAL_CI_UPP  ='Approximate Upper Confidence
                        Limit'
   ALPHA='ALPHA'
   BOOT_MIN ='Minimum Resampled Estimate'
   BOOT_MAX ='Maximum Resampled Estimate'
   BOOT_N ='Number of Resamples';
run;
proc print data=BOOTMASSZAHL noobs label;
  id STAT NAME;
run;
```

Die Ergebnisse des obigen SAS-Programms sind in der Abb. 7 zusammengefasst. Ein Vergleich der Abb. 2 und 7 zeigt die Übereinstimmung der Resultate der Bootstrap-Analysen.

Beispiel mit %BOOTSTRAP
Bootstrap Analysis uniform resampling, N_sample=10

Statistic	Name Variable	Observed Statistic	Number of Resamples	Minimum Resampled Estimate	Maximum Resampled Estimate	Bootstrap Mean	Approximate Standard Error
MEAN	X	0.22354	200	-0.42527	0.88835	0.22722	0.25813

Statistic	Name Variable	Approximate Bias	Bias-Corrected Statistic	ALPHA	Approximate Lower Confidence Limit	Approximate Upper Confidence Limit
MEAN	X	.003688310	0.21985	0.05	-0.28608	0.72578

Abbildung 7: Ergebnisausdruck der Bootstrap-Analyse

Mit `PROC UNIVARIATE` kann die 'Mittelwert-Verteilung' ebenso wie mit `PROC MEANS` durch statistische Maßzahlen beschrieben werden. Bei Verwendung des `HISTOGRAM Statements` in `PROC UNIVARIATE` kann zusätzlich ein Histogramm (s. Abb. 10) für das arithmetische Mittel erstellt werden.

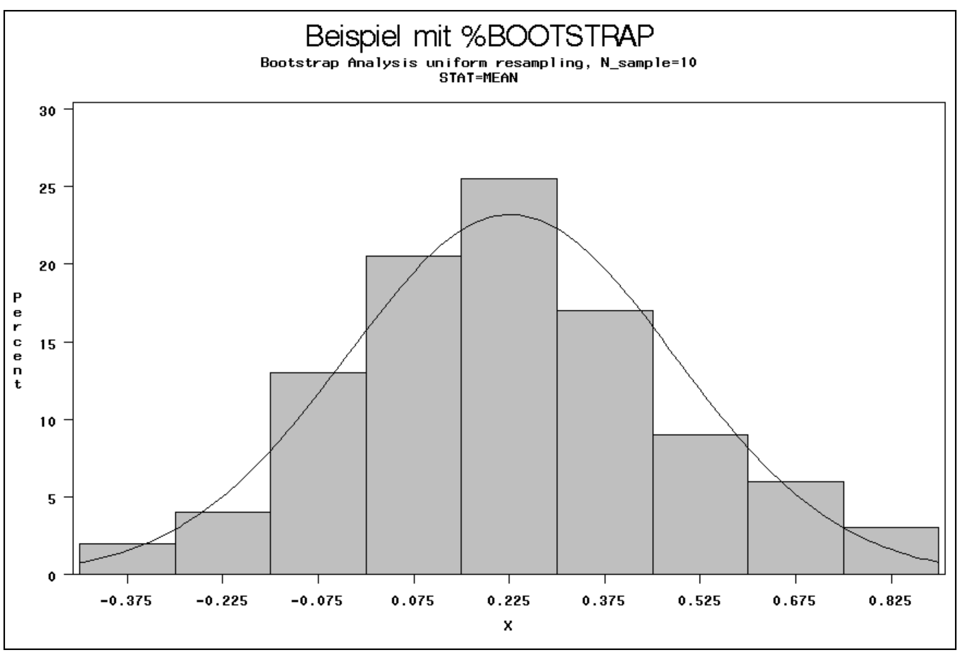


Abbildung 8: Histogramm für das arithmetische Mittel

Literatur

1. Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. **7**, 1-26.
2. Efron, B. & Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability. New York: Chapman & Hall.
3. Gleason, J. R. (1988). Algorithms for Balanced Bootstrap Simulations. American Statistician, 42, 263-266.
4. Johnson, N. L.; Kotz, S. (1972). Distributions in Statistics: Continuous Multivariate Distributions. J. Wiley, New York.
5. Rey, W. J. J. (1983). Introduction to Robust and Quasi-Robust Statistical Methods. Springer-Verlag Berlin, Heidelberg, New York, Tokyo.
6. SAS Institute Inc. (1999). SAS Macro Language: Reference, Version 8, Cary, NC: SAS Institute Inc.
7. SAS Institute Inc. (1999). SAS/IML User's Guide, Version 8, Cary, NC: SAS Institute Inc.
8. SAS Institute Inc. (1999). SAS/STAT® User's Guide, Version 8, Cary, NC: SAS Institute Inc.
9. Tuchscherer, A.; Rudolph, P. E.; Jäger, B.; Tuchscherer, M. (1999). Ein SAS-Makro zur Erzeugung multivariat normalverteilter Zufallsgrößen. In: Proceedings der 3. Konferenz der SAS-Anwender in Forschung und Entwicklung, Ed. Ortseifen, C., Heidelberg, 293-306.
10. Tuchscherer, A.; Rudolph, P. E.; Jäger, B.; Tuchscherer, M. (2000). Erzeugung nichtnormaler multivariater Zufallsgrößen mit SAS. In: Proceedings der 4. Konferenz der SAS-Anwender in Forschung und Entwicklung, Eds. Bödecker, R.-H.; Hollenhorst, M. S., Gießen, 235-265.

Anhang

Macro zum Erzeugen von Bootstrap-Stichproben

```
/*********************************************************
 * Macro zum Erzeugen von M Bootstrap-Stichproben jeweils *
 * vom Umfang N_RESAMPLE aus der Ausgangsstichprobe       *
 * SAMPLE                                                 *
 *********************************************************/

%macro Bootstrap  (
                  M,          /* Anzahl der Bootstrap-
                                 Stichproben */
                  N_RESAMPLE, /* Umfang der Bootstrap-
                                 Stichproben; Kann nur bei
                                 BALANCE=0 verschieden vom
                                 Umfang der Ausgangsstich-
                                 probe sein! */
                  START,      /* Startwert für den Zufalls-
                                 zahlengenerator */
                  SAMPLE,     /* SAS-Datei der Ausgangs-
                                 stichprobe */
                  BALANCE,    /* BALANCE=1: balanciertes
                                 Resampling
                                 BALANCE=0: gleichverteil-
                                 tes Resampling */
                  BOOT,       /* SAS-Datei der Bootstrap-
                                 Stichproben */
                  NAMES       /* Variablenbezeichnung in
                                 BOOT z.B.:
                              {'N_BOOT' 'NR_Obs' 'X' 'Y'}*/
                  );
proc iml;
   use &SAMPLE; read all into SP;
   ns=ncol(SP);         /* Anzahl der Variablen in der
                           Ausgangsstichprobe */
   nz=nrow(SP);         /* Umfang der Ausgangsstichprobe */
   BAL=0;
   if &BALANCE then BAL=&BALANCE;
   nr=nz;
   if &N_RESAMPLE then nr=&N_RESAMPLE;
   nstart=0;
   if &START then nstart=&START;
   ny=0;

/*********************************************************/

   if BAL=0 then do;    /* Beginn uniform resampling */

   Y=j(&m*nr,ns+2,0);
```

```
   b=j(nz,1,1/nz);
   Do i=2 to nz;
   b[i]=b[i-1]+b[i];
   end;

   Do lb=1 to &M;
     Do i=1 to nr;
       ny=ny+1;
       Z=UNIFORM(nstart);
       k=0;
       Do l=1 to nz;
          k=k+1;
          if z<=b[l] then goto Zeile;
       end;
Zeile: Y[ny,1]=lb;
       Y[ny,2]=k;
       Do j=1 to ns;
          Y[ny,j+2]=SP[k,j];
       end;
     end;
   end;

   end;                 /* Ende uniform resampling */

/***********************************************************/

   else do;             /* Beginn balanced resampling */

   if nr ^= nz then do;
          print 'BALANCED RESAMPLING';
          print 'N_RESAMPLE verschieden vom Umfang der
                 Ausgangsstichprobe!';
          print 'N_RESAMPLE wurde auf den Umfang der
                 Ausgangsstichprobe korrigiert.';
   end;

   Y=j(&m*nz,ns+2,0);

   _C=j(nz,1,&M);
   _P=j(nz,1,0);
   Do i=1 to nz;
     _P[i]=i;
   end;
   _k=nz;         /* Anzahl der für Bootstrap-Stichproben
                     noch zur Verfügung stehenden Zellen */
   _jbig=_k;      /* Index der größten Zelle von _C */
   _cbig=&M;      /* _cbig >= _C[_j] */

   Do lb=1 to &M;
     Do i=1 to nz;
       ny=ny+1;
```

```
      Do until (_s<=_C[_j]);
        _j=ceil(ranuni(&START)*_k);  /* Auswahl eines
                                        Elements der
                                     Ausgangsstichprobe */
        _s=ceil(ranuni(&START)*_cbig);  /* Akzeptanz
                                        dieses Elements */
      end;
      Y[ny,1]=lb;
      Y[ny,2]=_P[_j];
      Do j=1 to ns;
         Y[ny,j+2]=SP[_P[_j],j];
      end;
      _C[_j]=_C[_j]-1;
           if _j=_jbig then do;
              _a=floor((&M-lb-_k)/_k);
              if _cbig-_c[_j]>_a then do;
                 do _ii=1 to _k;
                    if _c[_ii]>_c[_jbig] then _jbig=_ii;
                 end;
                 _cbig=_c[_jbig];
              end;
           end;
           if _c[_j]=0 then do;
              if _jbig=_k then _jbig=_j;
              _p[_j]=_p[_k];
              _c[_j]=_c[_k];
              _k=_k-1;
           end;
    end;
  end;

  end;                 /* Ende balanced resampling */

/*********************************************************/

  create &BOOT from Y [colname=&names];
  append from Y;
  quit;
%mend Bootstrap;
```

Beispiele für den Aufruf des Macros:

```
/*********************************************************
 * Erzeugen von 300 Bootstrap-Stichproben jeweils vom     *
 * Umfang der Ausgangsstichprobe                          *
 * mit zufälligem Startwert des Zufallszahlengenerators   *
 * (START=0: Systemzeit) und gleichverteiltem Resampling  *
 *********************************************************/

%Bootstrap(300,N_RESAMPLE,START,STICHPROBE,BALANCE,
           BOOTSTICH,{'N_BOOT' 'NR_Obs' 'X' 'Y'});




/*********************************************************
 * Erzeugen von 200 Bootstrap-Stichproben jeweils vom     *
 * Umfang der Ausgangsstichprobe                          *
 * mit START=123(fester Startwert: 123) des Zufallszahlen-*
 * generators und balanciertem Resampling (BALANCE=1)     *
 *                                                        *
 * Hinweis:  N_RESAMPLE=25 wird auf den Umfang der        *
 * Ausgangsstichprobe korrigiert, wenn dieser verschieden *
 * von 25 ist!                                            *
**********************************************************/


%Bootstrap(200,25,123,STICHPROBE,1,BOOTSTICH,{'N_BOOT'
           'NR_Obs' 'X' 'Y'});
```