# Häufigkeitsdiagramme

PROC UNIVARIATE
PROC GCHART

```
PROC GCHART <DATA=sasdatei>;
  VBAR variablenliste </Optionen>;
       /* vertikales Histogramm */
  HBAR var.list </Optionen>;
       /* horizontales Histogramm */
  PIE var.list </Optionen>; /* Kreisdiagr
  STAR var.list </Optionen>; /* Sterndiag
  BLOCK var.list </Optionen>;
       /* 3 dim. Balkendiagramm */
RUN;
```

# Häufigkeitsdiagramme

## Optionen (1)

VBAR3D, HBAR3D, PIE3D anstelle von VBAR, HBAR, PIE liefern schönere Bilder.

DISCRETE Zusammenfassung von Ausprägungen wird unterdrückt, d.h. für jeden Wert wird eine Säule erzeugt.

| | |
|---|---|
| LEVELS = anzahl | gewünschte Anzahl Säulen |
| TYPE = FREQ | Häufigkeiten (Standard) |
| = PERCENT | Prozente |
| = CFREQ | kum. Häufigkeiten |
| = CPERCENT | kum. Prozente |
| = SUM | Summen (nur mit SUMVAR) |
| SUMVAR = anzahl | Anzahl ist bereits aufsummierte Häufigkeit |

# Häufigkeitsdiagramme

## Optionen (2)

MIDPOINTS = Mittelpunkte der Balken.
Balken haben alle die gleiche Breite!

GROUP= Gruppierungsvariable

SUBGROUP= Gruppierungsvariable, gemeinsame Auswertung

PATTERNID=Musterzuordnung
Vergleiche die PATTERN-Anweisung

```
Descr_Gchart_1a.sas
Descr_Gchart_1b.sas
Descr_Gchart_3.sas    3a,3b
Descr_Gchart_1.sas
```

# Häufigkeitsdiagramme

## Design der Diagramme

**PATTERNxn C= V=**

| | |
|---|---|
| C, COLOR | Farbe: blue,cyan,red,black...<br>black ist Voreinstellung |
| V, VALUE | Wert: star,plus point,... |
| x | Muster:<br>$X_n$: schraffiert<br>$S_n$: Solid<br>$R_n$: ///<br>$L_n$: $\backslash\backslash\backslash$ |
| n | 1-5: Dichte des Musters. |

# Histogramme und Dichteschätzung



Auch Prozedur UNIVARIATE liefert Histogramme

```
PROC UNIVARIATE;
  HISTOGRAM varname </Optionen>;
RUN;
```

Sie liefert auch Tabellen von Histogrammen;

```
PROC UNIVARIATE;
  CLASS Klassenvariablen;
  HISTOGRAM varname </Optionen>;
RUN;
```

```
Descr_Plot_Kuehl.sas
Desc_ZweidimHisto_Heroin.sas
```

# Histogramme und Dichteschätzung
## Optionen



| | |
|---|---|
| CBARLINE= | Farbe des Histogramms |
| WBARLINE= | Dicke der Histogrammlinien |
| L= | Linientyp (Standard: 1, solid) |
| MIDPOINTS= | wie bei GPLOT |
| KERNEL | Nichtparametr. Dichteschätzung |
| COLOR= | Farbe der Dichtekurve |
| NORMAL | Parametrische Dichteschätzung (Normalverteilung) |
| GAMMA | Parametrische Dichteschätzung (Gammaverteilung) |

# Parametrische Dichteschätzung



Vorgabe: Modell, z.B. Normalverteilung oder Gammaverteilung

Lediglich die Parameter werden geschätzt.

```
PROC UNIVARIATE;
  HISTOGRAM varn/normal gamma;/*Parametri
  HISTOGRAM varn/kernel; /*Nichtparametri
RUN;
```

# Parametrische Dichteschätzung



Vorgabe: Modell, z.B. Normalverteilung oder Gammaverteilung
Lediglich die Parameter werden geschätzt.

```
PROC UNIVARIATE;
  HISTOGRAM varn/normal gamma;/*Parametri
  HISTOGRAM varn/kernel; /*Nichtparametri
RUN;
```

**Frage: Wie wird geschätzt?**

bei Normalverteilung ist das klar: $\overline{X}$ und $s^2$ sind optimale Schätzungen für $\mu$ und $\sigma^2$.

# Parametrische Dichteschätzung



Vorgabe: Modell, z.B. Normalverteilung oder Gammaverteilung
Lediglich die Parameter werden geschätzt.

```
PROC UNIVARIATE;
  HISTOGRAM varn/normal gamma;/*Parametri
  HISTOGRAM varn/kernel; /*Nichtparametri
RUN;
```

**Frage: Wie wird geschätzt?**

bei Normalverteilung ist das klar: $\overline{X}$ und $s^2$ sind optimale Schätzungen für $\mu$ und $\sigma^2$.

Wie findet man (gute) Schätzungen bei anderen Verteilungen?

# Schätzmethoden



## Momentenmethode

Man drückt den zu schätzenden Parameter durch die Momente, z.B. $\mathbf{E}(X)$, $\mathbf{E}(X^2)$, aus.
Dann werden die Momente durch die empirischen Momente, hier $\overline{X}$, $\frac{1}{n}\sum X_i^2$ ersetzt.

## Maximum-Likelihood-Schätzung

Es wird der Schätzwert für den unbekannten Parameter ermittelt, bei dem die Beobachtungen am meisten für diesen Parameter sprechen (most likely).

## Normalverteilung $\mathcal{N}(\mu, \sigma^2)$

$\overline{X}$ und $\hat{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2$ sind Momentenschätzungen für $\mu$ und $\sigma^2$. Sie sind auch ML-schätzungen für $\mu$ und $\sigma^2$.

# Maximum-Likelihood-Schätzung
## $X_i \sim \mathcal{N}(\mu, 1)$ unabhängig



Likelihood: $L_n := f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$, die gemeinsame Dichtefunktion der $X_i$.

$$\begin{aligned}
L_n(\mu) &= \prod_{i=1}^{n} f_{X_i}(x_i) \qquad \text{(Unabhängigkeit)} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2} \\
\ln L_n(\mu) &= -n \ln(\sqrt{2\pi}) + \sum_{i=1}^{n} \Big(-\frac{(x_i-\mu)^2}{2}\Big) \\
\frac{\partial L_n(\mu)}{\partial \mu} &= \sum_{i=1}^{n} (x_i - \mu)
\end{aligned}$$

Nullsetzen liefert: $\hat{\mu} = \overline{X}$.

$K(t)$ eine Kernfunktion,

$$\begin{aligned} \int K(t)\,dt &= 1, \quad \int tK(t)\,dt = 0, \\ \int t^2K(t)\,dt &= 1, \quad \int K^2(t)\,dt < \infty \end{aligned}$$

# Nichtparametrsiche Dichteschätzung

## Überlagerung der Daten mit einer (Dichte-) Funktion

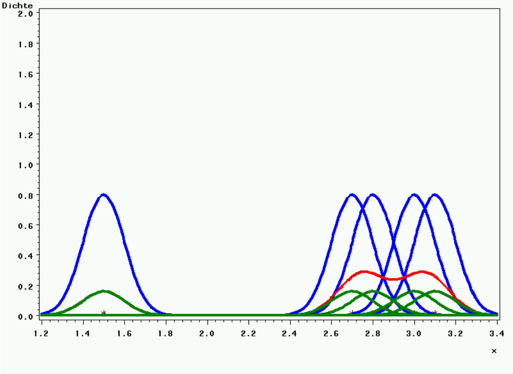
### $K(t)$ eine Kernfunktion,

$$\int K(t)\,dt = 1, \quad \int tK(t)\,dt = 0,$$
$$\int t^2K(t)\,dt = 1, \quad \int K^2(t)\,dt < \infty$$

### Dichteschätzung oder Dichtefunktionsschätzung.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)$$

$x_1, \ldots, x_n$: die Beobachtungen.
$h$: ein sogenannter Glättungsparameter.

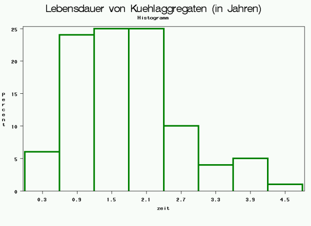


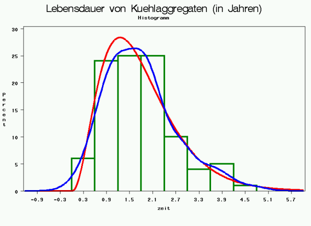

`Descr_Dichteschaetzung.sas`

# Dichteschätzung, Beispiel



## Kühlaggregate




Histogramm

Parametrische Dichteschätzung (Gamma)

Nichtparametrische Dichteschätzung