

PROC MEANS

zum Berechnen statistischer Maßzahlen
(für **quantitative** Merkmale)

Allgemeine Form:

```
PROC MEANS DATA=name Optionen ;  
  VAR variablenliste ;  
  CLASS vergleichsvariable ;  
RUN ;
```

Beispiel und Beschreibung der Programm-Statements:

```
TITLE1 'Statistische Maßzahlen' ;  
TITLE2 '-----' ;  
PROC MEANS DATA=beispiel MIN MAX MEAN MEDIAN STD MAXDEC=2 ;  
  VAR groesse gewicht ;  
  CLASS geschlecht ;  
RUN ;
```

```
TITLE1 'Statistische Maßzahlen' ;  
TITLE2 '-----' ;  
PROC MEANS DATA=beispiel MIN MAX MEAN MEDIAN STD MAXDEC=2 ;  
  VAR alter ;  
  CLASS rauchen alkoholg ;  
RUN ;
```

Die Prozedur beginnt mit PROC endet mit RUN ;

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

MEANS 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 der letzten Variablen.

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

vergleichsvariable ist der Name des **qualitativen** Merkmals, nach dessen Ausprägungen die Maßzahlen unterschieden werden sollen.

SAS berechnet im Beispiel die statistischen Maßzahlen Minimum (**MIN**), Maximum (**MAX**), Mittelwert (**MEAN**), Median (**MEDIAN**) und Standardabweichung (**STD**) mit jeweils zwei Nachkommastellen (**MAXDEC=2**), getrennt nach Männern und Frauen. Ohne Angabe ausgewählter Maßzahlen und einer Vorgabe von Dezimalstellen gibt SAS die statistischen Maßzahlen N (Anzahl Beobachtungen abzüglich fehlender Werte), Mittelwert, Standardabweichung, Minimum und Maximum mit jeweils sieben Nachkommastellen aus.

Mit dem **CLASS**-Statement definieren Sie eine Gruppenvariable. SAS berechnet die statistischen Maßzahlen für jede Ausprägung dieses Merkmals. **Die Gruppenvariable muss qualitativ sein.**

Hinweis! Geben Sie im **CLASS**-Statement mehr als ein qualitatives Merkmal an, sortiert SAS die berechneten statistischen Maßzahlen innerhalb des 1. Merkmals nach dem 2., innerhalb davon nach dem 3., usw. .

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 folgende Informationen:

Statistische Masszahlen

The MEANS Procedure

Geschlecht	Obs	Variable	Minimum	Maximum	Mean	Median	Std Dev
weiblich	62	Groesse	156.00	184.00	169.66	169.50	7.11
		Gewicht	49.00	83.00	60.92	60.00	8.54
maennlich	31	Groesse	167.00	196.00	182.84	183.00	6.81
		Gewicht	60.00	130.00	79.65	80.00	13.57

Statistische Masszahlen

The MEANS Procedure

Analysis Variable : **Alter**

Rauchen	Alkoholg	N Obs	Minimum	Maximum	Mean	Median	Std Dev
ja	ja	4	24.00	27.00	25.75	26.00	1.50
	nein	9	21.00	27.00	24.89	25.00	2.47
nein	ja	21	21.00	27.00	24.71	25.00	2.57
	nein	37	21.00	27.00	24.41	24.00	2.25

Beschreibung des Outputs:

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

Variable zeigt die Namen der verwendeten Merkmale, im Beispiel: **Groesse** und **Gewicht**. **Geschlecht** bzw. **Rauchen** und **Alkoholg** sind die Namen der Vergleichsmerkmale, SAS gibt hier die Ausprägungen **weiblich** und **männlich** bzw. **ja** und **nein** aus.

N Obs ist die Anzahl der von SAS eingelesenen Beobachtungen.

Minimum = kleinster Wert.

	weiblich	männlich		
Merkmal Groesse:	156,00 (cm)	167,00 (cm)		
Merkmal Gewicht:	49,00 (kg)	60,00 (kg)		
	Rauchen ja Alkohol ja	Rauchen ja, Alkohol nein	Rauchen nein, Alkohol ja	Rauchen nein, Alkohol nein
Merkmal Alter:	24 (Jahre)	21 (Jahre)	21 (Jahre)	21 (Jahre)

Maximum = größter Wert.

	weiblich	männlich		
Merkmal Groesse:	184,00 (cm)	196,00 (cm)		
Merkmal Gewicht:	83,00 (kg)	130,00 (kg)		
	Rauchen ja Alkohol ja	Rauchen ja, Alkohol nein	Rauchen nein, Alkohol ja	Rauchen nein, Alkohol nein
Merkmal Alter:	27 (Jahre)	27 (Jahre)	27 (Jahre)	27 (Jahre)

Mean = Mittelwert (arithmetisches Mittel).

Der Mittelwert ist ein Lagemaß, das nur bei quantitativen Merkmalen erlaubt ist. Er wird im Gegensatz zum Median stark von Ausreißern beeinflusst und sollte deshalb nur bei normalverteilten Merkmalen berechnet werden.

	weiblich	männlich		
Merkmal Groesse:	169,66 (cm)	182,84 (cm)		
Merkmal Gewicht:	60,92 (kg)	79,65 (kg)		
	Rauchen ja Alkohol ja	Rauchen ja, Alkohol nein	Rauchen nein, Alkohol ja	Rauchen nein, Alkohol nein
Merkmal Alter:	25,75 (Jahre)	24,89 (Jahre)	24,71 (Jahre)	24,41 (Jahre)

Median = mittlerer Wert der sortierten Stichprobe.

Der empirische Median ist ein Lagemaß, das die nach ihrer Größe geordnete Stichprobe in 2 Hälften teilt. Bei ungeradem Stichprobenumfang ist der Median der mittlere Wert; bei geradem Stichprobenumfang der Durchschnitt aus den beiden mittleren Werten.

	weiblich	männlich		
Merkmal Groesse:	169,50 (cm)	183,00 (cm)		
Merkmal Gewicht:	60,00 (kg)	80,00 (kg)		
	Rauchen ja Alkohol ja	Rauchen ja, Alkohol nein	Rauchen nein, Alkohol ja	Rauchen nein, Alkohol nein
Merkmal Alter:	26 (Jahre)	25 (Jahre)	25 (Jahre)	24 (Jahre)

Std Dev = Standardabweichung.

Wurzel aus der Varianz. Bevorzugtes Streuungsmaß bei quantitativ stetigen, symmetrisch verteilten Daten. Die Standardabweichung darf nur bei quantitativen Merkmalen berechnet werden. Bei normalverteilten Daten liegen etwa 2/3 der Messwerte innerhalb des Intervalls $\mu \pm \sigma$ und etwa 95 % innerhalb des Intervalls $\mu \pm 2\sigma$.

	weiblich	männlich		
Beispiel Groesse:	7,11 (cm)	6,81 (cm)		
Beispiel Gewicht:	8,54 (kg)	13,57 (kg)		
	Rauchen ja Alkohol ja	Rauchen ja, Alkohol nein	Rauchen nein, Alkohol ja	Rauchen nein, Alkohol nein
Merkmal Alter:	1,5 (Jahre)	2,47 (Jahre)	2,57 (Jahre)	2,25 (Jahre)

Beschreibung nützlicher Optionen:

- **Option N:** SAS gibt die der Berechnung zugrunde liegende Anzahl aus:

```
TITLE1 'Statistische Maßzahlen' ;
TITLE2 '-----' ;
PROC MEANS DATA=beispiel N MIN MAX MEAN MEDIAN MAXDEC=2 ;
  VAR groesse gewicht ;
  CLASS rauchen ;
RUN ;
```

- ⇒ SAS gibt zusätzlich zu den Maßzahlen aus, wie viele Beobachtungen der Berechnung zugrunde liegen:

Statistische Masszahlen

The MEANS Procedure

Rauchen	N Obs	Variable	N	Minimum	Maximum	Mean	Median
ja	13	Groesse	13	157.00	193.00	173.54	172.00
		Gewicht	13	50.00	106.00	70.00	71.00
nein	58	Groesse	58	156.00	196.00	172.88	173.00
		Gewicht	57	49.00	88.00	65.04	63.00

Im Beispiel fehlt eine Gewichtsangabe.

- Option **MAXDEC=zahl**: SAS gibt statt sieben Nachkommastellen nur so viel aus wie **zahl** angibt, z.B. MAXDEC=2 ⇒ 2 Nachkommastellen:

Beispiel ohne Vorgabe für die Anzahl von Nachkommastellen:

```
TITLE1 'Statistische Maßzahlen' ;
TITLE2 '-----' ;
PROC MEANS DATA=beispiel ;
    VAR groesse gewicht ;
    CLASS rauchen ;
RUN ;
```

⇒ SAS gibt sieben Nachkommastellen aus:

Statistische Masszahlen

The MEANS Procedure

Rauchen	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
ja	13	Groesse	13	173. 5384615	9.2070207	157.0000000	193.0000000
		Gewicht	13	70.0000000	13.8383525	50.0000000	106.0000000
nein	58	Groesse	58	172. 8793103	8.8199884	156.0000000	196.0000000
		Gewicht	57	65.0350877	10.9999430	49.0000000	88.0000000

Beispiel mit Vorgabe für die Anzahl von Nachkommastellen:

```
TITLE1 'Statistische Maßzahlen' ;
TITLE2 '-----' ;
PROC MEANS DATA=beispiel MAXDEC=2 ;
    VAR groesse gewicht ;
    CLASS rauchen ;
RUN ;
```

⇒ SAS gibt nur noch zwei Nachkommastellen aus:

Statistische Masszahlen

The MEANS Procedure

Rauchen	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
ja	13	Groesse	13	173. 54	9.21	157.00	193.00
		Gewicht	13	70.00	13.84	50.00	106.00
nein	58	Groesse	58	172. 88	8.82	156.00	196.00
		Gewicht	57	65.04	11.00	49.00	88.00

- Weitere statistische Kenngrößen, die man als Optionen angeben kann:

VAR = Varianz: mittlere quadratische Abweichung der Stichprobenwerte vom Mittelwert.

RANGE = Spannweite: Maximum - Minimum.

CV = Variationskoeffizienten: Quotient von Standardabweichung und arithmetischem Mittel.

SKEWNESS = Schiefe: Verschiebung der Merkmalsausprägungen nach links oder rechts vom Mittelwert aus gesehen.

Q1 = 1. Quartil und **Q3** = 3. Quartil, **P10** = 10. Perzentil und **P90** = 90. Perzentil.

CLM berechnet die obere bzw. untere Grenze des 95%-Konfidenzintervall bei $\alpha=5\%$ für den Mittelwert. Innerhalb dieses Bereichs liegt der Erwartungswert der Grundgesamtheit mit 95%-iger Wahrscheinlichkeit.