Mathematik II für Biologen

Beschreibende Statistik – Eindimensionale Daten

Stefan Keppeler

24. April 2009

Eberhard Karls
Universität
Tübingen

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

Und außerdem...

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)

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

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

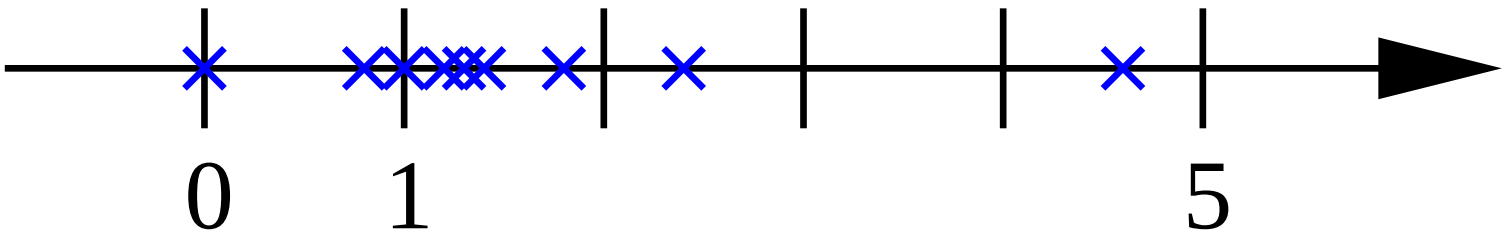
  besser:
  - Falls es x_i mit $|x_i - \overline{x}| > 3s_x$ gibt, so entferne das x_i mit dem größten $|x_i - \overline{x}|$.
  - Berechne $\overline{x}$ und s_x neu.
  - Wiederhole bis alle Werte im $3s_x$-Intervall liegen.
  - Entfernte Werte sind mögliche Ausreißer.
- empfehlenswert, da robust:
  x_i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

im **Beispiel:**

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

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

Eindimensionales Streudiagramm für unser Beispiel



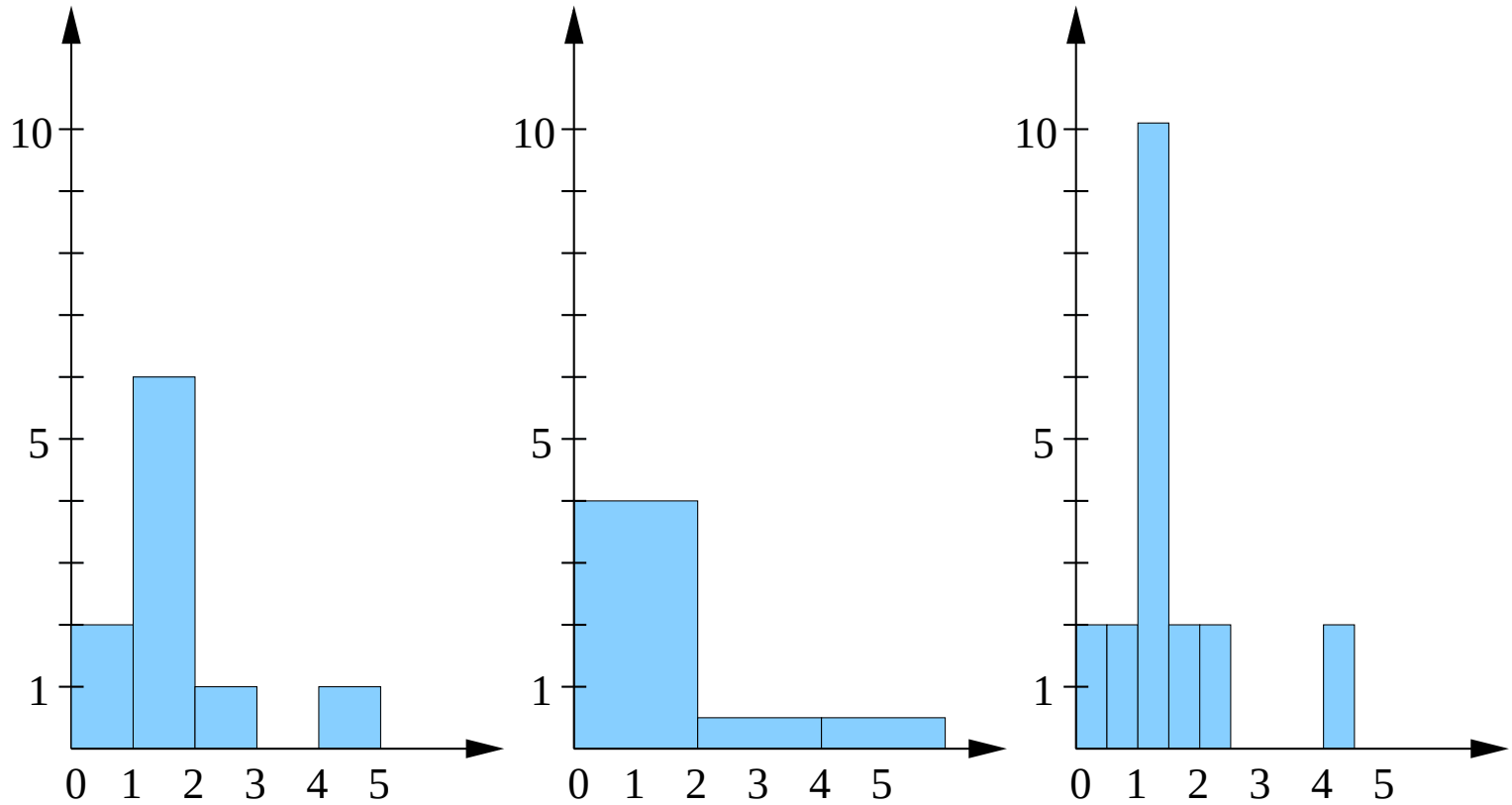
Zerlegung von x_i in Stamm- und Blattanteil, z.B.

- 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

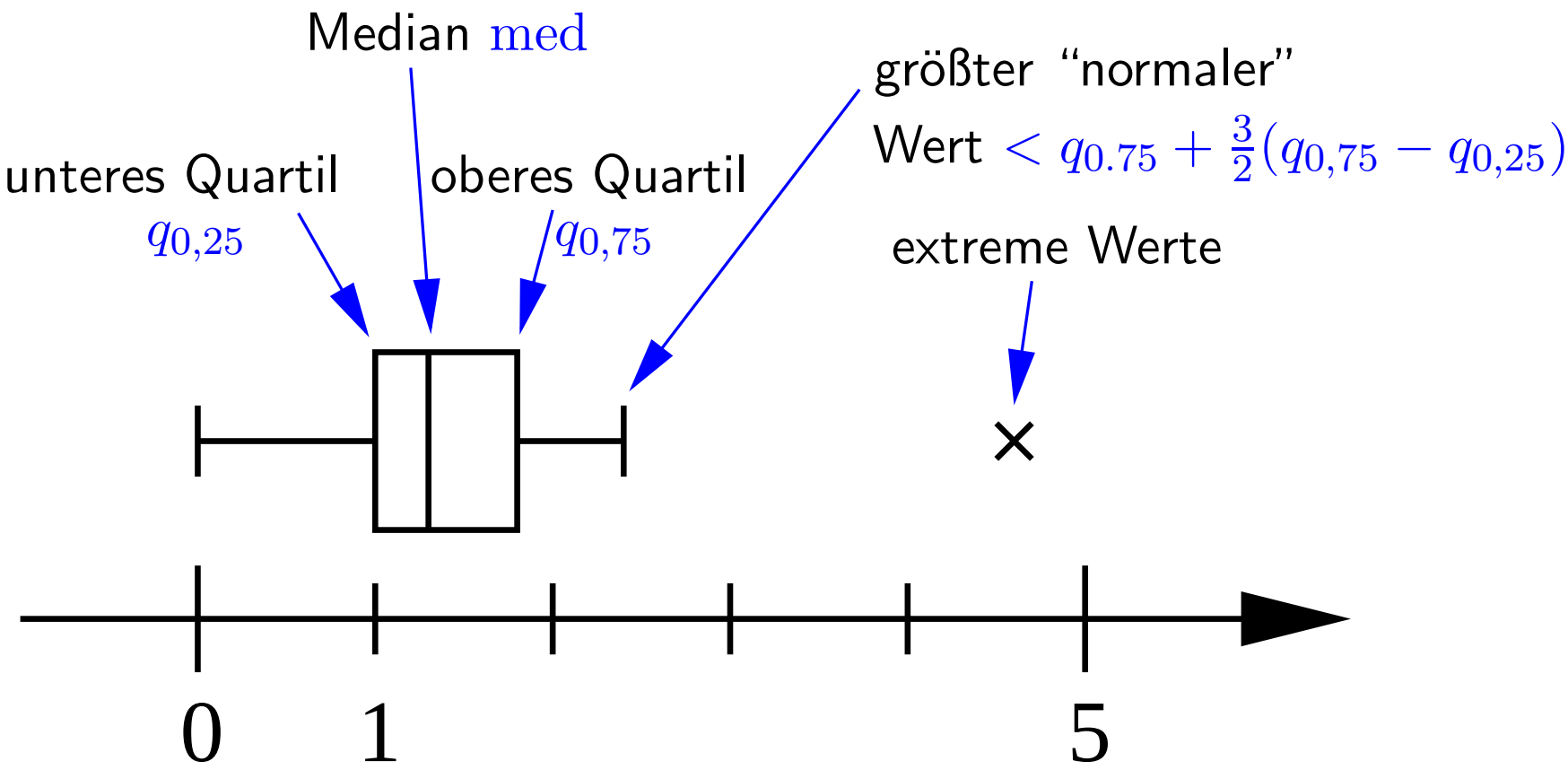
Histogramme ("Drehe Stamm- und Blattdiagramm") für Beispiel

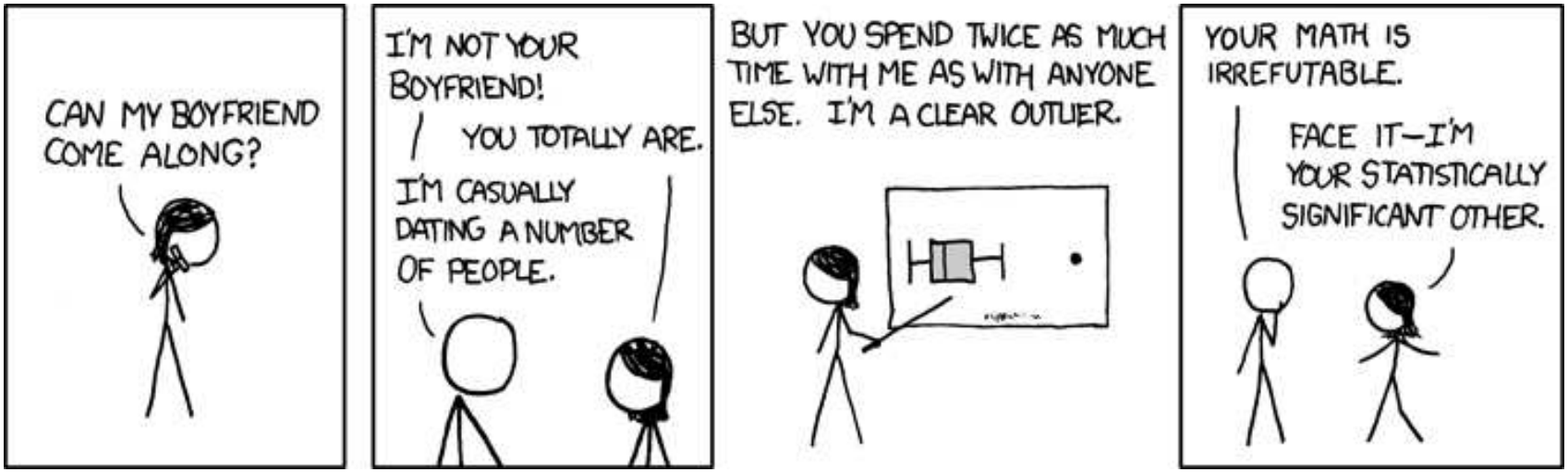


Klassenbreite: 1 2 0,5

Fläche ist poportional zur Häufigkeit, nicht die Höhe!

Boxplot für unser Beispiel:





http://xkcd.com/539

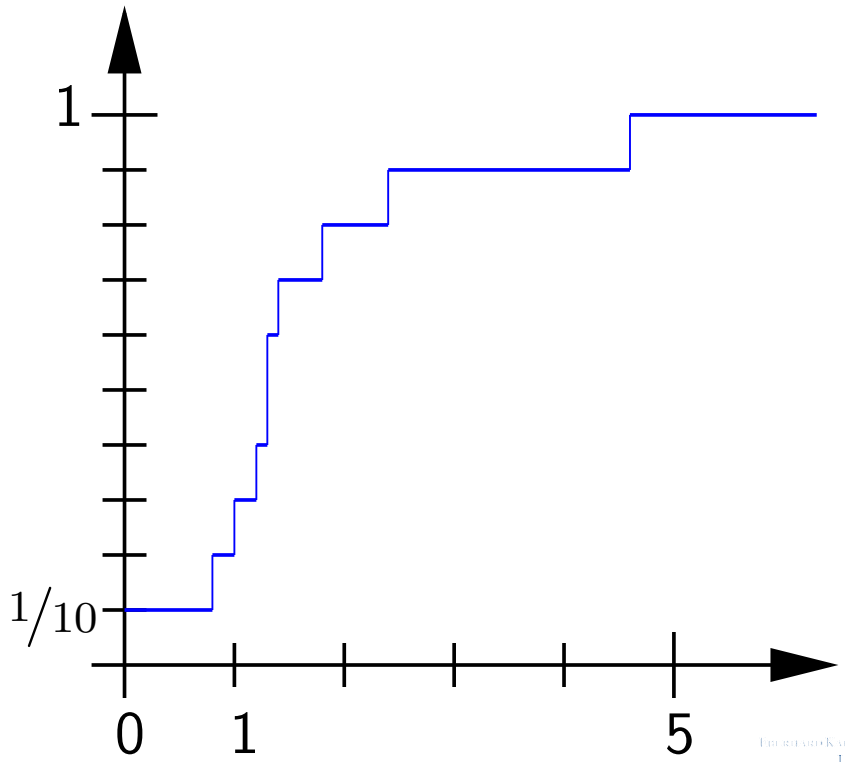
empirische kumulative Verteilungsfunktion $F : \mathbb{R} \to [0, 1]$

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

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

im Beispiel $\longrightarrow$

(senkrechte Linien gehören streng genommen nicht mit dazu)



. . . glauben Sie nicht alles!

Number of partners	Men (age-group [years])				Women (age-group [years])			
	16–24	25–34	35–44	All	16–24	25–34	35–44	All
Lifetime								
0	19·6%	3·5%	1·8%	7·2%	17·7%	0·9%	0·9%	5·3%
1	14·9%	8·4%	10·7%	11·0%	18·1%	16·2%	20·8%	18·3%
2	8·2%	7·2%	7·2%	7·5%	11·1%	10·8%	10·9%	10·9%
3–4	16·6%	14·3%	13·4%	14·6%	17·1%	19·7%	21·5%	19·6%
5–9	21·0%	25·2%	28·3%	25·2%	21·5%	29·8%	26·6%	26·5%
10+	19·7%	41·4%	38·7%	34·6%	14·6%	22·7%	19·4%	19·2%
Mean (SD)	6·9 (13·1)	13·6 (23·1)	16·0 (52·4)	12·7 (35·2)	5·0 (7·6)	7·3 (9·7)	6·8 (10·8)	6·5 (9·7)
Median (99th percentile)	3 (63)	7 (100)	7 (120)	6 (100)	3 (30)	5 (40)	4 (49)	4 (39)
Weighted, Unweighted bases*	1492, 1211	2092, 1759	1990, 1691	5573, 4661	1439, 1433	2017, 2486	1935, 2356	5390, 6275
Past 5 years								
0	20·6%	5·2%	4·4%	9·0%	18·2%	2·2%	3·8%	7·0%
1	17·2%	39·4%	64·7%	42·4%	24·4%	59·0%	75·2%	55·6%
2	10·5%	14·0%	11·2%	12·1%	13·9%	15·2%	11·9%	13·7%
3–4	18·2%	17·2%	10·6%	15·1%	18·3%	13·3%	6·2%	12·1%
5–9	19·5%	14·6%	6·2%	12·9%	16·0%	7·8%	2·3%	8·0%
10+	14·1%	9·6%	2·9%	8·4%	9·2%	2·5%	0·6%	3·6%
Mean (SD)	5·3 (10·7)	4·2 (8·6)	2·2 (3·8)	3·8 (8·2)	3·8 (6·7)	2·2 (3·1	1·5 (3·9	2·4 (4·6)
Median (99th per centile)	3 (41)	2 (30)	1 (19)	1 (30)	2 (27)	1 (15)	1 (6)	1 (19)
Weighted, unweighted bases*	1480, 1200	2082, 1751	1960, 1669	5522, 4620	1424, 1422	2008, 2474	1915, 2332	5346, 6228

All percentages are of column weighted base. *Bases vary from totals in table 1 due to item non-response.

Table 2: **Distribution of numbers of heterosexual partners over lifetime and in the past 5 years by gender and age-group: Natsal 2000**