

Bootstrap mit SAS

P. E. Rudolph*, A. Tuchscherer*, B. Jäger, M. Tuchscherer***

* Forschungsinstitut für die Biologie landwirtschaftlicher
Nutztiere Dummerstorf



** Institut für Biometrie und Medizinische Informatik
Ernst-Moritz-Arndt-Universität Greifswald



Gliederung:

1

Einleitung

2

Bootstrap-Macros im SAS-Programm JACKBOOT.SAS

3

Ein Bootstrap-Macro mit SAS-IML

4

Literatur

1

Einleitung

2

3

4

Gottfried August Bürger (1786): Wunderbare Reisen zu Wasser und zu Lande, Feldzüge und lustige Abenteuer des Freiherrn von Münchhausen

1



2

3

4

Rudolf Erich Raspe (London 1785):
 The Surprising Adventures of Baron
 Munchhausen



Bootstrap in der Statistik:

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. **7**, 1-26.

1

Rechenintensive Methoden mittels Resimulation bzw. Resampling z. B. zur

2

- Bestimmung des Standardfehlers einer Statistik (eines Schätzers)

3

- Prüfung einer Statistik auf Abweichung von der Erwartung unter einer gewissen Hypothese

4

Bootstrapschema

Ausgangs-
stichprobe

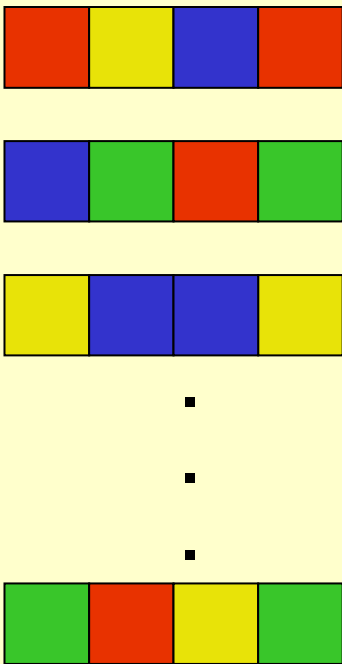
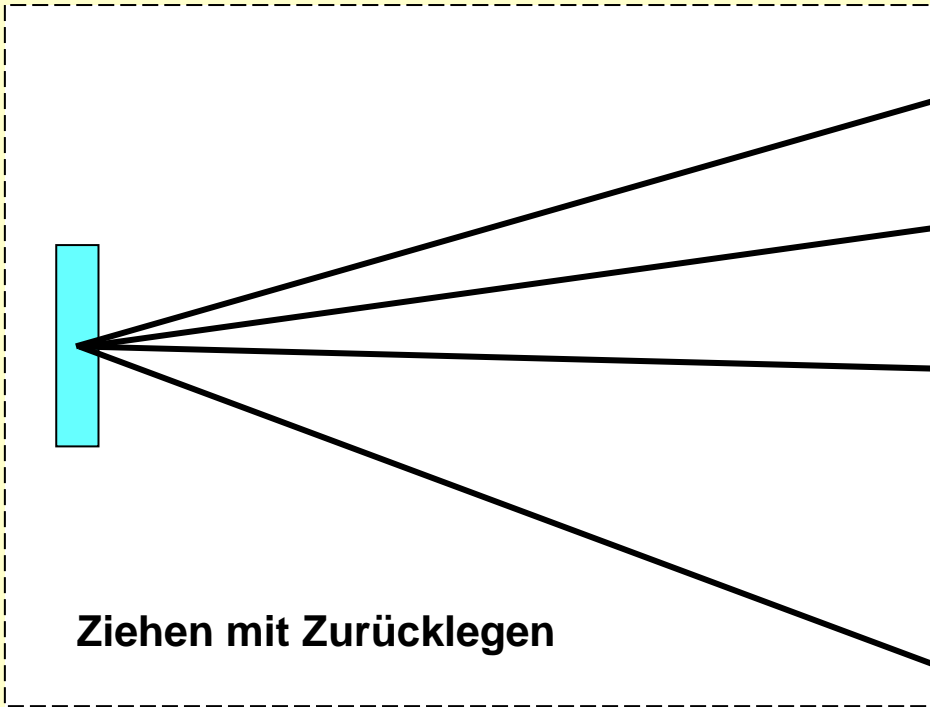
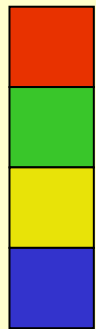
Bootstrap-
Stichproben

1

2

3

4



Die Verteilung einer Statistik kann durch die aus den Bootstrap-Stichproben ermittelte empirische Verteilung dieser Statistik approximiert werden.

