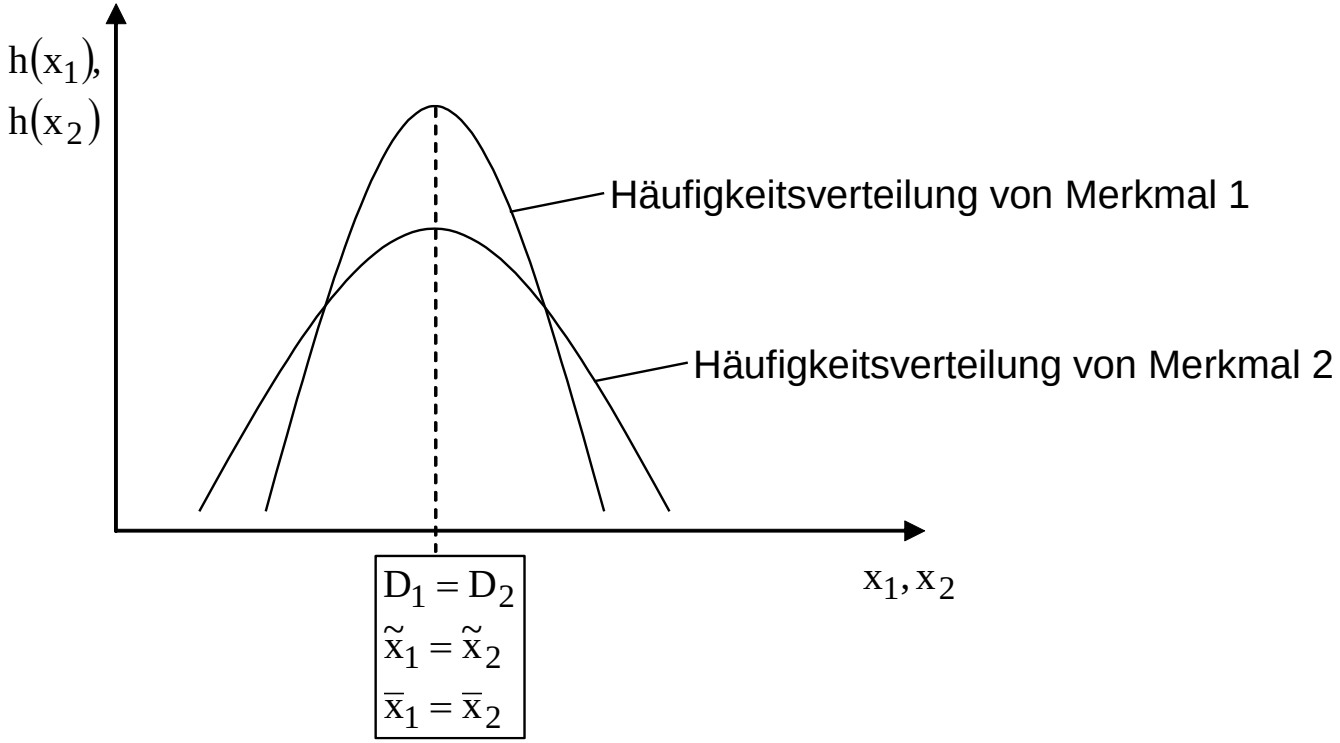
5. Streuung und Schiefe

Konzept der Streuung:

- Streuen Merkmalswerte gering, dann repräsentiert der Mittelwert eine Häufigkeitsverteilung besser als bei einer weiten Streuung.
- Daher sollten Streuungsmaße ergänzend zu Mittelwerten angegeben werden.
- Zudem: Spezielle Anwendungen (z.B. bei Qualitäts- und Risikobewertungen)

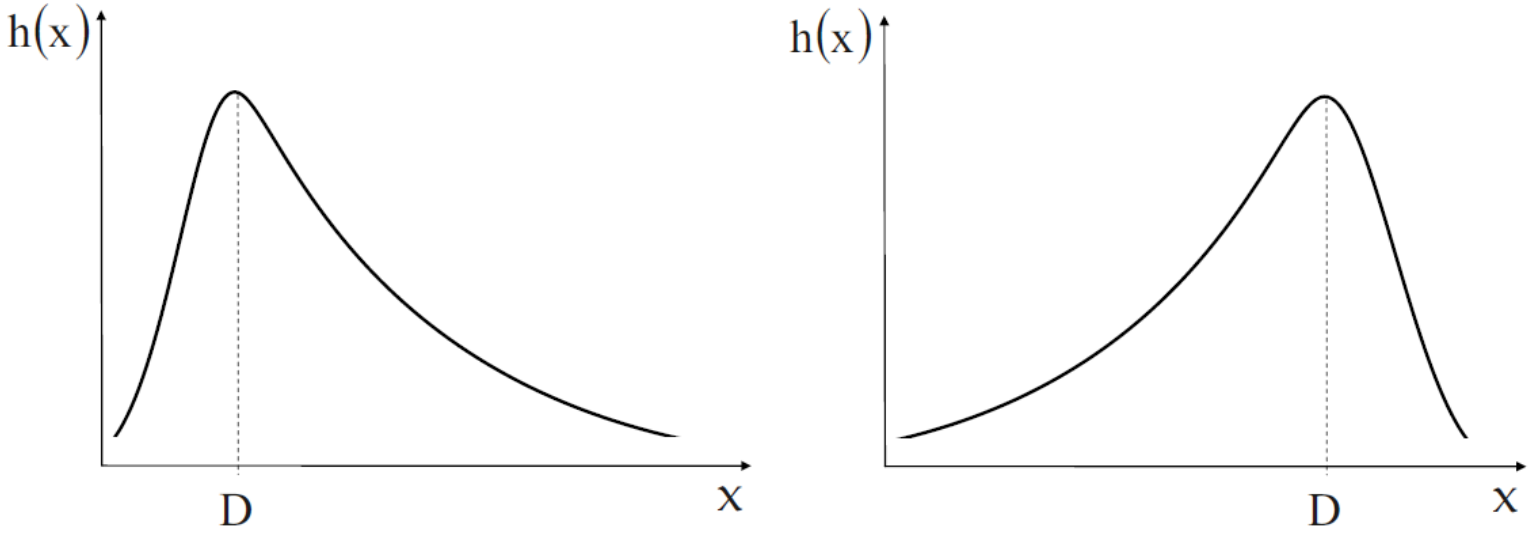
Beispiel: Symmetrische Häufigkeitsverteilungen mit gleichem Zentrum, bei unterschiedlicher Streuung



Konzept der Schiefe:

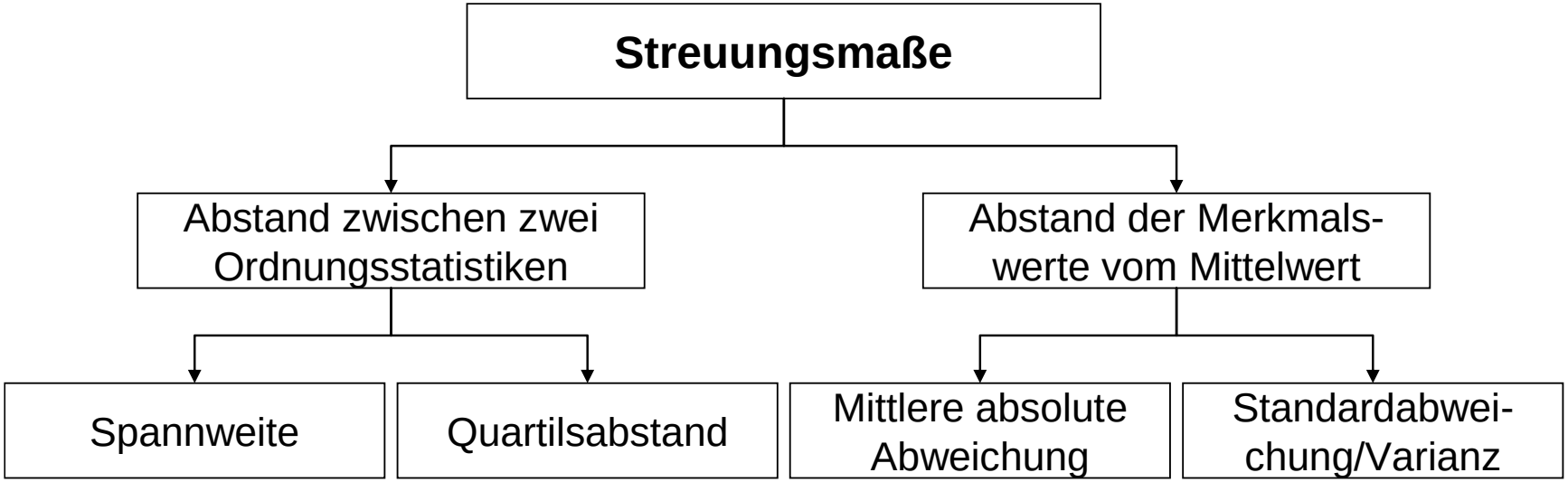
- Häufigkeitsverteilungen können bei gleichem Mittelwert und gleicher Streuung eine unterschiedliche Gestalt besitzen.
- Gleichförmige Verteilung der Merkmalswerte um den Mittelwert → symmetrische Verteilung
- Andernfalls → asymmetrisch oder schief

Beispiel: Rechts- und linksschiefe Verteilung



a) rechtsschief b) linksschief

5.1 Streuungsmaße

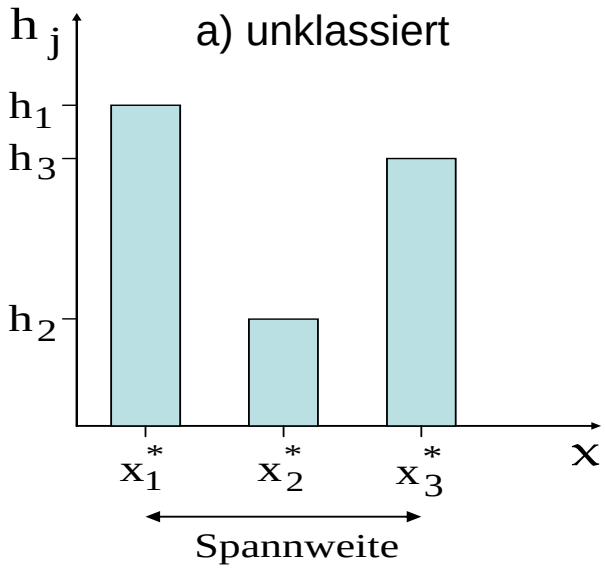


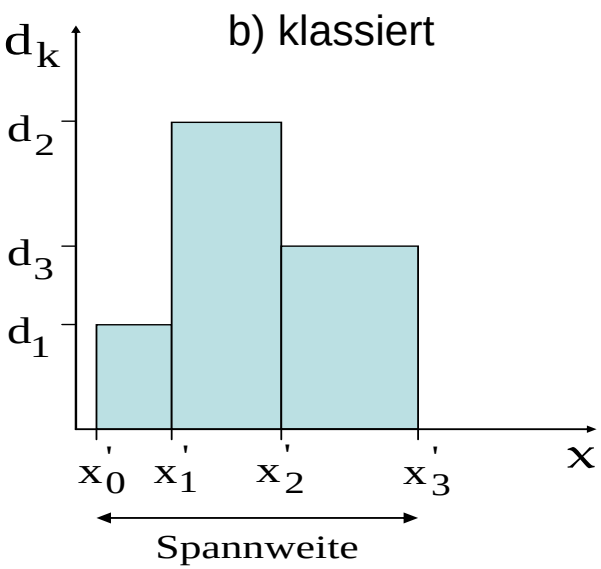
- Messung des Ausmaßes der Streuung, nicht der Richtung
- Voraussetzung: **metrische Skala**
- Streuungsmaße ≥ 0

5.1.1 Spannweite

- Die **Spannweite** (range) gibt an, in welchem Bereich sich alle Einzelwerte befinden.
- Bei Einzelwerten: $R = x_{(n)} - x_{(1)}$
- Bei einer unklassierten Häufigkeitsverteilung: $R = x_m^* - x_1^*$
- Bei einer klassierten Häufigkeitsverteilung: $R = x_p' - x_0'$

Beispiel: Spannweite bei Häufigkeitsverteilungen

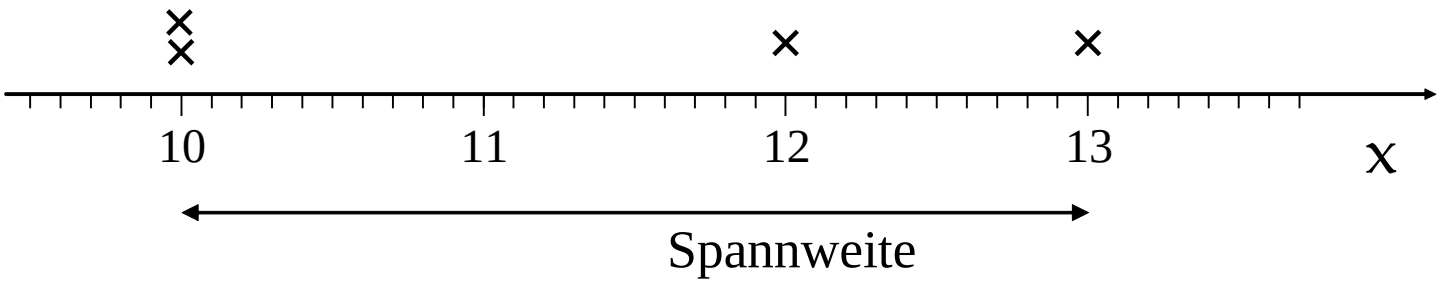




Beispiel (Fortsetzung): **Callcenter**

Für die eingehenden Anrufe pro Minute im Callcenter mit $x_1 = 10$, $x_2 = 12$, $x_3 = 13$, $x_4 = 10$ soll die Spannweite berechnet werden:

- Geordnete Reihe: $x_{(1)} = 10$, $x_{(2)} = 10$, $x_{(3)} = 12$, $x_{(4)} = 13$
- $R = x_{(n)} - x_{(1)} = x_{(4)} - x_{(1)} = 13 - 10 = 3$



Interpretation: Die gesamte Streubreite beträgt 3, da sich die eingehenden Anrufe über den Bereich von 10 bis 13 erstrecken.

Beispiel: Aktienkurse

Der Mindestkurs einer Aktie beträgt 495€, während der Höchstkurs bei 526€ liegt. Die Spannweite (Schwankungsbreite des Kurses) nimmt dann folgenden Wert an:

$$R = x_{(n)} - x_{(1)} = 526 - 495 = 31[€]$$

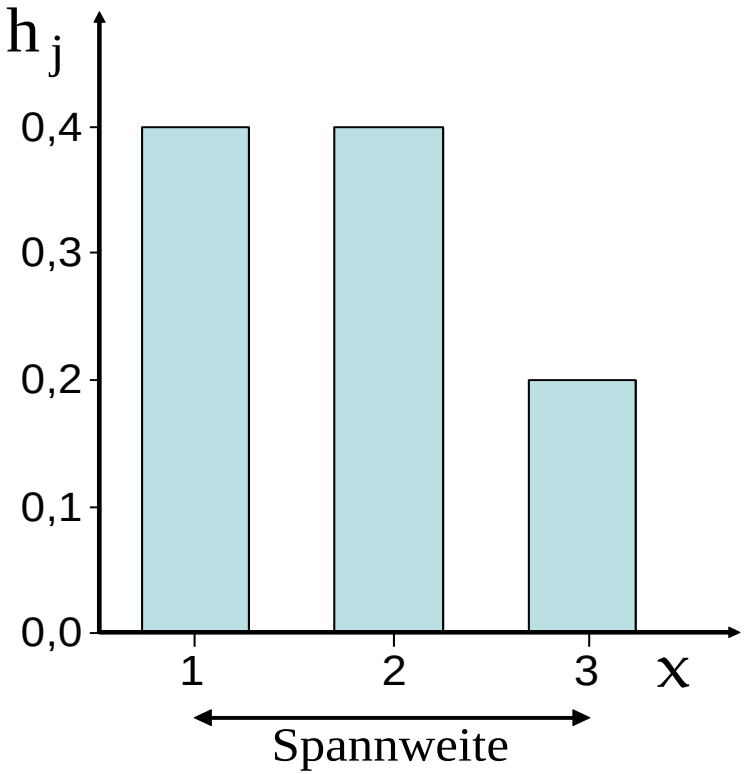
Beispiel (Fortsetzung): Haushaltsgröße

Bei der Häufigkeitsverteilung des Merkmals „Haushaltsgröße“ erhält man die Spannweite als Differenz zwischen größter und kleinster Merkmalsausprägung:

Berechnung der Spannweite	Grafische Darstellung

j	x_j^*	n_j	h_j
1	1	2	0,4
2	2	2	0,4
3	3	1	0,2
∑		n = 5	1

$$R = x_m^* - x_1^* = x_3^* - x_1^* = 3 - 1 = 2$$

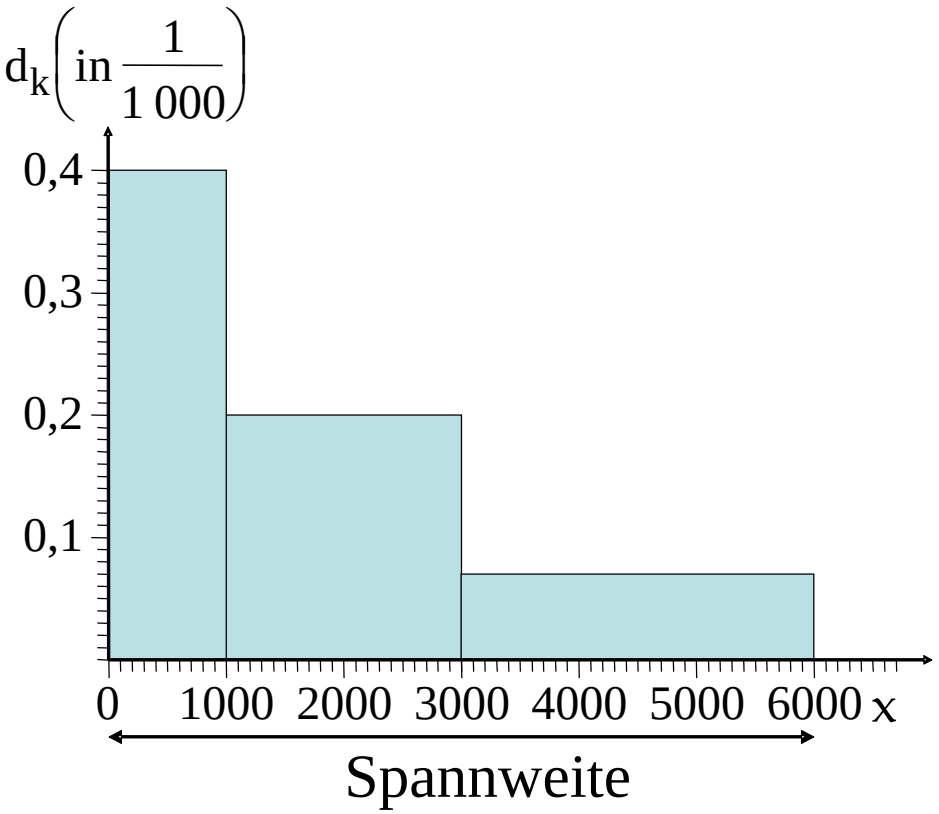


Beispiel (Fortsetzung): Einkommen

Berechung der Spannweite	Grafische Darstellung

k	von über… bis zu …	h_k
1	0 – 1000	0,4
2	1000 – 3000	0,4
3	3000 – 6000	0,2
∑		1

$$R = x'_p - x'_0 = x'_3 - x'_0$$
$$= 6000 - 0 = 6000[€]$$



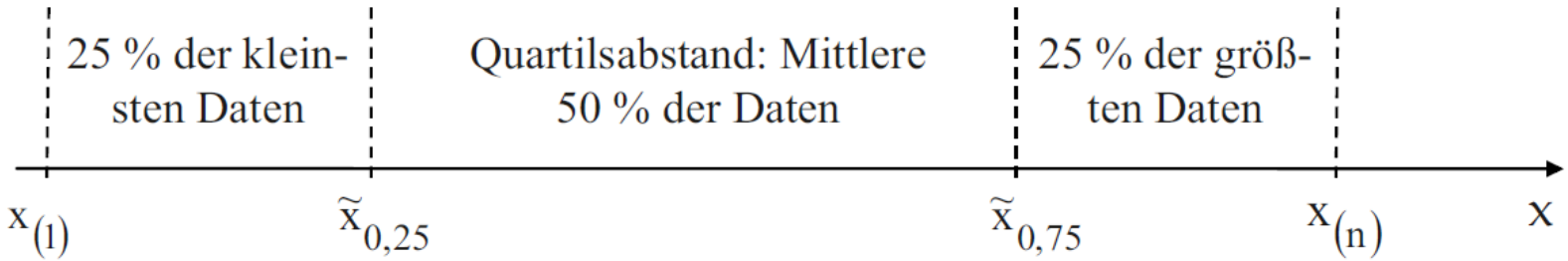
Eigenschaften der Spannweite:

- Starker Einfluss von Ausreißern
- Ausnutzung eines geringen Teils der Informationen

5.1.2 Quartilsabstand und Boxplot

- Der **Quartilsabstand** gibt an, in welchem Bereich sich die mittleren 50 % der Einzelwerte, die der Größe nach geordnet sind, befinden.
- Der Quartilsabstand berechnet sich als Differenz zwischen dem dritten Quartil und dem ersten Quartil:

$$Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$$



Beispiel (Fortsetzung): Einkommen (I)

Aus den Daten der Einkommensverteilung soll der Quartilsabstand bestimmt werden:

k	von über… bis zu …	h_k	H_k
1	0 – 1000	0,4	0,4
2	1000 – 3000	0,4	0,8
3	3000 – 6000	0,2	1
∑		1	

- 1. Quartil: $k^* = 1$, da die H_k in der ersten Klasse $p = 0{,}25$ durchlaufen.

$$\tilde{x}_{0,25} = x'_{k^*-1} + \frac{0{,}25 - H_{k^*-1}}{h_{k^*}} \cdot b_{k^*} = x'_0 + \frac{0{,}25 - H_0}{h_1} \cdot b_1$$
$$= 0 + \frac{0{,}25 - 0}{0{,}4} \cdot 1000 = 625[€]$$

- 3. Quartil: $k^* = 2$, da die H_k in der zweiten Klasse $p = 0{,}75$ durchlaufen.

$$\tilde{x}_{0,75} = x'_{k^*-1} + \frac{0{,}75 - H_{k^*-1}}{h_{k^*}} \cdot b_{k^*} = x'_1 + \frac{0{,}75 - H_1}{h_2} \cdot b_2$$
$$= 1000 + \frac{0{,}75 - 0{,}4}{0{,}4} \cdot 2000 = 2750[€]$$

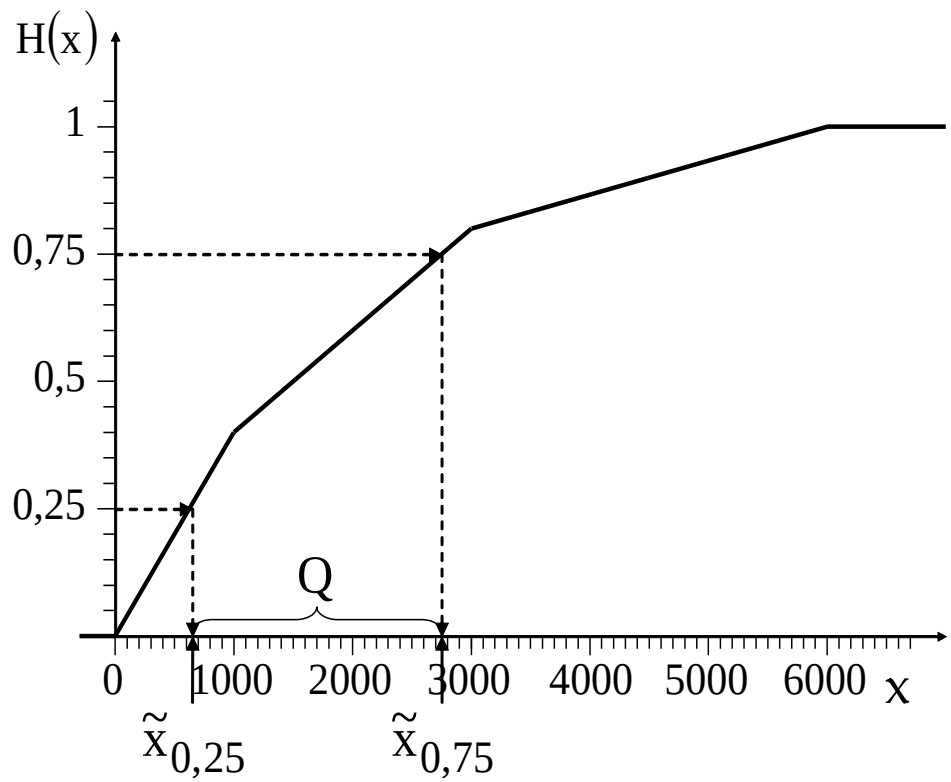
Beispiel (Fortsetzung): Einkommen (II)

- Quartilsabstand:

$$Q = \tilde{x}_{0,75} - \tilde{x}_{0,25} = 2750 - 625 = 2125[€]$$

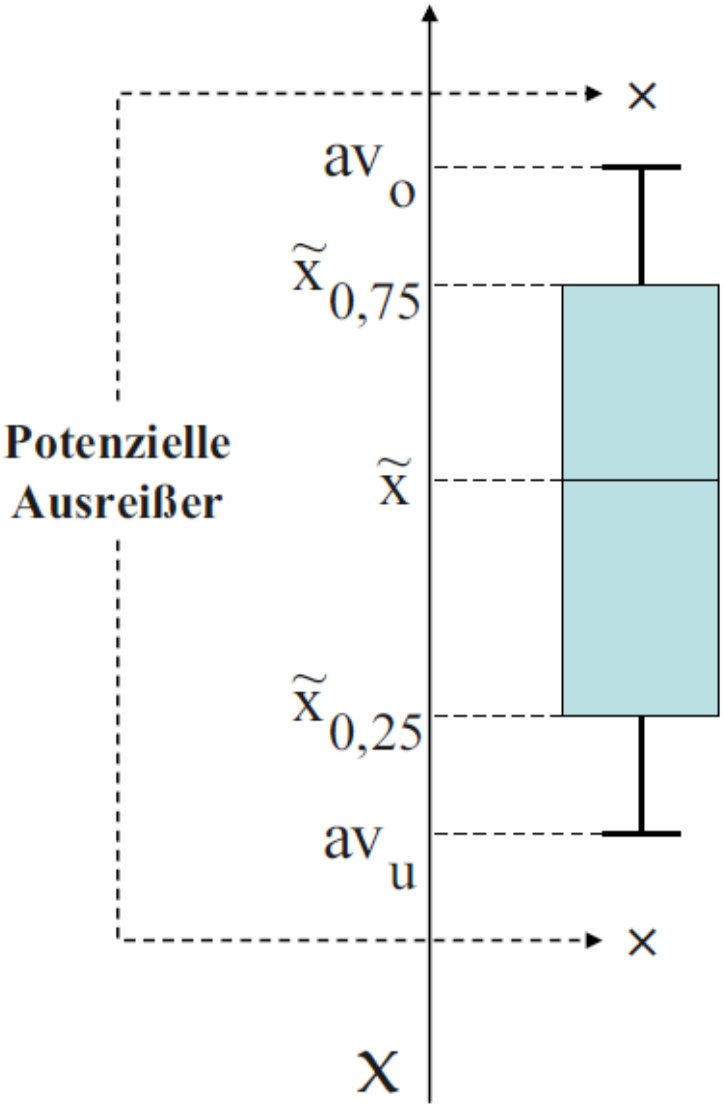
Kumulierte relative Häufigkeiten	Grafische Darstellung

k	von über… bis zu …	h_k	H_k
1	0 – 1000	0,4	0,4
2	1000 – 3000	0,4	0,8
3	3000 – 6000	0,2	1
∑		1	



Boxplot (I)

- Ein **Boxplot** liefert visuell übersichtliche Informationen über Lage, Streuung, Schiefe sowie potenzielle Ausreißer



Boxplot (II)

- An den äußeren Enden werden die „whiskers“ („Fühler“) berechnet. Diese liegen um das 1,5-fache des Quartilsabstands vom ersten bzw. dritten Quartil entfernt:

$$w_u = \tilde{x}_{0,25} - 1{,}5 \cdot Q \qquad \text{und} \qquad w_o = \tilde{x}_{0,75} + 1{,}5 \cdot Q$$

- Die „adjacent values“ – „benachbarte Werte“ von den „whiskers“ – sind die beiden äußersten Merkmalswerte, die sich gerade noch in dem geschlossenen Intervall, abgegrenzt durch die „whiskers“, befinden:

$$[w_u; w_o]$$

- Im ersten Schritt ist eine geordnete Reihe zu bilden. Der untere „adjacent value“ ist der kleinste Merkmalswert, der größer oder gleich dem unteren „whisker“ ist:

$$av_u = min\{x_{(i)} | x_{(i)} \geq w_u\}$$

- Als oberen „adjacent value“ weist man den größten Merkmalswert aus, der den oberen „whisker“ nicht übersteigt:

$$av_o = max\{x_{(i)} | x_{(i)} \leq w_o\}$$

- Alle Merkmalswerte, die außerhalb der „adjacent values“ liegen, sind potenzielle Ausreißer.

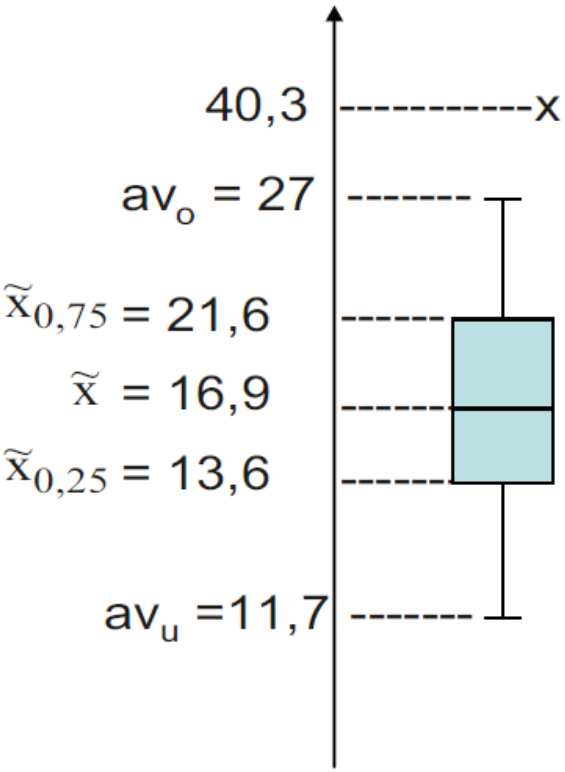
Beispiel (Fortsetzung): Pharmakonzerne

Für das Merkmal Umsatz wurde ein Median von 16,9 Mrd. $ berechnet. Darüber hinaus ergaben sich für das erste und dritte Quartil die Werte 13,6 Mrd. $ und 21,6 Mrd. $. Um einen Boxplot der Umsätze der Pharmakonzerne zu zeichnen, müssen noch der Quartilsabstand, die „whiskers“, die „adjacent values“ und potenzielle Ausreißer bestimmt werden.

- Quartilsabstand: $Q = \tilde{x}_{0,75} - \tilde{x}_{0,25} = 21{,}6 - 13{,}6 = 8\ [Mrd.€]$
- Die „whiskers“:
  - Unterer „whisker“: $w_u = \tilde{x}_{0,25} - 1{,}5 \cdot Q = 13{,}6 - 1{,}5 \cdot 8 = 1{,}6$
  - Oberer „whisker“: $w_o = \tilde{x}_{0,75} + 1{,}5 \cdot Q = 21{,}6 + 1{,}5 \cdot 8 = 33{,}6$
- Diese Werte begrenzen das Intervall, in dem die „adjacent values“ liegen:
  - Unterer „adjacent value“ (kleinster Merkmalswert, der aber trotzdem noch größer oder gleich $w_u = 1{,}6$ ist): $av_u = 11{,}7$
  - Oberer „adjacent value“ (größter Merkmalswert, der $w_o = 33{,}6$ nicht überschreitet): $av_o = 27$
- Ausreißer: Der größte Merkmalswert bildet **nicht** den oberen „adjacent value“, da $x_{(10)} = 40{,}3 > w_o$

Potenzielle Ausreißer sind Merkmalswerte, die kleiner als der untere „adjacent value“ oder größer als der obere „adjacent value“ sind. Im Beispiel liegt ein potenzieller Ausreißer vor: $x_{(10)} = 40{,}3$ ist ein potenzieller Ausreißer, da er größer als $av_o = 27$ ist.

