Lagemaße als Lösung eines Optimierungsproblems

Gegeben sei die Urliste $x_1, \ldots, x_n$ eines intervallskalierten Merkmals X, die zu einer Zahl a^* zusammengefasst werden soll. Man könnte sagen, der beste Wert a^* ist derjenige, der den Gesamtabstand zwischen a^* und den Daten minimiert.

Misst man den Abstand

quadratisch	$(x - a^*)^2$	so ergibt sich	$a^* = \bar{x}$
linear durch den Absolutbetrag	$\lvert x - a^* \rvert$	so ergibt sich	$a^* = x_{med}$

Für alle anderen Werte $a \in \mathbb{R}$ gilt:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \leq \sum_{i=1}^{n} (x_i - a)^2,$$

$$\sum_{i=1}^{n} \mid x_i - x_{med} \mid \leq \sum_{i=1}^{n} |x_i - a|.$$

3.1.5 Geometrisches Mittel

Sei $T := \{0, \ldots, n\}$ eine Menge von Zeitpunkten und $B(i) =: b_i$ ein zum Zeitpunkt i erhobenes Merkmal, z.B. das Bruttosozialprodukt.

Für $i = 1, \ldots, n$ heißt

$$x_i = \frac{b_i}{b_{i-1}}$$

der i-te *Wachstumsfaktor* und

$$r_i = \frac{b_i - b_{i-1}}{b_{i-1}} = x_i - 1$$

die i-te *Wachstumsrate*.

Man nennt

$$\bar{x}_{geom} := \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}}$$

das geometrische Mittel der Wachstumsfaktoren $x_1, \ldots, x_n$.

Beispiel : Wirtschaftwachstum gemessen zu drei Zeitpunkten.

i	0		1		2
b_i	1000		1500		750
x_i		1.5		0.5	
r_i		0.5		-0.5	

Geometrisches Mittel der Wachstumsfaktoren:

$$\bar{x}_{geom} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = (x_1 \cdot x_2)^{\frac{1}{2}} = \sqrt{1.5 \cdot 0.5} = \sqrt{0.75} \approx 0.8660$$

Bemerkungen:

- Es gilt $$b_n = b_0 \cdot (\bar{x}_{geom})^n$$
  d.h. $\bar{x}_{geom}$ ist tatsächlich ein durchschnittlicher Wachstumsfaktor, nämlich derjenige Wert, der sich aus b_n und b_0 ergäbe, wenn zu allen Zeitpunkten konstantes Wachstum geherrscht hätte.

- Das geometrische Mittel kann auch zur Prognose (unter der Stabilitätsannahme, dass das durchschnittliches Wachstum gleich bleibt) verwendet werden:

$$b_{n+q} = b_n \cdot (\bar{x}_{geom})^q, \quad q \in \mathbb{N}.$$

- Logarithmieren liefert:

$$\ln \bar{x}_{geom} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i.$$

  Das geometrische Mittel ist also ein arithmetisches Mittel auf der logarithmierten Skala.

- Man kann zeigen:

$$\bar{x}_{geom} \leq \bar{x}$$

  i.A. würde also die Angabe von $\bar{x}$ zu hohe Wachstumsraten vortäuschen.

3.1.6 Harmonisches Mittel

Beispiel: Die Entfernung von A nach B sei 99 km. Herr K. humpelt von A nach B mit konstant 1 km/h und fährt zurück mit konstant 99 km/h. Wie groß ist seine Durchschnittsgeschwindigkeit?

Die naive Lösung: 50 km/h ist falsch!

$$\text{Durchschnittsgeschwindigkeit} = \frac{\text{zurückgelegter Weg}}{\text{Zeit}} = \frac{198 \text{ km}}{100 \text{ h}} = 1.98 \text{ km/h}$$

Allgemein: Sei $x_1, \ldots, x_n$ mit $x_i \neq 0$ für alle i die Urliste eines verhältnisskalierten Merkmals X. Dann heißt

