Mathematik II für Biologen

Beschreibende Statistik – Eindimensionale Daten

Stefan Keppeler

16. April 2010

Prolog

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 Streudigramm – Dotplot
- Stamm- und Blattdiagramm
- Histogramm
- Boxplot
- Empirische (kumulative) Verteilungsfunktion

SPIEGEL ONLINE WISSENSCHAFT

NACHRICHTEN | VIDEO | THEMEN | FORUM | ENGLISH | DER SPIEGEL | SPIEGEL TV | ABO | SHOP

Home | Politik | Wirtschaft | Panorama | Sport | Kultur | Netzwelt | Wissenschaft | einestages | UniSPIEGEL | SchulSPIEGEL | Reise | Auto

Nachrichten > Wissenschaft > Mensch > Psychologie

Login | Registrierung

THEMA

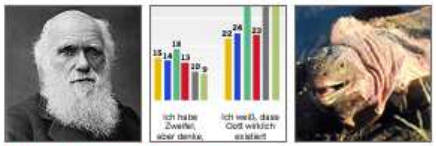
Psychologie

Alle Artikel und Hintergründe

01.03.2010

Drucken | Senden | Feedback | Merken

FOTOSTRECKE



Darwins Evolutionslehre: Wie die Natur das Leben lenkt

QUIZ

CENSEAM / Parkinson

Quiz: Die gemeinsten Streiche der Natur

Evolutionsquiz: Sind Sie Darwinist?

Tierischer Stimmentest: Roooo- roooooo-roooooooooaaaah!

MEHR AUF SPIEGEL ONLINE

Religion versus Evolution: Wie die Sünde in die Welt

Intelligenz und Evolution

Konservative haben geringeren IQ

dpa

Je intelligenter Menschen sind, umso eher sind sie bereit, sich auf Neues einzulassen. Konservative und religiöse Menschen haben hingegen einen geringeren Intelligenzquotienten. Psychologen glauben, dass man das Phänomen evolutionsbiologisch erklären kann.

http://www.spiegel.de/wissenschaft/mensch/0,1518,680956,00.html

"Die Gruppe der Nichtreligiösen hatte mit 103 den höchsten Intelligenzquotienten, die Strenggläubigen kamen auf einen mittleren IQ von 97 - das ist ein minimaler, aber nachweisbarer Unterschied. Ein IQ von 100 entspricht der durchschnittlichen Intelligenz der gesamten Bevölkerung.

"In der National Longitudinal Study of Adolescent Health, deren Daten die Londoner Forscher nutzten, wurde auch nach der politischen Überzeugung der Jugendlichen gefragt. Jene, die sich als 'very liberal' einstuften, was im Deutschen einer linken und linksliberalen Haltung entspricht, erreichten einen IQ von 106. Wer sich als 'sehr konservativ' charakterisierte, hatte hingegen nur einen IQ von 95, schreiben die Forscher im Fachblatt *Social Psychology Quarterly*."

Was heißt "nachweisbar"?

dx.doi.org/10.1177/0190272510361602

Social Psychology Quarterly
Vol. 73, No. 1, 33–57

DOI: 10.1177/0190272510361602
http://spq.sagepub.com

Why Liberals and Atheists Are More Intelligent

SATOSHI KANAZAWA
London School of Economics and Political Science

"(...) converted to the IQ metric, with a mean of 100 and a standard deviation of 15."

Mittelwert
Standardabweichung

"The differences in mean adolescent intelligence by adult political ideology is highly statistically significant ($F_{(4,13053)} = 83.6327$, $p < .00001$)."