Mit diesen Informationen kann der Boxplot gezeichnet werden:



5.1.3 Mittlere absolute Abweichung

- Die **mittlere absolute Abweichung** vom arithmetischen Mittel gibt an, wie stark die Merkmalswerte durchschnittlich vom arithmetischen Mittel abweichen.
- Verwendung der einfach Abweichungen $(x_i - \bar{x})$ ist ungeeignet, da sich die positiven und negativen Abweichungen aufgrund der Schwerpunkteigenschaft des arithmetischen Mittels ausgleichen:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

- Man kommt zur mittleren absoluten Abweichung, indem man die einfachen Abweichungen $(x_i - \bar{x})$ durch die absoluten Abweichungen $(|x_i - \bar{x}|)$ ersetzt.
- Mittlere absolute Abweichung von $\bar{x}$ bei Einzelwerten:

$$d_{\bar{x}} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \bar{x}|$$

- Mittlere absolute Abweichung von $\bar{x}$ bei unklassierter Häufigkeitsverteilung:

$$d_{\bar{x}} = \frac{1}{n} \cdot \sum_{j=1}^{m} \left|x_j^* - \bar{x}\right| \cdot n_j = \sum_{j=1}^{m} \left|x_j^* - \bar{x}\right| \cdot h_j$$

Beispiel: Dividenden

Ein Unternehmen hat an die Aktionäre in den letzten vier Jahren folgende Dividenden (in €) ausgezahlt: $x_1 = 35$, $x_2 = 20$, $x_3 = 30$, $x_4 = 15$. Wie groß ist die durchschnittliche Abweichung vom arithmetischen Mittel?

- Arithmetisches Mittel:

$$\bar{x} = \frac{1}{4} \cdot \sum_{i=1}^{4} x_i = \frac{1}{4} \cdot (35 + 20 + 30 + 15) = \frac{1}{4} \cdot 100 = 25$$

- Durchschnittliche absolute Abweichung:

$$\begin{aligned} d_{\bar{x}} &= \frac{1}{4} \cdot \sum_{i=1}^{4} |x_i - \bar{x}| = \frac{1}{4}(|35 - 25| + |20 - 25| + |30 - 25| + |15 - 25|) \\ &= \frac{1}{4}(|10| + |-5| + |5| + |-10|) = \frac{1}{4}(10 + 5 + 5 + 10) \\ &= \frac{1}{4} \cdot 30 = 7{,}5[€] \end{aligned}$$

- Die Dividenden weichen also durchschnittlich um 7,5 € vom arithmetischen Mittel mit 25 € ab.

Beispiel (Fortsetzung): Haushaltsgröße

Für das Merkmal „Haushaltsgröße“ soll die mittlere absolute Abweichung vom arithmetischen Mittel berechnet werden:

Häufigkeitstabelle	Berechnung von $d_{\bar{x}}$

j	x_j^*	n_j	h_j
1	1	2	0,4
2	2	2	0,4
3	3	1	0,2
∑		n = 5	1

- Arithmetisches Mittel:

$$\bar{x} = \sum_{j=1}^{3} x_j^* \cdot h_j = 1 \cdot 0{,}4 + 2 \cdot 0{,}4 + 3 \cdot 0{,}2 = 1{,}8$$

- Durchschnittliche absolute Abweichung:

$$\begin{aligned} d_{\bar{x}} &= \sum_{j=1}^{3} \left| x_j^* - \bar{x} \right| \cdot h_j \\ &= |1 - 1{,}8| \cdot 0{,}4 + |2 - 1{,}8| \cdot 0{,}4 + |3 - 1{,}8| \cdot 0{,}2 \\ &= 0{,}8 \cdot 0{,}4 + 0{,}2 \cdot 0{,}4 + 1{,}2 \cdot 0{,}2 \\ &= 0{,}64 \end{aligned}$$

5.1.4 Varianz und Standardabweichung

- Die **Varianz** (s^2):
  - Entspricht der durchschnittlichen Abweichungsquadratsumme (quadrierte Einheit der Merkmalswerte z.B. €², m², %²)
  - Konstruktionsprinzip: Verwendung der quadrierten Abweichungen $(x_i - \bar{x})^2$
  - Berechnung bei Einzelwerten:

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

- Die **Standardabweichung** ($s = \sqrt{s^2}$):
  - Entspricht der Wurzel aus der durchschnittlichen Abweichungsquadrat-summe (originäre Einheit der Merkmalswerte z.B. €, m, %)
  - Weist eine durchschnittliche Abweichung der Merkmalswerte vom arithmetischen Mittel aus
  - Andere Durchschnittsbildung als bei der mittleren quadratischen Abweichung. Es gilt: $s \geq d_{\bar{x}}$

Beispiel (Fortsetzung): Dividenden (I)

Zur Berechnung der Varianz vom Merkmal „Dividende“ (x_1 = 35, x_2 = 20, x_3 = 30, x_4 = 15) wird das bereits ermittelte arithmetische Mittel von 25 benötigt:

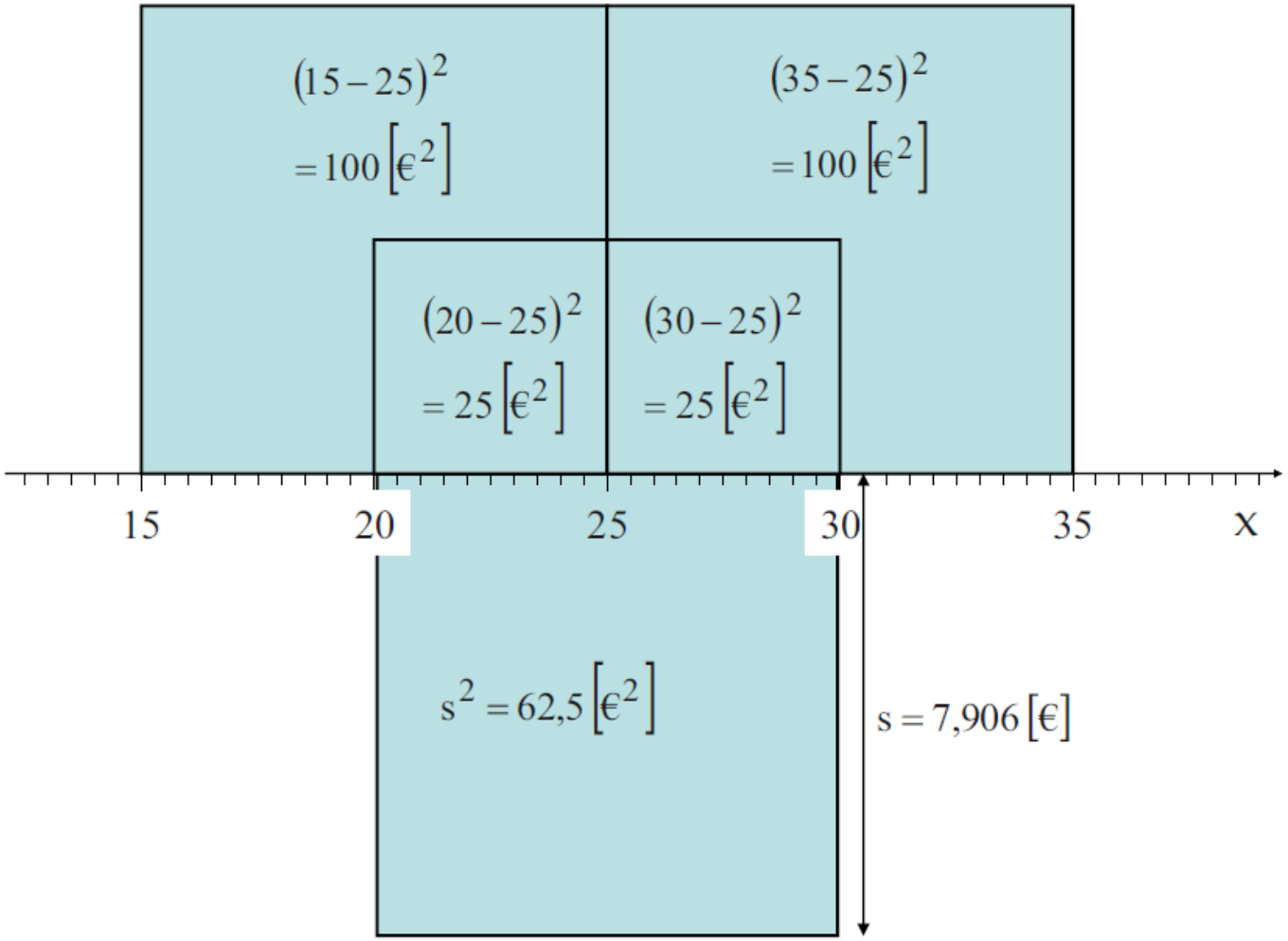
$$
\begin{aligned}
s^2 &= \frac{1}{4} \cdot \sum_{i=1}^{4} (x_i - \bar{x})^2 \\
&= \frac{1}{4} [(35-25)^2 + (20-25)^2 + (30-25)^2 + (15-25)^2] \\
&= \frac{1}{4} [(10)^2 + (-5)^2 + (5)^2 + (-10)^2] \\
&= \frac{1}{4} \cdot 250 = 62{,}5[€^2]
\end{aligned}
$$

Standardabweichung: $s = \sqrt{s^2} = \sqrt{62{,}5} = 7{,}91[€]$

Interpretation: Die Dividende weicht durchschnittlich um 7,91 € von der durchschnittlichen Dividende ab.

Beispiel (Fortsetzung): Dividenden (II)

Die Einzelwerte sind in € gemessen, während die Varianz in der Einheit €^{2} vorliegt. Die Varianz lässt sich als Durchschnitt der Abweichungsquadrate interpretieren:



Beispiel (Fortsetzung): Marktforschung

Als weiteres Beispiel sollen Varianz und Standardabweichung der Preise von Heimwerkerartikeln bestimmt werden (Einzelwerte: 310, 320, 330, 380, 400, 410, 420, 430; arithmetisches Mittel: $\bar{x} = 375$):

$$
\begin{aligned}
s^2 &= \frac{1}{8}\sum_{i=1}^{8}(x_i - \bar{x})^2 \\
&= \frac{1}{8}[(310-375)^2 + (320-375)^2 + (330-375)^2 + (380-375)^2 \\
&\quad + (400-375)^2 + (410-375)^2 + (420-375)^2 + (430-375)^2] \\
&= \frac{1}{8}16200 = 2025[€^2]
\end{aligned}
$$

Standardabweichung: $s = \sqrt{s^2} = \sqrt{2025} = 45[€]$

Varianzverschiebungssatz:

- Alternative Berechnungsformel zur Berechnung der Varianz
- Bildung der Abweichungsquadrate hier nicht nötig (Herleitung s. Lehrbuch)
- Formel:

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

oder

$$s^2 = \overline{x^2} - \bar{x}^2 \quad \text{mit} \quad \overline{x^2} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2$$

Beispiel (Fortsetzung): Dividende

$$s^2 = \frac{1}{4} \cdot \sum_{i=1}^{4} x_i^2 - \bar{x}^2 = \frac{1}{4} \cdot (35^2 + 20^2 + 30^2 + 15^2) - 25^2$$
$$= \frac{1}{4} \cdot 2750 - 25^2 = 687{,}5 - 625 = 62{,}5[€^2]$$

Berechnung bei unklassierter Häufigkeitsverteilung

- Originäre Rechenformel:

$$s^2 = \frac{1}{n} \cdot \sum_{j=1}^{m} \left(x_j^* - \bar{x}\right)^2 \cdot n_j = \sum_{j=1}^{m} \left(x_j^* - \bar{x}\right)^2 \cdot h_j$$

- Mit dem Verschiebungssatz:

$$s^2 = \frac{1}{n} \cdot \sum_{j=1}^{m} \left(x_j^*\right)^2 \cdot n_j - \bar{x}^2 = \sum_{j=1}^{m} \left(x_j^*\right)^2 \cdot h_j - \bar{x}^2$$

Beispiel (Fortsetzung): Haushaltsgröße (I)

