

## Weitere Lagemaße: Quantile/Perzentile I

- Für jeden Median  $x_{\text{med}}$  gilt: Mindestens 50% der Merkmalswerte sind kleiner gleich  $x_{\text{med}}$  und ebenso mindestens 50% größer gleich  $x_{\text{med}}$ .
- Verallgemeinerung dieser Eigenschaft auf beliebige Anteile geläufig, also auf Werte, zu denen mindestens ein Anteil  $p$  kleiner gleich und ein Anteil  $1 - p$  größer gleich ist, sog.  **$p$ -Quantilen** (auch  **$p$ -Perzentile**)  $x_p$ .
- Mediane sind dann gleichbedeutend mit 50%-Quantilen bzw. 0.5-Quantilen, es gilt also insbesondere bei eindeutigen Medianen

$$x_{\text{med}} = x_{0.5} .$$

## Weitere Lagemaße: Quantile/Perzentile II

### Definition 3.4 (Quantile/Perzentile, Quartile)

Sei  $X$  ein (mindestens) ordinalskaliertes Merkmal auf der Menge der vorstellbaren Merkmalsausprägungen  $M$  mit den Merkmalswerten  $x_1, \dots, x_n$ .

Für  $0 < p < 1$  heißt jeder Wert  $x_p \in M$  mit der Eigenschaft

$$\frac{\#\{i \in \{1, \dots, n\} \mid x_i \leq x_p\}}{n} \geq p \quad \text{und} \quad \frac{\#\{i \in \{1, \dots, n\} \mid x_i \geq x_p\}}{n} \geq 1 - p$$

**$p$ -Quantil** (auch  **$p$ -Perzentil**) von  $X$ . Man bezeichnet spezieller das 0.25-Quantil  $x_{0.25}$  als **unteres Quartil** sowie das 0.75-Quantil  $x_{0.75}$  als **oberes Quartil**.

## Weitere Lagemaße: Quantile/Perzentile III

- $p$ -Quantile kann man auch mit der emp. Verteilungsfunktion  $F$  bestimmen:
- Mit der Abkürzung

$$F(x - 0) := \lim_{\substack{h \rightarrow 0 \\ h > 0}} F(x - h), \quad x \in \mathbb{R},$$

für linksseitige Grenzwerte empirischer Verteilungsfunktionen  $F$  ist  $x_p$  ist genau dann ein  $p$ -Quantil, wenn gilt:

$$F(x_p - 0) \leq p \leq F(x_p)$$

- Spezieller ist  $x_p$  genau dann ein  $p$ -Quantil, wenn
  - ▶ bei Vorliegen der exakten Häufigkeitsverteilung  $r$  und Verteilungsfunktion  $F$

$$F(x_p) - r(x_p) \leq p \leq F(x_p),$$

- ▶ bei Verwendung der approximativen Verteilungsfunktion  $F$  bei klassierten Daten (wegen der Stetigkeit der Approximation!)

$$F(x_p) = p$$

gilt.

## Weitere Lagemaße: Quantile/Perzentile IV

- Genauso wie der Median muss ein  $p$ -Quantil nicht eindeutig bestimmt sein.
- Bei stetigen Merkmalen kann Eindeutigkeit *zum Beispiel* durch die gängige Festlegung

$$x_p = \begin{cases} x_{(\lfloor n \cdot p \rfloor + 1)} & \text{für } n \cdot p \notin \mathbb{N} \\ \frac{1}{2} \cdot (x_{(n \cdot p)} + x_{(n \cdot p + 1)}) & \text{für } n \cdot p \in \mathbb{N} \end{cases}$$

erreicht werden, wobei  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  die gemäß der vorgegebenen Ordnung sortierte Urliste ist und mit  $\lfloor y \rfloor$  für  $y \in \mathbb{R}$  die größte ganze Zahl kleiner gleich  $y$  bezeichnet wird.