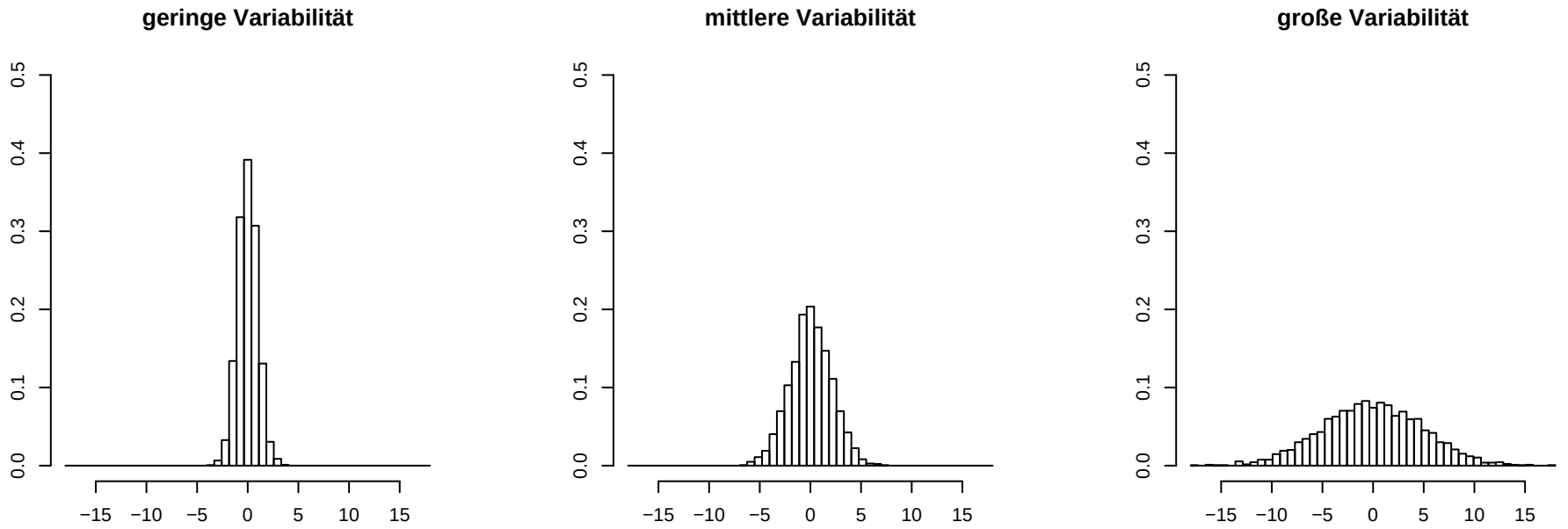
$$\bar{x}_{har} := \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}}$$

das harmonische Mittel der Werte $x_1, \ldots, x_n$.

3.2 Streuungsmaße

Eine Verteilung ist durch die Angabe von einem oder mehreren Lagemaßen nur unzureichend beschrieben.

Beispiel: Häufigkeitsverteilungen mit gleicher zentraler Tendenz:



Streuungsmaße beantworten Fragen wie

- Wie groß ist die durchschnittliche Abweichung vom Mittelwert?
- Über welchen Bereich erstrecken sich die Beobachtungen?
- Wie stark schwanken die Beobachtungen?

Bemerkung:

Von Streuung im eigentlichen Sinne kann man nur bei mindestens intervallskalierten Daten sprechen, da nur dort Abstände interpretierbar sind.

3.2.1 Varianz und Standardabweichung

Sei $x_1, \ldots, x_n$ die Urliste eines intervallskalierten Merkmals X. Dann heißt

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

die (empirische) Varianz oder Stichprobenvarianz und

$$\tilde{s}_X := \sqrt{\tilde{s}_X^2}$$

die empirische Streuung, Stichprobenstreuung oder Standardabweichung von X.

Bemerkungen:

- Die Varianz misst die durchschnittliche quadrierte Abweichung vom Mittelwert.
- Vorsicht: Der Begriff Streuung wird in einem doppelten Sinne gebraucht: Allgemein als Phänomen generell („wir suchen nach Maßzahlen zur Beschreibung der Streuung der Daten“), andererseits als eine bestimmte Maßzahl für das Problem.
- Durch das Quadrieren tragen negative und positive Abweichungen vom Mittelwert gleichermaßen zur Varianz bei.
- Die Varianz besitzt im Vergleich zum Merkmal X die quadrierte Einheit. Sie ist daher unanschaulicher zu interpretieren, besitzt aber andererseits viele mathematische Vorzüge.
- Die Standardabweichung wird in der gleichen Einheit gemessen wie X.

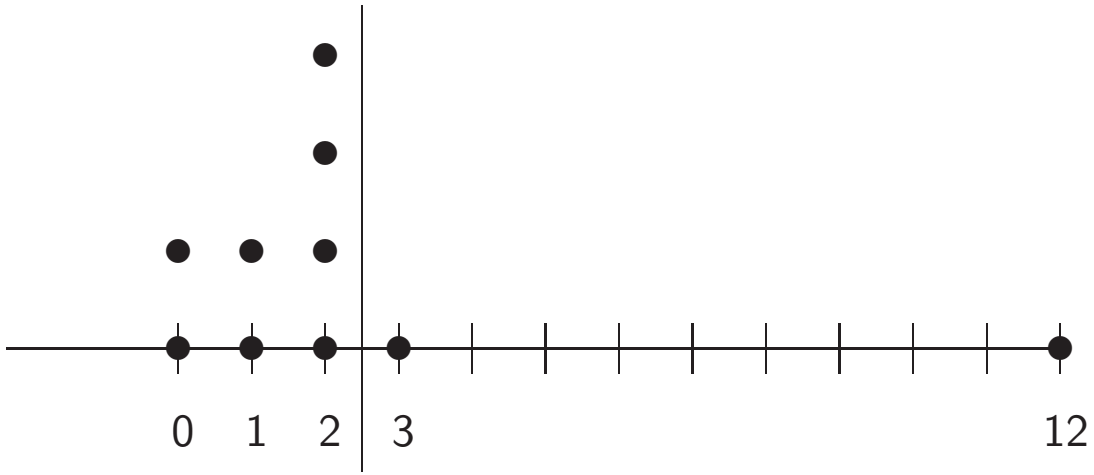
- Sind die Ausprägungen $a_1, \ldots, a_k$ mit (relativer) Häufigkeitsverteilung $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$ gegeben, so gilt

$$\begin{aligned} \tilde{s}_X^2 &= \frac{1}{n} \sum_{j=1}^{k} h_j (a_j - \bar{x})^2 = \\ &= \sum_{j=1}^{k} f_j (a_j - \bar{x})^2. \end{aligned}$$

- Ist aus dem Kontext klar ersichtlich welches Merkmal betrachtet wird, so lässt man das X in der Notation auch häufig weg, schreibt also einfach $\tilde{s}^2$ und $\tilde{s}$.

Beispiel: Statistikbücher.

	Häufigkeiten
$a_1 = 0$	$h_1 = 2$
$a_2 = 1$	$h_2 = 2$
$a_3 = 2$	$h_3 = 4$
$a_4 = 3$	$h_4 = 1$
$a_5 = 12$	$h_5 = 1$



Mittelwert $\bar{x}$=2.5

$x_{0.25}$=1 Median x_{med} =2 $x_{0.75}$=2

Berechnung der Varianz über die ursprüngliche Formel:

$$
\begin{aligned}
\tilde{s}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \\
&= \frac{1}{10}((0-2.5)^2 + (0-2.5)^2 + (1-2.5)^2 + (1-2.5)^2 \\
&\quad +(2-2.5)^2 + (2-2.5)^2 + (2-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (12-2.5)^2) \\
&= \frac{108.5}{10} = 10.85
\end{aligned}
$$

Berechnung über die Häufigkeitsverteilung:

j	a_j	h_j	$(a_j - \bar{x})$	$(a_j - \bar{x})^2$	$(a_j - \bar{x})^2 h_j$
1	0	2	0-2.5=-2.5	6.25	12.5
2	1	2	-1.5	2.25	4.5
3	2	4	-0.5	0.25	1
4	3	1	0.5	0.25	0.25
5	12	1	9.5	90.25	90.25
Summe		10			$n\tilde{s}^2 = 108.5$

$\Longrightarrow \tilde{s}^2 = 10.85$

Standardabweichung:

$$\tilde{s} = \sqrt{10.85} \approx 3.29 \qquad \text{(Einheit: Bücher)}$$

Transformationen: Wie ändert sich die Varianz bei linear affiner Transformation eines Merkmals?

Satz 3.7.

Sei $x_1, \ldots, x_n$ die Urliste eines mindestens intervallskalierten Merkmals X mit $\tilde{s}_X > 0$ und $y_1, \ldots, y_n$ die zugehörige Urliste des Merkmals $Y = a \cdot X + b$.
Dann gilt

$$\tilde{s}_Y^2 = a^2 \cdot \tilde{s}_X^2$$

und

$$\tilde{s}_Y = |a| \cdot \tilde{s}_X.$$

Bemerkungen:

- Die additive Konstante b spielt keine Rolle. Diese bewirkt lediglich eine Verschiebung der Werte, ändert aber nicht die Form und damit das Streuverhalten.
- Vorfaktoren sind bei der Varianz $quadratisch$ „herauszuziehen“.
- Eine spezielle Transformation, die sogenannte $Standardisierung$, ist der Übergang zum Merkmal Z mit

$$z_i := \frac{x_i - \bar{x}}{\tilde{s}_X}.$$

Z besitzt arithmetisches Mittel 0 und (empirische) Varianz 1.

Verschiebungssatz: Es gilt

$$\tilde{s}_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \overline{x^2} - (\bar{x})^2.$$

Achtung (sehr häufige Fehlerquelle):

Der Verschiebungssatz ist sehr bequem zum Berechnen der Varianz. Allerdings können bei sehr großen Ausprägungen beim Verwenden von Taschenrechnern starke Rundungsfehler auftreten, die das Ergebnis eventuell verfälschen.

Beispiel: Statistikbücher.

Berechne die empirische Varianz mit Hilfe des Verschiebungssatzes.

Person i	Anzahl Bücher: X x_i	x_i^2
1	0	0
2	2	4
3	1	
4	2	
5	2	
6	3	
7	0	
8	12	
9	1	
10	2	
Summe	25	
Mittelwert		

$$\tilde{s}_X^2 = \overline{x^2} - (\bar{x})^2 = 17.1 - (2.5)^2 = 10.85, \qquad \tilde{s}_X = 3.29$$

Varianzzerlegung / Streuungszerlegung: Varianz bei geschichteten Daten.

Zur Erinnerung: Daten liegen oft in Schichten vor.
Beispiel: Daten über Einkommensverteilung geschichtet nach Bundesland.
Bei der Berechnung von $\bar{x}$ waren die einzelnen Besetzungszahlen sehr wichtig.

Schicht	$1, \ldots, l, \ldots, z$	
Besetzungszahlen	$n_1, \ldots, n_l, \ldots, n_z;$	$\sum_{l=1}^{z} n_l = n$
Mittelwerte	$\bar{x}_1, \ldots, \bar{x}_l, \ldots, \bar{x}_z$	
Varianzen	$\tilde{s}_1^2, \ldots, \tilde{s}_l^2, \ldots, \tilde{s}_z^2$	

Für das arithmetische Mittel gilt

$$\bar{x} = \frac{1}{n} \sum_{l=1}^{z} n_l \bar{x}_l.$$

Seien nun

$$\tilde{s}^2_{innerhalb} := \frac{1}{n}\sum_{l=1}^{z} n_l \tilde{s}^2_l$$

sowie

$$\tilde{s}^2_{zwischen} := \frac{1}{n}\sum_{l=1}^{z} n_l (\bar{x}_l - \bar{x})^2$$

- $\tilde{s}^2_{innerhalb}$: durchschnittliche Varianz $innerhalb$ der Schichten
- $\tilde{s}^2_{zwischen}$: Varianz der Durchschnittswerte $zwischen$ den Schichten

Wann ist

- $\tilde{s}^2_{zwischen} = 0$?
- $\tilde{s}^2_{innerhalb} = 0$?

Wie setzt sich die Gesamtvarianz aus den beiden Bestandteilen zusammen?

Varianzzerlegung Es gilt

Gesamtvarianz	=	Varianz in. d. Schichten	+	Varianz zw. d. Schichten
$\tilde{s}^2$	=	$\tilde{s}^2_{innerhalb}$	+	$\tilde{s}^2_{zwischen}$

Bemerkungen:

- Im Detail gilt also mit den Urlisten $\{x_{1l}, x_{2l}, \ldots, x_{n_l l}\}$ in Schicht $l, l = 1, \ldots, z$,

$$\frac{1}{n}\sum_{l=1}^{z}\left(\sum_{i=1}^{n_l}(x_{il}-\bar{x})^2\right) = \frac{1}{n}\sum_{l=1}^{z}\sum_{i=1}^{n_l}(x_{il}-\bar{x}_l)^2 + \frac{1}{n}\sum_{l=1}^{z} n_l(\bar{x}_l-\bar{x})^2.$$

- Diese Zerlegungsmöglichkeit gilt *nur für Varianzen*, nicht aber für andere Streuungsmaße. Letztendlich ist sie der Grund für die Beliebtheit der Varianz – trotz anderer Unannehmlichkeiten. Deshalb sollte man eher von der Varianzzerlegung als von der Streuungszerlegung sprechen.
- Bei vielen Verfahren werden Varianzzerlegungen betrachtet; dies ist ein ganz grundlegendes Prinzip in der Statistik.

Interpretation der Varianzzerlegung

Beispiel: Einkommen der einzelnen Bundesländer:

Die Wichtigkeit (Erklärungskraft) der schichtbildenden Variable kann bewertet werden: je größer $\tilde{s}^2_{zwischen}$ im Vergleich zu $\tilde{s}^2$ bzw. $\tilde{s}^2_{innerhalb}$ ist, desto „mehr Variation“ wird durch die Schichtungsvariable erklärt.

Was bedeutet: $\tilde{s}^2_{zwischen}$ groß im Vergleich zu $\tilde{s}^2_{innerhalb}$?

Korrigierte empirische Varianz:

Neben der empirischen Varianz existiert noch eine alternative Definition der Varianz, die korrigierte empirische Varianz.

Sei $x_1, \ldots, x_n$ die Urliste eines intervallskalierten Merkmals X. Dann heißt

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

die *korrigierte empirische Varianz* oder *korrigierte Stichprobenvarianz von* X.

Bemerkungen:

- Der Sinn des Vorfaktors $\frac{1}{n-1}$ wird erst in Statistik II deutlich: s_X^2 hat theoretisch schönere Eigenschaften als $\tilde{s}_X^2$.
- Für großen Stichprobenumfang n nähern sich s_X^2 und $\tilde{s}_X^2$ an, weil dann $n - 1 \approx n$.
- Auch für die korrigierte Varianz gilt die Aussage zu linearen Transformationen, d.h. ist $x_1, \ldots, x_n$ die Urliste eines mindestens intervallskalierten Merkmals X mit $s_X > 0$ und $y_1, \ldots, y_n$ die zugehörige Urliste des Merkmals $Y = a \cdot X + b$. Dann gilt

$$s_Y^2 = a^2 \cdot s_X^2.$$

3.2.2 Weitere Streuungsmaße

Variationskoeffizient:

Definition 3.8.

Ist $\bar{x} > 0$, so heißt die Größe

$$v_X := \frac{\tilde{s}_X}{\bar{x}}$$

Variationskoeffizient des Merkmals X.

Bemerkungen:

- Gemessen wird hier die Streuung relativ zum Mittelwert. Insbesondere ist v_X dimensionslos.
- Der Variationskoeffizient erlaubt beispielsweise auch den Vergleich der Streuung von Preisen, die in verschiedenen Währungen gemessen wurden.

Inter-Quartils-Abstand:

Sind $x_{0.25}$ und $x_{0.75}$ das obere und das untere Quartil eines Merkmals, so heißt

$$d_{QX} := x_{0.75} - x_{0.25}$$

der *Interquartilsabstand*.

Median-Absolute-Deviation:

Der Median der Werte $|x_i - x_{med}|$, $i = 1, \ldots, n$ heißt Median-Absolute-Deviation von X (MAD_X).

Spannweite:

Die Größe

$$R_X := x_{(n)} - x_{(1)}$$

heißt *Spannweite* von X.

Bemerkungen

- Alle betrachteten Streuungsmaße sind nur für (mindestens) intervallskalierte Merkmalse sinnvoll definiert, da sie auf Abständen (typischerweise dem Abstand der Beobachtungen zu einem Lagemaß) beruhen.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind die gebräuchlichsten Streuungsmaße.
- $\tilde{s}^2$, $\tilde{s}$, s^2, s sind sehr empfindlich gegenüber Ausreißern! Das Gleiche gilt für die Spannweite R. MAD und d_Q hingegen entstammen der sogenannten robusten Statistik, die sich um ausreißerresistente Methoden bemüht.
- Gilt $x_1 = x_2 = \ldots = x_n$, so weisen alle Streungsmaße den Wert 0 auf. Mit Ausnahme von d_Q und MAD gilt auch die Umkehrung: Sind die Steuungsmaße (außer eben d_Q und MAD) $= 0$, so sind alle Werte der Urliste gleich.
- Eine häufige Ursache für Verwirrung und Missverständnisse liegt darin, dass der Begriff „Streuung“ in der Statistik in einem doppelten Sinn gebraucht wird:
  – in einem allgemeinen Sinn: Streuung als Phänomen („Die Daten streuen stark“).
  – in einem speziellen Sinn: als $eine$ Maßzahl für dieses Phänomen.

Beispiel: Statistikbücher

Ausprägungen	h_j
0	2
1	2
2	4
3	1
12	1
$\sum$	10

$v_X = \tilde{s}_X/\bar{x} = 3.29/2.5 = 1.316$ (dimensionslos, normiert)

$x_{0.25} = 1 \quad x_{0.75} = 2 \qquad \Rightarrow \quad d_{QX} = 1$

$R_X = 12$

3.3 Box-Plot

Ziele:

- einfache Darstellung von Verteilungen und ihren Kennzahlen
- Identifikation von potentiellen Ausreißern
  $\Rightarrow$ nicht ausreißeranfällige Meßzahlen verwenden.

Idee:

i) markiere den Median

ii) symbolisiere Lage der „mittleren Werte“ durch eine Box

iii) wie weit reichen „weitere nicht atypische“ Werte

iv) identifiziere potentielle Ausreißer: atypische (ungewöhnlich große, ungewöhnlich kleine) Werte, die genauerer Untersuchung bedürfen

zu ii) wähle mittlere 50%: Box hat Länge $dQ = x_{0.75} - x_{0.25}$

zu iii) als „nicht atypisch“ gelten alle Werte, die nicht weiter als $1.5dQ$ von der Box entfernt sind

Also bestimme:

- $x_{0.25}, \quad x_{0.50}, \quad x_{0.75}$.
- Interquartilsabstand: $d_{QX} = x_{0.75} - x_{0.25}$
- Zäune z_u, z_o, die am kleinsten bzw. größten Datenpunkt im Bereich $x_{0.25} - 1.5 \cdot d_{QX}$; $x_{0.75} + 1.5 \cdot d_{QX}$ liegen.
- Ausserhalb der Zäune werden $alle$ Punkte eingezeichnet; sie sind ausreißerverdächtig.

Bemerkungen:

- Der Box-Plot gibt einen kompakten Überblick über die Form der Verteilung (Zentrale Tendenz, Variabilität, Schiefe, extreme Werte).

- Vorsicht bei der Anwendung von Software! Vor allem außerhalb der Box sind auch andere Darstellungen üblich (z.B. Zäune immer bis $x_{(1)}$ und $x_{(n)}$). Manchmal z.B. Unterscheidung zwischen
  - Ausreißern ($1.5 \cdot d_{QX}$ bis $3 \cdot d_{QX}$ von Rändern der Box entfernt) und
  - Extremwerten (mehr als $3 \cdot d_{QX}$ vom Rand entfernt).

  Oft wird der Median durch einen dicken Punkt ausgedrückt.

Box-Plots können auch zum graphischen Vergleich von Verteilungen verwendet werden: