Beschreibende Statistik

Beispiel
Treibstoffverbrauch von Neuwagen
mit Gewicht unter 1200 kg in l/100 km:

Rang = 1

$i =$	1	2	3	4	5	6	7	8	9	10
$x_i =$	7.0	10.8	9.8	7.6	8.8	9.9	8.3	9.4	5.4	6.8
$i =$	11	12	13	14	15	16	17	18	19	20
$x_i =$	9.4	8.0	9.1	8.2	7.5	8.2	10.2	7.7	6.4	9.3
$i =$	21	22	23	24	25	26	27	28	29	30
$x_i =$	8.7	10.0	8.0	7.8	5.9	6.6	9.2	9.6	8.4	9.1
$i =$	31	32	33	34	35	36	37 = n			
$x_i =$	7.3	8.0	10.4	6.8	7.3	10.5	8.9			

Rang = 5.5

Rang = 2

Stichproben-grösse

aus: Autobeilage des Tages-Anzeigers vom 6.3.89
(zitiert nach W. Stahel: Statistische Datenanalyse)

nach Größe sortieren $\longrightarrow$ geordnete Werte:

$i =$	1	2	3	4	5	6	7	8	9	10
$x_{(i)} =$	5.4	5.9	6.4	6.6	6.8	6.8	7.0	7.3	7.3	7.5
$i =$	11	12	13	14	15	16	17	18	19	20
$x_{(i)} =$	7.6	7.7	7.8	8.0	8.0	8.0	8.2	8.2	8.3	8.4
$i =$	21	22	23	24	25	26	27	28	29	30
$x_{(i)} =$	8.7	8.8	8.9	9.1	9.1	9.2	9.3	9.4	9.4	9.6
$i =$	31	32	33	34	35	36	37			
$x_{(i)} =$	9.8	9.9	10.0	10.2	10.4	10.5	10.8			

$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ geordnete Stichprobe

Stichprobengrösse

Messung der freigesetzten Wärme beim Übergang von Eis zu Wasser:

Rang 2 Rang 1 n =

Beobachtung	1	2	3	4	5	6	7	8	9	10	11	12	13
Methode A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03	80.02	80.00	80.02
Methode B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97					

Methode A der Grösse nach geordnet: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

i	1	2	3	4	5	6	7	8	9	10	11	12	13
$x_{(i)}$	79.97	79.98	80.00	80.02	80.02	80.02	80.03	80.03	80.03	80.04	80.04	80.04	80.05

Rang 5

90%-Quantil:

$0.9 \cdot 13 = 11.7$

7.1. Kennzahlen

arithmetisches Mittel:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$ "Lagemass"

empirische Varianz / Standardabweichung:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$s = \sqrt{s^2}$$ "Streuungsmass"

geordnete Werte

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

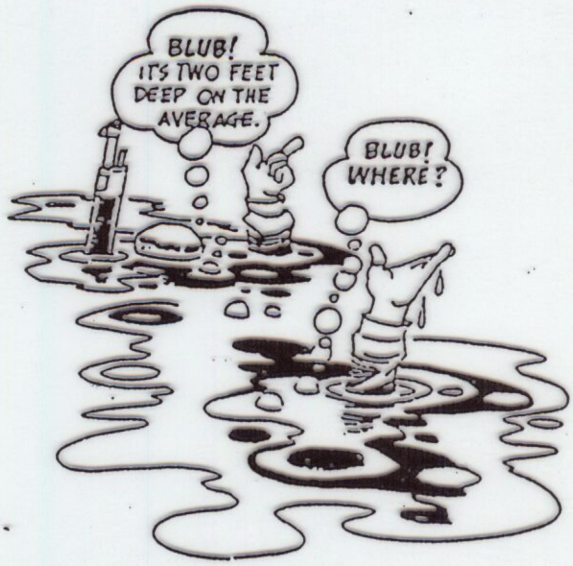
empirisches α-Quantil ($0 < \alpha < 1$):

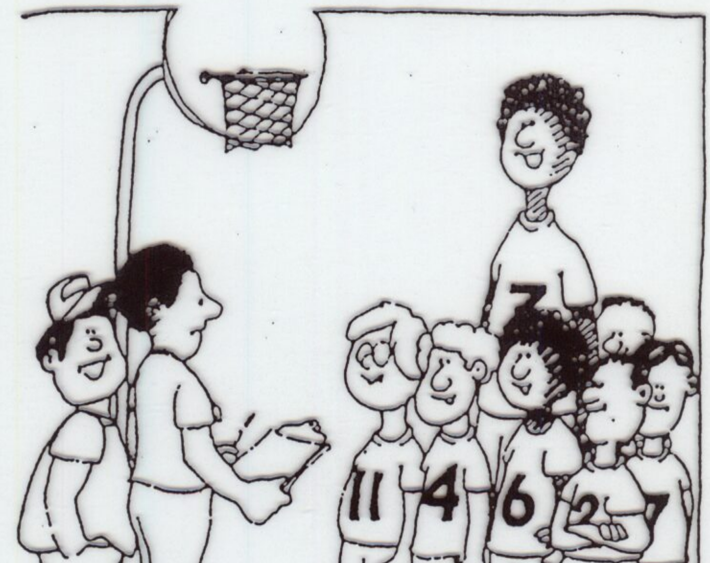
$x_{(k)}$ wobei $k > \alpha n$ und $k \in \mathbb{N}$ minimal

$\rightsquigarrow$ approximativ $\alpha \cdot 100\,\%$ Daten $\leq$ emp. α-Quantil

$(1-\alpha) \cdot 100\,\%$ " $>$ " "

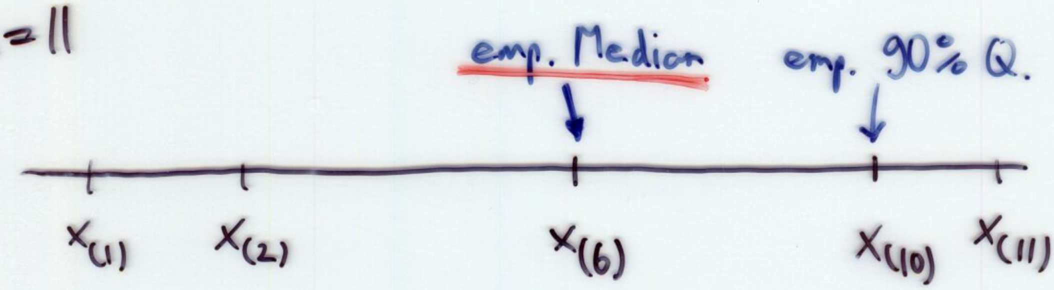
falls $\alpha n \in \mathbb{N}$: emp. α-Quantil $= \frac{1}{2}\left(x_{(\alpha n)} + x_{(\alpha n+1)}\right)$





»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

$n = 11$



emp. 90% Quantil: $0.9 \cdot 11 = 9.9$

$\longrightarrow x_{(10)}$

emp. 50% Quantil:

empirischer Median "Lagemass"

"mittlere Beobachtung"

unteres Quartil: emp. 25% Quantil

oberes " : " 75% "

Quartilsdifferenz = oberes Quartil − unteres Quartil

"Streuungsmass"

Bem: emp. Median ist robustes Lagemass

7.2 Grafische Darstellungen

empirische kumulative Vert.-funktion

$$\hat{F}(x) = \frac{1}{n} \overbrace{\# \{ i ; x_i \le x \}}^{\text{Anzahl}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[x_i \le x]}$$

$$\mathbb{1}_{[x_i \le x]} = \begin{cases} 1 & \text{falls } x_i \le x \\ 0 & \text{sonst} \end{cases}$$

- $\hat{F}(\cdot)$ ist eine kum. Vert.fkt.
- springt an den Stellen x_i um $\frac{1}{n}$ (oder Vielfaches davon wenn mehrere gleiche Daten)

Histogramm:

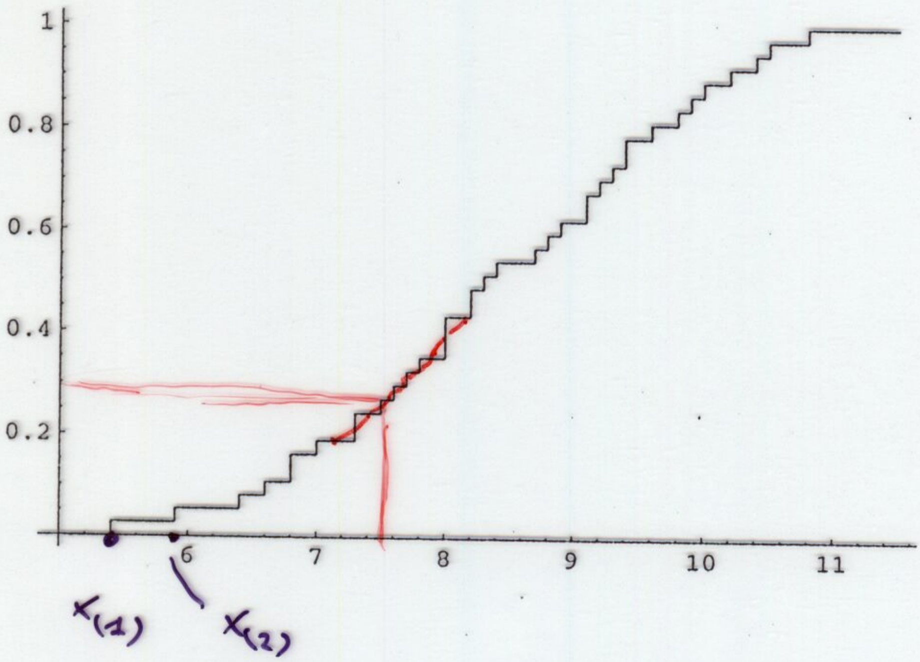
"Fläche proportional zu Häufigkeit"

Empirische kumulative Verteilungsfunktion

Die *empirische* kumulative Verteilungsfunktion ist jene Treppenfunktion, die bei $x_{(i)}$ von $(i-1)/n$ auf i/n springt $(i = 1, \ldots, n)$:

$$\widehat{F}(x) = \frac{1}{n} \#\{i : x_{(i)} \leq x\}$$

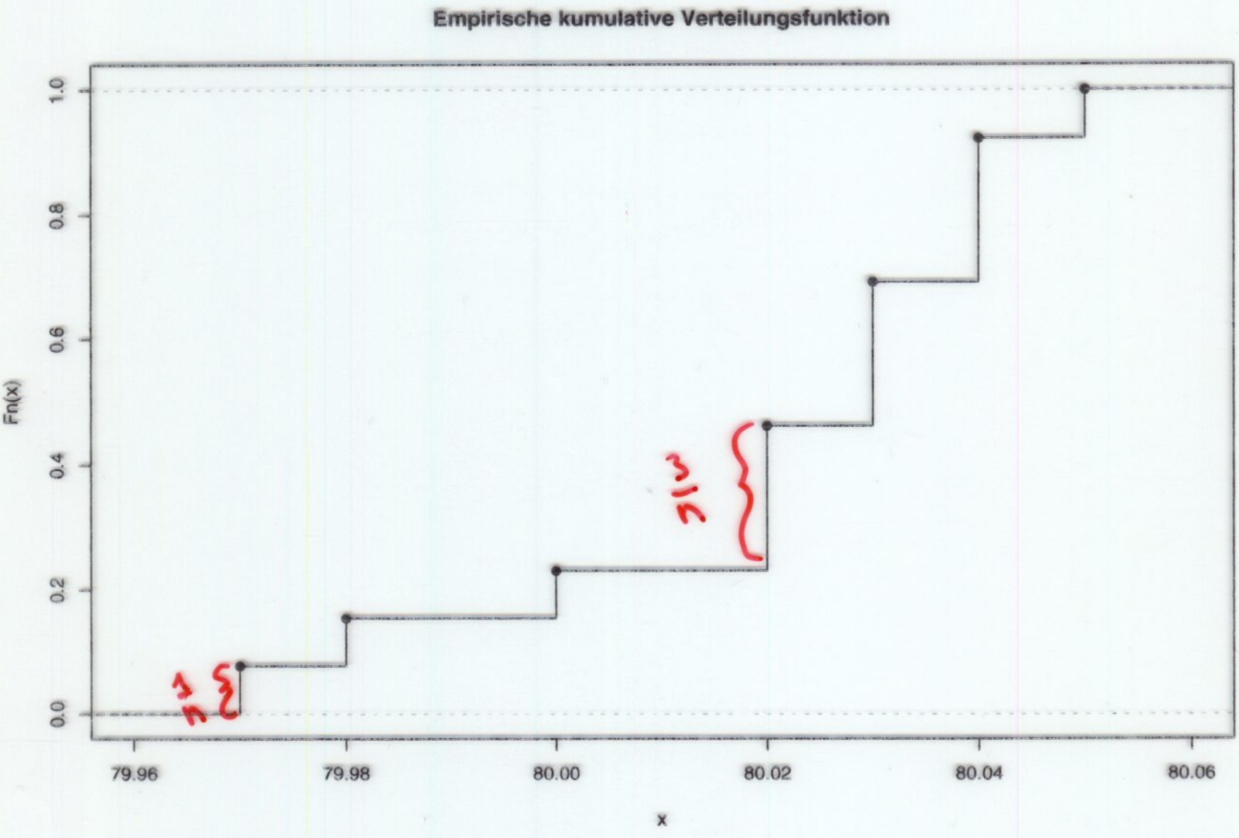
Treibstoffverbrauch



Empirische kumulative Verteilungsfunktion

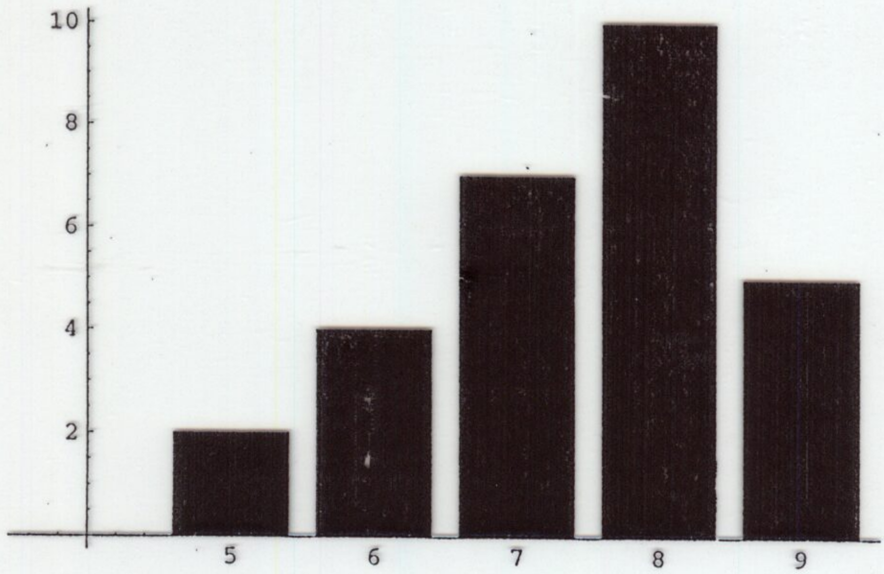
Die *empirische* kumulative Verteilungsfunktion ist jene Treppenfunktion, die bei $x_{(i)}$ von $(i-1)/n$ auf i/n springt $(i = 1, \ldots, n)$:

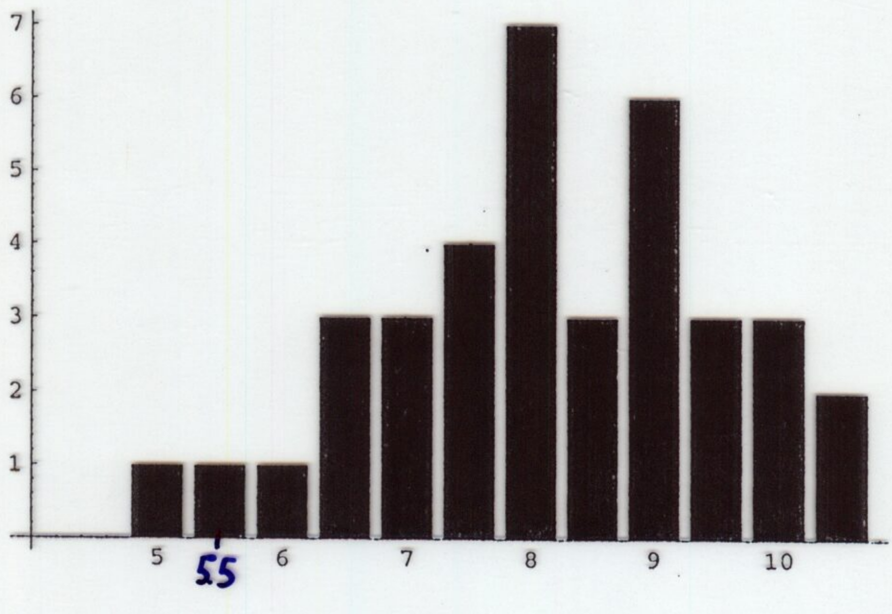
$$\hat{F}(x) = \frac{1}{n} \#\{i : x_{(i)} \leq x\}$$

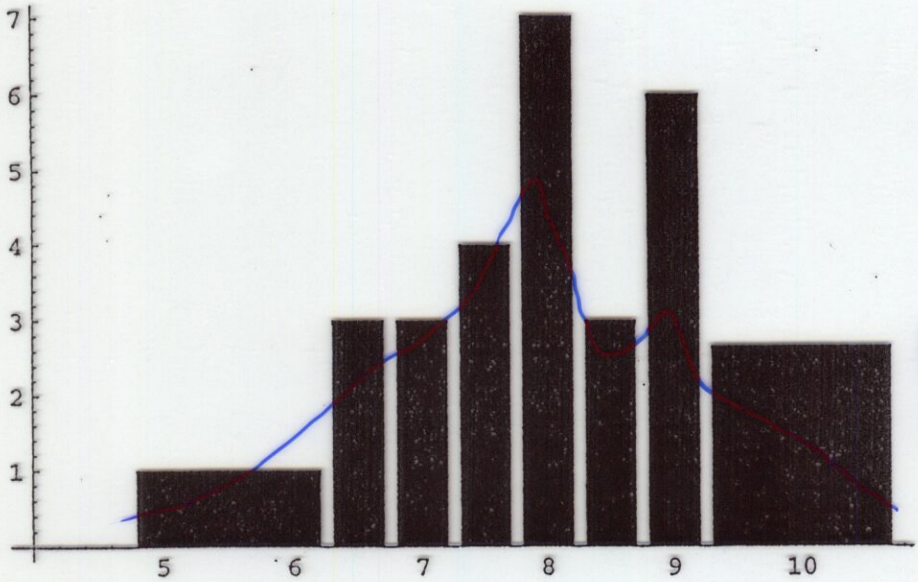


Histogramme

(Säulen- oder Balkendarstellung)







Häufigkeit $\propto$ Fläche

Die Balkenhöhe ist proportional zu Häufigkeit / Klassenlänge.

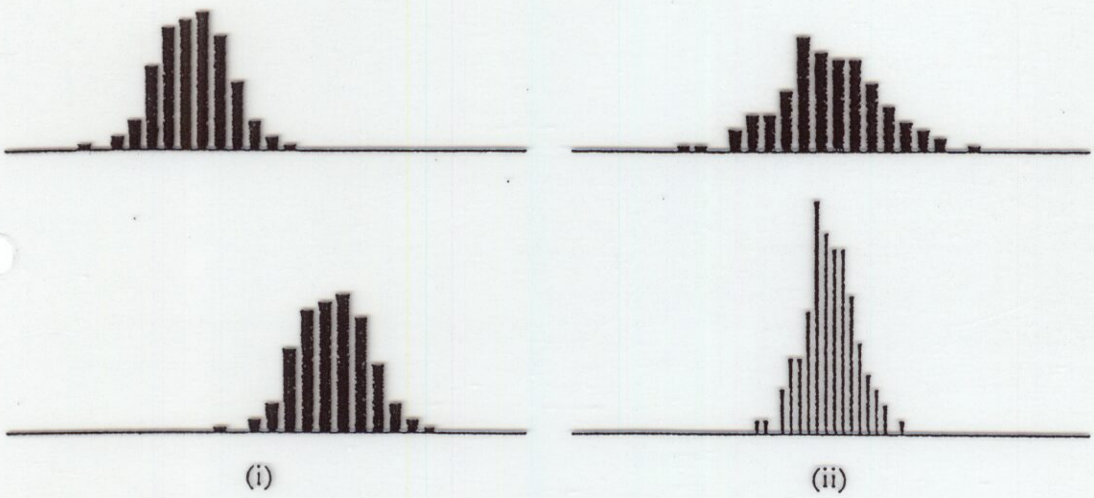
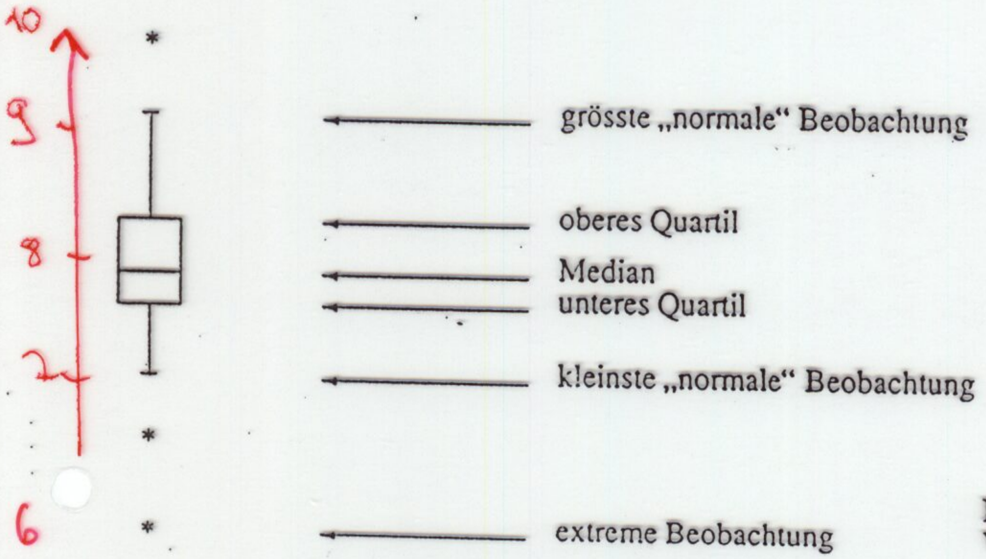


Bild 2.3.a Histogramme mit unterschiedlicher Lage (i) und mit unterschiedlicher Streuung (ii)

(Stahel)

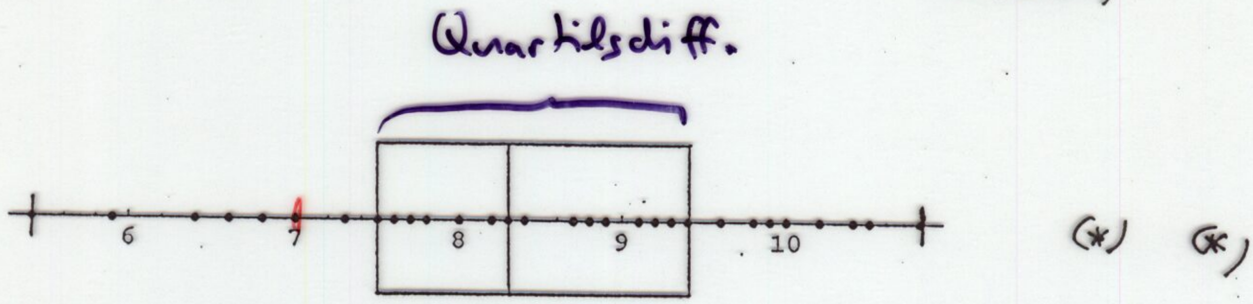
Boxplot



höchstens
1.5 · Quartilsdifferenz
von oberem Quartil
entfernt

Bild 2.5.d (i)
Was der Box-Plot zeigt

(Stahel)



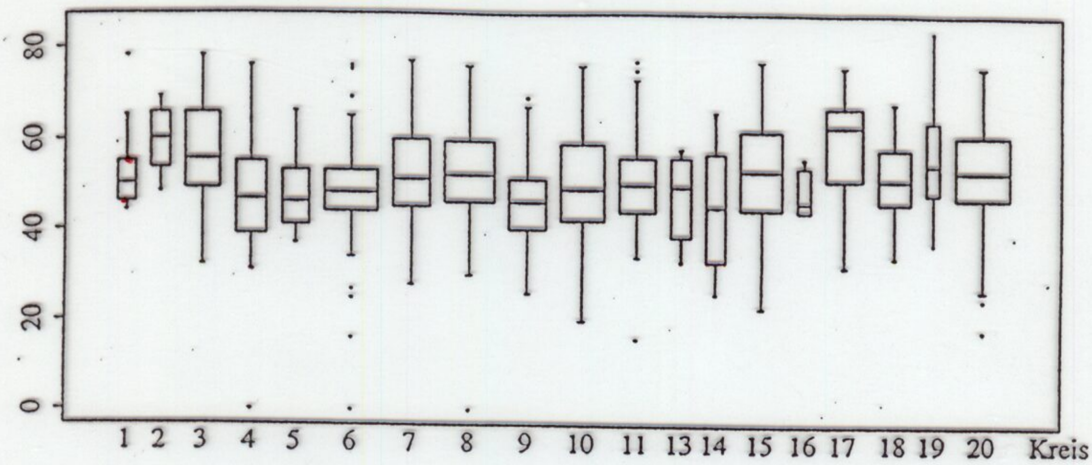


Bild 2.5.d (ii) Box-Plots zum Vergleich von Gruppen: NO_2-Wochenmittel für die 20 Basler Stadtkreise

(Stahel)

Boxplot

- Kennzahlen für Lage und Streuung ablesbar
- Multimodalität (mehrere peaks beim Histogramm) nicht erkenntlich
- geeignet um versch. Gruppen zu vergleichen