- Zum Beispiel ist für die (bereits sortierte) Urliste

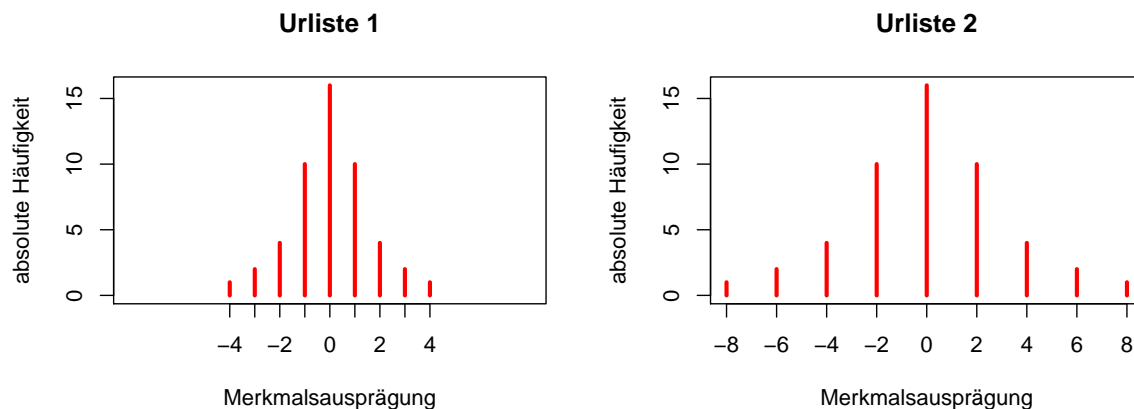
$$6.77, 7.06, 8.84, 9.98, 11.87, 12.18, 12.7, 14.92$$

der Länge  $n = 8$  das 0.25-Quantil  $x_{0.25}$  wegen  $n \cdot p = 8 \cdot 0.25 = 2 \in \mathbb{N}$  nicht eindeutig bestimmt, sondern alle Werte  $x_{0.25} \in [7.06, 8.84]$  sind 0.25-Quantile. Die eindeutige Festlegung nach obiger Konvention würde dann die „Auswahl“  $x_{0.25} = \frac{1}{2} (7.06 + 8.84) = 7.95$  treffen.

## Streuungsmaße I

- Verdichtung der Merkmalswerte auf einen Lageparameter als einzige Kennzahl recht unspezifisch.
- Starke Unterschiede trotz übereinstimmender Lagemaße möglich:

Stabdiagramme zu Urlisten mit identischem Mittelwert, Modus, Median



## Streuungsmaße II

- Bei kardinalskalierten Merkmalen: zusätzliche Kennzahl für Variation bzw. Streuung der Merkmalswerte von Interesse
- Ähnlich wie bei Lagemaßen: verschiedene Streuungsmaße gängig
- Allen Streuungsmaßen gemeinsam: Bezug zu „Abstand“ zwischen Merkmalswerten
- *Ein* möglicher Abstand: (Betrag der) Differenz zwischen Merkmalswerten

## Streuungsmaße III

### Definition 3.5 (Spannweite, IQA, mittlere abs. Abweichung)

Seien  $x_1, \dots, x_n$  die Urliste zu einem kardinalskalierten Merkmal  $X$ ,  $x_{med}$  der Median und  $x_{0.25}$  bzw.  $x_{0.75}$  das untere bzw. obere Quartil von  $X$ .

Dann heißt

- 1  $SP := \left( \max_{i \in \{1, \dots, n\}} x_i \right) - \left( \min_{i \in \{1, \dots, n\}} x_i \right) = x_{(n)} - x_{(1)}$  die **Spannweite** von  $X$ ,
- 2  $IQA := x_{0.75} - x_{0.25}$  der **Interquartilsabstand (IQA)** von  $X$ ,
- 3  $MAA := \frac{1}{n} \sum_{i=1}^n |x_i - x_{med}|$  die **mittlere absolute Abweichung** von  $X$ .

