SAS Makro UNISTATS 2.0 – Ein universelles Werkzeug

Heinrich Stürzl
Siemens Healthcare Diagnostics Products GmbH
Emil-von-Behring-Str. 76
35041 Marburg

Cornelius Gutenbrunner
Siemens Healthcare Diagnostics Products GmbH
Emil-von-Behring-Str. 76
35041 Marburg

Zusammenfassung

Das SAS Makro UNISTATS liefert sämtliche PROC UNIVARIATE Statistiken in übersichtlicher Form, d.h. in einer Zeile pro Analysevariable und By-Gruppe (ähnlich wie PROC MEANS) und erstellt eine entsprechende SAS Tabelle für die Ausgabe und Weiterverarbeitung. Die Analysevariablen lassen sich auch über Variablenlisten wie x1-x10, a—z, abc:, _numeric_ angeben und die zu verarbeitenden Beobachtungen der Eingabe Tabelle können über eine Where-Bedingung gefiltert werden.

Mit der neuen Version 2.0 können beliebige Perzentile z.B. 97,5 % mit den Perzentil Definitionen von SAS und Microsoft Excel (!) berechnet werden. Außerdem neu sind die robusten Statistiken von PROC UNIVARIATE, die Ausgabe von Variablenlabels, und eine Syntaxhilfe mit %unistats(?).

Das Makro zeichnet sich aus durch

- Universelle Einsetzbarkeit
- Hohe Flexibilität
- Leichte Bedienbarkeit
- Benutzerfreundlichkeit
- Effiziente Programmierung

UNISTATS ist unter der GNU General Public Licence als Open Source frei verfügbar. (http://www.opensource.org/licenses/gpl-license.php)

Schlüsselwörter: Univariate Statistik, Deskriptive Statistik, PROC UNIVARIATE, PROC MEANS, Perzentil, Quantil, Quartil, Median, Excel

1 Einleitung

Das Makro UNISTATS entstand 1997/98 in der Mailingliste SAS-L durch Zusammenarbeit der SAS Programmierer Ian Whitlock, Jozsef Vitrai, Michael Friendly und Heinrich Stürzl; Also zu einer Zeit, als PROC MEANS noch keinen Median berechnen konnte und die Ausgabe von PROC UNIVARIATE sehr unhandlich war. Der Wunsch war, die Mächtigkeit der Prozedur UNIVARIATE mit dem übersichtlichen Ausgabe Layout der Prozedur MEANS zu kombinieren, indem die Analysevariablen in Zeilen und die Statistiken in Spalten dargestellt werden. Außerdem sollten die Ergebnisse in dieser Form als SAS Tabelle für die weitere Verarbeitung gespeichert werden können.

Die Version 1.0 erfüllte diese Anforderungen und erlaubte auch SAS übliche Variablenlisten wie x1-x10 oder _numeric_ für die Analysevariablen, sowie By-Gruppen-Verarbeitung und eine Where-Bedingung. Es konnten alle Standardstatistiken berechnet werden, die über „Statistical Keywords“ in PROC UNIVARIATE verfügbar sind. Dazu zählten auch die Standardperzentile MEDIAN, Q1, Q3, P1, P5, P10, P90, P95, P99. Andere Perzentile waren nicht zu bekommen. Hierfür musste man doch noch die Prozedur UNIVARIATE mit der Option PCTLPTS verwenden.

Mit der Version 2.0 wird dieser Mangel beseitigt, indem **PCTLPTS** und **PCTLDEF** unterstützt werden. Damit lassen sich mit UNISTATS **beliebige Perzentile mit den Perzentil-Definitionen von SAS und Excel** berechnen. Neben der Standarddefinition PCTLDEF=5 stehen die übrigen SAS Definitionen 1 bis 4 und die von Microsoft Excel, die mit keiner der fünf SAS Definitionen übereinstimmt zur Verfügung. Letzteres geht über den Funktionsumfang von PROC UNIVARIATE hinaus und ist auch sonst in der SAS Software nicht enthalten.

Außerdem neu mit der Version 2.0:

- Die Variablenlabel der Analysevariablen werden ausgegeben (falls vorhanden).
- Makro bietet eine Syntaxhilfe im Log mit **`%UNISTATS(?)`**. Siehe Beispiel 2.

2 Beschreibung von UNISTATS 2.0

2.1 Eingabeparameter

Das Makro kommt mit wenigen Einstellungen aus. Im einfachsten Fall gibt man nur die SAS Tabelle (data set) an und erhält für jede enthaltene numerische Variable die als Standard eingestellten Statistiken N MEDIAN MEAN STD CV MIN MAX.
Selbstverständlich kann diese Voreinstellung geändert werden, so dass andere Kennzahlen berechnet werden, wenn nichts anderes angegeben ist.

Beispiel 1: Standardstatistiken für alle numerischen Variablen

```
%UNISTATS(data=sashelp.shoes);
```

Listenausgabe im Output Fenster (verkürzt dargestellt):

Name	Label	Pctl Def	N	Median	Mean
Stores	Number of Stores	5	395	10	11.65
Sales	Total Sales	5	395	38912	85700.17
Inventory	Total Inventory	5	395	118849	250898.86
Returns	Total Returns	5	395	1438	2967.32

Die Dokumentation im Log zeigt sämtliche Makroparameter und ihre aktuell verwendeten Einstellungen. Darin werden bereits alle Voreinstellungen der nicht explizit ver-

wendeten Makroparameter sichtbar. Dies kann auch als Kopiervorlage für künftige Makroaufrufe dienen.

```
********** Start Macro UNISTATS Version 2.0 ********
%unistats (
  , data=sashelp.shoes
  , vars=Stores Sales Inventory Returns
  , by=
  , where=
  , stats=N MEDIAN MEAN STD CV MIN MAX
  , pctlpts=
  , pctldef=5
  , out=work.stat1
  , print=Y
  )
********** End Macro UNISTATS Version 2.0 **********
```

Hinweis: Falls keine Tabelle angegeben wird, wird die zuletzt verwendete Tabelle (&syslast) verwendet.

Beispiel 2: Syntaxhilfe im Log (Englisch)

```
%UNISTATS(?);
```

Dies führt zu folgender Kurzbeschreibung der Makroparameter im Log mit Beispiel:

```
********** Start Macro UNISTATS Version 2.0 **********
Macro UNISTATS makes proc univariate statistics more convenient
presenting one row for each analysis variable and by-group

%unistats (
  , data=   input data set. Default: &syslast

  , vars=   analysis variable(s) e.g. x1 x3 x5 | x1-x10 | x: |
            _numeric_ (Default)

  , by=     by-variable(s) (optional)

  , where=  where condition for input data set (optional)

  , stats=  any statistic keyword(s) of proc univariate
            Default: N MEDIAN MEAN STD CV MIN MAX
            N NOBS NMISS SUM MEDIAN MEAN VAR STD CV MIN MAX Q1 Q3 P1
             P5 P10 P90 P95 P99 NORMAL PROBN T PROBT MSIGN PROBM
             SIGNRANK PROBS STDMEAN USS CSS SKEWNESS KURTOSIS SUMWGT
```

```
              RANGE QRANGE MODE GINI MAD QN SN STD_GINI STD_MAD STD_QN
              STD_QRANGE STD_SN

  , pctlpts=user-defined percentile(s) (optional)

  , pctldef=definition for computing percentiles e.g. 1-5 | Excel
            Default: 5
            Excel: according to MS Excel function QUANTIL(Matrix,
             Alpha)

  , out=    Output data set containing one observation for each
             analysis variable and by-group
            Default=work.stat1

  , print=  Printing Option. Default=Y
  )

Example: Calculating N MEDIAN MEAN STD CV MIN MAX and 2.5 97.5
   percentile for all numerical variables of sashelp.class data set
-------------------------------------------------------------------
%unistats (
  , data=sashelp.class
  , vars=_numeric_
  , by=
  , where=
  , stats=N MEDIAN MEAN STD CV MIN MAX
  , pctlpts=2.5 97.5
  , pctldef=5
  , out=work.stat1
  , print=y
  )
********** End Macro UNISTATS Version 2.0 **********
```

Im Folgenden werden die Makroparameter im Einzelnen beschrieben.