statistisch signifikant

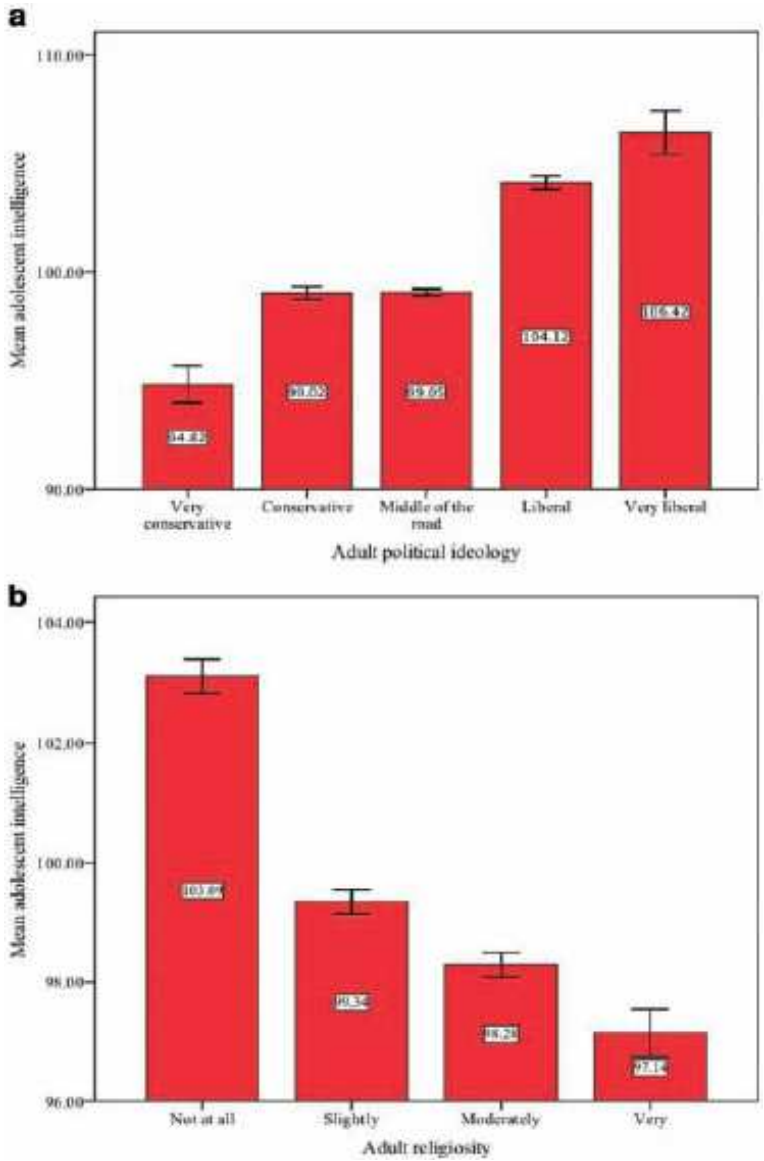


Figure 1. Mean Adolescent Intelligence by Political Ideology and Religiosity Add Health Data, Wave III (2001-2002). Error bars indicate standard error of the mean.

Standardabweichung
des Mittelwerts?

Stichprobe: $x_1, x_2, \ldots, x_n$

- 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), $n = 10$

$$1{,}2 \quad 2{,}4 \quad 1{,}3 \quad 1{,}3 \quad 0{,}0 \quad 1{,}0 \quad 1{,}8 \quad 0{,}8 \quad 4{,}6 \quad 1{,}4$$

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

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

- $x_{(k)} = k$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)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$$

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

- Median $\mathrm{med}(x_1, \ldots, x_n) = \mathrm{med}$

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

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

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

Vergleich von $\overline{x}$ und med:

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

Verallgemeinerung des Medians:

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

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

Genauer:

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

- Median $=$ $0{,}5$-Quantil: $\mathrm{med} = q_{1/2}$
- 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

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

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

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

$s \approx 1{,}23$

Oft (nicht immer) gilt (Faustregel):

- Ungefähr $2/3$ der Daten liegen zwischen $\overline{x} - s_x$ und $\overline{x} + s_x$
- Abweichungen von $\overline{x}$ um bis zu $2s_x$ sind durchaus möglich. (ca. 95% der Daten zwischen $\overline{x} \pm 2s_x$)
- 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)

  $$\mathrm{MAD} = \mathrm{med}\Big(\left|x_1 - \mathrm{med}(x_1,\ldots,x_n)\right|, \ldots, \left|x_n - \mathrm{med}(x_1,\ldots,x_n)\right| \Big)$$

  sehr robust

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

Ausreißer: “verdächtig große/kleine Werte”

mögliche Gründe:

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

Methoden zur Erkennung potentieller Ausreißer:

- poplär, wenig robust:
  x_i ist Ausreißer, falls $|x_i - \overline{x}| > 3s_x$ (oder $> 4s_x$)

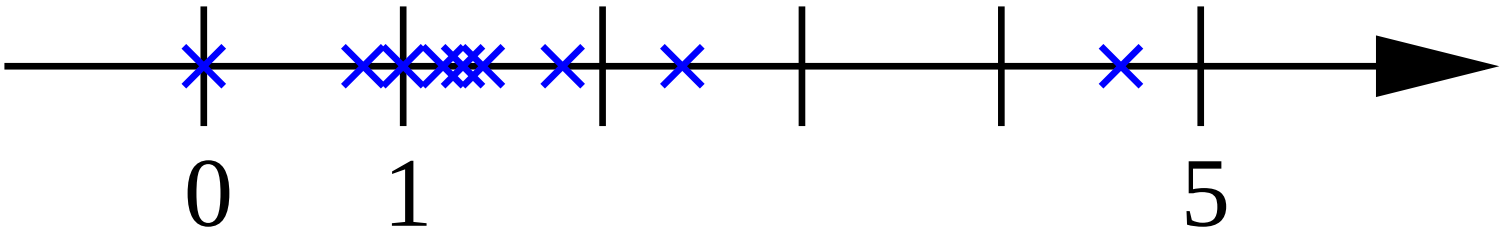
  besser:
  - Falls es x_i mit $|x_i - \overline{x}| > 3s_x$ gibt, so entferne das x_i mit dem größten $|x_i - \overline{x}|$.
  - Berechne $\overline{x}$ und s_x neu.
  - Wiederhole bis alle Werte im $3s_x$-Intervall liegen.
  - Entfernte Werte sind mögliche Ausreißer.
- empfehlenswert, da robust:
  x_i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

im **Beispiel:**

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

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

Eindimensionales Streudiagramm für unser Beispiel



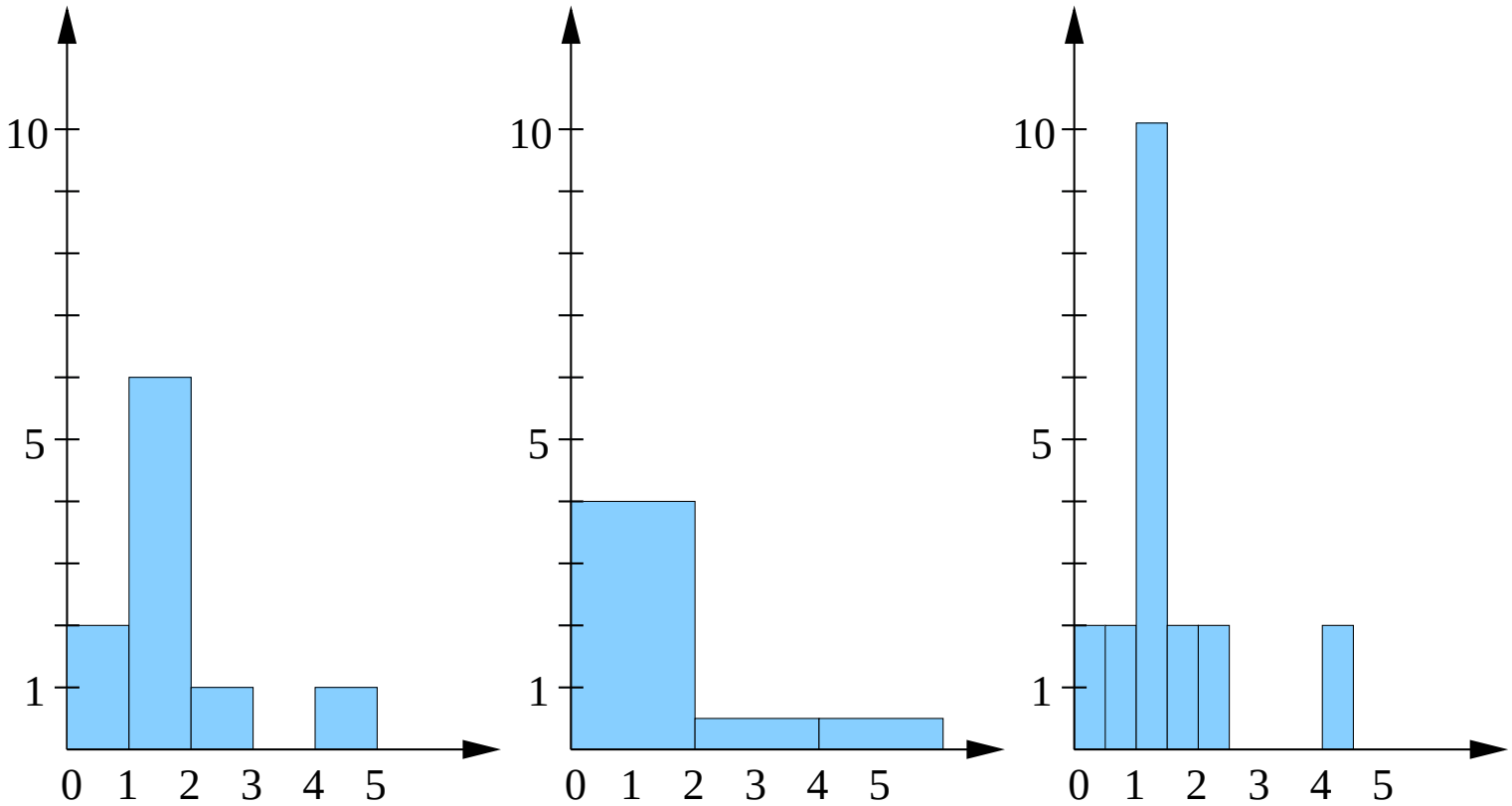
Zerlegung von x_i in Stamm- und Blattanteil, z.B.

- 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

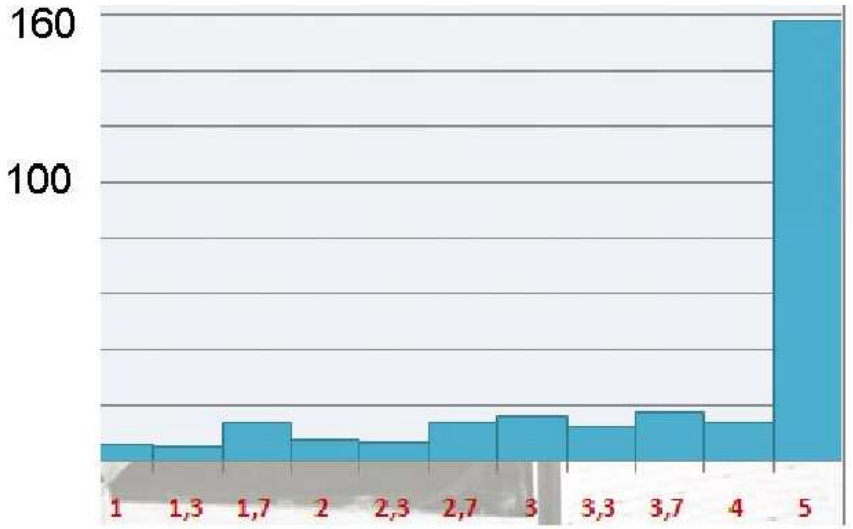
Histogramme ("Drehe Stamm- und Blattdiagramm") für Beispiel

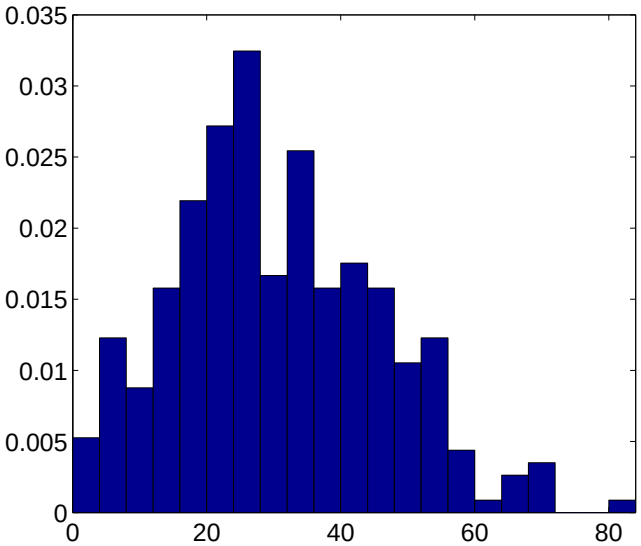


Klassenbreite: 1 2 0,5

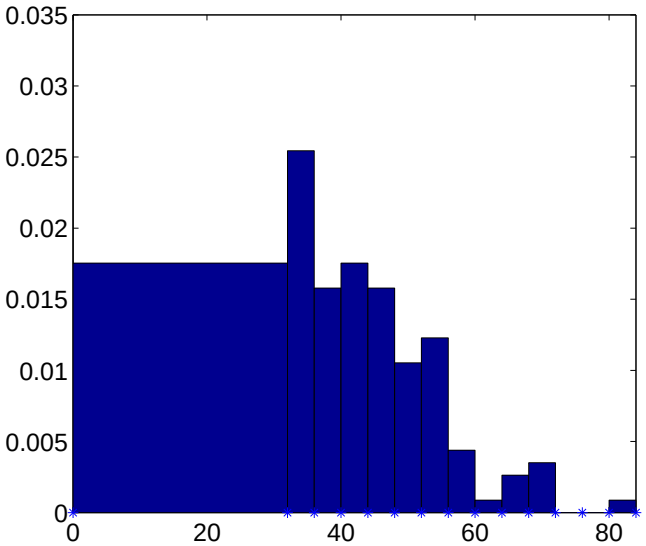
Fläche ist poportional zur Häufigkeit, nicht die Höhe!

aus dem Forum (WS 09/10)

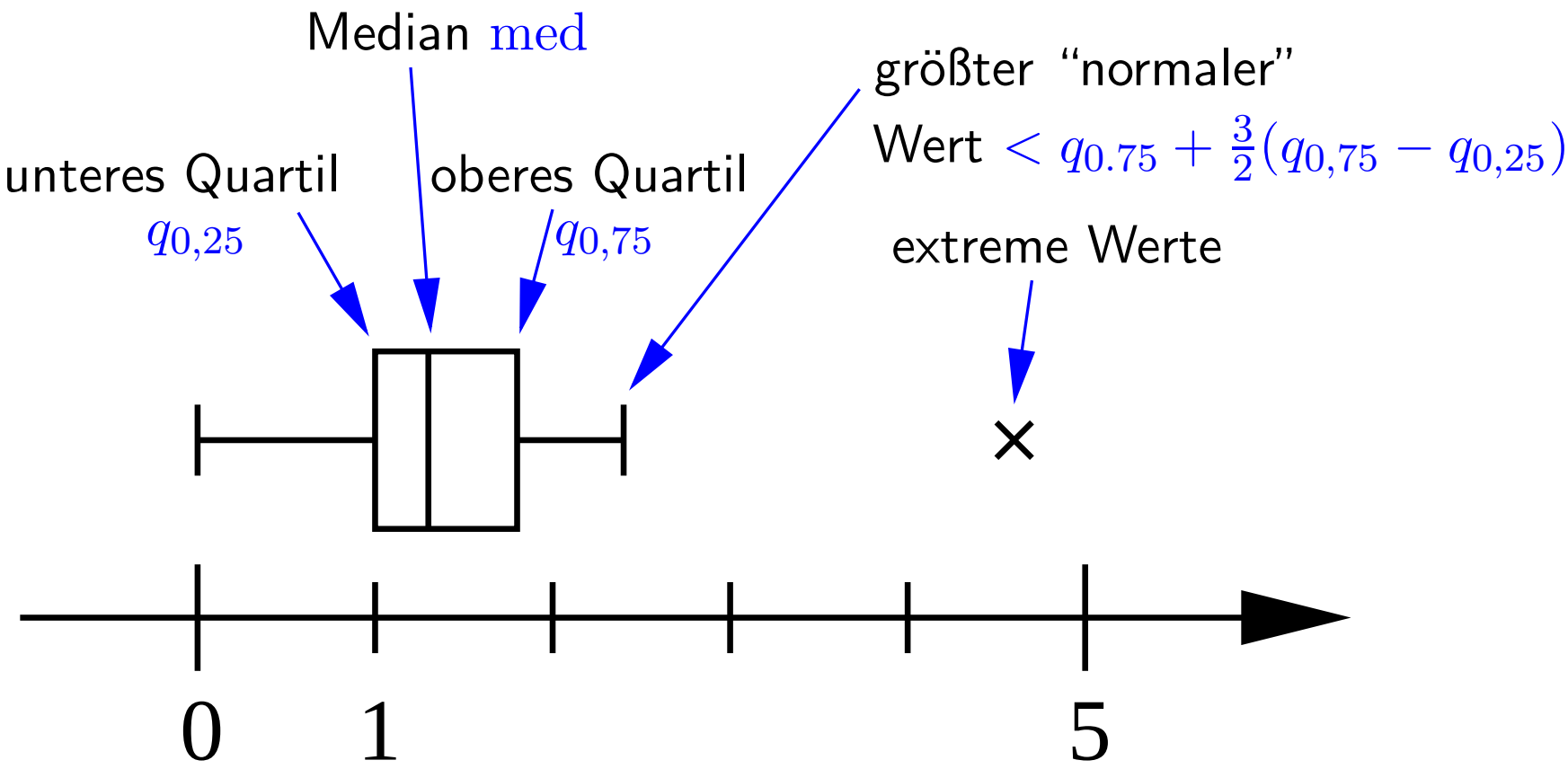


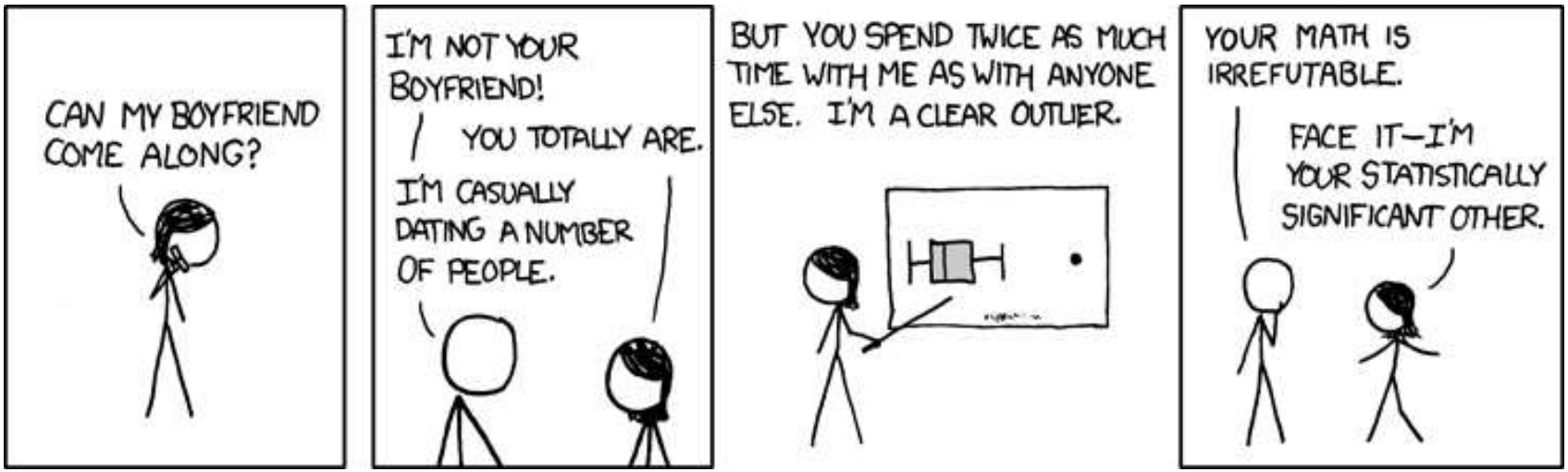


Alternativen



Boxplot für unser Beispiel:





http://xkcd.com/539

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

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

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

im Beispiel $\longrightarrow$

(senkrechte Linien gehören streng genommen nicht mit dazu)