$\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ Stichprobe nach einer Zufallsgröße mit der Verteilung F

1

$s(\mathbf{x})$ Stichprobenfunktion (Statistik)

2

$\hat{\vartheta} = s(\mathbf{x})$ Schätzer für einen unbekanntem Parameter der Verteilung F

3

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:

4

1. Aus der Ausgangsstichprobe

$$\mathbf{x} = \{x_1, x_2, \dots, x_n\}$$

1

werden unabhängig voneinander m Bootstrap-Stichproben

2

$$\mathbf{x}_j^* = \{x_{j1}^*, x_{j2}^*, \dots, x_{jn}^*\} \quad , \quad j = 1, \dots, m$$

3

gezogen, wobei die Elemente der Bootstrap-Stichproben durch zufällige Auswahl von Elementen der Ausgangsstichprobe mit der Wahrscheinlichkeit $1/n$ mit Zurücklegen gebildet werden.

4

1

2. Für die m Bootstrap-Stichproben werden die Stichprobenfunktionen

2

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

3

und damit eine empirische Verteilung der Stichprobenfunktion s berechnet.

4

3. Eine Bootstrap-Schätzung für den Standardfehler $se_F(\hat{\vartheta})$ des Schätzers $\hat{\vartheta} = s(\mathbf{x})$ ergibt sich damit aus der Standardabweichung von

1

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

2

als

$$se_{boot} = \left[\frac{1}{m-1} \sum_{j=1}^m (\hat{\vartheta}_j^* - \bar{\hat{\vartheta}}^*)^2 \right]^{\frac{1}{2}}$$

3

mit

4

$$\bar{\hat{\vartheta}}^* = \frac{1}{m} \sum_{j=1}^m \hat{\vartheta}_j^*$$

1

2

Bootstrap-Macros im SAS-Programm JACKBOOT.SAS

3

4

<http://ftp.sas.com/techsup/download/stat/jackboot.sas>

1

Zur Durchführung von Bootstrap-Analysen sind im SAS-Programm `jackboot.sas` die beiden Macros `%BOOT` und `%BOOTCI` verfügbar.

2

Zunächst ist allerdings ein Macro `%ANALYZE` zu schreiben, das die gewünschten Statistiken berechnet.

3

Dies geschieht in der Regel unter Verwendung geeigneter SAS-Prozeduren.

4

1

```
%macro analyze(data=,out=);
```

SAS-Prozedur zur Berechnung der Stichprobenfunktion

```
%if &syserr=0 %then %do;
```

```
  data step(s)
```

```
%end;
```

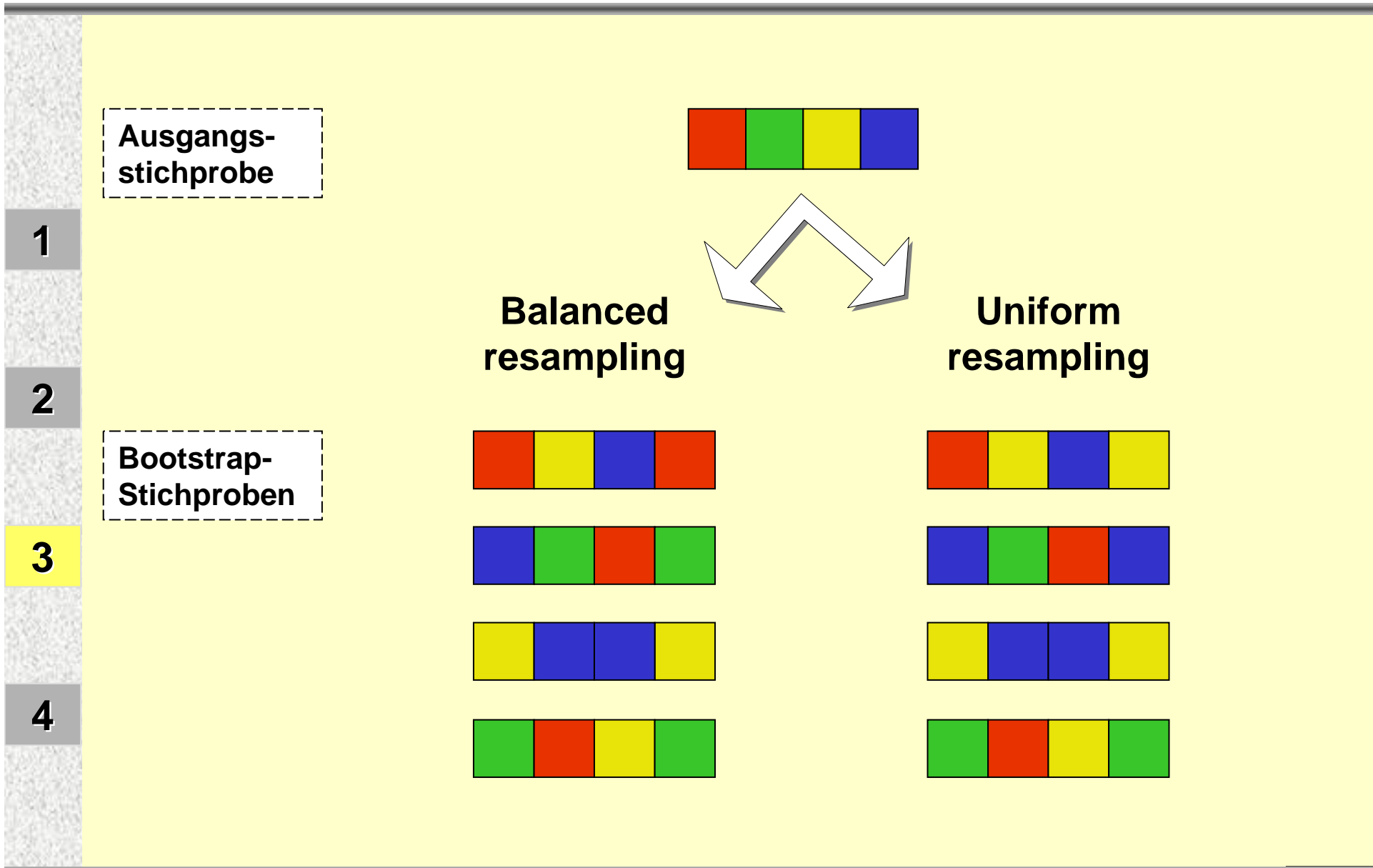
```
%mend;
```

2

```
%macro boot( data=,  
             samples=200,  
             residual=, equation=,  
             size=,  
             balanced=,  
             random=0,  
             stat=_numeric_,  
             id=,  
             biascorr=1,  
             alpha=.05,  
             print=1,  
             chart=1 );
```

3

4



werden in %BOOT aufgerufen:

```
%macro bootby /* Uniform bootstrap resampling */
```

oder:

```
%macro bootbal /* Balanced bootstrap resampling */
```

```
* Gleason, J.R. (1988) "Algorithms for balanced bootstrap  

  simulations," American Statistician, 42, 263-266;
```

```
%macro bootse /* Bootstrap estimates of  

  standard error,  

bias, and  

normal confidence intervals */
```

wird nach %BOOT aufgerufen (Nichtnormalität):

```
%macro bootci /* Bootstrap percentile-based confidence  

intervals.*/
```

2.1 Bootstrap-Schätzung für den Standardfehler des Mittelwertes

1. Aus einer vorliegenden konkreten Stichprobe

1
$$\mathbf{x} = \{x_1, x_2, \dots, x_n\}$$

werden zunächst m Bootstrap-Stichproben

2
$$\mathbf{x}_j^* = \{x_{j1}^*, x_{j2}^*, \dots, x_{jn}^*\} \quad j = 1, \dots, m$$

gezogen.

