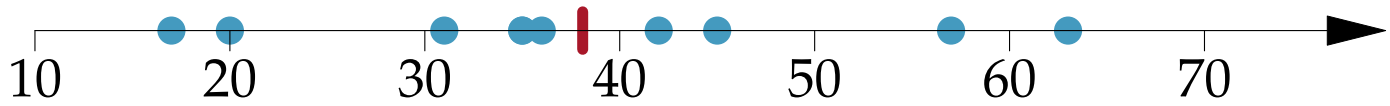
Motivation Streuungsmaße (1)

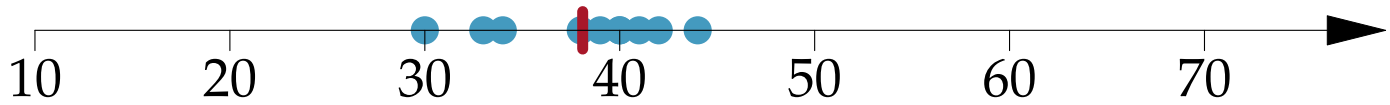
- Lagemaße: beschreiben das Zentrum einer Verteilung durch einen numerischen Wert
- weitere Fragen:
  - Liegen die Daten nahe am Zentrum?
  - Über welchen Bereich erstrecken sich die Beobachtungen?
  - Enthalten sie möglicherweise Ausreißer?
- Streuungsmaße: erfassen die Variabilität in den Beobachtungen

Motivation Streuungsmaße (2)

- Daten A:



- Daten B:



Streuungsmaße

- Eine sinnvolle Interpretation von Lagemaßen ist nur in Zusammenhang mit der Streuung innerhalb der Daten möglich.
- Streuungsmaße vermitteln eine Vorstellung davon, wie repräsentativ die zentralen Lagemaße für die Daten sind.

Spannweite

- **Spannweite** (oder **Range**): Abstand zwischen dem größten und dem kleinsten Beobachtungswert
  - Es gilt:

$$\begin{aligned} R &= x_{(n)} - x_{(1)} \\ &= x_{\max} - x_{\min} \end{aligned}$$

- nur für intervall- und verhältnisskalierte Merkmale berechenbar
- bei ordinalskalierten Merkmalen Angabe von $x_{(1)}$ und $x_{(n)}$
- basiert auf Extremwerten; sehr empfindlich gegenüber Ausreißern

Interquartilsabstand

- Der **Interquartilsabstand (IQR)** ist die Distanz zwischen den 25 %- und 75 %-Quantilen.
  - Es gilt:

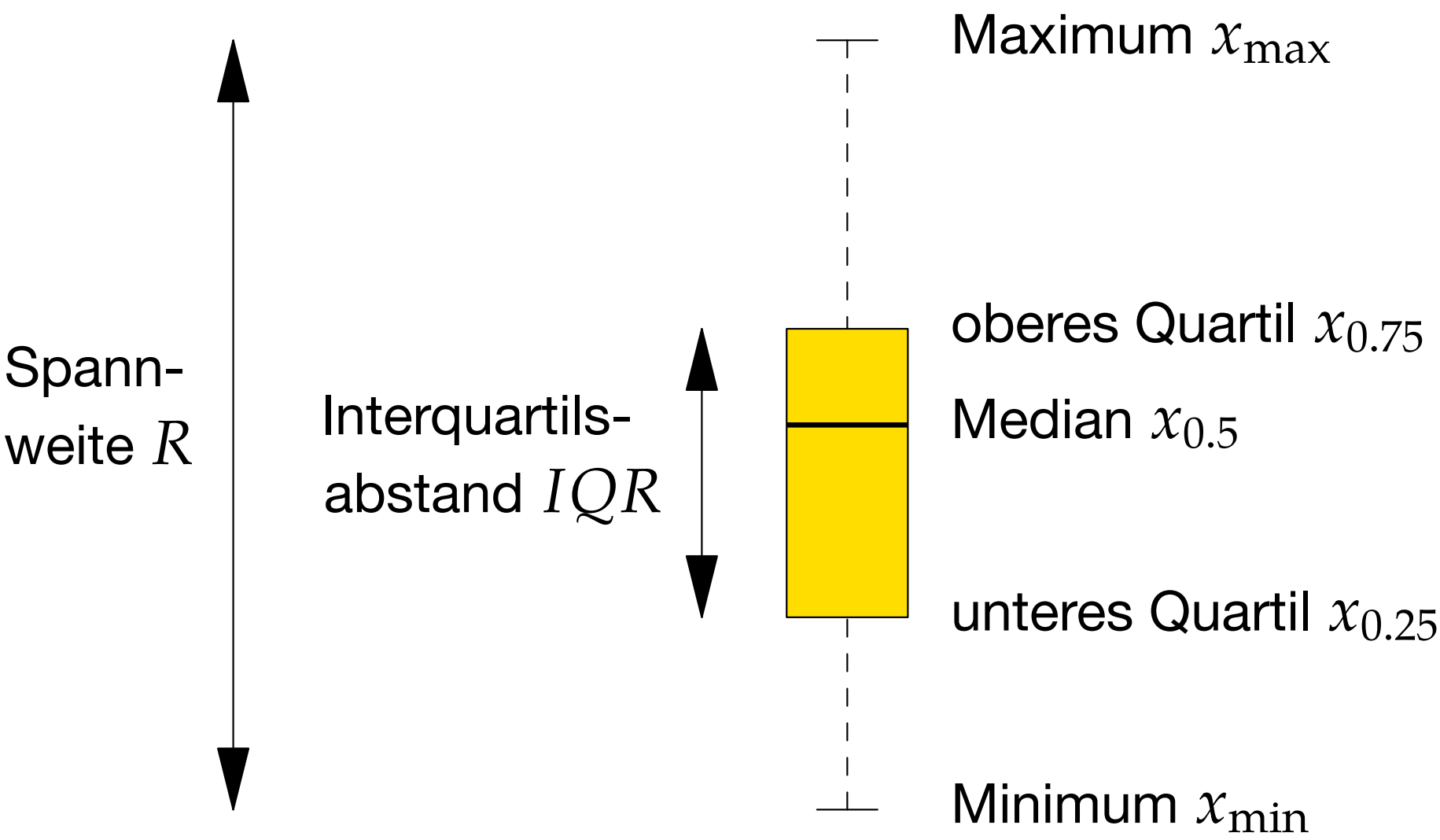
$$IQR = x_{0.75} - x_{0.25}$$

- Ist nur für intervall- und verhältnisskalierte Merkmale berechenbar.
- Bei ordinalskalierten Merkmalen erfolgt die Angabe von $x_{0.25}$ und $x_{0.75}$.
- Ist nicht empfindlich gegenüber Ausreißern.

Box-Plot (1)

- Komprimierte Visualisierung einer Verteilung
  - $x_{0.25}$ ist der Anfang der Schachtel („box“), $x_{0.75}$ das Ende der Schachtel und IQR die Länge der Schachtel.
  - Der Median wird durch einen Punkt oder Strich in der Box markiert.
  - Zwei Linien („whiskers“) außerhalb der Box gehen bis zu $x_{\min}$ und $x_{\max}$.

Box-Plot (2)


Maximum $x_{\max}$
oberes Quartil $x_{0.75}$
Median $x_{0.5}$
unteres Quartil $x_{0.25}$
Minimum $x_{\min}$
Spann-
weite R
Interquartils-
abstand IQR

Bsp. Säulendiagr. und Box-Plots (1)

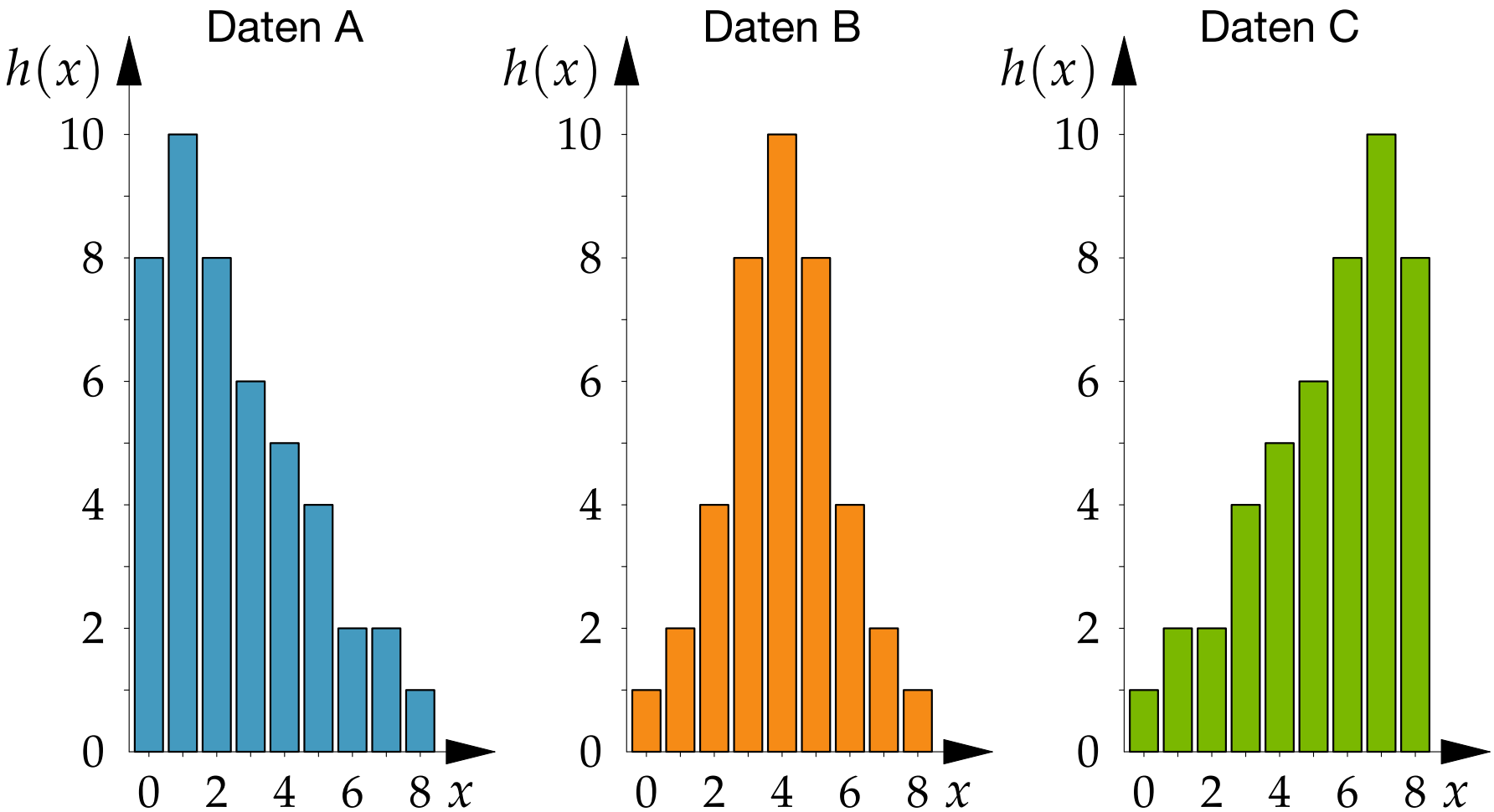


Abb.: Vergleichende Darstellung der Daten A, B und C mittels Säulendiagrammen

Bsp. Säulendiagr. und Box-Plots (2)

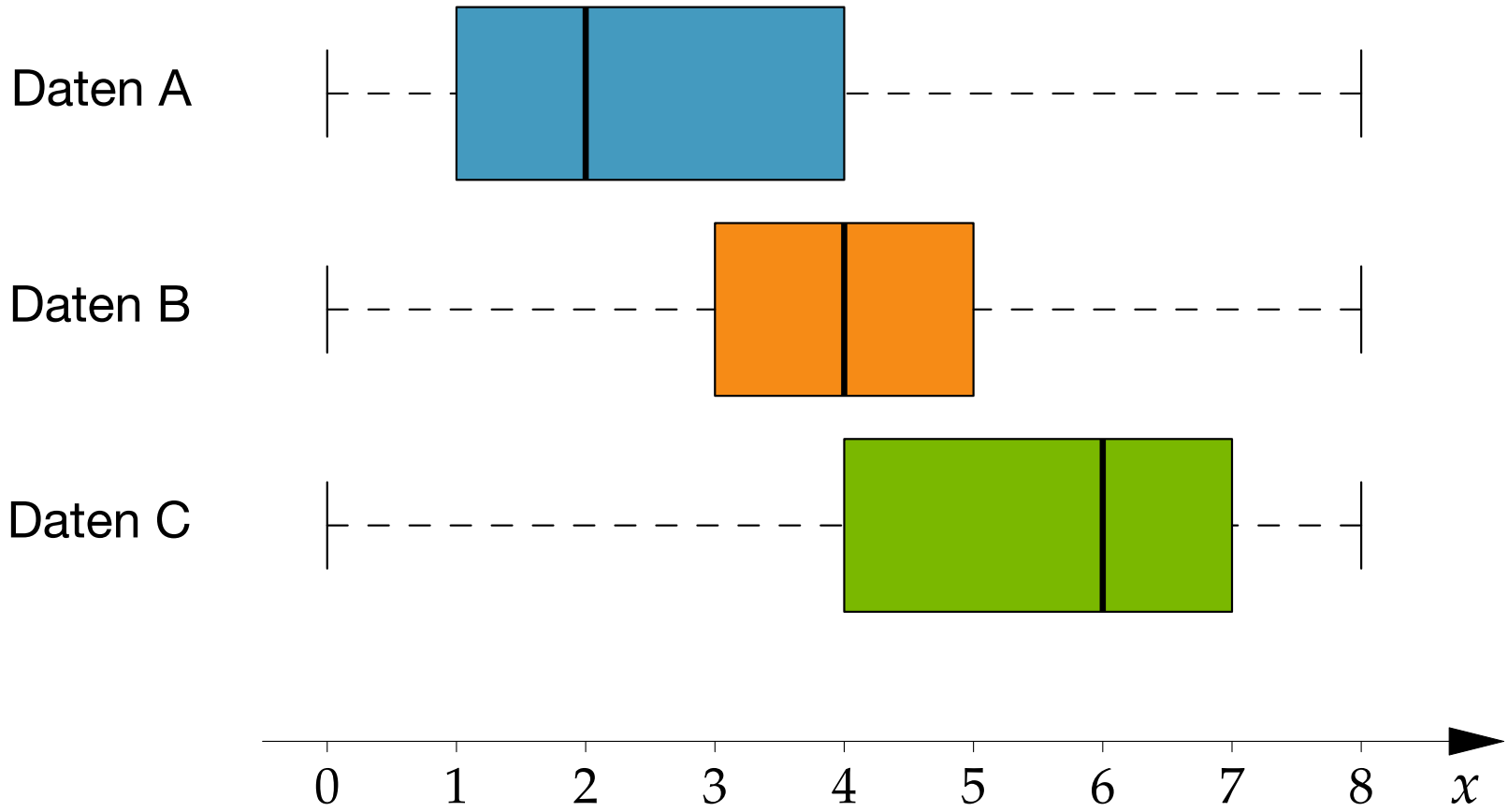
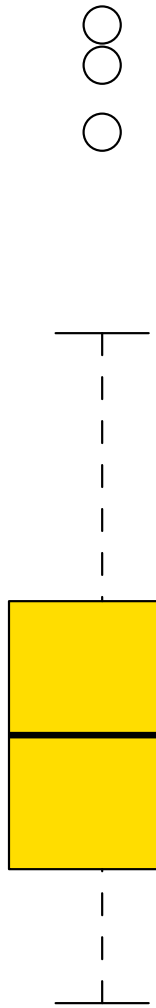


Abb.: Vergleichende Darstellung der Daten A, B und C mittels Box-Plots

Bem. Box-Plot

- Definition für die „whiskers“ des Box-Plots nicht einheitlich
- Definition nach John W. Tukey:
  - Länge der „whiskers“ wird maximal auf das 1.5-fache des Interquartilsabstands ($1.5 \times IQR$) beschränkt. Dabei endet der „whisker“ jedoch nicht genau nach dieser Länge, sondern bei dem Wert aus den Daten, der noch innerhalb dieser Grenze liegt.
  - Werte außerhalb der „whiskers“ werden separat in das Diagramm eingetragen.

Box-Plot nach John W. Tukey



Ausreißer

oberer „whisker“

oberes Quartil $x_{0.75}$

Median $x_{0.5}$

unteres Quartil $x_{0.25}$

unterer „whisker“

Abb.: Modifizierter Box-Plot

Bsp. Box-Plot nach John W. Tukey (1)

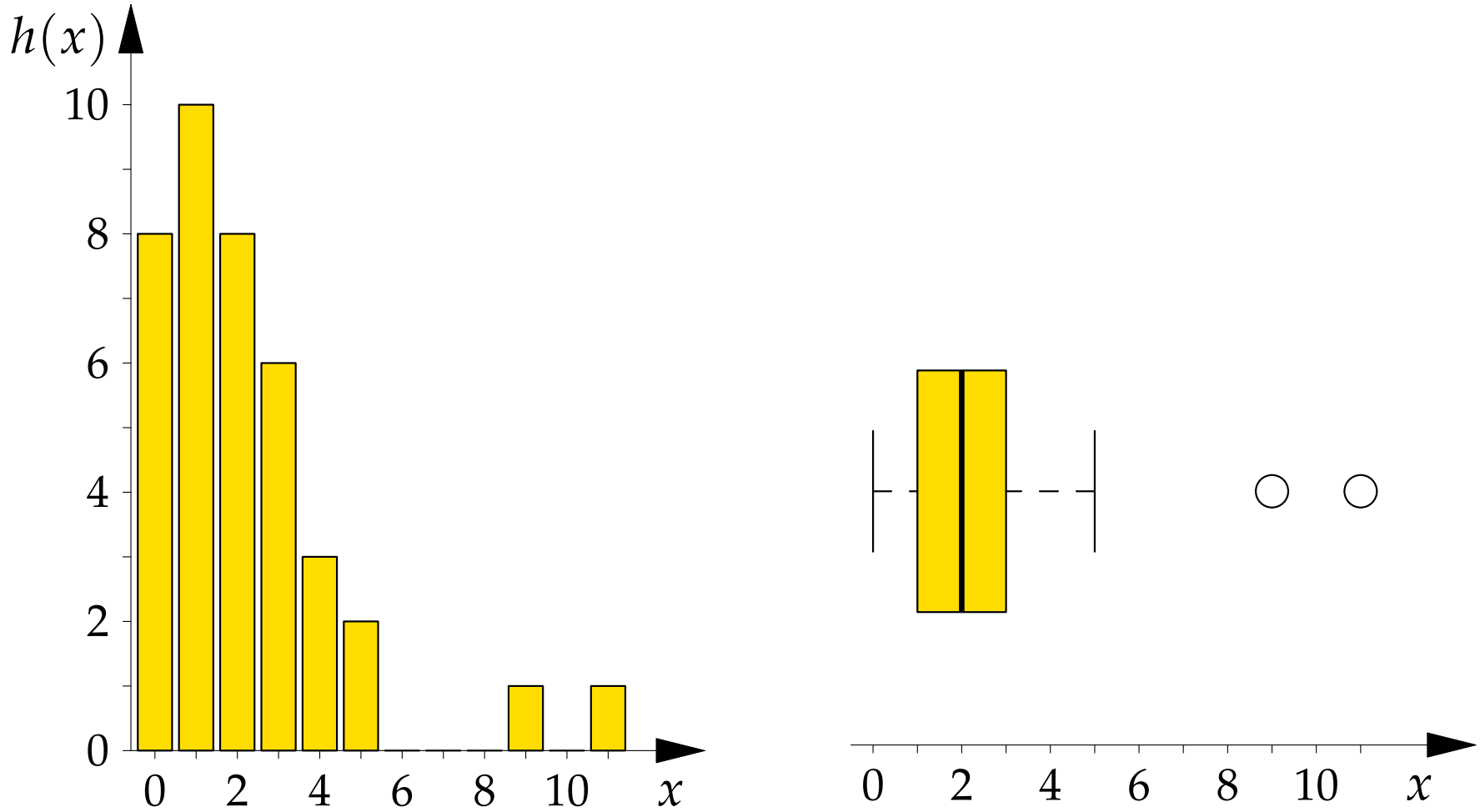


Abb.: Säulendiagramm und Box-Plot der selben Daten

Bsp. Box-Plot nach John W. Tukey (2)

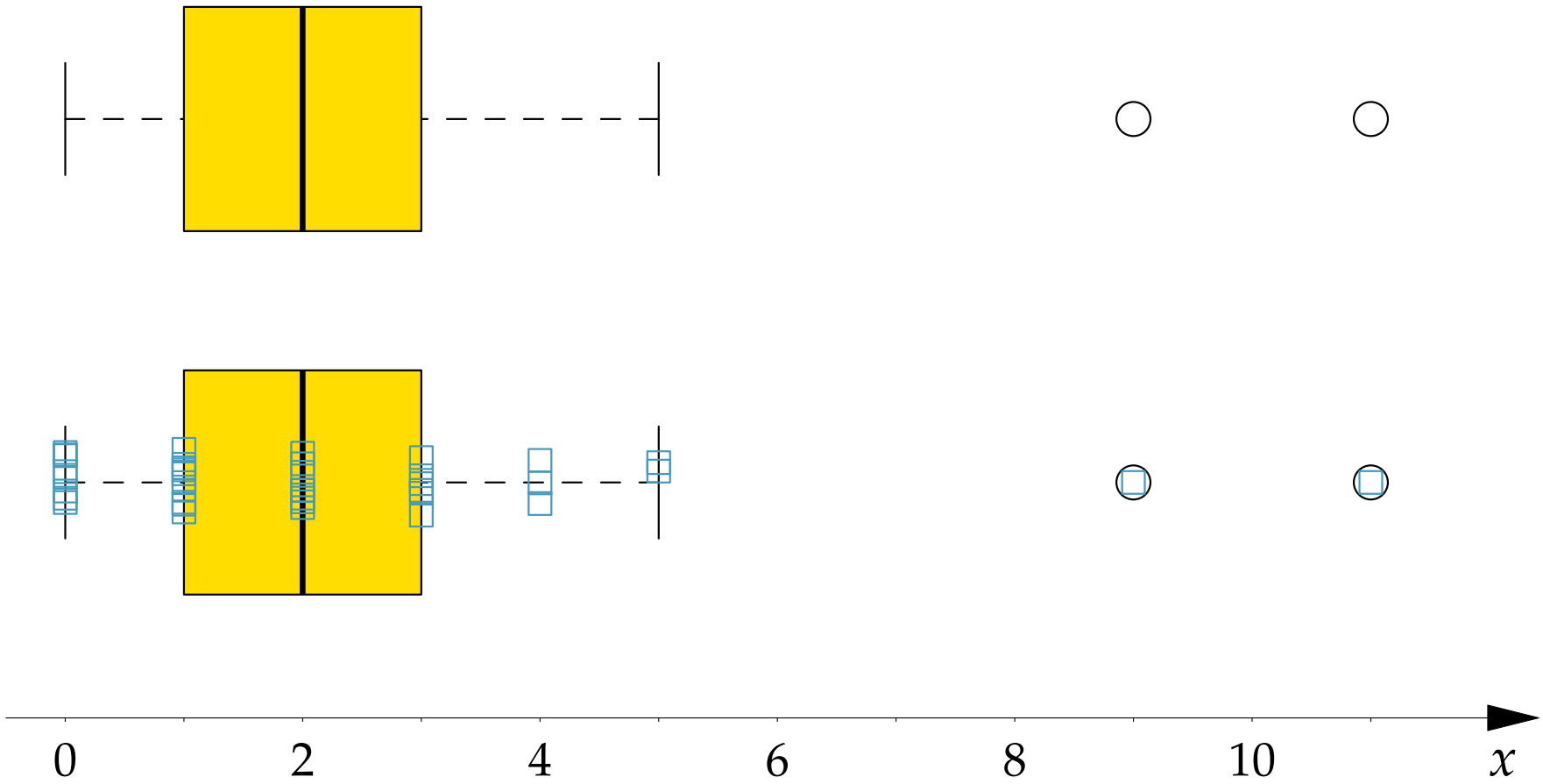


Abb.: Box-Plot nach John W. Tukey ohne (oben) und mit (unten) Rohdaten (Quadrate)

Empirische Varianz (1)

- **empirische Varianz**: mittlere quadrierte Abweichung der Beobachtungswerte vom arithmetischen Mittel
  - Für die Werte $x_1, \ldots, x_n$ gilt:

$$\tilde{s}^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- nur für intervall- und verhältnisskalierte Merkmale geeignet
- empfindlich gegenüber Ausreißern

Empirische Varianz (2)

- Für die Häufigkeitsdaten gilt:

$$\tilde{s}^2 = \frac{1}{n}\left[(a_1 - \bar{x})^2 h_1 + \ldots + (a_k - \bar{x})^2 h_k\right]$$
$$= \frac{1}{n}\sum_{j=1}^{k}\left(a_j - \bar{x}\right)^2 h_j$$

bzw.

$$\tilde{s}^2 = (a_1 - \bar{x})^2 f_1 + \ldots + (a_k - \bar{x})^2 f_k$$
$$= \sum_{j=1}^{k}\left(a_j - \bar{x}\right)^2 f_j$$

Bsp. Empirische Varianz

i	Maßkrug Bier	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1	-1.2	1.44
2	0	-2.2	4.84
3	3	0.8	0.64
4	4	1.8	3.24
5	4	1.8	3.24
6	2	-0.2	0.04
7	0	-2.2	4.84
8	3	0.8	0.64
9	0	-2.2	4.84
10	5	2.8	7.84
Σ	22	0	31.60

arithmetisches Mittel:

$\bar{x} = 2.2$ Maßkrug Bier

empirische Varianz:

$\tilde{s}^2 = \frac{31.60}{10}$ Maßkrug Bier^2

$= 3.16$ Maßkrug Bier^2

Varianz (1)

- **(Stichproben-) Varianz**: *praktisch* die mittlere quadrierte Abweichung der Beobachtungswerte vom arithmetischen Mittel
  - Für die Werte $x_1, \ldots, x_n$ gilt:

$$s^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right]$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- nur für intervall- und verhältnisskalierte Merkmale geeignet
- empfindliche Reaktion auf extreme Werte

Varianz (2)

- Für die Häufigkeitsdaten gilt:

$$s^2 = \frac{1}{n-1}\left[(a_1 - \bar{x})^2 h_1 + \ldots + (a_k - \bar{x})^2 h_k\right]$$
$$= \frac{1}{n-1}\sum_{j=1}^{k}(a_j - \bar{x})^2 h_j$$

bzw.

$$s^2 = \frac{n}{n-1}(a_1 - \bar{x})^2 f_1 + \ldots + (a_k - \bar{x})^2 f_k$$
$$= \frac{n}{n-1}\sum_{j=1}^{k}(a_j - \bar{x})^2 f_j$$

Bsp. Varianz

i	Maßkrug Bier	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1	-1.2	1.44
2	0	-2.2	4.84
3	3	0.8	0.64
4	4	1.8	3.24
5	4	1.8	3.24
6	2	-0.2	0.04
7	0	-2.2	4.84
8	3	0.8	0.64
9	0	-2.2	4.84
10	5	2.8	7.84
Σ	22	0	31.60

arithmetisches Mittel:

$\bar{x} = 2.2$ Maßkrug Bier

empirische Varianz:

$\tilde{s}^2 = \frac{31.60}{10}$ Maßkrug Bier^2

$= 3.16$ Maßkrug Bier^2

Varianz:

$s^2 = \frac{31.60}{9}$ Maßkrug Bier^2

$= 3.51$ Maßkrug Bier^2

Bem. Varianz

- In der induktiven Statistik wird die Varianz s^2 bevorzugt (weil erwartungstreu).
- Die Varianz s^2 besitzt als Dimension das Quadrat der Dimension der Beobachtungen x_i .
  - z. B. cm^2 bei Längenmessungen in cm

Standardabweichung

- Die **(Stichproben-) Standardabweichung**
  - ist die positive Quadratwurzel aus der Varianz.
    - Es gilt

$$s = +\sqrt{s^2}\,.$$

  - beschreibt, wie stark die Beobachtungswerte um den Mittelwert $\bar{x}$ streuen.
  - besitzt die gleiche Dimension wie die Beobachtungswerte x_i und der Mittelwert $\bar{x}$.
  - reagiert empfindlich gegenüber Ausreißern.

Bsp. Standardabweichung

i	Maßkrug Bier	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1	-1.2	1.44
2	0	-2.2	4.84
3	3	0.8	0.64
4	4	1.8	3.24
5	4	1.8	3.24
6	2	-0.2	0.04
7	0	-2.2	4.84
8	3	0.8	0.64
9	0	-2.2	4.84
10	5	2.8	7.84
Σ	22	0	31.60

arithmetisches Mittel:

$\bar{x} = 2.2$ Maßkrug Bier

Varianz:

$s^2 = \frac{31.60}{9}$ Maßkrug Bier^2

$= 3.51$ Maßkrug Bier^2

Standardabweichung:

$s = \sqrt{s^2}$

$= 1.9$ Maßkrug Bier

Bem. s^2 und s

- Varianz s^2 und Standardabweichung s sind lageunabhängig, d. h. sie bleiben unverändert, wenn die Beobachtungen x_i um einen konstanten Wert c verkleinert oder vergrößert werden.

Variationskoeffizienten (1)

- Der **Variationskoeffizient** (auch: relative Standardabweichung)
  - ist die Standardabweichung dividiert durch das arithmetische Mittel.
    - Es gilt

$$v = \frac{s}{\bar{x}}, \quad \text{alle } x_i \geq 0 \text{ und } \bar{x} > 0\,.$$

  - ist ein relatives, dimensionsloses Streuungsmaß mit dem Mittelwert als Einheit.
  - reagiert empfindlich gegenüber Ausreißern.

Variationskoeffizienten (2)

- Liefert die Antwort auf die Frage: „Wie stark ist die Streuung relativ zur mittleren Größe der Datenwerte?“
- Eignet sich besonders zum Vergleich der Streuungen verschiedener Messreihen, deren Merkmalsausprägungen sich hinsichtlich der Größenordnung unterscheiden.
- Das Maximum des Variationskoeffizienten beträgt $\sqrt{n}$. Der **normierte Variationskoeffizient**

$$v^* = \frac{v}{\sqrt{n}}$$

nimmt Werte zwischen 0 und 1 an und ist empfindlich gegenüber Ausreißern.

Gini-Simpson-Index

- Der **Gini-Simpson-Index** ist eine Maßzahl für die Streuung nominaler Merkmale.
  - Es gilt

$$v_G = \sum_{j=1}^{k} f_j \left(1 - f_j\right) \ .$$

- Er ist Null, wenn alle Beobachtungen in eine Kategorie fallen (maximale Konzentration).
- Die größte Streuung in den beobachteten Häufigkeiten eines Merkmals mit k Merkmalsausprägungen ist durch $(k-1)/k$ gegeben.
- Er ist robust gegenüber Ausreißern.

Übersicht Lage- u. Streuungsmaße

Skalenart	Lagemaße	Streuungsmaße
Nominalskala	Modus (x_{mod})	Gini-Simpson-Index (v_G)
Ordinalskala	x_{mod}, Median (x_{med}), p-Quantil (x_p)	v_G, Spannweite (R), Quartilsabstand (IQR)
Intervallskala	x_{mod}, x_{med}, x_p, arithmetischer Mittelwert ($\bar{x}$)	R, IQR, Standardabweichung (s), Varianz (s^2)
Verhältnisskala	x_{mod}, x_{med}, x_p, $\bar{x}$, geometrischer Mittelwert ($\bar{x}_{\text{geom}}$)	R, IQR, s, s^2, Variationskoeffizient (v und v^*)

3. Bivariate deskriptive Statistik

Mehrdimensionale Daten

- Es liegen multivariate Daten vor, wenn an jeder statistischen Einheit gleichzeitig mehrere Merkmale $X_1, X_2, \ldots, X_g$ erhoben werden.

	Merkmal			
i	X_1	X_2	$\ldots$	X_g
1	$x_{1,1}$	$x_{1,2}$	$\ldots$	$x_{1,g}$
2	$x_{2,1}$	$x_{2,2}$	$\ldots$	$x_{2,g}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{n,1}$	$x_{n,2}$		$x_{n,g}$

Kernfragen d. deskriptiven Statistik (1)

- bei univariaten Daten :
  - Wo sind die Ausprägungen des Merkmals zentriert?
  - Wie streuen sie? Gibt es womöglich Ausreißer?
  - Welche Form hat ihre Verteilung?
  - Wie kann ich alles übersichtlich darstellen?
- bei multivariaten Daten noch zusätzlich:
  - Gibt es Zusammenhänge zwischen den Ausprägungen verschiedener Merkmale?

Kernfragen d. deskriptiven Statistik (2)

- Werden für jede statistische Einheit nur zwei Merkmale X und Y gemeinsam beobachtet, so handelt es sich um bivariate Daten.
- Fragestellungen sind dann:
  - $X \leftrightarrow Y$: Wie stark hängen X und Y zusammen? (Korrelation)
  - $X \rightarrow Y$: Wie beeinflusst X das (Ziel-) Merkmal Y? (Regression)

Zwei diskrete Merkmale

- X und Y seien diskret, wobei sie nur relativ wenige Ausprägungen aufweisen.
- Das Skalenniveau von X und Y ist beliebig. Benutzt wird nur das Nominalskalenniveau der Merkmale, auch wenn diese ein höheres Meßniveau besitzen.

Bsp. Bivariate Daten (1)

i	Geschlecht	Blutgruppe
1	weiblich	AB
2	weiblich	0
3	männlich	B
4	weiblich	A
⋮	⋮	⋮
129	weiblich	B
130	männlich	A

Bsp. Bivariate Daten (2)

- gemeinsame Häufigkeitsverteilung für Geschlecht und Blutgruppe, $n = 130$

Geschlecht \ Blutgruppe	0	A	AB	B	Summe
männlich	29	24	1	6	60
weiblich	27	29	5	9	70
Summe	56	53	6	15	130

Kontingenztafel (1)

- **Kontingenztafel:** Gibt die gemeinsame Verteilung der Merkmale X und Y in absoluten Häufigkeiten wieder.
  - Seien X und Y Merkmale mit den möglichen Ausprägungen

    $a_1, \ldots, a_k$ für X ,
    $b_1, \ldots, b_m$ für Y .

  - Es folgt Bestimmung der Häufigkeiten

    $$h_{ij} = h(a_i, b_j)$$

    der möglichen Kombinationen (a_i, b_j),
    $i = 1, \ldots, k$, $j = 1, \ldots, m$.

Kontingenztafel (2)

- Die Häufigkeitstabelle besitzt die Form:

Merkmal X \ Merkmal Y	b_1	$\dots$	b_m	
a_1	h_{11}	$\dots$	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	$\dots$	h_{2m}	$h_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	h_{k1}	$\dots$	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	$\dots$	$h_{\cdot m}$	n

- Kontingenztafeln werden durch die Zeilen- und Spaltensummen ergänzt.

Kontingenztafel (3)

- Zeilensummen bzw. Randhäufigkeiten des Merkmals X:

$$h_{i\cdot} = h_{i1} + \ldots + h_{im}\,, \quad i = 1, \ldots, k$$

- Spaltensummen bzw. Randhäufigkeiten des Merkmals Y:

$$h_{\cdot j} = h_{1j} + \ldots + h_{kj}\,, \quad j = 1, \ldots, m$$

Bsp. Absolventenfeier-Daten

i	Maßkrug Bier	Geschlecht	Einstiegs-gehalt	Noten im Studium
1	1	männlich	45 000	gut
2	0	weiblich	46 000	gut
3	3	männlich	38 000	schlecht
4	4	männlich	42 000	mittel
5	4	weiblich	47 000	mittel
6	2	weiblich	42 000	gut
7	0	weiblich	41 000	gut
8	3	männlich	45 000	schlecht
9	0	männlich	40 000	mittel
10	5	männlich	42 142	mittel

Kontingenztafel rel. Häufigkeiten (1)

- Die Häufigkeitstabelle besitzt die Form:

Merkmal X \ Merkmal Y	b_1	$\dots$	b_m	
a_1	f_{11}	$\dots$	f_{1m}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	f_{k1}	$\dots$	f_{km}	$f_{k\cdot}$
	$f_{\cdot 1}$	$\dots$	$f_{\cdot m}$	1

- Dabei bezeichnet

$$f_{ij} = h_{ij}/n$$

die relative Häufigkeit der Kombination (a_i, b_j).

Kontingenztafel rel. Häufigkeiten (2)

- Die Häufigkeitstabelle besitzt die Form:

Merkmal X \ Merkmal Y	b_1	$\dots$	b_m	
a_1	f_{11}	$\dots$	f_{1m}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	f_{k1}	$\dots$	f_{km}	$f_{k\cdot}$
	$f_{\cdot 1}$	$\dots$	$f_{\cdot m}$	1

- Dabei bezeichnen

  $f_{1\cdot} = \sum_{j=1}^{m} f_{ij} = h_{i\cdot}/n$ die Randhäufig. von X und

  $f_{\cdot 1} = \sum_{i=1}^{k} f_{ij} = h_{\cdot j}/n$ die Randhäufig. von Y.

Bsp. Daten zur Arbeitslosigkeit (1)

- Häufigkeitstabelle zur ausbildungsspezifischen Dauer der Arbeitslosigkeit für männliche Deutsche (Aus: Fahrmeir u. a. 2016)

	Kurzzeit-arbeitslosigkeit	mittelfristige Arbeitslosigkeit	Langzeit-arbeitslosigkeit	
Keine Ausbildung	86	19	18	123
Lehre	170	43	20	233
Fachspez. Ausbildung	40	11	5	56
Hochschulabschluss	28	4	3	35
	324	77	46	447

- Zusammenhang zwischen X und Y aus den gemeinsamen Häufigkeiten h_{ij} bzw. f_{ij} schwer ersichtlich

Bedingte relative Häufigkeiten (1)

- relative Häufigkeiten bezogen auf die Zeilen- oder Spaltensummen
- Verteilung des einen Merkmals für einen festgehaltenen Wert des zweiten Merkmals

Bsp. Daten zur Arbeitslosigkeit (2)

- Für festgehaltenes Ausbildungsniveau ($X = a_i$) erhält man die relative Verteilung über die Dauer der Arbeitslosigkeit (Aus: Fahrmeir u. a. 2016).

	Kurzzeit-arbeitslosigkeit	mittelfristige Arbeitslosigkeit	Langzeit-arbeitslosigkeit	
Keine Ausbildung	0.699	0.154	0.147	1
Lehre	0.730	0.184	0.086	1
Fachspez. Ausbildung	0.714	0.197	0.089	1
Hochschulabschluss	0.800	0.114	0.086	1

- Verteilung der Dauer der Arbeitslosigkeit für die Subpopulationen „Keine Ausbildung“, „Lehre“, usw. wird deutlich.

Bedingte relative Häufigkeiten (2)

- Wählt man $X = a_i$ fest, ergibt sich die bedingte Häufigkeitsverteilung von Y unter der Bedingung $X = a_i$ (abgekürzt $Y|X = a_i$) durch

$$f_Y(b_1|a_i) = \frac{h_{i1}}{h_{i\cdot}}, \quad \ldots \quad, f_Y(b_m|a_i) = \frac{h_{im}}{h_{i\cdot}}.$$

- Wählt man $Y = b_j$ fest, ergibt sich die bedingte Häufigkeitsverteilung von X unter der Bedingung $Y = b_j$ (abgekürzt $X|Y = b_j$) durch

$$f_X(a_1|b_j) = \frac{h_{1j}}{h_{\cdot j}}, \quad \ldots \quad, f_X(a_k|b_j) = \frac{h_{kj}}{h_{\cdot j}}.$$