

Mathematik II für Biologen  
Beschreibende Statistik – Eindimensionale Daten

Stefan Keppeler

16. April 2008

## Stichproben

Geordnete Stichprobe – Rang

## Kennzahlen

Maße für die mittlere Lage der Daten

Robustheit

Quantile

Maße für die Streuung der Daten

## Ausreißer

Erkennung potentieller Ausreißer

## Graphische Darstellung

Eindimensionales Streudiagramm – Dotplot

Stamm- und Blattdiagramm

Histogramm

Boxplot

Empirische (kumulative) Verteilungsfunktion



**Stichprobe:**  $x_1, x_2, \dots, x_n$

- ▶ Daten
- ▶ Messergebnisse
- ▶ Ansammlung von Zahlen

Stichprobenumfang:  $n$

Historisches **Beispiel:** (1905)

Schlafverlängerung durch Medikament B gegenüber Medikament A

- ▶  $x_i$  = Schlafverlängerung bei Testperson  $i$  (in h),  $n = 10$

1,2 2,4 1,3 1,3 0,0 1,0 1,8 0,8 4,6 1,4

- ▶ also  $x_1 = 1,2$ ,  $x_4 = 1,3$  etc.
- ▶ i.A. nicht geordnet

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

- ▶  $x_{(k)}$  =  $k$ ter Wert in der geordneten Stichprobe
- ▶  $k$  heißt Rang

Im obigen **Beispiel**:

Rang $k$	1	2	3	4	5	6	7	8	9	10
$x_{(k)}$	0,0	0,8	1,0	1,2	1,3	1,3	1,4	1,8	2,4	4,6

- ▶ Der Rang von 2,4 ist 9.
- ▶ Der Rang von 1,3 ist 5,5 (oder: 5 und 6).

- **Durchschnitt** (Mittelwert, arithmetisches Mittel)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

im Beispiel:  $\bar{x} = \frac{1}{10} (1,2 + 2,4 + \dots + 1,4) = 1,58$

- **Median**  $\text{med}(x_1, \dots, x_n) = \text{med}$

$$\text{med} = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2} \left( x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right) & \text{falls } n \text{ gerade} \end{cases}$$

also  $\#\{x_i : x_i < \text{med}\} = \#\{x_i : x_i > \text{med}\}$

im Beispiel:  $\text{med} = \frac{1}{2} (x_{(5)} + x_{(6)}) = \frac{1}{2} (1,3 + 1,3) = 1,3$

## Vergleich von $\bar{x}$ und $med$ :

- ▶ Falls 4,6 durch 460 ersetzt wird (“Kommafehler”), ändert sich  $\bar{x}$  drastisch!); dagegen bleibt  $med$  unverändert.
- ▶ Der Median  $med$  ist robuster als  $\bar{x}$ .

## Verallgemeinerung des Medians:

Sei  $0 < \alpha < 1$ . Das  $\alpha$ -Quantil,  $q_\alpha$  teilt die Stichprobe (ungefähr) im Verhältnis  $\alpha$  zu  $1 - \alpha$ , d.h.

$$\frac{\#\{x_i : x_i < q_\alpha\}}{n} \approx \alpha$$

Genauer:

$$q_\alpha = \begin{cases} x_{(k)} & \text{mit } k = \alpha n + \frac{1}{2}, \text{ gerundet, falls } \alpha n \notin \mathbb{Z} \\ \frac{1}{2}(x_{\alpha n} + x_{\alpha n+1}), & \text{falls } \alpha n \in \mathbb{Z} \end{cases}$$

- ▶ Median = 0,5-Quantil:  $\text{med} = q_{1/2}$
- ▶ unteres Quartil = 0,25-Quantil:  $q_{0,25}$
- ▶ oberes Quartil = 0,75-Quantil:  $q_{0,75}$

im **Beispiel**:  $q_{0,25} = x_{(3)} = 1,0$  und  $q_{0,75} = x_{(8)} = 1,8$



## (empirische) Varianz

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

## (empirische) Standardabweichung: $s = s_x := \sqrt{s^2}$

im **Beispiel**:  $s^2 = \frac{1}{9} \left( (1,2 - 1,58)^2 + \dots + (1,4 - 1,58)^2 \right) \approx 1,51$   
 $s \approx 1,23$

Oft (nicht immer) gilt (Faustregel):

- ▶ Ungefähr  $2/3$  der Daten liegen zwischen  $\bar{x} - s_x$  und  $\bar{x} + s_x$
- ▶ Abweichungen von  $\bar{x}$  um bis zu  $2s_x$  sind durchaus möglich.  
(ca. 95% der Daten zwischen  $\bar{x} \pm 2s_x$ )
- ▶ Abweichungen der Daten um mehr als  $3s_x$  ( $4s_x$ ) treten selten (fast nie) auf.





Weitere Streumaße neben  $s_x$

- ▶ **Quartilsdifferenz:**  $q_{0,75} - q_{0,25}$   
im Beispiel:  $1,8 - 1,0 = 0,8$

- ▶ **Medianabweichung:** (median absolute deviation)

$$\text{MAD} = \text{med}\left(|x_1 - \text{med}(x_1, \dots, x_n)|, \dots, |x_n - \text{med}(x_1, \dots, x_n)|\right)$$

sehr robust

im Beispiel:  $\text{MAD} = 0,4$

**Ausreißer:** “verdächtig große/kleine Werte”

mögliche Gründe:

- ▶ Fehler (Mess-, Abschreib-, Versuchs-, ...)
- ▶ falsche Erwartungen (falsches Modell)
- ▶ seltenes Ereignis beobachtet

## Methoden zur Erkennung potentieller Ausreißer:

- ▶ populär, wenige robust:

$x_i$  ist Ausreißer, falls  $|x_i - \bar{x}| > 3s_x$  (oder  $> 4s_x$ )

besser:

- ▶ Falls es  $x_i$  mit  $|x_i - \bar{x}| > 3s_x$  gibt, so entferne das  $x_i$  mit dem größten  $|x_i - \bar{x}|$ .
  - ▶ Berechne  $\bar{x}$  und  $s_x$  neu.
  - ▶ Wiederhole bis alle Werte im  $3s_x$ -Intervall liegen.
  - ▶ Entfernte Werte sind mögliche Ausreißer.
- ▶ empfehlenswert, da robust:  
 $x_i$  ist Ausreißer, falls  $|x_i - \text{med}| > 5 \text{MAD}$

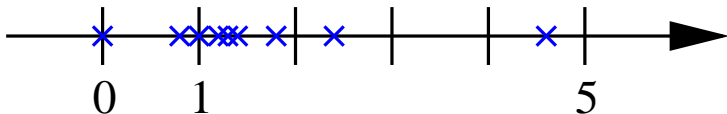
im **Beispiel**:

$\bar{x} \pm 3s_x$ :  $[-2,1, 5,3] \rightsquigarrow$  keine Ausreißer

$\text{med} \pm 5 \text{MAD}$ :  $[-0,7, 3,3] \rightsquigarrow x_9 = 4,6$  möglicher Ausreißer



## Eindimensionales Streudiagramm für unser Beispiel



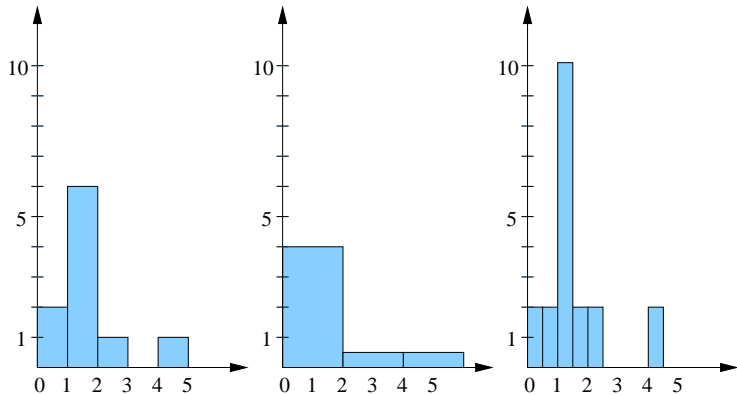
Zerlegung von  $x_i$  in Stamm- und Blattanteil, z.B.

- ▶ 1,3 in Stamm 1 und Blatt 3 und  
1,8 in Stamm 1 und Blatt 8
- ▶ oder  
1,3 in Stamm 1 und Blatt 3 und  
1,8 in Stamm 1+ und Blatt 3
- ▶ etc.

Stamm	Blätter
0	0 8
1	2 3 3 0 8 4
2	4
3	
4	6

Stamm	Blätter
0	0
0+	3
1	2 3 3 0 4
1+	3
2	4
2+	
3	
3+	
4	6

## Histogramme (“Drehe Stamm- und Blatttdiagramm”) für Beispiel



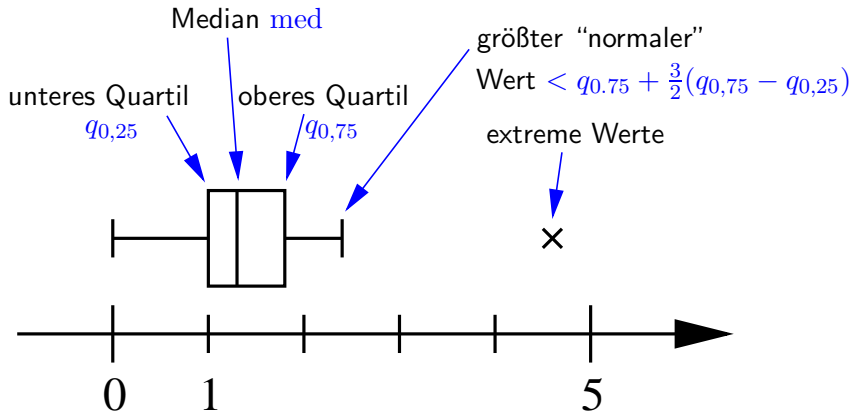
Klassenbreite: 1

2

0,5

Fläche ist proportional zur Häufigkeit, nicht die Höhe!

Boxplot für unser Beispiel:



empirische kumulative Verteilungsfunktion  $F : \mathbb{R} \rightarrow [0, 1]$

$$F(x) = \frac{\#\{x_i : x_i \leq x\}}{n}$$

Stufe der Höhe  $\frac{1}{n}$  bei jedem Wert.

im Beispiel  $\rightarrow$

(senkrechte Linien gehören streng  
genommen nicht mit dazu)