3

2. Berechnung der Mittelwerte aus den Bootstrap-Stichproben

4
$$\bar{x}_j^* = \frac{1}{n} \sum_{i=1}^n x_{ji}^* \quad j = 1, \dots, m \quad \longrightarrow \quad \bar{\mathbf{x}}^* = \{\bar{\mathbf{x}}_1^*, \bar{\mathbf{x}}_2^*, \dots, \bar{\mathbf{x}}_m^*\}$$

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

von $\bar{\mathbf{x}}^* = \{ \bar{\mathbf{x}}_1^*, \bar{\mathbf{x}}_2^*, \dots, \bar{\mathbf{x}}_m^* \}$:

1

2

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

3

4

2.2 Beispiel

Tabelle 1: Stichprobenelemente

1

Stichprobe

2

$$\mathbf{x} = \{x_1, x_2, \dots, x_{10}\}$$

3

aus einer Normalverteilung mit
 Erwartungswert **0**
 und Varianz **1**
 in der temporären SAS-Datei:

4

STICH1

x	
-0.32659	x_1
1.54244	x_2
0.25903	x_3
-0.55583	x_4
0.77872	x_5
-0.65520	x_6
-0.02210	x_7
-0.77265	x_8
0.72423	x_9
1.26331	x_{10}

Macro %ANALYZE für die Statistik 'MEAN'

%macro analyze(data=,out=);

%VARDEF

```
proc means noprint data=&data vardef=DF ;
  output out=&out (drop=_type_ _freq_
                  rename=(_STAT_=STAT)) ;
```

```
var X ;
  %bystmt ;
```

**ID = STAT
 in %BOOT**

```
run;
%if &syserr=0 %then %do;
  data &out;
    set &out;
    where STAT='MEAN' ;
  run;
%end;
```

%mend ;

1

2

3

4

Aufruf des Macros %BOOT :

```
title1 'Beispiel mit JACKBOOT';  
title2 'Bootstrap Analysis uniform resampling';  
%boot(data= STICH1,  
       samples=200,  
       balanced=0,  
       id=stat,  
       random=123);
```

1

2

3

4

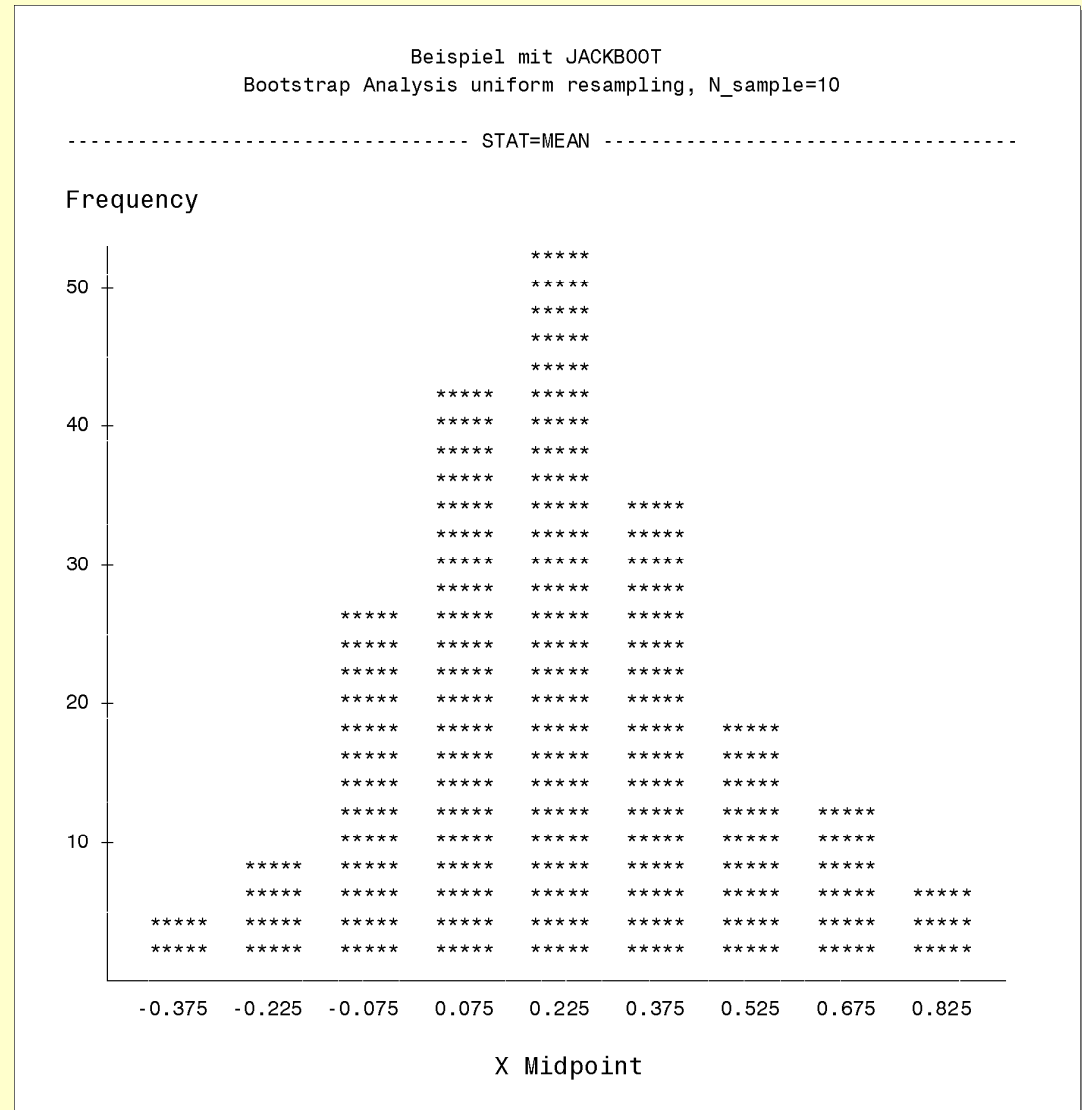
1

2

3

4

Abbildung 1:
 Histogramm für das
 arithmetische Mittel



1

Beispiel mit JACKBOOT
 Bootstrap Analysis uniform resampling, N_sample=10

2

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

3

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

4

Abbildung 2: Statistische Maßzahlen

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
<hr/>				
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
<hr/>				
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
<hr/>				
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
<hr/>				
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

Tabelle 4: Mittelwerte von 1000 Wiederholungen der Bootstrap-Analysen von Tabelle 2

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

1

2

3

4

1

2

3

Ein Bootstrap-Macro mit SAS-IML

4

Vorgehensweise zur Realisierung von Bootstrap-Analysen:

1

Erzeugung einer vorgegebenen Anzahl von Bootstrap-Stichproben aus der Ausgangsstichprobe mit dem **Macro Bootstrap**

2

3

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

4

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

3.1 Das SAS-Macro %BOOTSTRAP

1

Voraussetzung für die Anwendung des Macros ist das Vorliegen der SAS-Datei `SAMPLE`, die die Ausgangsstichprobe enthält.

2

3

Die Anwendung dieses Macros erfordert keine Kenntnisse der Macro-Programmierung. Man muß nur die Parameter beim Aufruf des Macros setzen.

4

```
%macro Bootstrap(M, N_RESAMPLE, SAMPLE, START,  

  BALANCE, BOOT, NAMES)
```

M

Anzahl der Bootstrap-Stichproben

1

N_RESAMPLE

Umfang der Bootstrap-Stichproben

Kann nur bei BALANCE=0 verschieden vom
 Umfang der Ausgangsstichprobe sein!

2

START

Startwert für den Zufallszahlengenerator bei der
 Resampling-Prozedur:

3

START=0: zufälliger Startwert durch Systemzeit
 START=G: fester Startwert mit der positiven ganzen
 Zahl G

4

SAMPLE

temporäre SAS-Datei der Ausgangsstichprobe
 mit der/den Variablen der Ausgangsstichprobe
 (z.B.: X, Y)

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

BALANCE

Parameter zum Einstellen der Resampling-Prozedur:

BALANCE=1: balanciertes Resampling

BALANCE=0: gleichverteiltes Resampling

BOOT

temporäre 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 Ausgangstichprobe (z.B.: X, Y)

NAMES

Variablenbezeichnung in BOOT

z.B.: {'N_BOOT' 'NR_Obs' 'X' 'Y'}

1

2

3

4

Aufruf des Macros z. B.:

1

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

M
 N_RESAMPLE
 START
 SAMPLE
 BALANCE
 BOOT
 NAMES

Uniform
 resampling

2

3

```
%Bootstrap( 200 ,
            10 ,
            123 ,
            STICHPROBE ,
            1 ,
            BOOTSTICH ,
            { 'N_BOOT' 'NR_Obs' 'X' } ) ;
```

M
 N_RESAMPLE
 START
 SAMPLE
 BALANCE
 BOOT
 NAMES

Balanced
 resampling

4

3.2 Beispiel

Erzeugen der 200 Bootstrap-Stichproben

```

title1 'Beispiel mit %BOOTSTRAP';
title2 'Bootstrap Analysis uniform';

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

Abbildung 3:
 Erzeugte Bootstrap-Stichproben

3000	N_BOOT	NR_Obs	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
18	2	9	0.7242
19	2	2	1.5424
20	2	9	0.7242
21	3	7	-0.0221
22	3	8	-0.7726
23	3	8	-0.7726
24	3	4	-0.5558
25	3	6	-0.6552
26	3	10	1.2633
27	3	1	-0.3266
28	3	7	-0.0221
29	3	7	-0.0221
30	3	4	-0.5558

Berechnung des Mittelwerts der Ausgangsstichprobe

1

```
proc means noprint data=STICH1 vardef=DF ;
  output out=OUTSTICH1 (drop=_type_ _freq_
                        rename=( _STAT_=STAT) ) ;

  var X ;
```

2

```
run ;
data OUTSTICH1 ;
  set OUTSTICH1 ;
  where STAT= 'MEAN' ;
```

3

```
run ;
```

4

Abbildung 4:
 Mittelwert der Ausgangsstichprobe

The screenshot shows a SAS output window titled 'WORK.OUTSTICH1'. It displays a table with the following data:

2	Nom	Int			
1	STAT	X			
1	MEAN	0.2235			

Berechnung der 200 Mittelwerte der Bootstrap-Stichproben

1

```
proc means noprint data=BOOTSTICH
  output out=OUTBOOT1 (drop=_type
                    rename=(_s
```

2

```
  var X ;
  by N_BOOT;
```

```
run;
```

3

```
data OUTBOOT1;
  set OUTBOOT1;
  where STAT='MEAN' ;
run;
```

4

Abbildung 5:
 Mittelwerte der Bootstrap-Stichproben

200		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	
■	18	18 MEAN	0.2394	
■	19	19 MEAN	0.2921	
■	20	20 MEAN	-0.3594	
■	21	21 MEAN	0.2297	
■	22	22 MEAN	0.4437	

Statistische Maßzahlen der Verteilung des Mittelwertschätzers

```

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;
  
```

The screenshot shows a SAS output window titled 'WORK.STATBOOTMEAN'. It displays a table with 5 rows of statistical measures for the variable X, grouped by STAT. The measures include Mean, N, Minimum, Maximum, and Standard Deviation.

STAT	MEASURE	VALUE
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 Mittelwerte der Bootstrap-Stichproben

Data steps ...

1

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

2

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

3

4

```

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;
```

1

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

2

Name	Observed	Number of Resampled 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

3

Name	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

4

Abbildung 7: Ergebnisausdruck der Bootstrap-Analyse

... und Grafik mit PROC UNIVARIATE

1

```
PROC UNIVARIATE Data=WORK.OUTBOOT1
                ALL
                VARDEF=df;
```

2

```
HISTOGRAM X / normal (noprint color=black )
                cbarline=black cfill=LIGR ;
```

3

```
OUTPUT OUT=OUTUNIVAR N=BOOT_N
                MIN=BOOT_MIN
                MAX=BOOT_MAX
                MEAN=BOOT_MEAN ;
```

4

```
VAR X ;
by STAT;
Run ;
```

1

2

3

4

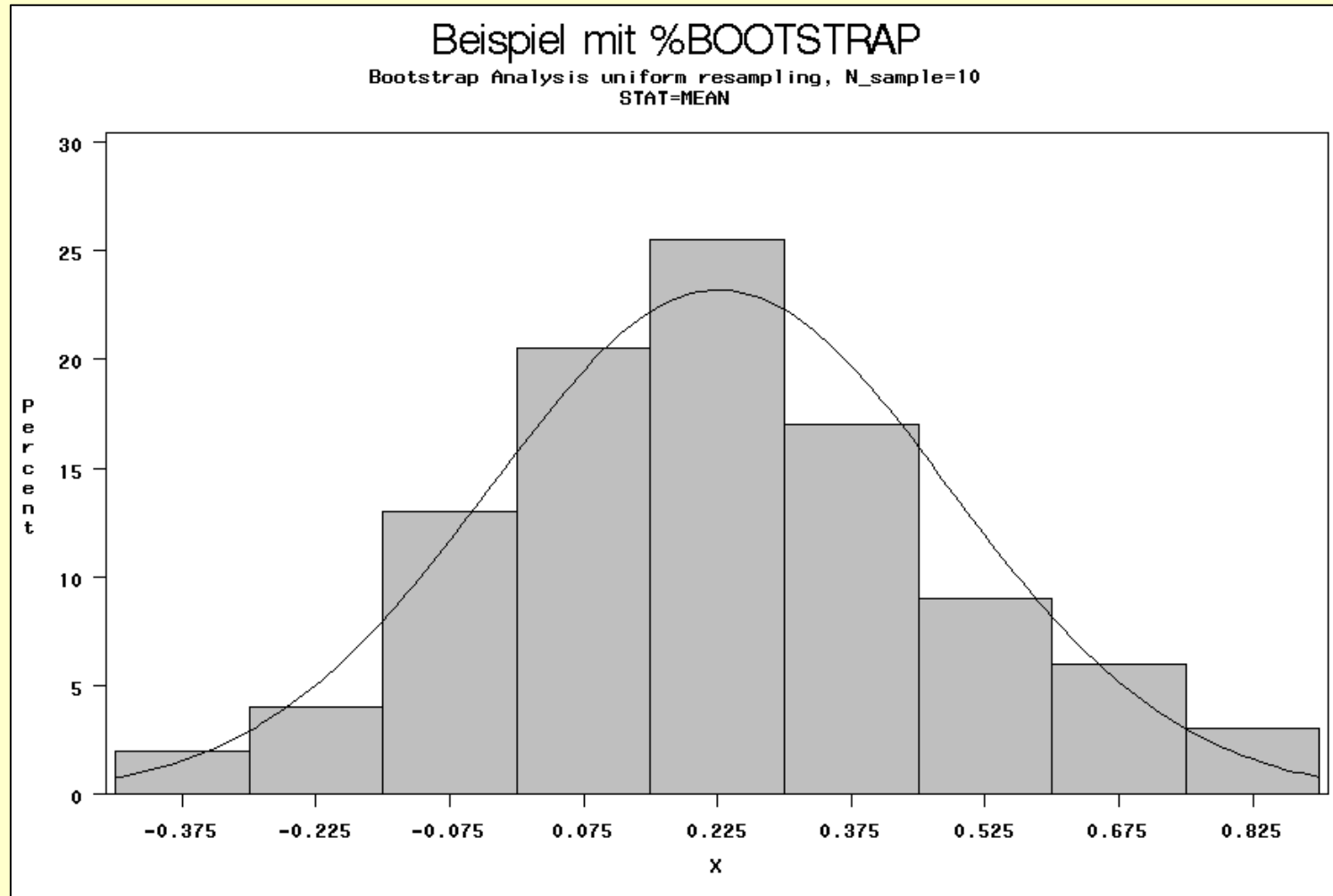


Abbildung 8: Histogramm für das arithmetische Mittel

1

2

3

4

Literatur



1

- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* **7**, 1-26.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability. New York: Chapman & Hall.
- Gleason, J. R. (1988). Algorithms for Balanced Bootstrap Simulations. *American Statistician*, **42**, 263-266.
- Johnson, N. L.; Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* J. Wiley, New York.

2

- Rey, W. J. J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods.* Springer-Verlag Berlin, Heidelberg, New York, Tokyo.
- SAS Institute Inc. (1999). *SAS Macro Language: Reference, Version 8,* Cary, NC: SAS Institute Inc.

3

- SAS Institute Inc. (1999). *SAS/IML User's Guide, Version 8,* Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1999). *SAS/STAT® User's Guide, Version 8,* Cary, NC: SAS Institute Inc.

4

- 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.
- 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.