## Streuungsmaße IV

- Die Betragstriche in Teil 1 und 2 von Definition 3.5 fehlen, da sie überflüssig sind.
- Um Eindeutigkeit in Teil 2 von Definition 3.5 zu erhalten, sind die für kardinalskalierte Merkmale üblichen Konventionen zur Berechnung von Quantilen aus Folie 76 anzuwenden.
- Verwendung von  $\bar{x}$  statt  $x_{med}$  in Teil 3 von Definition 3.5 prinzipiell möglich, aber: Beachte Folie 72!
- Weiterer möglicher Abstand: Quadrate der Differenzen zwischen Merkmalswerten

## Streuungsmaße V

### Definition 3.6 (empirische Varianz, empirische Standardabweichung)

Seien  $x_1, \dots, x_n$  die Urliste zu einem kardinalskalierten Merkmal  $X$ ,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  das arithmetische Mittel von  $X$ . Dann heißt

- 1  $s^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$  die **(empirische) Varianz** von  $X$ ,
- 2 die (positive) Wurzel  $s = \sqrt{s^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$  die **(empirische) Standardabweichung** von  $X$ .

## Streuungsmaße VI

- Empirische Varianz bzw. Standardabweichung sind die gebräuchlichsten Streuungsmaße.
- Standardabweichung  $s$  hat dieselbe Dimension wie die Merkmalswerte, daher i.d.R. besser zu interpretieren als Varianz.
- Für Merkmale mit positivem Mittelwert  $\bar{x}$  als relatives Streuungsmaß gebräuchlich: **Variationskoeffizient**  $VK := \frac{s}{\bar{x}}$
- „Rechenregeln“ zur alternativen Berechnung von  $s$  bzw.  $s^2$  vorhanden.

### Satz 3.1 (Verschiebungssatz)

Seien  $x_1, \dots, x_n$  die Urliste zu einem kardinalskalierten Merkmal  $X$ ,  $\bar{x}$  das arithmetische Mittel und  $s^2$  die empirische Varianz von  $X$ . Dann gilt

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

## Streuungsmaße VII

- Mit der Schreibweise  $\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$  erhält man aus Satz 3.1 die kürzere Darstellung  $s^2 = \overline{x^2} - \bar{x}^2$ .
- Liegt zum Merkmal  $X$  die absolute Häufigkeitsverteilung  $h(a)$  bzw. die relative Häufigkeitsverteilung  $r(a)$  auf der Menge der Ausprägungen  $A = \{a_1, \dots, a_m\}$  vor, so kann  $s^2$  auch durch

$$s^2 = \frac{1}{n} \sum_{j=1}^m h(a_j) \cdot (a_j - \bar{x})^2 = \sum_{j=1}^m r(a_j) \cdot (a_j - \bar{x})^2$$

berechnet werden. (Berechnung von  $\bar{x}$  dann mit Häufigkeiten als  $\bar{x} = \frac{1}{n} \sum_{j=1}^m h(a_j) \cdot a_j = \sum_{j=1}^m r(a_j) \cdot a_j$ , siehe Bemerkung 3.1 auf Folie 67)

- Natürlich kann alternativ auch Satz 3.1 verwendet und  $\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$  mit Hilfe der Häufigkeitsverteilung durch

$$\overline{x^2} = \frac{1}{n} \sum_{j=1}^m h(a_j) \cdot a_j^2 = \sum_{j=1}^m r(a_j) \cdot a_j^2$$

berechnet werden.

## Empirische Varianz bei klassierten Daten

- Bei klassierten Daten: auch für empirische Varianz nur Approximation möglich.
- Analog zur Berechnung von  $s^2$  aus Häufigkeitsverteilungen:
  - ▶ Näherungsweise Berechnung von  $s^2$  aus Klassenmitten  $m_j$  und absoluten bzw. relativen Klassenhäufigkeiten  $h_j$  bzw.  $r_j$  der  $l$  Klassen als

$$s^2 = \frac{1}{n} \sum_{j=1}^l h_j \cdot (m_j - \bar{x})^2 \quad \text{mit} \quad \bar{x} = \frac{1}{n} \sum_{j=1}^l h_j \cdot m_j$$

bzw.