2.1.1 Input Parameter DATA

Optional. Name der zu verarbeitenden SAS Tabelle (data set) in der Form `libref.table`. Falls keine Tabelle angegeben wird, wird die zuletzt verwendete Tabelle (&syslast) verwendet.

Falls die Tabelle nicht existiert oder kein Zugriff möglich ist, bricht das Makro mit einer entsprechenden Fehlermeldung und dem Returncode UNISTATS_RC=1 ab.

2.1.2 Input Parameter VARS

Optional. Name der (numerischen) Analysevariable(n) in der verwendeten SAS Tabelle, deren Ausprägungen analysiert werden. Falls keine Variable angegeben wird, werden alle numerischen Variablen der Tabelle analysiert (maximal 99999).

Die Variablen können einzeln aufgezählt werden (durch Blank getrennt) oder verkürzt als Liste in einer der folgenden Formen angegeben werden:

- x1-x99 alle fortlaufend nummerierten Variablen x1, x2, …, x99
- aa--zz alle Variablen von aa bis zz einschließlich
- abc: alle Variablen, deren Name mit ‚abc' beginnt
- _numeric_ alle numerischen Variablen (Voreinstellung)

Alle Variablen müssen numerisch und in der Tabelle vorhanden sein. Andernfalls bricht das Makro mit einer entsprechenden Fehlermeldung und dem Returncode UNISTATS_RC=1 ab.

2.1.3 Input Parameter BY

Optional. Name der By-Variable(n), für deren Ausprägungen die Analyse gruppiert wird. Für jede By-Gruppe (Kombination der Ausprägungen der By-Variablen) wird für jede Analysevariable eine getrennte Analyse durchgeführt und in einer separaten Zeile gespeichert und ausgegeben. Falls nichts angegeben wird, erfolgt die Analyse ungruppiert für alle Beobachtungen der Input Tabelle.

Die Ergebnisse werden nach den By-Gruppen aufsteigend sortiert ausgegeben. Die Input Tabelle muss nicht nach den By-Variablen sortiert sein und bleibt unverändert.

Alle By-Variablen müssen in der Tabelle vorhanden sein. Andernfalls bricht das Makro mit einer entsprechenden Fehlermeldung und dem Returncode UNISTATS_RC=1 ab.

Beispiel 3: By-Gruppen Verarbeitung

```
%UNISTATS(data=sashelp.shoes, by=region subsidiary);
```

Listenausgabe im Output Fenster (verkürzt dargestellt):

```
---------------------- Region=Western Europe Subsidiary=Paris -----
                                  Pctl
Name                 Label        Def     N      Median          Mean

Stores       Number of Stores       5      8        10.0         10.75
Sales        Total Sales            5      8     90195.5      77734.63
Inventory    Total Inventory        5      8    170326.5     183730.75
Returns      Total Returns          5      8      2380.0       2415.50

---------------------- Region=Western Europe Subsidiary=Rome ------
                                  Pctl
Name                 Label        Def     N      Median          Mean

Stores       Number of Stores       5      8         8.0          9.38
Sales        Total Sales            5      8     32246.5      35258.00
Inventory    Total Inventory        5      8     83259.0     129232.25
Returns      Total Returns          5      8      1204.5       1300.88
```

2.1.4 Input Parameter WHERE

Optional. Where-Bedingung, die vor der Analyse auf die Input Tabelle angewendet wird. Damit lässt sich die Analyse auf eine Teilmenge der Beobachtungen einschränken, die eine bestimmte Bedingung erfüllen. Falls nichts angegeben wird, erfolgt die Analyse für alle Beobachtungen der Input Tabelle.

Die angegebene Bedingung muss eine gültige SAS Where-Bedingung sein, die sich auf die Input Tabelle anwenden lässt. Andernfalls bricht das Makro mit einer entsprechenden Fehlermeldung und dem Returncode UNISTATS_RC=1 ab.

Beispiel 4: Einschränkung der Analyse auf eine Teilmenge der Daten

```
%UNISTATS(data=sashelp.shoes, where=region="Western Europe");

%UNISTATS(data=sashelp.shoes, where=region=:"A");
```

beginnt mit "A" → Africa, Asia

```
%UNISTATS(data=sashelp.shoes, where=region CONTAINS "Europe");
```

enthält "Europe" → Eastern Europe, Western Europe

```
%UNISTATS(data=sashelp.shoes,
          where=region="Asia" and Returns > 1000);
```

2.1.5 Input Parameter STATS

Optional. Liste der Statistiken, die berechnet werden. Falls nichts angegeben wird, werden die vordefinierten Statistiken berechnet (N MEDIAN MEAN STD CV MIN MAX). Diese Voreinstellung kann bei der Definition des Makroparameters im Makro Quelltext einfach geändert werden.

Die Statistiken werden in derselben Reihenfolge ausgegeben wie sie angegeben werden.

Zulässig ist eine beliebige Menge (mindestens eines) der folgenden statistischen Schlüsselwörter von PROC UNIVARIATE oder das Schlüsselwort **_ALL_**:

Tabelle 1: Deskriptive Statistiken

CSS	Corrected sum of squares
CV	Coefficient of variation
KURTOSIS	Kurtosis
MAX	Largest value
MEAN	Sample mean
MIN	Smallest value
MODE	Most frequent value
N	Sample size
NMISS	Number of missing values
NOBS	Number of observations
RANGE	Range
SKEWNESS	Skewness
STD	Standard deviation
STDMEAN	Standard error of the mean
SUM	Sum of the observations
SUMWGT	Sum of the weights
USS	Uncorrected sum of squares
VAR	Variance

Tabelle 2: Quantile/Perzentile

P1	1st percentile
P5	5th percentile
P10	10th percentile
Q1	Lower quartile (25th percentile)
MEDIAN	Median (50th percentile)
Q3	Upper quartile (75th percentile)
P90	90th percentile
P95	95th percentile
P99	99th percentile
QRANGE	Interquartile range (Q3 - Q1)

Tabelle 3: Robuste Statistiken (neu ab Version 2.0)

GINI	Gini's mean difference
MAD	Median absolute difference about the median
QN	Q_n, alternative to MAD
SN	S_n, alternative to MAD
STD_GINI	Gini's standard deviation
STD_MAD	MAD standard deviation
STD_QN	Q_n standard deviation
STD_QRANGE	Interquartile range standard deviation
STD_SN	S_n standard deviation

Tabelle 4: Hypothesentests

MSIGN	Sign statistic
NORMAL | NORMALTEST	Test statistic for normality
SIGNRANK	Signed rank statistic
PROBM	Probability of a greater absolute value for the sign statistic
PROBN	Probability value for the test of normality
PROBS	Probability value for the signed rank test
PROBT	Probability value for the Student's t test
T	Statistic for the Student's t test

Im Fall des Schlüsselworts _ALL_ werden alle genannten Statistiken berechnet.

Für Details zu den Statistiken siehe die Dokumentation von PROC UNIVARIATE.

Beispiel 5: Auswahl gewünschter Statistiken

```
%UNISTATS(data=sashelp.shoes, stats=nobs nmiss n mean median mode);

%UNISTATS(data=sashelp.shoes, stats=_all_);
```

2.1.6 Input Parameter PCTLPTS

Optional. Liste beliebiger Perzentile (percentile points), die berechnet werden. Die Perzentile werden wie bei PROC UNIVARIATE definiert und durch Blank getrennt. Sie dürfen maximal zwei Nachkommastellen besitzen.[2]

Die Standard Perzentile P1 P5 P10 Q1 MEDIAN Q3 P90 P95 P99 können auch angefordert werden mit PCTLPTS=1 5 10 25 50 75 90 95 99.

[2] Diese Einschränkung kommt von PROC UNIVARIATE, wo für die Namen der erzeugten Variablen von PCTLPTS maximal zwei Nachkommastellen möglich sind. Das bedeutet auch, dass Perzentile, die sich erst ab der dritten Nachkommastellen unterscheiden grundsätzlich nicht berechnet werden können. Beispiel: 90,567 und 90,568. In diesem Fall erscheint ein Warnhinweis im Log, dass eine gleichnamige Variable schon vorhanden ist.

Die Perzentile werden nach den anderen Statistiken ausgegeben.

Beispiel 6: 2,5 und 97,5 % Perzentile gemäß Standard Definition

```
%UNISTATS(data=sashelp.shoes, stats=n min max, pctlpts=2.5 97.5);
```

Name	Label	Pctl Def	N	Min	Max	Percentile 2.5%	Percentile 97.5%
Stores	Number of Stores	**5**	395	1	41	1	31
Sales	Total Sales	**5**	395	325	1298717	936	434496
Inventory	Total Inventory	**5**	395	374	2881005	3384	1147300
Returns	Total Returns	**5**	395	10	57362	35	16833

2.1.7 Input Parameter PCTLDEF

Optional. Perzentil-Definition, die bei der Berechnung der Standard Perzentile P1 P5 P10 Q1 MEDIAN Q3 P90 P95 P99 QRANGE und den über PCTLPTS angeforderten Perzentilen verwendet wird. Falls nichts angegeben wird, wird die in SAS verwendete Voreinstellung **PCTLDEF=5** verwendet. Diese Voreinstellung kann bei der Definition des Makroparameters im Makro Quelltext einfach geändert werden. Die verwendete Definiton wird in der Variable _PctlDef_ ausgegeben.

Mögliche Einstellungen sind:

- PCTLDEF=5 SAS Standard-Definition (Voreinstellung seit Version 6)
- PCTLDEF=**4** SAS Definition[3] (Voreinstellung bis SAS Version 5)
- PCTLDEF=3 SAS Definition
- PCTLDEF=2 SAS Definition
- PCTLDEF=1 SAS Definition
- PCTLDEF=EXCEL Gemäß der Excel Funktion QUANTIL(Matrix, Alpha)

Im Fall **PCTLDEF=EXCEL** gelten folgende Einschränkungen, weil die Prozedur UNIVARIATE hierfür mit modifizierten Perzentilen separat aufgerufen werden muss. Andernfalls bricht das Makro mit einer entsprechenden Fehlermeldung und dem Returncode UNISTATS_RC=1 ab.

1. Die Standard-Perzentile P1 P5 P10 Q1 MEDIAN Q3 P90 P95 P99 und QRANGE sind im Makroparameter STATS nicht zulässig, weil sie nicht gemäß Excel berechnet werden. Stattdessen muss PCTLPTS=1 5 10 25 50 75 90 95 99 verwendet werden.
2. Es darf nur eine Analysevariable und keine By-Variable angegeben werden. Alternativ mehrere Makroaufrufe für jede Analysevariable und jede By-Gruppe

[3] Die Definition PCTLDEF=4 wird auch von der Statistik Software JMP, SPSS und Minitab verwendet und der IFCC und ICSH empfohlen. Siehe Kapitel 3 „Vergleich der Perzentil-Definitionen“

mit einer entsprechenden Where-Bedingung verwenden und die Ausgabetabellen anschließend mit SET zusammenführen. Siehe Beispiel 8.

Für die genauere Unterscheidung der verschiedenen Perzentil-Definitionen siehe weiter unten Kapitel 3 „Vergleich der Perzentil-Definitionen“.

Beispiel 7: 2,5 und 97,5 % Perzentile gemäß MS Excel

```
%UNISTATS(data=sashelp.shoes,  stats=n  min  max,  pctlpts=2.5  97.5,

          pctldef=excel, vars=returns);
```

Name	Label	Pctl Def	N	Min	Max	Percentile 2.5%	Percentile 97.5%
returns	Total Returns	EXCEL	395	10	57362	35	15881

Beispiel 8: Excel Perzentile Q1, Median, Q3 für By-Gruppen

```
%UNISTATS(data=sashelp.class,  stats=n  min  max,  pctlpts=25  50  75,

     pctldef=excel, vars=age, where=sex="F", out=stat1, print=n);
%UNISTATS(data=sashelp.class,  stats=n  min  max,  pctlpts=25  50  75,

     pctldef=excel, vars=age, where=sex="M", out=stat2, print=n);
data all;
  set stat1 stat2;
run;
```

2.1.8 Input Parameter OUT

Optional. Name der Ausgabetabelle (data set) in der Form `libref.table`, in der die Ergebnisse gespeichert werden. Falls nichts angegeben wird, wird **`work.stat1`** verwendet. Diese Voreinstellung kann bei der Definition des Makroparameters im Makro Quelltext einfach geändert werden.

Die Ausgabetabelle enthält eine Beobachtung pro By-Gruppe und Analysevariable und eine Variable für jede berechnete Statistik. Diese Variablen sind nach dem jeweiligen statistischen Schlüsselwort benannt und haben ein kurzes Label. Darüber hinaus sind folgende Variablen enthalten.

- VName "Name" $32: Name der Analyse Variable
- VLabel "Label" $256: Label der Analyse Variable (ggfs. leer)
- _PctlDef_ "PctlDef" $5: Verwendete Perzentil-Definition
- Ggfs. die verwendeten By-Variablen

Die Ausgabetabelle ist aufsteigend nach den verwendeten By-Variablen sortiert.

Beispiel 9: Ausgewählte Statistiken mit By-Gruppen

```
%UNISTATS(data=sashelp.shoes,
          stats=nobs nmiss n mean median mode,
          by=region subsidiary);
```

Struktur der Ausgabetabelle work.stat1:

#	Variable	Type	Len	Label
1	VName	Char	32	Name
2	VLabel	Char	256	Label
3	_PctlDef_	Char	5	PctlDef
4	Region	Char	25	
5	Subsidiary	Char	12	
6	nobs	Num	8	N total
7	nmiss	Num	8	N missing
8	n	Num	8	N
9	mean	Num	8	Mean
10	median	Num	8	Median
11	mode	Num	8	Mode

2.1.9 Input Parameter PRINT

Optional. Schalter, ob die Ausgabetabelle direkt angezeigt werden soll. Falls nichts angegeben wird, wird die Tabelle mit PROC PRINT und der LABEL Option ggfs. nach By-Gruppen sortiert angezeigt.

Mögliche Einstellungen sind:

- PRINT=Y|y Anzeige der Ausgabetabelle (Voreinstellung)
- Andernfalls findet keine Anzeige statt (Ausgabetabelle wird trotzdem erstellt)

2.2 Technische Hinweise

Das Makro UNISTATS 2.0 erfordert mindestens **SAS 9.1**. Es wurde von uns ausschließlich mit SAS 9.1.3 SP4 unter Windows getestet, ist aber vermutlich auch unter anderen Betriebssystemen lauffähig.

Das Makro liefert in der globalen Makrovariable **UNISTATS_RC** die folgenden Returncodes, die von aufrufenden Programmen abgefragt werden können.

0 OK
1 Unvollständige oder falsche Parameter. Abbruch der Berechnung.
2 Laufzeitfehler.

Die Plot Optionen von PROC UNIVARIATE werden nicht unterstützt.

Das Makro verwendet einen Positionsparameter und ansonsten Keywordparameter (data, vars, by, where, stats, pctlpts, pctldef, out, print). Der Positionsparameter dient zum Aufruf der Syntaxhilfe in der Form `%unistats(?)`.

Die Funktionsweise des Makros lässt sich wie folgt beschreiben:

- Überprüfung der Input Parameter auf Plausibilität
- Erstellen einer temporären Kopie der Input Tabelle
- Ggfs. Where-Bedingung darauf anwenden
- Ggfs. diese Tabelle sortieren bei der Verwendung von BY-Gruppen
- Aufruf von PROC UNIVARIATE mit den entsprechenden Einstellungen und Ausgabe als Data Set
- Data Set transponieren und die Ausgabetabelle erstellen
- Ggfs. Anzeige der Ausgabetabelle mit PROC PRINT

3 Vergleich der Perzentil-Definitionen

Die Berechnung der Perzentile bzw. Quantile ist nicht standardisiert. Die verschiedenen Definitionen können durchaus zu unterschiedlichen Werten führen! Die Unterschiede sind abhängig von der Verteilung der Daten und umso größer, je kleiner der Stichprobenumfang ist. Der Median ist bei den Definitionen PCTLDEF 5 und 4, sowie der von Microsoft Excel identisch. Die Definitionen PCTLDEF 1, 2, 3 sind asymmetrisch.

Tabelle 5: Perzentil Vergleich für die Zahlen von 1 bis 8

PctlDef	1%	2,5%	5%	10%	Q1	Median	Q3	90%	95%	97,5%	99%
5	1	1	1	1	2,5	4,5	6,5	8	8	8	8
4	1	1	1	1	2,25	4,5	6,75	8	8	8	8
EXCEL	1,07	1,175	1,35	1,7	2,75	4,5	6,25	7,3	7,65	7,825	7,93
3	1	1	1	1	2	4	6	8	8	8	8
2	1	1	1	1	2	4	6	7	8	8	8
1	1	1	1	1	2	4	6	7,2	7,6	7,8	7,92

Tabelle 6: Perzentil Vergleich für die Zahlen von 1 bis 30

PctlDef	1%	2,5%	5%	10%	Q1	Median	Q3	90%	95%	97,5%	99%
5	1	1	2	3,5	8	15,5	23	27,5	29	30	30
4	1	1	1,55	3,1	7,75	15,5	23,25	27,9	29,45	30	30
EXCEL	1,29	1,725	2,45	3,9	8,25	15,5	22,75	27,1	28,55	29,275	29,71
3	1	1	2	3	8	15	23	27	29	30	30
2	1	1	2	3	8	15	22	27	28	29	30
1	1	1	1,5	3	7,5	15	22,5	27	28,5	29,25	29,7

Im Folgenden werden die drei relevanten Definitionen PCTLDEF 5 und 4, sowie die von Microsoft Excel kurz dargestellt.

Sei *n* die Anzahl der non-missing Werte $x_1, x_2, .., x_n$, in aufsteigend sortierter Reihenfolge. So gilt für das p-te Perzentil *y* in Abhängigkeit vom Stichprobenumfang *n*:

3.1 SAS PCTLDEF=5

Rangzahl $R = n\,p/100 = j + g$, wobei j der ganzzahlige Anteil und g der Dezimalteil von R ist.

$$y = \frac{1}{2}\left(x_j + x_{j+1}\right) \quad \text{wenn } g = 0$$
$$y = x_{j+1} \quad \text{wenn } g > 0$$

Beispiele für $x_1 = 1, x_2 = 2, ..., x_n = n$:

n=8, p=75: $R = 8 \cdot 0,75 = 6,0$ mit j=6 und g=0,0 → $y = \frac{1}{2}\left(x_6 + x_7\right) = 6,5$

n=30, p=75: $R = 30 \cdot 0,75 = 22,5$ mit j=22 und g=0,5 → $y = x_{23} = 23$

Die SAS Perzentil Definition PCTLDEF=5 ist ab SAS Version 6 voreingestellt (bis zur SAS Version 5 war PCTLDEF=4 die Voreinstellung).

3.2 SAS PCTLDEF=4

Rangzahl $R = (n+1)\,p/100 = j + g$, wobei j der ganzzahlige Anteil und g der Dezimalteil von R ist.

$$y = (1-g)\,x_j + g x_{j+1}$$

mit $y = x_1$, wenn $p/100 < 1/(n+1)$

und $y = x_n$, wenn $p/100 > n/(n+1)$.

Beispiele für $x_1 = 1, x_2 = 2, ..., x_n = n$:

n=8, p=75:

$R = 9 \cdot 0,75 = 6,75$ mit j=6 und g=0,75. → $y = 0,25 \cdot x_6 + 0,75 \cdot x_7 = 6,75$

n=30, p=75:

$R = 31 \cdot 0,75 = 23,25$ mit j=23 und g=0,25 → $y = 0,75 \cdot x_{23} + 0,25 \cdot x_{24} = 23,25$

Die SAS Perzentil Definition PCTLDEF=4 war die Voreinstellung bis zur SAS Version 5 (ab SAS Version 6 gilt die Voreinstellung PCTLDEF=5) und wird heute auch von **JMP, SPSS** und **Minitab** verwendet (andere Definitionen sind dort nicht wählbar). Diese Perzentil-Definition wird außerdem auch von der **IFCC** und **ICSH** empfohlen.[4]

[4] IFCC 1987/9 Vol. 25, 645-656, Seite 650

3.3 Excel Perzentil-Definition

Rangzahl $R = 1 + (n-1)\,p/100 = j + g$, wobei j der ganzzahlige Anteil und g der Dezimalteil von R ist

$$y = (1-g)\,x_j + g\,x_{j+1}\,.$$

Beispiele für $x_1 = 1, x_2 = 2, \ldots, x_n = n$:

n=8, p=75:

$R = 1 + 7 \cdot 0,75 = 6,25$ mit j=6 und g=0,25. → $y = 0,75 \cdot x_6 + 0,25 \cdot x_7 = 6,25$

n=30, p=75:

$R = 1 + 29 \cdot 0,75 = 22,75$ mit j=22 und g=0,75 → $y = 0,25 \cdot x_{22} + 0,75 \cdot x_{23} = 22,75$

Um dies mit PROC UNIVARIATE zu bestimmen, kann man das modifizierte Perzentil $P_{\mathrm{mod}} = \big(p(n-1)+100\big)\big/(n+1)$ gemäß PCTLDEF=4 berechnen.

Die Perzentil-Definition von Microsoft Excel (QUANTIL(Matrix, Alpha) Funktion) scheint auch von anderen Statistik Tools verwendet zu werden wie z.B. **S-Plus, StarOffice Calc** (ohne Gewähr, da wir dies nicht überprüft haben).

Die Ergebnisse mit PCTLDEF=EXCEL wurden mit Excel 2002 und 2003 unter Windows XP verglichen und stimmen exakt überein. Als Testfälle dienten die in den Tabellen 5 und 6 gezeigten Daten.

4 Makroquelltext

Der Makroquelltext von UNISTATS 2.0 steht unter der GNU General Public Licence als Open Source frei zur Verfügung. Details dazu sind zu finden unter:
http://www.opensource.org/licenses/gpl-license.php

Er kann bei Redscope heruntergeladen werden (Anmeldung erforderlich):
http://www.redscope.org
Makro-Sammlung **de.unihd***

Literatur

[1] The UNIVARIATE Procedure: Calculating Percentiles, SAS Procedures Guide, SAS Institute.

[2] International Federation of Clinial Chemistry (IFCC) and International Commitee for Standardization in Heamatology (ICSH), Approved Recommendation (1987) on the Theory of Reference Values: Part 5: Statistical Treatment of Collected Reference Values. Determination of Reference Limits; J. Clin. Chem. Clin. Biochem. Vol 25, 1987, pp. 645-656