Für das Merkmals „Haushaltsgröße“ liegt folgende Häufigkeitstabelle vor ($\bar{x} = 1{,}8$):

j	x_j^*	n_j	h_j
1	1	2	0,4
2	2	2	0,4
3	3	1	0,2
∑		n = 5	1

Beispiel (Fortsetzung): Haushaltsgröße (II)

Berechnung der Varianz unter Verwendung der originären Berechnungsformel:

$$s^2 = \frac{1}{5} \cdot \sum_{j=1}^{3} \left(x_j^* - \bar{x}\right)^2 \cdot n_j = \frac{1}{5} \cdot [(1 - 1{,}8)^2 \cdot 2 + (2 - 1{,}8)^2 \cdot 2 + (3 - 1{,}8)^2 \cdot 1]$$
$$= \frac{1}{5}(1{,}28 + 0{,}08 + 1{,}44) = \frac{1}{5} \cdot 2{,}8 = 0{,}56$$

oder:

$$s^2 = \sum_{j=1}^{3} \left(x_j^* - \bar{x}\right)^2 \cdot h_j = [(1 - 1{,}8)^2 \cdot 0{,}4 + (2 - 1{,}8)^2 \cdot 0{,}4 + (3 - 1{,}8)^2 \cdot 0{,}2]$$
$$= 0{,}56$$

Beispiel (Fortsetzung): Haushaltsgröße (III)

Berechnung der Varianz mit Hilfe des Varianzverschiebungssatzes:

$$s^2 = \frac{1}{5} \cdot \sum_{j=1}^{3} \left(x_j^*\right)^2 \cdot n_j - \bar{x}^2 = \frac{1}{5}(1^2 \cdot 2 + 2^2 \cdot 2 + 3^2 \cdot 1) - 1{,}8^2$$
$$= \frac{1}{5} \cdot 19 - 3{,}24 = 3{,}8 - 3{,}24 = 0{,}56$$

oder:

$$s^2 = \sum_{j=1}^{3} \left(x_j^*\right)^2 \cdot h_j - \bar{x}^2 = (1^2 \cdot 0{,}4 + 2^2 \cdot 0{,}4 + 3^2 \cdot 0{,}2) - 1{,}8^2$$
$$= 0{,}4 + 1{,}6 + 1{,}8 - 3{,}24 = 0{,}56$$

Die Standardabweichung beträgt:

$$s = \sqrt{0{,}56} = 0{,}748$$

Die mittlere Abweichung der einzelnen Haushaltsgrößen von der durchschnittlichen Haushaltsgröße beträgt also 0,748.

Berechnung bei klassierten Häufigkeitsverteilungen (I)

- Varianzzerlegung: Gesamtvarianz = externe Varianz + interne Varianz:

$$s^2 = s_{ext}^2 + s_{int}^2$$

- Die **externe Varianz** ist als Streuung zwischen den Klassen zu interpretieren. Sie gibt den Anteil der Gesamtvarianz wieder, der durch die Klassierung erklärt wird.

- Die externe Varianz wird dadurch ermittelt, dass anstelle der Einzelwerte die Klassenmittelwerte eingesetzt werden:

$$\begin{aligned} s_{ext}^2 &= \frac{1}{n}\left[(\bar{x}_1 - \bar{x})^2 + \cdots + (\bar{x}_1 - \bar{x})^2 + \cdots + \left(\bar{x}_p - \bar{x}\right)^2 + \cdots + \left(\bar{x}_p - \bar{x}\right)^2\right] \\ &= \frac{1}{n}\left[(\bar{x}_1 - \bar{x})^2 \cdot n_1 + (\bar{x}_2 - \bar{x})^2 \cdot n_2 + \cdots + \left(\bar{x}_p - \bar{x}\right)^2 \cdot n_p\right] \end{aligned}$$

bzw.

$$s_{ext}^2 = \frac{1}{n} \cdot \sum_{k=1}^{p} (\bar{x}_k - \bar{x})^2 \cdot n_k = \sum_{k=1}^{p} (\bar{x}_k - \bar{x})^2 \cdot h_k$$

- Die externe Varianz stellt somit einen gewogenen Durchschnitt der Abweichungsquadrate des Klassenmittelwertes $\bar{x}_k$ vom arithmetischen Mittel $\bar{x}$ dar.

Berechnung bei klassierten Häufigkeitsverteilungen (II)

- Zur Berechnung der **internen Varianz** werden die Klassenvarianzen herangezogen:

$$s_k^2 = \frac{1}{n_k} \cdot \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2$$

- Speziell wird ein (gewogenes) arithmetisches Mittel der p Klassenvarianzen berechnet:

$$s_{int}^2 = \frac{1}{n} \cdot \sum_{k=1}^{p} s_k^2 \cdot n_k = \sum_{k=1}^{p} s_k^2 \cdot h_k$$

Beispiel (Fortsetzung): Marktforschung (I)

Wie groß ist die Varianz der Verkaufspreise auf Basis der klassierten Daten?

Zuerst wird die externe Varianz berechnet, die einen Großteil der Streuung erfasst. Hierzu können die absoluten oder relativen Klassenhäufigkeiten herangezogen werden:

k	von über… bis zu …	Einzelwerte x_{ik}	n_k	h_k	$\bar{x}_k$
1	300 – 350	310; 320; 330	3	0,375	$\bar{x}_1 = \frac{1}{3} \cdot (310 + 320 + 330) = 320$
2	350 – 400	380; 400	2	0,250	$\bar{x}_2 = \frac{1}{2} \cdot (380 + 400) = 390$
3	400 – 450	410; 420; 430	3	0,375	$\bar{x}_3 = \frac{1}{3} \cdot (410 + 420 + 430) = 420$
∑			n = 8	1	

Beispiel (Fortsetzung): Marktforschung (II)

Externe Varianz (mit absoluten Klassenhäufigkeiten):

$$
\begin{aligned}
s_{ext}^2 &= \frac{1}{n} \cdot \sum_{k=1}^{p} (\bar{x}_k - \bar{x})^2 \cdot n_k = \frac{1}{8} \cdot \sum_{k=1}^{3} (\bar{x}_k - 375)^2 \cdot n_k \\
&= \frac{1}{8} \cdot [(320 - 375)^2 \cdot 3 + (390 - 375)^2 \cdot 2 + (420 - 375)^2 \cdot 3] \\
&= \frac{1}{8} \cdot (9075 + 450 + 6075) = \frac{1}{8} \cdot 15600 = 1950[€^2]
\end{aligned}
$$

Externe Varianz (mit relativen Klassenhäufigkeiten):

$$
\begin{aligned}
s_{ext}^2 &= \sum_{k=1}^{p} (\bar{x}_k - \bar{x})^2 \cdot h_k = \sum_{k=1}^{3} (\bar{x}_k - 375)^2 \cdot h_k \\
&= [(320 - 375)^2 \cdot 0{,}375 + (390 - 375)^2 \cdot 0{,}250 + (420 - 375)^2 \cdot 0{,}375] \\
&= 1134{,}375 + 56{,}25 + 759{,}375 = 1950[€^2]
\end{aligned}
$$

Beispiel (Fortsetzung): Marktforschung (III)

Zur Bestimmung der internen Varianz (Reststreuung) werden die Klassenvarianzen benötigt, die in der folgenden Tabelle ausgewiesen sind:

k	s_k^2
1	$s_1^2 = \frac{1}{3}[(310-320)^2 + (320-320)^2 + (330-320)^2] = \frac{1}{3} \cdot 200 = 66{,}667$
2	$s_2^2 = \frac{1}{2}[(380-390)^2 + (400-390)^2] = \frac{1}{2} \cdot 200 = 100$
3	$s_3^2 = \frac{1}{3}[(410-420)^2 + (420-420)^2 + (430-420)^2] = \frac{1}{3} \cdot 200 = 66{,}667$

Damit erhält man die folgende interne Varianz:

$$s_{int}^2 = \frac{1}{8} \cdot \sum_{k=1}^{3} s_k^2 \cdot n_k = \frac{1}{8}(66{,}667 \cdot 3 + 100 \cdot 2 + 66{,}667 \cdot 3) = 75[€^2]$$

$$s_{int}^2 = \sum_{k=1}^{3} s_k^2 \cdot h_k = 66{,}667 \cdot 0{,}375 + 100 \cdot 0{,}250 + 66{,}667 \cdot 0{,}375 = 75[€^2]$$

Die Gesamtvarianz beträgt $s^2 = s_{ext}^2 + s_{int}^2 = 1950 + 75 = 2025[€^2]$. Das Ergebnis stimmt mit dem aus den Einzelwerten berechneten s^2 überein. Somit ergibt sich auch hier wiederum eine Standardabweichung in Höhe von 45[€].

Unvollständige Informationen:

Daten liegen allein in klassierter Form vor (Einzelwerte und Klassenmittelwerte unbekannt):

- Ersetzen der Klassenmittelwerte $\bar{x}_k$ durch die Klassenmitten m_k
- Keine Berücksichtigung der Restvarianz (interne Varianz)
- Approximative Gesamtvarianz:

$$s_m^2 = \frac{1}{n} \cdot \sum_{k=1}^{p} (m_k - \bar{x}_m)^2 \cdot n_k = \sum_{k=1}^{p} (m_k - \bar{x}_m)^2 \cdot h_k$$

Beispiel (Fortsetzung): Marktforschung (I)

Nehmen wir an, Einzelwerte und Klassenmittelwerte wären für das Merkmal „Verkaufspreise“ unbekannt:

k	von über… bis zu …	n_k	h_k	m_k
1	300 – 350	3	0,375	$m_1 = \frac{1}{2}(300 + 350) = 325$
2	350 – 400	2	0,250	$m_2 = \frac{1}{2}(350 + 400) = 375$
3	400 – 450	3	0,375	$m_3 = \frac{1}{2}(400 + 450) = 425$
∑		n = 8	1	

Das approximative arithmetische Mittel lautet dann:

$$\bar{x}_m = \frac{1}{n} \cdot \sum_{k=1}^{3} m_k \cdot n_k = \frac{1}{8} \cdot (325 \cdot 3 + 375 \cdot 2 + 425 \cdot 3) = \frac{1}{8} \cdot 3000 = 375[€]$$

$$\bar{x}_m = \sum_{k=1}^{3} m_k \cdot h_k = 325 \cdot 0{,}375 + 375 \cdot 0{,}250 + 425 \cdot 0{,}375 = 375[€]$$

Beispiel (Fortsetzung): Marktforschung (II)

Man erhält damit folgende approximative Varianz, die sich von dem genauen Wert aus dem vorherigen Beispiel unterscheidet:

$$s_m^2 = \frac{1}{n} \cdot \sum_{k=1}^{p} (m_k - \bar{x}_m)^2 \cdot n_k = \frac{1}{8} \cdot \sum_{k=1}^{3} (m_k - 375)^2 \cdot n_k$$
$$= \frac{1}{8} \cdot [(325 - 375)^2 \cdot 3 + (375 - 375)^2 \cdot 2 + (425 - 375)^2 \cdot 3]$$
$$= \frac{1}{8} \cdot (7500 + 0 + 7500) = \frac{1}{8} \cdot 15000 = 1875[€^2]$$

$$s_m^2 = \sum_{k=1}^{p} (m_k - \bar{x}_m)^2 \cdot h_k = \sum_{k=1}^{3} (m_k - 375)^2 \cdot h_k$$
$$= [(325 - 375)^2 \cdot 0{,}375 + (375 - 375)^2 \cdot 0{,}250 + (425 - 375)^2 \cdot 0{,}375]$$
$$= 937{,}5 + 0 + 937{,}5 = 1875[€^2]$$

Hieraus ergibt sich die folgende approximative Standardabweichung:

$$s_m = \sqrt{s_m^2} = \sqrt{1875} = 43{,}30[€]$$

Beispiel (Fortsetzung): Arbeitslosigkeit

Für das Merkmal „Dauer der Arbeitslosigkeit“ wurden Klassenmitten und das approximative arithmetische Mittel bereits berechnet:

k	von über… bis zu …	n_k	h_k	m_k	$m_k \cdot h_k$
1	0 – 3	300	0,30	½(0+3) = 1,5	1,5·0,3 = 0,450
2	3 – 6	250	0,25	½(3+6) = 4,5	4,5·0,25 = 1,125
3	6 – 12	250	0,25	½(6+12) = 9	9·0,25 = 2,250
4	12 – 24	200	0,20	½(12+24) = 18	18·0,2 = 3,600
∑		n = 1000	1		$\bar{x}_m = 7{,}425$

Die approximative Varianz beträgt:

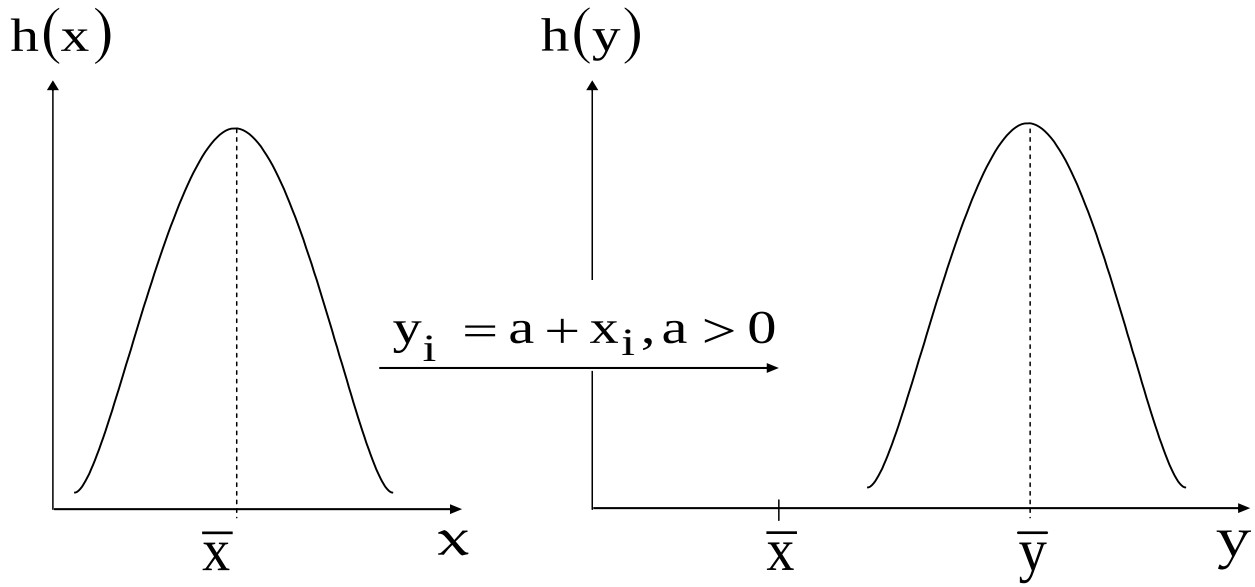
k	n_k	h_k	$(m_k - \bar{x}_m)^2 \cdot n_k = (m_k - 7{,}425)^2 \cdot n_k$	$(m_k - \bar{x}_m)^2 \cdot h_k$
1	300	0,30	10531,6875	10,5317
2	250	0,25	2138,9063	2,1389
3	250	0,25	620,1563	0,6202
4	200	0,20	22366,1250	22,3661
∑	n = 1000	1	35656,875	$s_m^2 = 35{,}657[Monate^2]$

Hieraus erhält man die approximative Standardabweichung in Höhe von:

$$s_m = \sqrt{s_m^2} = \sqrt{35{,}657} = 5{,}971[Monate]$$

Eigenschaften von Varianz und Standardabweichung

- Lineartransformation: $y_i = a + b \cdot x_i$
- Lineare Transformation mit b=1



Die Varianz ist invariant gegenüber einer Verschiebung des Nullpunktes. Wenn sich alle Merkmalswerte um den Faktor b vervielfachen, erhöht sich die Varianz um den Faktor b^2 und die Standardabweichung um den Faktor $|b|$.

$$s_y^2 = b^2 \cdot s_x^2 \qquad (b \neq 0)$$

$$s_y = |b| \cdot s_x \qquad (b \neq 0)$$

Beispiel (Fortsetzung): Tarifverhandlungen (I)

Ein Unternehmer hat drei Angestellte, die 2000€, 2200€, 1500€ monatlich verdienen. Bei Tarifverhandlungen wird eine Lohnerhöhung von 3 % und ein Sockelbetrag von 50 € pro Monat vereinbart. Das arithmetische Mittel vor und nach der Lohnerhöhung wurde bereits berechnet ($\bar{x} = 1900$, $\bar{y} = 2007$). Wie hoch sind Varianz und Standardabweichung nach der Lohnerhöhung?

Erste Möglichkeit:

- Neue Einzelwerte:

$$y_1 = 50 + 1{,}03 \cdot 2000 = 2110$$

$$y_2 = 50 + 1{,}03 \cdot 2200 = 2316$$

$$y_3 = 50 + 1{,}03 \cdot 1500 = 1595$$

- Neue Varianz:

$$s_y^2 = \frac{1}{3} \cdot \sum_{i=1}^{3} (y_i - \bar{y})^2 = \frac{1}{3} \cdot [(2110 - 2007)^2 + (2316 - 2007)^2 + (1595 - 2007)^2]$$
$$= 91944{,}67[€^2]$$

- Neue Standardabweichung:

$$s_y = \sqrt{91944{,}67} = 303{,}22[€]$$

Beispiel (Fortsetzung): Tarifverhandlungen (II)

Zweite Möglichkeit:

- Alte Varianz:

$$s_x^2 = \frac{1}{3} \cdot \sum_{i=1}^{3} (x_i - \bar{x})^2 = \frac{1}{3} \cdot [(2000 - 1900)^2 + (2200 - 1900)^2 + (1500 - 1900)^2]$$
$$= 86666{,}67[€^2]$$

- Neue Varianz:

- Alte Standardabweichung:

$$s_x = \sqrt{86666{,}67} = 294{,}39[€]$$

- Neue Standardabweichung:

5.1.5 Variationskoeffizient

- Problematik bei **Vergleich der Streuung** zweier Merkmale: Abhängigkeit der behandelten absoluten Streuungsmaße von der Maßeinheit und Größenordnung der Merkmalswerte.
- Bei unterschiedlicher Maßeinheit und/oder unterschiedlicher Größenordnung erfolgt ein Vergleich von Streuungen daher auf der Basis eines **relativen Streuungsmaßes**.

Beispiel (Fortsetzung): Tarifverhandlungen

Die Angestelltengehälter liegen in € vor: 2000€, 2200€, 1500€. Wie hoch wäre die Varianz, wenn die Löhne in Dollar (Umrechnungskurs: 1,10$ = 1€) gemessen werden?

- Anzuwenden ist eine Lineartransformation (a = 0 und b = 1,10).
- Die Varianz erhöht sich hier von $s_x^2 = 86666{,}67[€^2]$ auf:

$$s_y^2 = b^2 \cdot s_x^2 = 1{,}10^2 \cdot 86666{,}67 = 104866{,}67[\$^2]$$

- Die Standardabweichung von $s_x = 294{,}39[€]$ erhöht sich auf:

$$s_y = |b| \cdot s_x = 1{,}10 \cdot 294{,}39 = 323{,}83[\$]$$

- Grund hierfür ist der Übergang von der Währungseinheit Euro (€) in die höher bewertete Währungseinheit Dollar ($).

Absolute und relative Streuungsmaße

Streuungsmaße

Absolut: Abhängig von den Einheiten, in denen Merkmale gemessen werden	**Relativ**: Unabhängig von den Einheiten, in denen Merkmale gemessen werden (**dimensionslos**); prozentual interpretierbar
Spannweite	Variationskoeffizient
Quartilsabstand	
Mittlere absolute Abweichung	
Varianz/Standardabweichung	

Variationskoeffizient:

$$v = \frac{s}{\bar{x}}$$

Variationskoeffizienten sind dimensionslos und prozentual interpretierbar

Beispiel (Fortsetzung): Dividende

Für das Merkmal „Dividende“ beträgt der Variationskoeffizient:

$$\mathrm{v} = \frac{\mathrm{s}}{\bar{\mathrm{x}}} = \frac{\sqrt{62{,}5}}{25} = 0{,}316$$

Die Dividenden weichen im Mittel um 31,6 % von ihrem Durchschnittswert ab.

Beispiel: Elektronikgeräte

In den Elektronikgeschäften einer Stadt sind in einem Monat CD-Player zu einem Durchschnittspreis von 200€ bei einer Standardabweichung von 100€ abgesetzt worden. Bei Farbfernsehgeräten ist dagegen ein Durchschnittspreis von 1000€ bei einer Standardabweichung von 100€ erzielt worden.
Obwohl die absolute Streuung bei beiden Verteilungen gleich groß ist, haben gleich große Preisdifferenzen bei CD-Playern und Farbfernsehgeräten doch eine unterschiedliche Bedeutung. So entspricht eine Preisdifferenz von 100€ vom mittleren Preisniveau nach unten bei den Schallplattenspielern einer Halbierung des Durchschnittspreises, wohingegen der gleiche Tatbestand bei den Farbfernsehgeräten nur eine 10%-ige Reduzierung des mittleren Preises bedeutet. Somit besitzt die Häufigkeitsverteilung der Preise von CD-Playern eine größere relative Streuung als diejenige der Preise von Farbfernsehgeräten. Für den Vergleich der Streuungen ist deshalb der Variationskoeffizient heranzuziehen.

CD-Player	Farbfernsehgeräte
$v = {}^{s}/_{\bar{x}} = {}^{100[€]}/_{200[€]} = 0{,}5$	$v = {}^{s}/_{\bar{x}} = {}^{100[€]}/_{1000[€]} = 0{,}1$

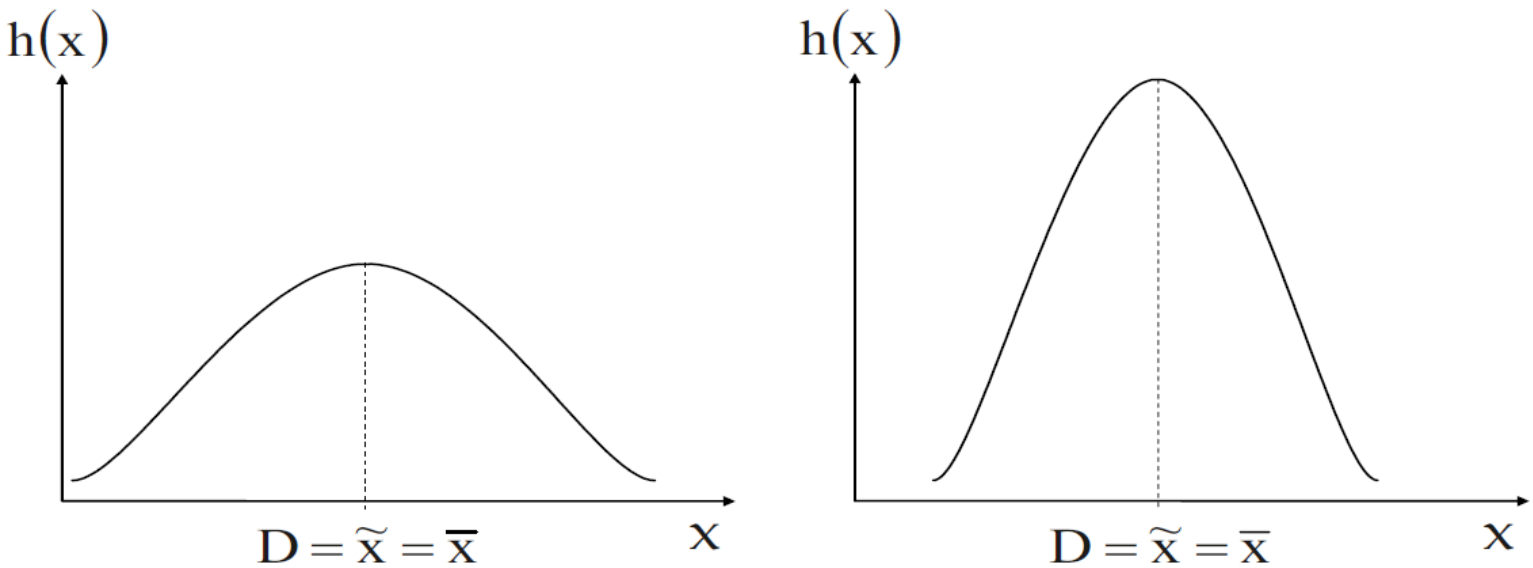
5.2 Schiefe

- Häufigkeitsverteilungen können bei gleichem Mittelwert und gleicher Varianz eine unterschiedliche Gestalt besitzen.
- Wenn sich die Merkmalswerte gleichförmig um den Mittelwert verteilen, liegt eine symmetrische Häufigkeitsverteilung vor.
- Andernfalls ist die Häufigkeitsverteilung asymmetrisch oder schief.
- Mit Hilfe eines Schiefemaßes sollen Richtung und Grad der Schiefe gemessen werden.
- Der Modus D gibt den Gipfel einer Häufigkeitsverteilung an.

Symmetrische Verteilung

Bei einer symmetrischen Verteilung stimmen Modus, Median und arithmetisches Mittel überein:

$$D = \tilde{x} = \bar{x} \quad \Rightarrow \quad \text{symmetrische Verteilung}$$



a) relativ flach

b) relativ steil

Asymmetrische Verteilung (I)

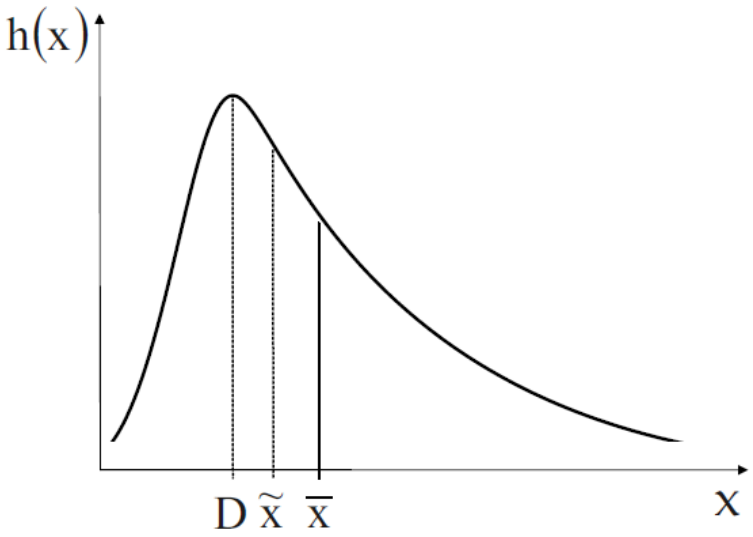
- Liegt eine asymmetrische Verteilung vor, dann sind die genannten Lagemaße nicht gleich groß.
- Bei einer **rechtsschiefen Häufigkeitsverteilung** ist der Median größer als der Modus, weil die 50 %-Trennmarke hinter dem Gipfel der Häufigkeitsverteilung liegt. Das arithmetische Mittel nimmt einen noch größeren Wert als der Median an, weil es die quadrierten Abweichungen und nicht wie der Median die einfachen Abweichungen minimiert. Somit werden bei seiner Berechnung Merkmalswerte, die weit vom Zentrum entfernt liegen, stärker berücksichtigt.

$$D < \tilde{x} < \bar{x} \quad \Rightarrow \quad \text{(linkssteile) rechtsschiefe Verteilung}$$

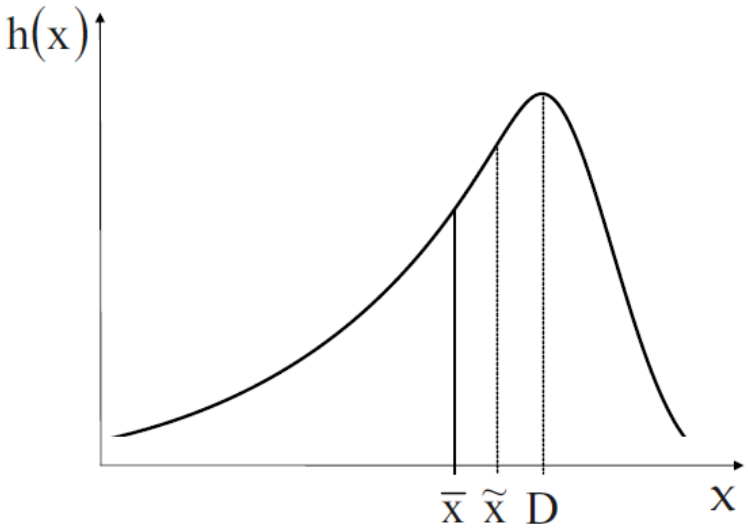
- Auf die gleiche Weise lässt sich begründen, dass bei einer **linksschiefen Häufigkeitsverteilung** das arithmetische Mittel am kleinsten und der Modus am größten sind:

$$D > \tilde{x} > \bar{x} \quad \Rightarrow \quad \text{(rechtssteile) linksschiefe Verteilung}$$

Asymmetrische Verteilung (II)



a) rechtsschief (linkssteil)



b) linksschief (rechtssteil)

Fechnersche Lageregel

- Diese Beziehungen werden auch als **Fechnersche Lageregel** bezeichnet.
- Weicht eine Häufigkeitsverteilung nur geringfügig von der Symmetrie ab, dann kann mit der Fechnerschen Lageregel möglicherweise keine Entscheidung über die Schiefe getroffen werden.
- Auf Basis der Fechnerschen Lageregel lassen sich zwei Schiefemaße konstruieren, die bei einem positiven Wert eine Rechtsschiefe und bei einem negativen Koeffizienten eine Linksschiefe ausweisen:

$$g_p = \frac{\bar{x} - D}{s} \quad \Rightarrow \quad \textbf{Schiefemaß von Pearson}$$

$$g_Y = \frac{3 \cdot (\bar{x} - \tilde{x})}{s} \quad \Rightarrow \quad \textbf{Schiefemaß von Yule}$$

Beispiel (Fortsetzung): Arbeitslosigkeit (I)

k	von über… bis zu …	h_k	b_k	d_k	H_k	m_k	$m_k \cdot h_k$
1	0 – 3	0,30	3	0,100	0,30	1,5	0,450
2	3 – 6	0,25	3	0,083	0,55	4,5	1,125
3	6 – 12	0,25	6	0,042	0,80	9	2,250
4	12 – 24	0,20	12	0,017	1	18	3,600
∑		1					$\bar{x}_m = 7{,}425$

- Modus:
  - $k^* = 1$ (erste Klasse weist die maximale Häufigkeitsdichte auf)
  - Klassierte Daten: Modus wird durch Klassenmitte der Modalklasse repräsentiert:

$$D = m_1 = 1{,}5$$

- Arithmetisches Mittel:

$$\bar{x}_m = \sum_{k=1}^{p} m_k \cdot h_k = 7{,}425$$

Beispiel (Fortsetzung): Arbeitslosigkeit (II)

k	von über… bis zu …	h_k	b_k	d_k	H_k	m_k	$m_k \cdot h_k$
1	0 – 3	0,30	3	0,100	0,30	1,5	0,450
2	3 – 6	0,25	3	0,083	0,55	4,5	1,125
3	6 – 12	0,25	6	0,042	0,80	9	2,250
4	12 – 24	0,20	12	0,017	1	18	3,600
∑		1					$\bar{x}_m = 7{,}425$

- Median:
  - $k^* = 2$ (die H_k durchlaufen in der zweiten Klasse die 50%-Marke)

$$\tilde{x} = x'_{k^*-1} + \frac{0{,}5 - H_{k^*-1}}{h_{k^*}} \cdot b_{k^*} = 3 + \frac{0{,}5 - 0{,}3}{0{,}25} \cdot 3 = 5{,}4$$

- Standardabweichung (siehe Folie 34):

$$s_m = \sqrt{s_m^2} = \sqrt{35{,}657} = 5{,}971$$

Beispiel (Fortsetzung): Arbeitslosigkeit (II)

k	von über… bis zu …	h_k	b_k	d_k	H_k	m_k	$m_k \cdot h_k$
1	0 – 3	0,30	3	0,100	0,30	1,5	0,450
2	3 – 6	0,25	3	0,083	0,55	4,5	1,125
3	6 – 12	0,25	6	0,042	0,80	9	2,250
4	12 – 24	0,20	12	0,017	1	18	3,600
∑		1					$\bar{x}_m = 7{,}425$

- Fechnersche Lageregel:

$$(D = 1{,}5) < (\tilde{x} = 5{,}4) < (\bar{x}_m = 7{,}425) \quad \Rightarrow \quad \text{rechtsschiefe Verteilung}$$

- Schiefemaß von Pearson:

$$g_p = \frac{\bar{x}_m - D}{s} = \frac{7{,}425 - 1{,}5}{5{,}971} = 0{,}992 \;\Rightarrow\; \text{rechtsschiefe Verteilung}$$

- Schiefemaß von Yule:

$$g_Y = \frac{3 \cdot (\bar{x}_m - \tilde{x})}{s} = \frac{3 \cdot (7{,}425 - 5{,}4)}{5{,}971} = 1{,}017 \;\Rightarrow\; \text{rechtsschiefe Verteilung}$$

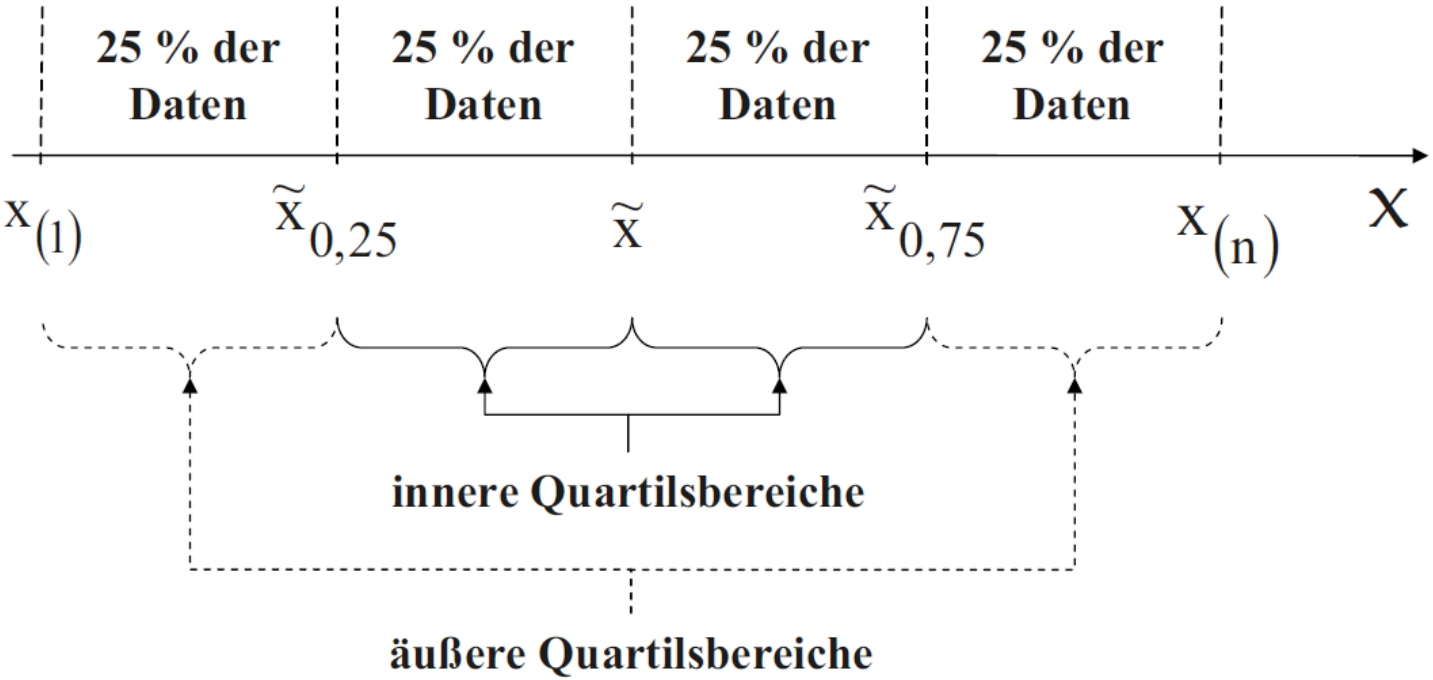
Quartilsregel (I)

- Neben den Mittelwerten können auch die Quartile zur Bestimmung der Schiefe einer Häufigkeitsverteilung herangezogen werden.
- Mit Hilfe der Quartile lassen sich Quartilsbereiche abgrenzen. Speziell wird eine Quartilsregel angewendet, die die beiden inneren und äußeren Quartilsbereiche miteinander vergleicht.
- Sind der rechte äußere und innere Quartilsbereich größer als die entsprechenden linken Quartilsbereiche, dann ist die Häufigkeitsverteilung rechts flacher als links, also rechtsschief oder linkssteil.
- Im umgekehrten Fall liegt eine linksschiefe oder rechtssteile Häufigkeitsverteilung vor.

$$\left(\tilde{x}_{0,25} - x_{(1)}\right) < \left(x_{(n)} - \tilde{x}_{0,75}\right) \text{ und } \left(\tilde{x} - \tilde{x}_{0,25}\right) < \left(\tilde{x}_{0,75} - \tilde{x}\right) \Rightarrow \text{rechtsschiefe Verteilung}$$

$$\left(\tilde{x}_{0,25} - x_{(1)}\right) > \left(x_{(n)} - \tilde{x}_{0,75}\right) \text{ und } \left(\tilde{x} - \tilde{x}_{0,25}\right) > \left(\tilde{x}_{0,75} - \tilde{x}\right) \Rightarrow \text{linksschiefe Verteilung}$$

Quartilsregel (II)

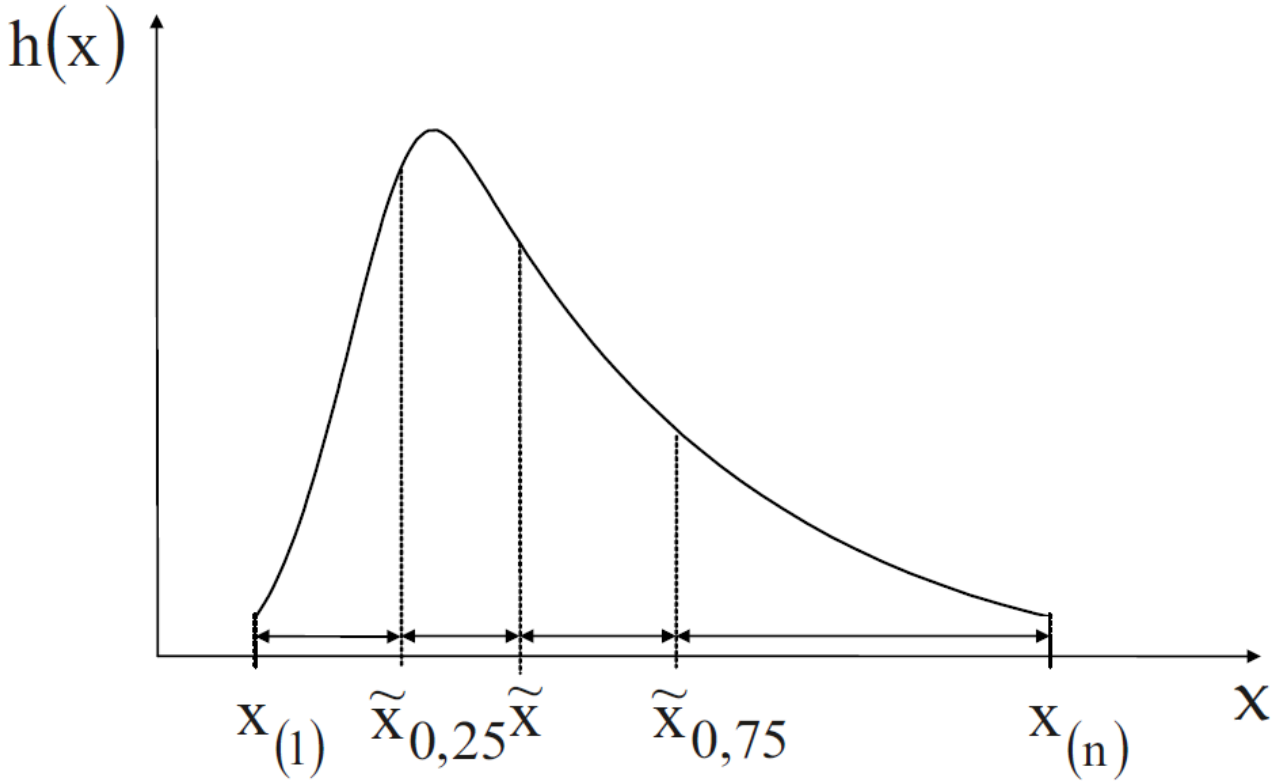


Die Quartilsregel versagt allerdings, wenn bei einem Vergleich der linke und beim anderen Vergleich der rechte Quartilsbereich größer ist. In diesem Fall kann keine Aussage über die Schiefe getroffen werden.

Quartilskoeffizient

Indem man die Differenz der beiden inneren Quartilsbereiche auf den Quartilsabstand bezieht, erhält man den Quartilskoeffizienten, der bei einem positiven (negativen) Wert auf eine rechtsschiefe (linksschiefe) Verteilung hinweist:

$$g_Q = \frac{\left(\tilde{x}_{0,75} - \tilde{x}\right) - \left(\tilde{x} - \tilde{x}_{0,25}\right)}{\tilde{x}_{0,75} - \tilde{x}_{0,25}}$$



Beispiel (Fortsetzung): Pharmakonzerne

Es liegen folgende Umsatzdaten von 10 Pharmakonzernen vor: 17,3; 16,6; 14,7; 27,0; 17,2; 21,6; 13,6; 40,3; 12,4; 11,7.

Das erste Quartil lautet $\tilde{x}_{0,25} = 13{,}6$, das zweite Quartil (= Median) $\tilde{x} = 16{,}9$ und das dritte Quartil $\tilde{x}_{0,75} = 21{,}6$.

- Quartilsregel: rechtsschiefe Verteilung, da

$$\left(\tilde{x}_{0,25} - x_{(1)} = 13{,}6 - 11{,}7 = 1{,}9\right) < \left(x_{(10)} - \tilde{x}_{0,75} = 40{,}3 - 21{,}6 = 18{,}7\right)$$

und

$$\left(\tilde{x} - \tilde{x}_{0,25} = 16{,}9 - 13{,}6 = 3{,}3\right) < \left(\tilde{x}_{0,75} - \tilde{x} = 21{,}6 - 16{,}9 = 4{,}7\right)$$

- Quartilskoeffizient: rechtsschiefe Verteilung, da

$$g_Q = \frac{\left(\tilde{x}_{0,75} - \tilde{x}\right) - \left(\tilde{x} - \tilde{x}_{0,25}\right)}{\tilde{x}_{0,75} - \tilde{x}_{0,25}} = \frac{(21{,}6 - 16{,}9) - (16{,}9 - 13{,}6)}{21{,}6 - 13{,}6} = 0{,}175$$

Schiefekoeffizient

- Eine vollständige Auswertung der in den Daten enthaltenen Informationen bietet der Schiefekoeffizient (Momentschiefe), der als Quotient aus dem dritten zentralen Moment und der in die dritte Potenz erhobenen Standardabweichung definiert ist:

$$g = \frac{m_3(\bar{x})}{s^3}$$

- Die zentralen Momente stellen eine Verallgemeinerung der Varianz dar. Sie werden bei Einzelwerten folgendermaßen bestimmt:

$$m_k(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

- Da das dritte zentrale Moment sensitiv auf weit entfernt von der Mitte der Verteilung liegende Beobachtungen reagiert, stellen sich in der Regel die gewünschten Effekte ein.
- Bei einer rechtsseitig langschweifigen Häufigkeitsverteilung überwiegen tendenziell die positiven Beiträge zum dritten zentralen Moment und bei einer linksseitig langschweifigen Verteilung dominieren die negativen Beiträge.
- Da das Vorzeichen von g nicht durch die Standardabweichung beeinflusst wird, nimmt der Schiefekoeffizient im ersteren Fall (Rechtsschiefe) einen positiven und im letzteren Fall einen negativen Wert an.

- Für symmetrische Verteilungen ist g gleich null.

Beispiel (Fortsetzung): Pharmakonzerne

- Arithmetisches Mittel:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 19{,}24$$

- Varianz und Standardabweichung:

$$s^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x}) = 68{,}026 \quad \text{und} \quad s = \sqrt{68{,}026} = 8{,}248$$

- Drittes zentrales Moment:

$$m_3(\bar{x}) = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^3 = 876{,}519$$

- Schiefekoeffizient:

$$g = \frac{m_3(\bar{x})}{s^3} = \frac{876{,}519}{8{,}248^3} = 1{,}562 \quad \Longrightarrow \quad \text{Rechtsschiefe}$$