$$s^2 = \sum_{j=1}^l r_j \cdot (m_j - \bar{x})^2 \quad \text{mit} \quad \bar{x} = \sum_{j=1}^l r_j \cdot m_j .$$

- ▶ Alternativ: Verwendung von Satz 3.1 mit

$$\bar{x} := \frac{1}{n} \sum_{j=1}^l h_j \cdot m_j = \sum_{j=1}^l r_j \cdot m_j$$

und

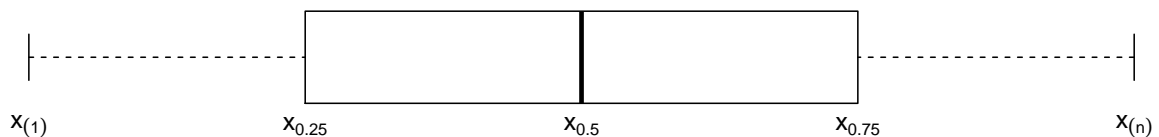
$$\overline{x^2} := \frac{1}{n} \sum_{j=1}^l h_j \cdot m_j^2 = \sum_{j=1}^l r_j \cdot m_j^2 .$$

## Box-and-whisker-Plot I

- Häufig von Interesse:  
Visueller Vergleich **eines** Merkmals für **verschiedene** statistische Massen
- Dazu nötig: Grafische Darstellung mit Ausdehnung (im Wesentlichen) nur in einer Dimension (2. Dimension für Nebeneinanderstellung der Datensätze)

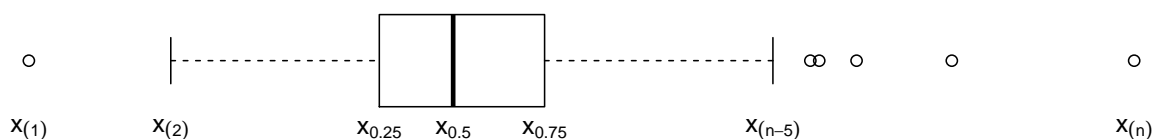
↪ **Box-and-whisker-Plot** oder kürzer **Box-Plot**:

Zur Urliste  $x_1, \dots, x_n$  eines kardinalskalierten Merkmals werden *im Prinzip* die 5 Kennzahlen  $x_{(1)}, x_{0.25}, x_{0.5}, x_{0.75}, x_{(n)}$  in Form eines durch  $x_{0.5}$  geteilten „Kästchens“ (Box) von  $x_{0.25}$  bis  $x_{0.75}$  und daran anschließende „Schnurrhaare“ (Whisker) bis zum kleinsten Merkmalswert  $x_{(1)}$  und zum größten Merkmalswert  $x_{(n)}$  dargestellt:



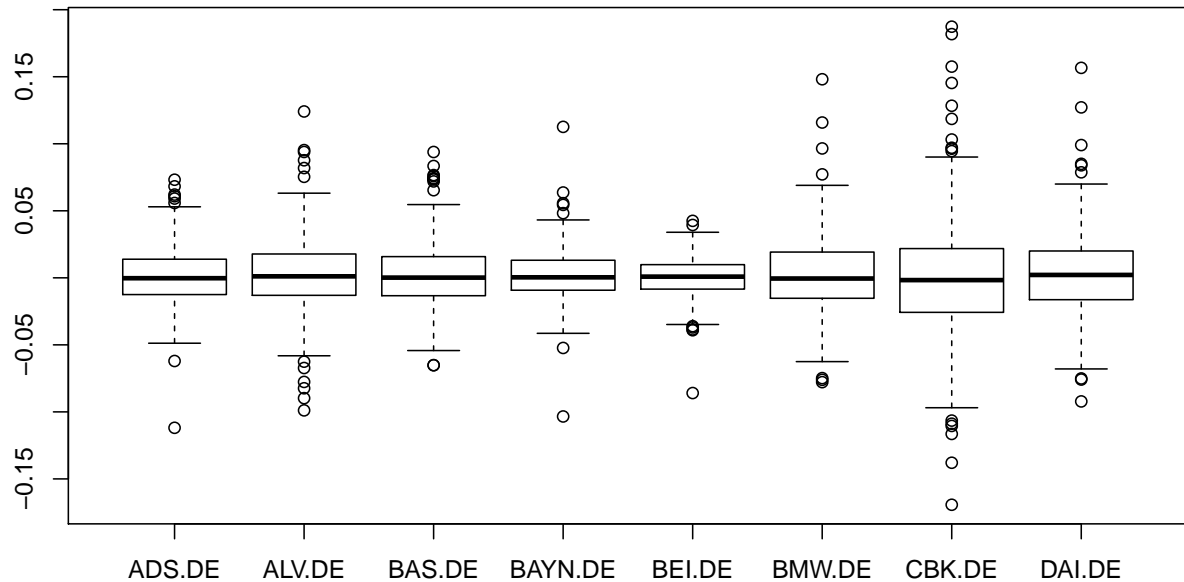
## Box-and-whisker-Plot II

- (Häufig auftretende!) Ausnahme:  
 $x_{(1)}$  und/oder  $x_{(n)}$  liegen weiter als der 1.5-fache Interquartilsabstand (IQA)  $x_{0.75} - x_{0.25}$  von der Box entfernt (also weiter als die 1.5-fache Breite der Box)
- ↪ Dann: Whiskers nur bis zu äußersten Merkmalswerten innerhalb dieser Distanz und separates Eintragen der „Ausreißer“, d.h. aller Urlisteneinträge, die nicht von der Box und den Whiskers abgedeckt werden.
- Beispiel mit „Ausreißern“:



## Box-and-whisker-Plot III

- Beispiel für Gegenüberstellung mehrerer Datensätze (Diskrete Tagesrenditen verschiedener DAX-Papiere)



## Symmetrie(-maß), Schiefe I

- Neben Lage und Streuung bei kardinalskalierten Merkmalen auch interessant: **Symmetrie** (bzw. Asymmetrie oder Schiefe) und **Wölbung**
- Ein Merkmal  $X$  ist symmetrisch (um  $\bar{x}$ ), wenn die Häufigkeitsverteilung von  $X - \bar{x}$  mit der von  $\bar{x} - X$  übereinstimmt.  
(Dabei ist mit  $X - \bar{x}$  das Merkmal mit den Urlistenelementen  $x_i - \bar{x}$  für  $i \in \{1, \dots, n\}$  bezeichnet, dies gilt analog für  $\bar{x} - X$ .)
- Symmetrie eines Merkmals entspricht also der Achsensymmetrie des zugehörigen Stabdiagramms um  $\bar{x}$ .
- Ist ein Merkmal nicht symmetrisch, ist die **empirische Schiefe** bzw. **empirische Skewness** ein geeignetes Maß für die Stärke der Asymmetrie.



# Symmetrie(-maß), Schiefe II

## Definition 3.7 (empirische Schiefe, Skewness)

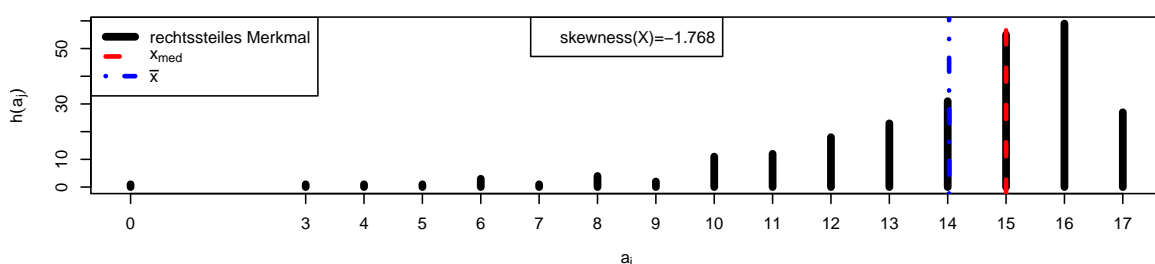
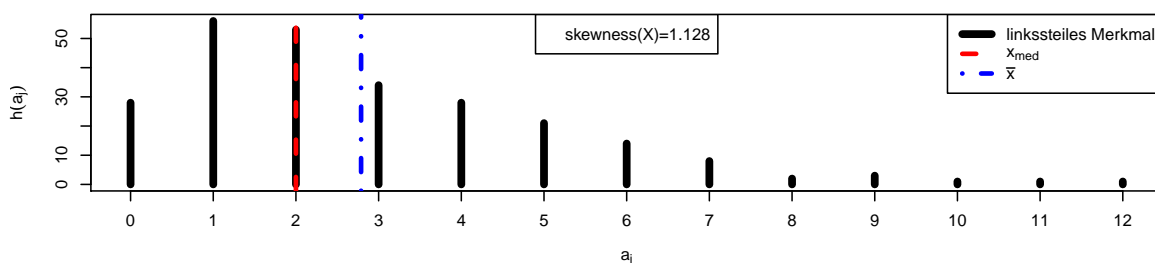
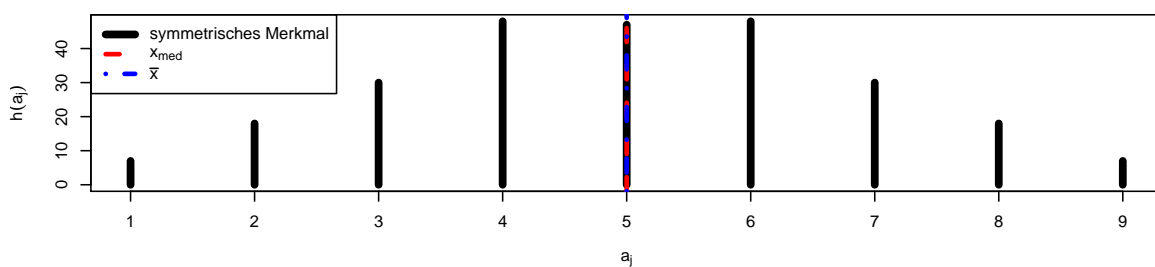
Sei  $X$  ein Merkmal mit der Urliste  $x_1, \dots, x_n$ . Dann heißt

$$\text{skewness}(X) := \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

mit  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  und  $s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$  die **empirische Schiefe (Skewness)** von  $X$ .

- Man kann zeigen:  $X$  symmetrisch  $\Rightarrow$   $\text{skewness}(X) = 0$
- $X$  heißt **linkssteil** oder **rechtsschief**, falls  $\text{skewness}(X) > 0$ .
- $X$  heißt **rechtssteil** oder **linksschief**, falls  $\text{skewness}(X) < 0$ .
- Für symmetrische Merkmale ist  $\bar{x}$  gleichzeitig Median von  $X$ , bei linkssteilen Merkmalen gilt *tendenziell*  $\bar{x} > x_{\text{med}}$ , bei rechtssteilen *tendenziell*  $\bar{x} < x_{\text{med}}$ .

### Beispiele für empirische Schiefe von Merkmalen



## Wölbungsmaß (Kurtosis) I

### Definition 3.8 (empirische Wölbung, Kurtosis)

Sei  $X$  ein Merkmal mit der Urliste  $x_1, \dots, x_n$ . Dann heißt

$$\text{kurtosis}(X) := \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4$$

mit  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  und  $s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$  die **empirische Wölbung (Kurtosis)** von  $X$ .

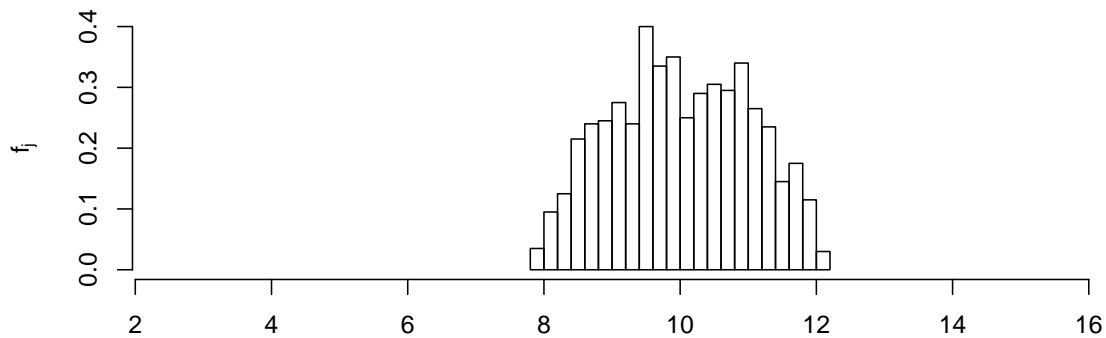
- Kurtosis misst bei Merkmalen mit *einem* Modalwert, wie „flach“ (kleiner Wert) bzw. „spitz“ (großer Wert) der „Gipfel“ um diesen Modalwert ist.

## Wölbungsmaß (Kurtosis) II

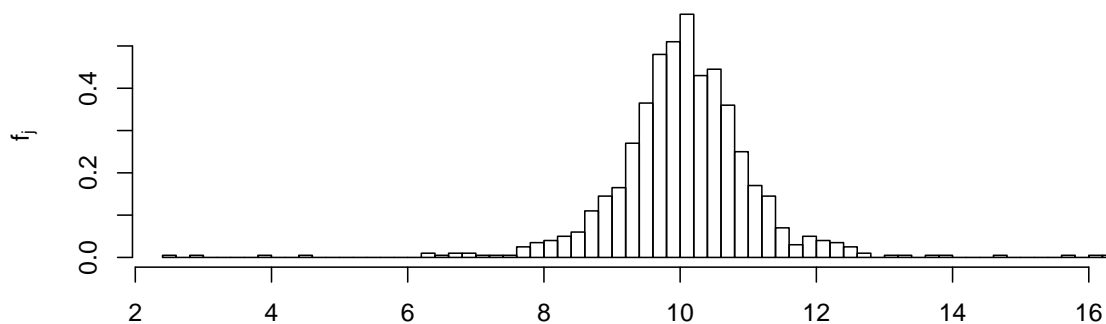
- Bei gleicher mittlerer quadratischer Abweichung vom Mittelwert ( $\rightsquigarrow$  Varianz) müssen Merkmale mit größerer emp. Kurtosis (mehr Werten in der Nähe des Gipfels) auch mehr weit vom Gipfel entfernte Merkmalswerte besitzen.
- Der Wert 3 wird als „normaler“ Wert für die empirische Kurtosis angenommen, Merkmale mit  $1 \leq \text{kurtosis}(X) < 3$  heißen platykurtisch, Merkmale mit  $\text{kurtosis}(X) > 3$  leptokurtisch.
- *Vorsicht:* Statt der Kurtosis von  $X$  wird oft die **Exzess-Kurtosis** von  $X$  angegeben, die der um den Wert 3 verminderten Kurtosis entspricht.

## Beispiele für Merkmale mit unterschiedlicher empirischer Kurtosis

### Merkmal mit kleiner empirischer Kurtosis (2.088)

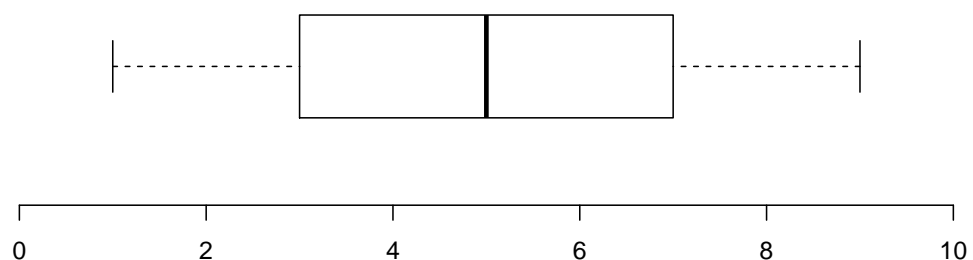


### Merkmal mit großer empirischer Kurtosis (12.188)



## Schiefe und Wölbung in grafischen Darstellungen I

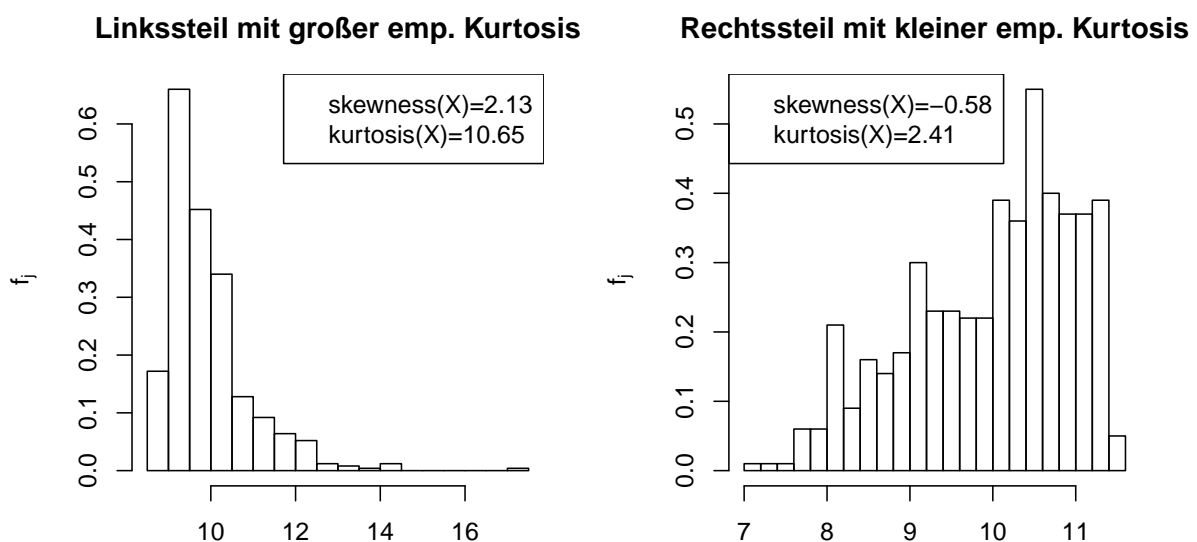
- Box-Plots lassen auch auf empirische Schiefe und Kurtosis schließen.
- Bei symmetrischen Merkmalen sind auch die Box-Plots symmetrisch.  
Beispiel: Box-Plot zur Urliste 1, 2, 3, 4, 5, 6, 7, 8, 9:



# Schiefe und Wölbung in grafischen Darstellungen II

- Bei linkssteilen Merkmalen hat *tendenziell* der rechte/obere Teil (rechter/oberer Teil der Box und rechter/oberer Whisker) eine **größere** Ausdehnung als der linke/untere Teil.
- Bei rechtssteilen Merkmalen hat *tendenziell* der rechte/obere Teil (rechter/oberer Teil der Box und rechter/oberer Whisker) eine **kleinere** Ausdehnung als der linke/untere Teil.
- Bei Merkmalen mit **großer** empirischer Kurtosis gibt es *tendenziell* **viele** „Ausreißer“, also separat eingetragene Merkmalswerte außerhalb der Whiskers (wenigstens auf einer Seite).
- Bei Merkmalen mit **kleiner** empirischer Kurtosis gibt es häufig **wenige** oder **gar keine** „Ausreißer“.

- Beispiele für Merkmale mit unterschiedlicher empirischer Schiefe/Kurtosis



- Zugehörige Box-Plots:

