PROC UNIVARIATE

zum Berechnen statistischer Maßzahlen, Prüfung auf Normalverteilung, Häufigkeitslisten für stetige Merkmale
(für **quantitative** Merkmale)

Allgemeine Form:

```
PROC UNIVARIATE DATA=name Optionen ;
    VAR variablenliste ;
RUN ;
```

Beispiel und Beschreibung der Programm-Statements:

```
TITLE1 ' PROC UNIVARIATE, Standardeinstellung' ;
TITLE2 '----------------------------------------' ;
PROC UNIVARIATE DATA=beispiel ;
    VAR gewicht ;
RUN ;
```

Die Prozedur beginnt mit PROC endet mit RUN ;

Das ; am Ende eines Befehls beendet eine Anweisung innerhalb der Prozedur.

UNIVARIATE ist der Name der Prozedur.

DATA = beispiel legt das zu verwendende Datenfile fest (im Beispiel beispiel).

VAR ist das gehörende Schlüsselwort und steht für die Liste der Merkmalsnamen.
Ohne VAR - Statement berechnet SAS die statistischen Maßzahlen für **alle numerischen** Merkmale in der Datendatei.

Die Aufzählung der Merkmalsnamen darf in einem VAR-Statement auch über mehrere Zeilen gehen. Das Semikolon kommt dann nach dem letzten Merkmal.

variablenliste wird mit den Namen der zu verarbeitenden Merkmale überschrieben. Trennen Sie die Merkmalsnamen mit Leerzeichen!!!

SAS berechnet im Beispiel die statistischen Maßzahlen, prüft die Verteilungsform der Daten und erstellt eine Häufigkeitsliste für die Merkmale *groesse* und *gewicht*.

Mit dem TITLE-Befehl stellen Sie den Ergebnissen Überschriften voran. Der Befehl ist nicht zwingend notwendig, bringt aber Struktur in die Auswertung. **Achtung!** - SAS übernimmt Titel in nachfolgende Prozeduren, wenn dort das TITLE-Statement fehlt.

Starten Sie die Programmzeilen aus dem Beispiel, zeigt SAS im Output-Fenster die Informationen auf der Rückseite:

PROC UNIVARIATE, Standardeinstellung

The UNIVARIATE Procedure
Variable: Gewicht_nach

① Moments

N	70	Sum Weights	70
Mean	65.9571429	Sum Observations	4617
Std Deviation	11.6313366	Variance	135.287992
Skewness	0.69716351	Kurtosis	0.61575427
Uncorrected SS	313859	Corrected SS	9334.87143
Coeff Variation	17.6346884	Std Error Mean	1.39021063

② Basic Statistical Measures

Location		Variability	
Mean	65.95714	Std Deviation	11.63134
Median	63.00000	Variance	135.28799
Mode	75.00000	Range	57.00000
		Interquartile Range	17.00000

③ Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	47.44399	Pr > |t|	<.0001
Sign	M	35	Pr >= |M|	<.0001
Signed Rank	S	1242.5	Pr >= |S|	<.0001

④ Quantiles (Definition 5)

Quantile	Estimate
100% Max	106.0
99%	106.0
95%	85.0
90%	81.5
75% Q3	74.0
50% Median	63.0
25% Q1	57.0
10%	50.5
5%	50.0
1%	49.0
0% Min	49.0

⑤ Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
49	59	85	5
49	28	85	19
50	71	87	12
50	60	88	6
50	49	106	2

⑥ Missing Values

Missing Value	Count	-----Percent Of----- All Obs	Missing Obs
.	1	1.41	100.00

Beschreibung der Ergebnisse auf der vorhergehenden Seite:

Die berechneten Maßzahlen beziehen sich auf die Gesamtzahl der Beobachtungen abzüglich der Datensätze mit fehlenden Werten.

Teil ① zeigt die Moments = statistische Masszahlen:

Maßzahl (englisch)	Maßzahl (deutsch)	nähere Erläuterung
N	Anzahl der Beobachtungen	
Mean	Mittelwert	
Std Deviation	Standardabweichung	Wurzel aus der Varianz
Skewness	Schiefe	Symmetrie der Verteilung: Schiefe = 0 ⇨ symmetrische Verteilung, Schiefe > 0 ⇨ rechtsschiefe Verteilung, Schiefe < 0 ⇨ linksschiefe Verteilung
Uncorrected SS	unkorrigierte Quadratsumme	Summe aller Wert im Quadrat
Coeff Variation	Variationskoeffizient: s / $\bar{x}$	Quotient von Standardabweichung und arithmetischem Mittel
SUM WEIGHTS	Summe der Gewichtungen für Beobachtungswerte.	bei gleicher Gewichtung (z.B. 1) entspricht dieser Wert N.
Sum Observations	Summe der Stichprobenwerte	
Variance	Varianz	mittlere quadratische Abweichung der Daten vom Mittelwert
Corrected SS	korrigierte Quadratsumme	Summe der Werte abzüglich Mittelwert im Quadrat
Std Error Mean	Standardfehler des Mittelwertes	
Kurtosis	Wölbung	Massenanhäufung an den Enden bzw. um den Mittelwert der Verteilung: Kurtosis=0⇨normalverteilte Daten, Kurtosis>0⇨Werte häufen sich in der Umgebung des Mittelwertes: Verteilung schmaler und steiler als die Normalverteilung Kurtosis<0⇨Verteilung flacher als die Normalverteilung

Teil ② gibt die Basic Statistical Measures = Lage- und Streuungsmaße aus:

zu den Lagemaßen (Location) gehören:		
Maßzahl (englisch)	**Maßzahl (deutsch)**	**nähere Erläuterung**
Mean	arithmetisches Mittel	
Median	Median	eilt die Stichprobe in zwei Hälften: die eine Hälfte der Daten ist höchstens so groß wie der Median und die andere Hälfte mindestens so groß
Mode	Modus	Ausprägung mit der größten Häufigkeit

zu den Streuungsmaßen (Variability) gehören:		
Maßzahl (englisch)	**Maßzahl (deutsch)**	**nähere Erläuterung**
Std Deviation	Standardabweichung	
Variance	Varianz	
Range	Spannweite	Maximum - Minimum
Interquartile Range	Interquartilsbereich	3. Quartil - 1. Quartil

Teil ③ enthält **Tests for Location = t-Test (verbundene Stichproben), Wilcoxon-Test (verbundenStichproben), Vorzeichentest:**

MuO=O enthält die Nullhypothese, dass der Median der Verteilung Null ist.

Unter **Test** zeigt SAS den Namen des durchgeführten Tests, unter **Statistic** die berechnete Testgröße und unter **p Value** die zugehörigen p-Werte:

Test		Testgröße	p-Wert
Students's t	t-Test für verbundene Stichproben	**t**	**Pr > |t|**
Sign	Vorzeichentest	**M**	**Pr >= |M|**
Signed Rank	Wilcoxon-Test für verbundene Stich-proben	**S**	**Pr >= |S|**

Teil ④ zeigt die **Quantile** der Stichprobe**:** Mit **Quantilen** teilt man eine nach der Größe der Werte sortierte Stichprobe in gleichmäßige Teile auf. Bei der Verwendung von Quartilen teilt man die Stichprobe in vier Teile, bei Dezilen in zehn Teile und bei Perzentilen in 100 Teile.

Unter **Quantiles (Definition 5)** führt SAS die Quantile auf, wobei unter der Überschrift **Quantile** das entsprechende Quantil und unter **Estimate** der zugehörige Wert aus der Stichprobe ausgegeben wird.

Wichtig sind hier das **100% Max**=Maximum, **75% Q3**=3. Quartil, **50% Median**=Median, **25% Q1**=1. Quartil und **0% Min**=Minimum.

Teil ⑤ zeigt Informationen zu den **Extremwerten** in der Stichprobe: Unter **Extreme Observations** listet SAS die fünf kleinsten (**Lowest**) und fünf größten (**Highest**) Werte auf, wobei jeweils die SAS-interne Nummer der Beobachtung (**Obs**) hinter dem Wert aufgeführt wird. Kommen Werte mehrfach vor, listet SAS diese entsprechend oft auf.

Teil ⑥ gibt Informationen zur Handhabung **fehlender Werte:** Unter **Missing Values** zeigt wie fehlende Werte dargestellt werden (**Missing Value**), wie viele fehlende Werte es gibt (**Count**), wie hoch der relative Anteil (**Percent of All Obs**) und die relative Summenhäufigkeit der fehlende Werte (**Percent of Missing Obs**) ist.

Beschreibung nützlicher Optionen:

- **Option FREQ** erstellt eine Häufigkeitsliste der einzelnen Ausprägungen:

```
TITLE1 'PROC UNIVARIATE, mit Haeufigkeitsliste' ;
TITLE2 '----------------------------------------------' ;
PROC UNIVARIATE DATA=beispiel FREQ ;
   VAR gewicht ;
RUN ;
```

⇨ SAS gibt zusätzlich zu den statistischen Maßzahlen eine Liste mit Häufigkeiten aus:

```
PROC UNIVARIATE, mit Haeufigkeitsliste
-------------------------------------

The UNIVARIATE Procedure
Variable:  Gewicht

                              Frequency Counts

                Percents                Percents                Percents
Value Count   Cell    Cum     Value Count   Cell    Cum     Value Count   Cell    Cum
   49     2    2.9    2.9        61     3    4.3   41.4        74     3    4.3   75.7
   50     5    7.1   10.0        62     3    4.3   45.7        75     6    8.6   84.3
   51     1    1.4   11.4        63     4    5.7   51.4        77     1    1.4   85.7
   52     1    1.4   12.9        65     1    1.4   52.9        79     1    1.4   87.1
   53     1    1.4   14.3        66     2    2.9   55.7        80     2    2.9   90.0
   55     4    5.7   20.0        67     1    1.4   57.1        83     2    2.9   92.9
   56     2    2.9   22.9        68     1    1.4   58.6        85     2    2.9   95.7
   57     3    4.3   27.1        70     4    5.7   64.3        87     1    1.4   97.1
   58     3    4.3   31.4        71     2    2.9   67.1        88     1    1.4   98.6
   59     2    2.9   34.3        72     1    1.4   68.6       106     1    1.4  100.0
   60     2    2.9   37.1        73     2    2.9   71.4
```

Spalte **`Value`** zeigt den Wert, Spalte **`Count`** die absolute Häufigkeiten, Spalte **`Cell`** die relativen Häufigkeiten und Spalte **`Cum`** die relativen Summenhäufigkeiten.

- **Option NORMAL**: SAS führt eine rechnerische Prüfung auf Normalverteilung durch. Für Stichproben mit weniger als 2000 Beobachtung verwendet man den p-Wert des Shapiro-Wilk-Tests:

```
TITLE1 ' PROC UNIVARIATE, mit Haeufigkeitsliste ' ;
TITLE2 '----------------------------------------------' ;
PROC UNIVARIATE DATA=beispiel NORMAL ;
   VAR gewicht ;
RUN ;
```

⇨ SAS gibt zusätzlich zu den statistischen Maßzahlen die folgenden Zeilen aus:

```
PROC UNIVARIATE, Pruefung auf Normalverteilung
----------------------------------------------

The UNIVARIATE Procedure
Variable:  Gewicht_nach

                  Tests for Normality
Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.951258    Pr < W      0.0084
Kolmogorov-Smirnov    D      0.11463    Pr > D      0.0223
Cramer-von Mises      W-Sq  0.112338    Pr > W-Sq   0.0792
Anderson-Darling      A-Sq   0.73228    Pr > A-Sq   0.0548
```

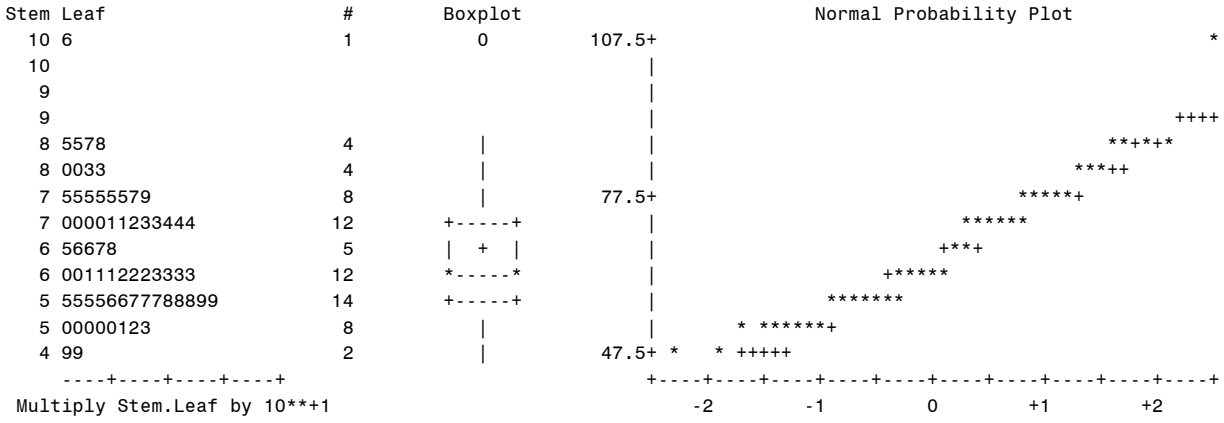
`W` = `0.951258`) ist die Testgröße, **`Pr < W`** = `0.0084` gibt den p-Wert an.

Ist **`Pr < W`** größer als 0,05, dann liegt eine Normalverteilung vor, nimmt Pr < W einen Wert kleiner oder gleich 0,05 an, kann man nicht von einer Normalverteilung ausgehen.

- **Option PLOT**: SAS führt eine rechnerische Prüfung auf Normalverteilung durch. Für Stichproben mit weniger als 2000 Beobachtung verwendet man den p-Wert des Shapiro-Wilk-Tests:

```
TITLE1 'PROC UNIVARIATE, mit Diagrammen' ;
TITLE2 '----------------------------------------' ;
PROC UNIVARIATE DATA=beispiel PLOT ;
   VAR gewicht ;
RUN ;
```

⇨ SAS gibt zusätzlich zu den statistischen Maßzahlen drei Diagramme aus:



Das **linke Diagramm** ist das Stamm-Blatt-Diagramm (**Stem Leaf**), das die empirische Häufigkeitsverteilung der Daten in Form eines vertikalen Diagramms zeigt: SAS trägt auf der x-Achse die absoluten Häufigkeiten und auf der y-Achse die in Klassen gefassten Beobachtungen. Die vertikale Achse bildet dabei den Stamm und die Werte dahinter die Blätter. Kombiniert man die Zahlenwerte aus Stamm und Blättern, erhält man die tatsächlichen Stichprobenwerte (4 9 entspricht 49, 13 0 wird zu 130). Die Anzahl der Zahlenwerte pro Zeile ergibt die Anzahl der Beobachtungen pro Klasse. SAS zeigt sie unter dem #-Zeichen an. Dieses Diagramm stellt die Verteilungsform der Daten dar.

Das **mittlere Diagramm** bezeichnet man als **Boxplot** (Box and Whiskers Plot). SAS beschreibt hier graphisch die Quartile. Die Höhe der Box ergibt sich aus dem Quartilabstand (**Q3** (= obere Linie) – **Q1** (= untere Linie)), das +-Zeichen innerhalb der Box stellt den Mittelwert und die gestrichelte Linie mit den Sternchen rechts und links (*----*) den Median dar. Die beiden senkrechten Linien oberhalb und unterhalb der Box nennt man whiskers. Sie laufen nach oben zum größten Wert und nach unten zum kleinsten Wert, erstrecken sich aber maximal 1,5 Quartilabstände nach oben bzw. unten. Werte, die 1,5 bis 3,0 Quartilabstände außerhalb der Box liegen, werden durch den Wert 0 und Werte, die noch weiter außerhalb der Box liegen mit einem * dargestellt (im Beispiel der Wert 130). Man nennt solche Werte Ausreißer (outliers). Stimmen Mittelwert und Median annähernd überein, kann man von einer Normalverteilung ausgehen.

Das **rechte Diagramm** zeigt den **Normal Probability Plot**, der Hinweise auf das Vorliegen einer Normalverteilung gibt. Hier werden die Quantile der Verteilung der Beobachtungen als Sterne (*) und die Quantile der Normalverteilung als Pluszeichen (+) dargestellt. Aus diesem Grund bezeichnet man diesen Plot auch als Quantil-Quantil (QQ)-Plot. Die +-Zeichen der Normalverteilung bilden eine gerade Linie, die von den * der tatsächlichen Verteilung überlagert werden. Je weniger +-Zeichen zu erkennen sind, desto mehr nähert sich die Verteilung der vorliegenden Daten an die Normalverteilung an.

- **Option NOPRINT** unterdrückt die Ausgabe der Ergebnisse der PROC UNIVARIATE im Output-Fenster.

- **Statement CLASS variablenname(n)** dient der Angabe einer Gruppenvariablen (qualitativ!!!), nach deren Ausprägungen die berechneten Maßzahlen und Testwerte unterschieden werden soll, vgl. PROC MEANS mit Gruppenvariable.

Weitere Optionen bzw. statistische Kenngrößen, die man als Optionen angegeben kann:

OUPUT OUT=datenfile schlüsselwort_für_kenngröße=name erzeugt eine SAS-Datei, deren Namen mit **OUT=** festgelegt wird und die die Ergebnisse für die mit dem Schlüsselwort spezifizierten Maßzahlen enthält.

Schlüsselwörter wichtiger Kennzahlen: **MAX**=Maximum, **MIN**=Minimum, **STD**=Standardabweichung, **VAR**=Varianz, **Q1**=1. Quartil, **Q3**=3. Quartil, **MSIGN**=Testgröße des Vorzeichen-Tests, **PROBM**=p-Wert des Vorzeichen-Tests, **SIGNRANK**=Testgröße des Wilcoxon-Tests für verbundene Stichproben, **PROBS**=p-Wert des Wilcoxon-Tests für verbundene Stichproben, **T**=Testgröße des t-Tests für verbundene Stichproben, **PROBT**=p-Wert des t-Tests für verbundene Stichproben, **PROBN**=p-Wert für den Shapiro-Wilk-Test:

```
TITLE1 'Pruefung auf Normalverteilung' ;
TITLE2 '----------------------------------' ;
PROC UNIVARIATE DATA=beispiel NORMAL NOPRINT ;
   VAR alter groesse gewicht ;
   OUTPUT OUT=shapiro PROBN = alterp  groessep  gewichtp ;
RUN ;
```

```
TITLE1 ' p-Werte der Pruefung auf Normalverteilung (Shapiro-Wilk)' ;
TITLE2 '-----------------------------------------------------------------------' ;
PROC PRINT DATA=shapiro ;
   VAR  alterp  groessep  gewichtp ;
RUN ;
```

⇨ SAS gibt die in der PROC UNIVARIATE berechneten und im Datenfile shapiro abgelegten p-Werte in Form einer Datenliste aus:

```
p-Werte der Pruefung auf Normalverteilung (Shapiro-Wilk)
-------------------------------------------------------

Obs          alterp        groessep        gewichtp
 1       4.3121013E-8     0.2473219702     0.0084428745
```

Der p-Wert für das Merkmal *Alter* lautet $4.3121013*10^{-8}$ (E steht für die Basis 10 und -8 für den Exponenten $^{-8}$), der für das Merkmal *Groesse* 0.2473219702 und der für das Merkmal *Gewicht* 0.0084428745. Hier wären also die beiden Merkmale *Alter* und *Gewicht* nicht normalverteilt, denn der p-Wert des Shapiro-Wilk-Test ist signifikant und das Merkmal *Groesse* normalverteilt, weil der p-Wert des Shapiro-Wilk-Test **nicht** signifikant ist.