Kumulierte Häufigkeiten

- Bei zumindest ordinalskalierten Merkmalen sind die Merkmalsausprägungen $a_1, \ldots, a_k$ mit $a_1 < \ldots < a_k$ geordnet.
- Deshalb kann man die absoluten und relativen Häufigkeiten sinnvoll kumulieren.
- Die j-te kumulierte absolute Häufigkeit ist
$$\sum_{i=1}^{j} h_i = h_1 + \ldots + h_j .$$
- Die j-te kumulierte relative Häufigkeit ist
$$\sum_{i=1}^{j} f_i = f_1 + \ldots + f_j .$$

Kumulierte Häufigkeitsverteilungen (1)

- **kumulierte absolute Häufigkeitsverteilung:** Für jede vorgegebene Zahl $x \in \mathbb{R}$ bestimmt man die Anzahl von Beobachtungswerten, die kleiner oder gleich x sind.
  - Mit den Ausprägungen $a_1 < \ldots < a_k$ und deren Häufigkeiten gilt:

$$H(x) = h(a_1) + \ldots + h(a_j) = \sum_{i \,|\, a_i \leq x} h_i$$

Dabei ist a_j die größte Ausprägung, für die noch $a_j \leq x$ gilt, so dass also $a_{j+1} > x$ ist.

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

Kumulierte Häufigkeitsverteilungen (2)

- **empirische Verteilungsfunktion**: Statt absoluter werden relative Häufigkeiten aufsummiert.
  - Es gilt

$$F(x) = f(a_1) + \ldots + f(a_j) = \sum_{i \,|\, a_i \leq x} f_i \,,$$

    wobei $a_j \leq x$ und $a_{j+1} > x$.
  - Dies liefert die Antwort auf die Fragestellung: „Welcher Anteil der Daten ist kleiner oder gleich einem interessierenden Wert x?“
  - *Empirisch* bedeutet aus konkreten Daten berechnet (vs. Verteilungsfunktion einer Zufallsvariable).

Bsp. Emp. Verteilungsfkt. Maßkrug Bier

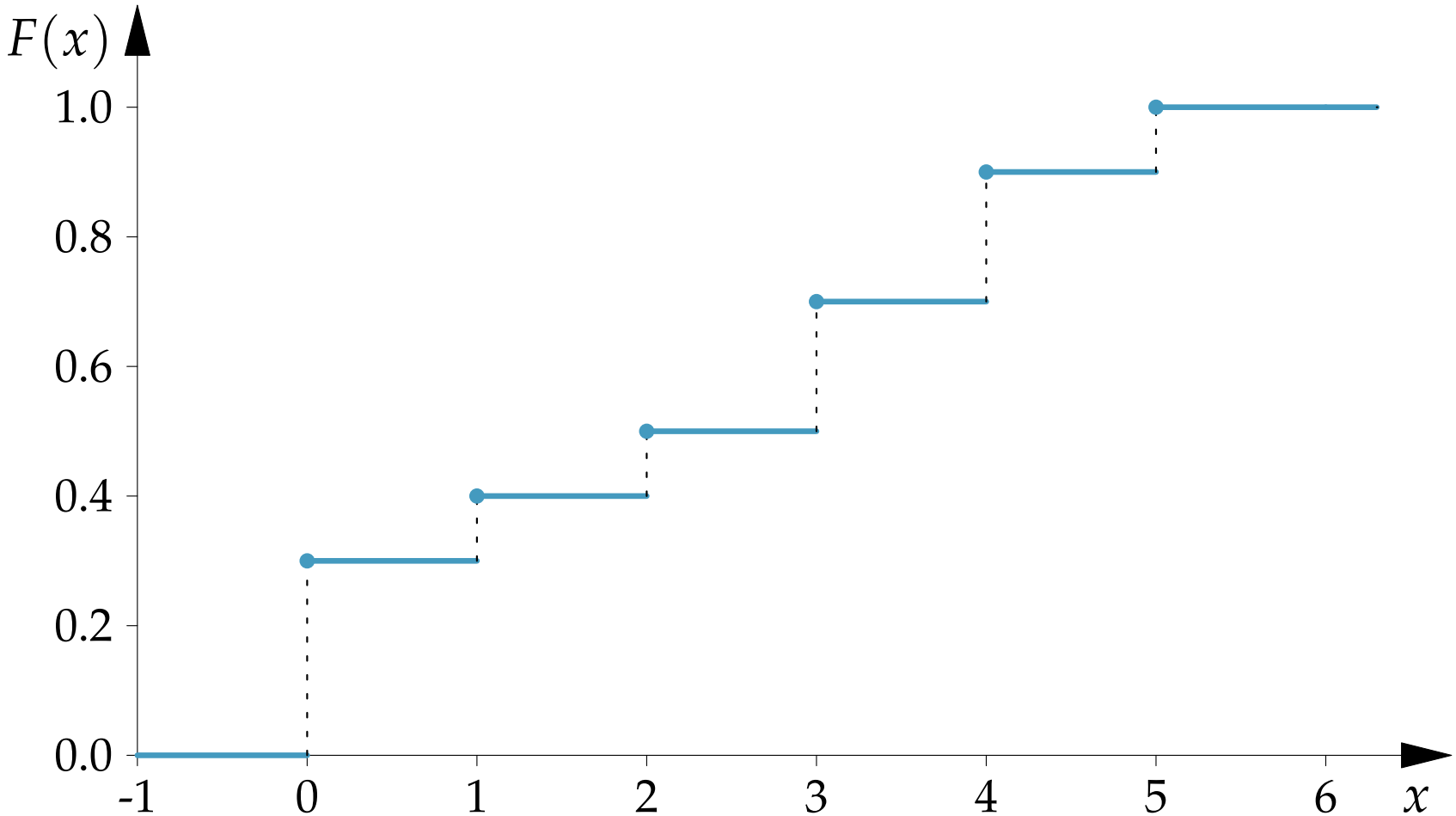

Abb.: Empirische Verteilungsfunktion des Merkmals Maßkrug Bier

Bem. Kum. Häufigkeitsverteilungen

- Kumulierte absolute Häufigkeitsverteilung und empirische Verteilungsfunktion sind monoton wachsende Treppenfunktionen, die an den Ausprägungen $a_1, \ldots, a_k$ um die entsprechende absolute bzw. relative Häufigkeit nach oben springen.
- An den Sprungstellen ist der obere Wert, d. h. die Treppenkante, der zugehörige Funktionswert.
- Die Funktionen $H(x)$ bzw. $F(x)$ sind gleich 0 für alle $x < a_1$ und gleich n bzw. 1 für $x \geq a_k$.

Bsp. Emp. Verteilungsfkt. BMI

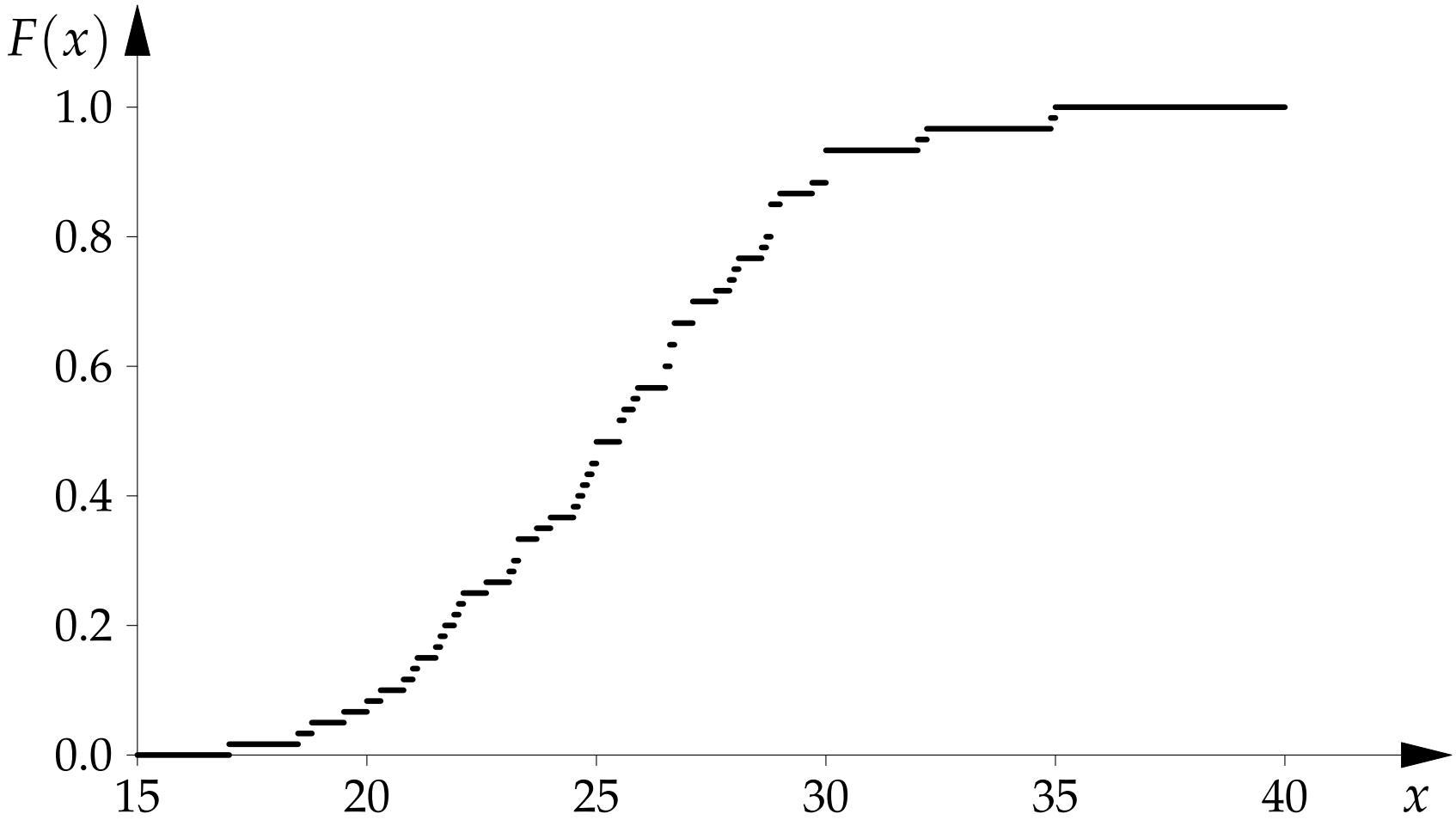

Abb.: Empirische Verteilungsfunktion zum Body-Mass-Index auf der Basis der kumulierten Einzelwerte, $n = 60$

Kennzahlen von Daten

- Maß- oder Kennzahlen sind die Verdichtung der statistischen Eigenschaften einer Datenmenge auf einen oder wenige Parameter.

Lagemaße

- Maßzahlen zur Lage beschreiben das Zentrum einer Verteilung durch einen numerischen Wert:
  - arithmetisches Mittel
  - Median
  - Modus
  - geometrisches Mittel
  - …

Arithmetisches Mittel (1)

- **arithmetisches Mittel:** Summe aller beobachteten Werte geteilt durch die Anzahl der Beobachtungen
  - Bestimmung aus der Urliste:

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - bei Häufigkeitsdaten mit Ausprägungen $a_1, \ldots, a_k$ und absoluten Häufigkeiten $h_1, \ldots, h_k$:

$$\bar{x} = \frac{1}{n}(a_1 h_1 + \ldots + a_k h_k) = \frac{1}{n}\sum_{j=1}^{k} a_j h_j$$

Arithmetisches Mittel (2)

- bei Häufigkeitsdaten mit Ausprägungen $a_1, \ldots, a_k$ und relativen Häufigkeiten $f_1, \ldots, f_k$:

$$\bar{x} = a_1 f_1 + \ldots + a_k f_k = \sum_{j=1}^{k} a_j f_j$$

Bem. Arithmetisches Mittel

- Das arithmetische Mittel
  - ist für intervall- und verhältnisskalierte Merkmale sinnvoll definiert.
  - reagiert empfindlich auf extreme Werte oder Ausreißer in den Daten (empfindliches Lagemaß).
  - entspricht dem (physikalischen) Schwerpunkt der Daten.

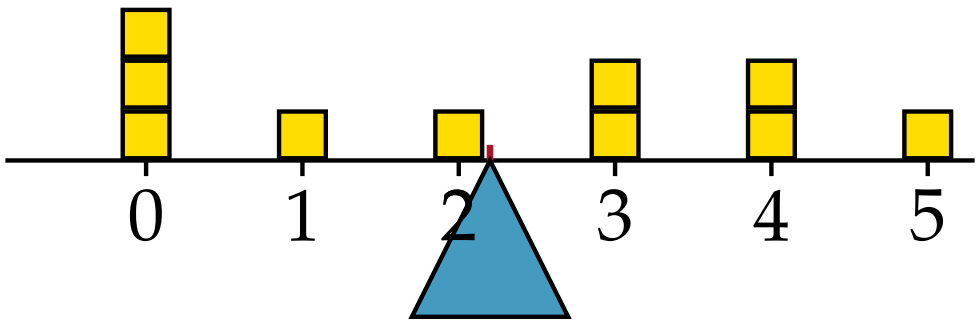

Abb.: Illustration des arithmetischen Mittels des Merkmals Maßkrug Bier

Bsp. Absolventenfeier-Daten (erweitert 1)

i	Maßkrug Bier	Geschlecht	Einstiegs-gehalt	Noten im Studium	Brezn
1	1	männlich	45 000	gut	0
2	0	weiblich	46 000	gut	1
3	3	männlich	38 000	schlecht	0
4	4	männlich	42 000	mittel	1
5	4	weiblich	47 000	mittel	0
6	2	weiblich	42 000	gut	1
7	0	weiblich	41 000	gut	0
8	3	männlich	45 000	schlecht	–
9	0	männlich	40 000	mittel	0
10	5	männlich	42 142	mittel	15

Bsp. Arithmetisches Mittel

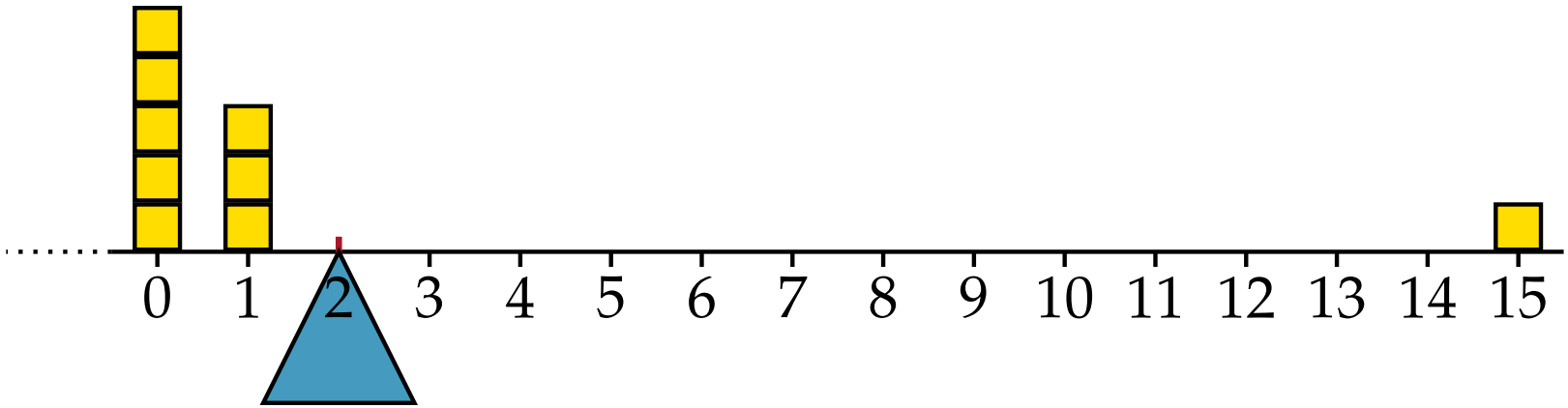

Abb.: Illustration des arithmetischen Mittels des Merkmals Brezn

Median

- **Median:** Wert der nach Größe sortierten Beobachtungen, der genau in der Mitte liegt
  - Ordnen der Werte $x_1, \ldots, x_n$ der Größe nach:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

  - Bestimmung aus der geordneten Urliste durch

$$x_{\text{med}} = \begin{cases} x_{(\frac{n+1}{2})} & \text{für } n \text{ ungerade} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{für } n \text{ gerade} \end{cases}$$

Bsp. Absolventenfeier-Daten (erweitert 2)

i	Maßkrug Bier	Geschlecht	Einstiegs-gehalt	Noten im Studium	Brezn
1	1	männlich	45 000	gut	0
2	0	weiblich	46 000	gut	1
3	3	männlich	38 000	schlecht	0
4	4	männlich	42 000	mittel	1
5	4	weiblich	47 000	mittel	0
6	2	weiblich	42 000	gut	1
7	0	weiblich	41 000	gut	0
8	3	männlich	45 000	schlecht	1
9	0	männlich	40 000	mittel	0
10	5	männlich	42 142	mittel	15

Bem. Median (1)

- Eigenschaften des Medians sind:
  - Mindestens $50\,\%$ der Daten sind kleiner oder gleich x_{med}.
  - Mindestens $50\,\%$ der Daten sind größer oder gleich x_{med}.
- Für ungerades n ist der Median x_{med} die mittlere Beobachtung der geordneten Urliste.
- Für gerades n ist der Median x_{med} das arithmetische Mittel der beiden in der Mitte liegenden Beobachtungen (Vereinbarung).

Bem. Median (2)

- Der Median
  - setzt ein mindestens ordinalskaliertes Merkmal voraus.
  - ist als robustes Lagemaß auch bei metrischen Größen sinnvoll.
  - ist einfacher interpretierbar als das arithmetische Mittel.

Bem. Median (3)

- Der Median ist als Lagemaß zu bevorzugen beim Vorliegen
  - nur weniger Messwerte,
  - asymmetrischer Verteilungen,
  - bei Verdacht auf Ausreißer (robustes Lagemaß).

Modus

- Der Modus x_{mod} ist
  - die Ausprägung mit der größten Häufigkeit.
  - bereits für nominalskalierte Merkmale geeignet.
  - eindeutig, falls die Häufigkeitsverteilung ein eindeutiges Maximum besitzt.
  - in der Darstellung durch Stab- oder Säulendiagramme die Ausprägung mit dem höchsten Stab.

Lageregeln

- Bei intervall- und verhältnisskalierten Merkmalen gilt:
  - symmetrische Verteilungen:
    $$x_{\text{mod}} \approx x_{\text{med}} \approx \bar{x}$$
  - linkssteile Verteilungen:
    $$x_{\text{mod}} < x_{\text{med}} < \bar{x}$$
  - rechtssteile Verteilungen:
    $$x_{\text{mod}} > x_{\text{med}} > \bar{x}$$
- Die Lageregeln sind in erster Linie für unimodale Verteilungen von Bedeutung.
- Je stärker sich $\bar{x}, x_{\text{med}}$ und x_{mod} unterscheiden, desto schiefer sind die Verteilungen.

Bsp. Lageregeln

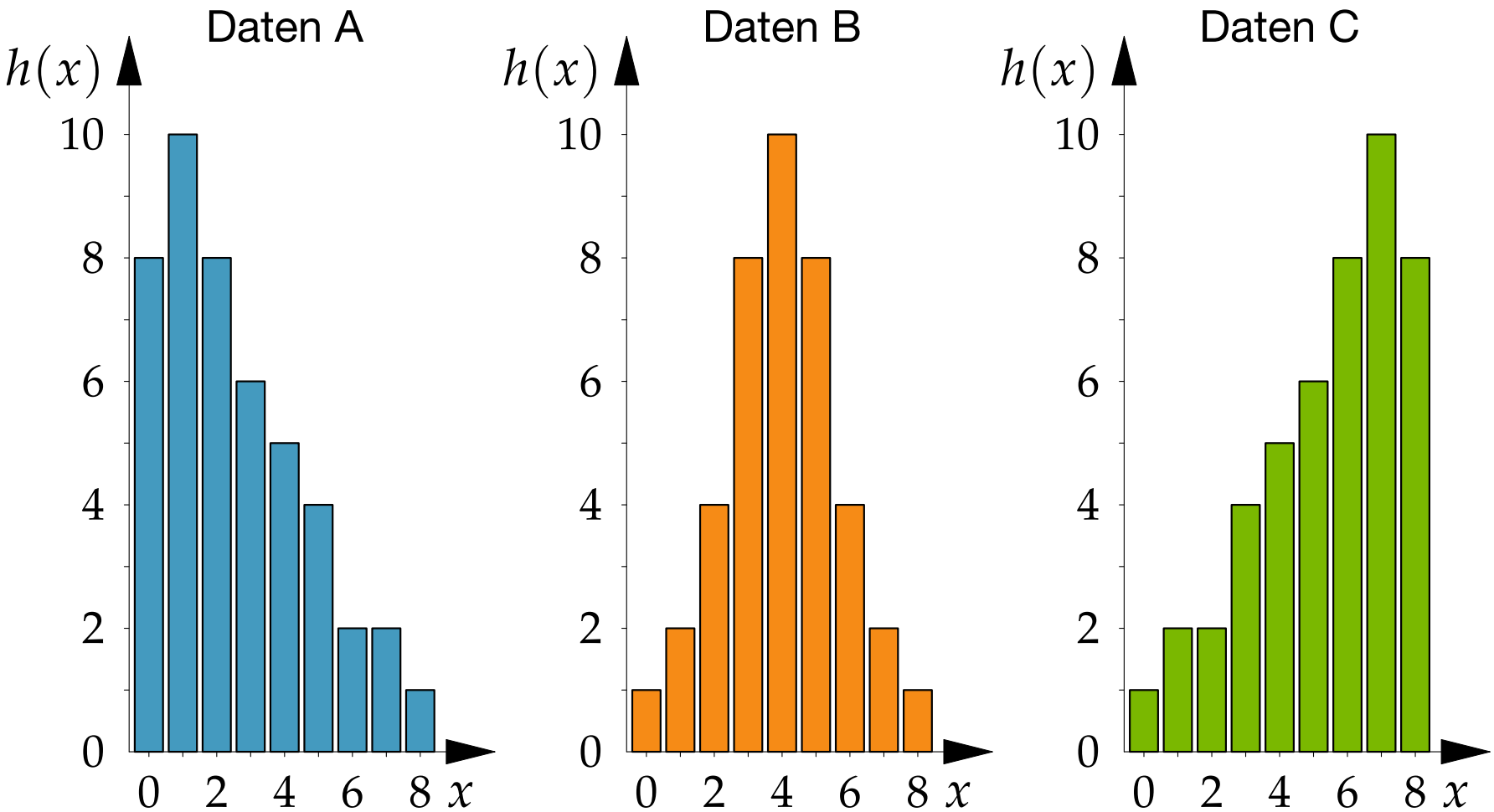

Abb.: Absolute Häufigkeitsverteilungen (Säulendiagramme)

Weitere Lagemaße

- Lagemaße für bestimmte Fragestellungen:
  - geometrisches Mittel
  - harmonisches Mittel
  - getrimmtes Mittel
  - ...

Bsp. Zinseszinsrechnung

- Ein Anfangskapital K_0 wird im ersten Jahr mit zwei Prozent, im zweiten Jahr mit sieben und im dritten Jahr mit fünf Prozent verzinst. Welcher über die drei Jahre konstante Zinssatz $p\,\%$ hätte zum Schluss das gleiche Endkapital ergeben?

Geometrisches Mittel

- **geometrisches Mittel:** n-te Wurzel aus dem Produkt aller beobachteten Werte (Faktoren)
  - Bestimmung für die Faktoren $x_1, \ldots, x_n$ durch

$$\bar{x}_{\text{geom}} = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

Bem. Geometrisches Mittel

- Geeignetes Mittelmaß für Größen, von denen das Produkt anstelle der Summe interpretierbar ist.
- Anwendung bei Merkmalsausprägungen die relative Änderungen darstellen: z. B. Wachstum, Zuwachsraten oder Produktionssteigerungen.
- Veränderungen sollten in jeweils gleichen zeitlichen Abständen gegeben sein.
- Veränderung über die Zeit in einigermaßen konstantem Verhältnis.
- Es gilt $\bar{x}_{\text{geom}} \leq \bar{x}$, wobei $\bar{x}_{\text{geom}} = \bar{x}$ genau dann gilt, wenn $x_1 = \ldots = x_n$ ist.

Quantile (1)

- **p-Quantil:** Wert x_p, der Daten so in zwei Teile trennt, dass etwa $p \cdot 100\,\%$ der Daten darunter und $(1-p) \cdot 100\,\%$ darüber liegen
  - Es muss gelten

$$\frac{\text{Anzahl } (x\text{-Werte } \leq x_p)}{n} \geq p$$

und

$$\frac{\text{Anzahl } (x\text{-Werte } \geq x_p)}{n} \geq 1-p\,.$$

Quantile (2)

- Damit gilt für das p-Quantil:
  - $x_p = x_{([np]+1)}$, wenn np nicht ganzzahlig
    - Dabei ist $[np]$ die zu np nächste kleinere ganze Zahl.
  - $x_p = \frac{1}{2}(x_{(np)} + x_{(np+1)})$, wenn np ganzzahlig
  - Mindestens $p \cdot 100\,\%$ der Daten sind kleiner oder gleich x_p und mindestens $(1-p) \cdot 100\,\%$ der Daten sind größer oder gleich x_p .

Bem. Quantile (1)

- Ein Quantil ist ein Lagemaß.
- Der Median ist das 50 %-Quantil $= x_{0.5}$.
- Das **untere Quartil** ist das 25 %-Quantil $= x_{0.25}$.
- Das **obere Quartil** ist das 75 %-Quantil $= x_{0.75}$.
- Das Quantil ist die Umkehrfunktion der empirischen Verteilungsfunktion.

Grafische Bestimmung Quantile

- Bestimmung des p-Quantils x_p aus der empirischen Verteilungsfunktion:
  - Im Abstand p zur x-Achse eine Horizontale eintragen.
  - Horizontale trifft auf ein senkrechtes Stück der Treppenfunktion: Der dazugehörige x-Wert ist das p-Quantil x_p .
  - Horizontale trifft auf ein waagerechtes Stück der Treppenfunktion: Der mittlere Wert der beiden Beobachtungen, die die Treppenstufe definieren, ist das p-Quantil x_p .

Bsp. Emp. Verteilungsfkt. Maßkrug Bier

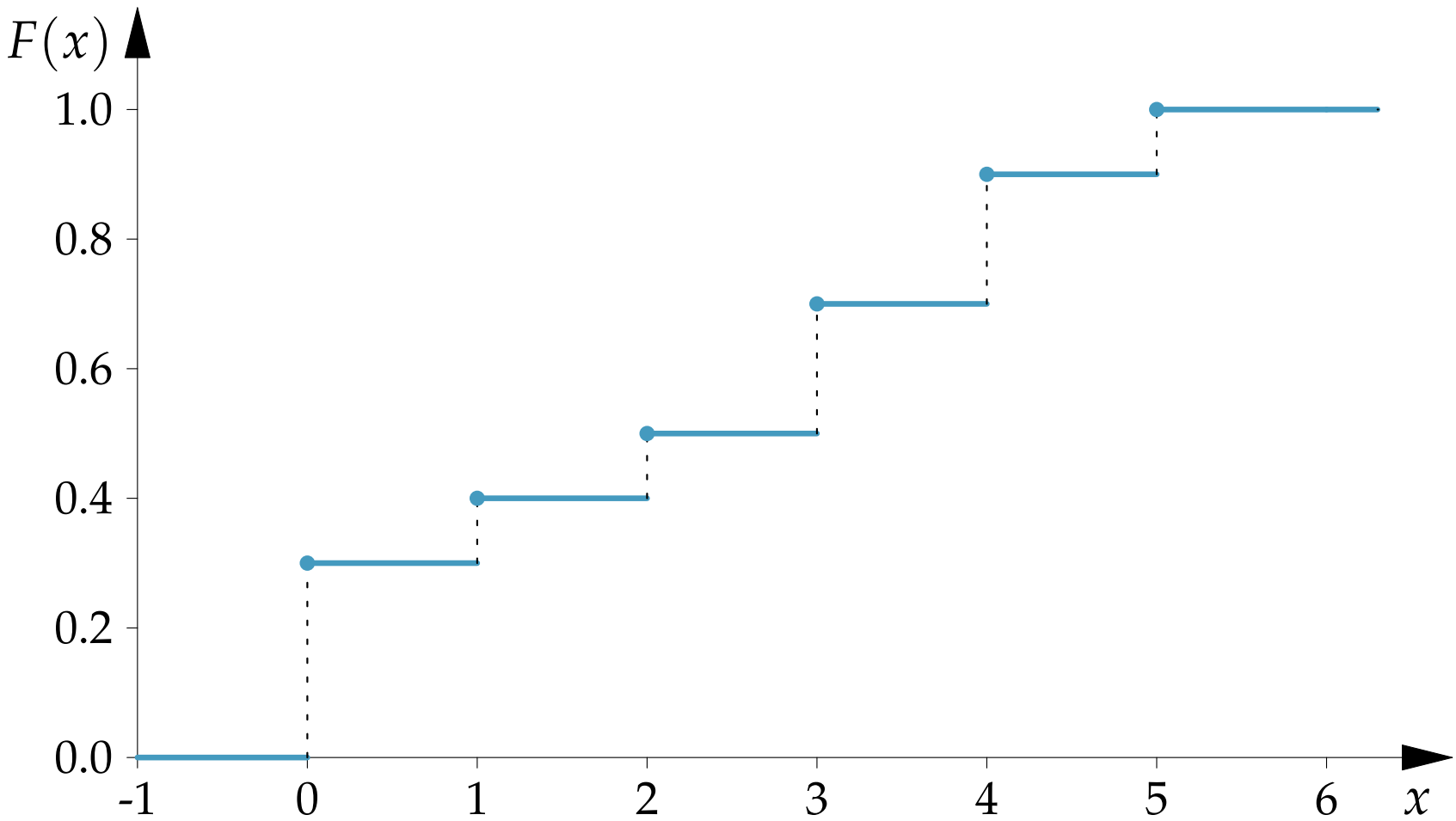

Abb.: Empirische Verteilungsfunktion des Merkmals Maßkrug Bier

Bem. Quantile (2)

- Berechnung des p-Quantils ist nicht einheitlich.
- In R stehen 9 Varianten zur Verfügung. In der Praxis unterscheiden sich diese aber kaum.

Fünf-Punkte-Zusammenfassung

- Die Fünf-Punkte-Zusammenfassung ist die komprimierte Information über eine Verteilung.
- Sie besteht aus

$$x_{\min} = x_{(1)}\,,\; x_{0.25}\,,\; x_{\text{med}}\,,\; x_{0.75}\,,\; x_{\max} = x_{(n)}\,.$$

- Die Quartile, das Minimum, das Maximum sowie der Median teilen den Datensatz in vier Teile, wobei jeder dieser Teile in etwa ein Viertel der Beobachtungswerte enthält.
- Visualisierung erfolgt durch den *Box-Plot*.