Definition 2.2

- Situation wie in Definition 2.1 (Merkmal X, mögliche Ausprägungen $a_1, ..., a_k$, Beobachtung von n Ausprägungen $x_1, ..., x_n$)
- X mindestens ordinal skaliert
- Die empirische Verteilungsfunktion $F_n(x)$ ist gleich der Summe der relativen Häufigkeiten aller Merkmalsausprägungen kleiner oder gleich x
- Formell:

$$F_n(x) = \sum_{a_i \leq x} h(a_i) \quad (x \in \mathbb{R})$$

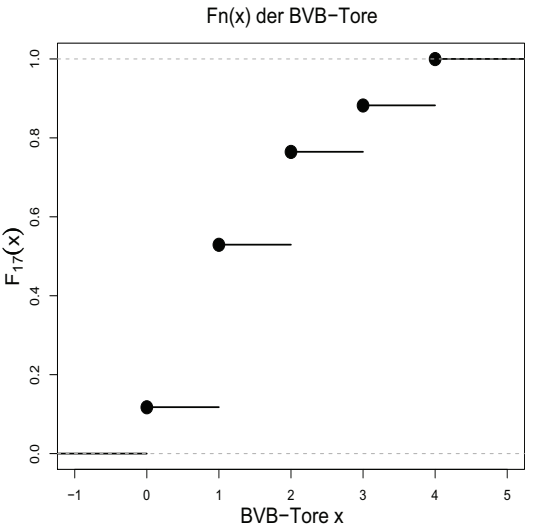
- $F_n(x)$ entspricht dem Anteil an Beobachtungen, die höchstens den Wert x haben

Beispiel 2.3

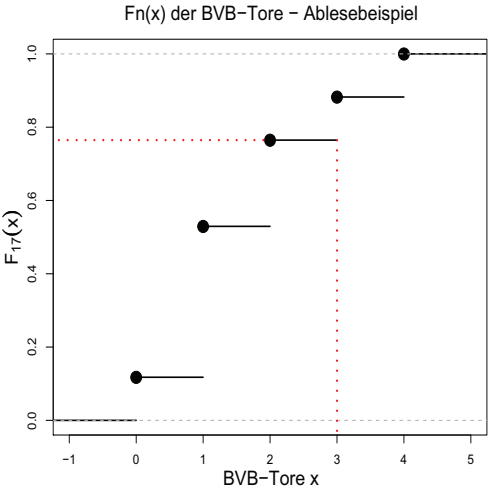
(BVB-Tore, vgl. die Beispiele 2.1 und 2.2)

$$F_{17}(x) = \begin{cases} 0 & \text{für} \quad x < 0 \\ h(0) = 2/17 & \text{für} \quad 0 \leq x < 1 \\ 2/17 + h(1) = 9/17 & \text{für} \quad 1 \leq x < 2 \\ 9/17 + h(2) = 13/17 & \text{für} \quad 2 \leq x < 3 \\ 13/17 + h(3) = 15/17 & \text{für} \quad 3 \leq x < 4 \\ 1 & \text{für} \quad x \geq 4 \end{cases}$$

Beispiel 2.3 (Fortsetzung)



Beispiel 2.3 (Fortsetzung)



$\rightarrow$ In ca. 80 Prozent der Spiele (genauer: in $F_{17}(2) \cdot 100 = 76,5$ Prozent) sind weniger als drei Tore gefallen

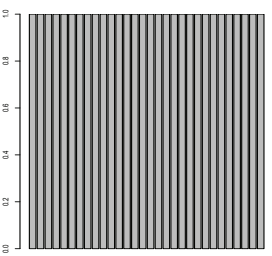
Bemerkung (Eigenschaften von $F_n(x)$)

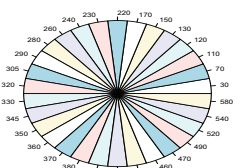
- $F_n(x) \in [0,1]$ für alle x
- $F_n(x)$ ist monoton nicht fallend
- $F_n(x)$ ist rechtsseitig stetig
- Es gilt:

$$\lim_{x \to -\infty} F_n(x) = 0 \quad \text{und} \quad \lim_{x \to \infty} F_n(x) = 1.$$

Beispiel 2.4

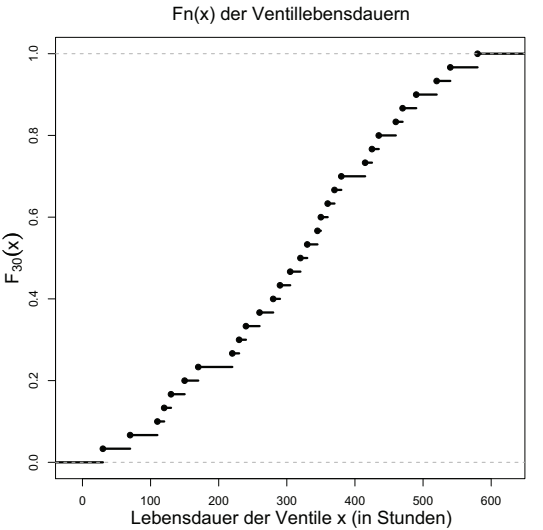
- Lebensdauer (in Betriebsstunden) von Ventilen in kunststoffverarbeitendem Betrieb, vgl. Bamberg et al. (2007)
- 110, 520, 490, 30, 120, 290, 370, 305, 415, 170, 280, 70, 540, 460, 260, 345, 150, 220, 435, 425, 470, 350, 130, 380, 230, 320, 360, 240, 330, 580
- 30 unterschiedliche Beobachtungen → Säulen/Kreisdiagramm bringen keinen Informationsgewinn





Beispiel 2.4 (Fortsetzung)

- Empirische Verteilungsfunktion konstruierbar



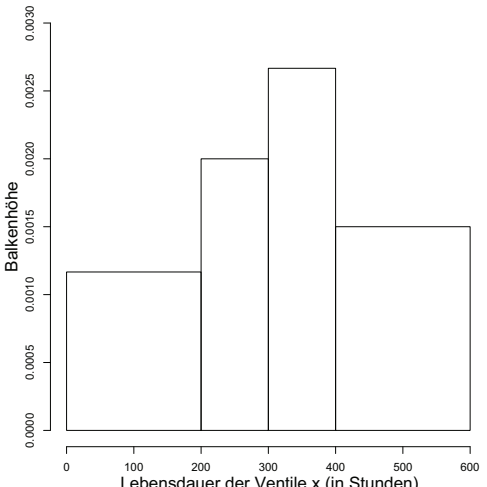
Beispiel 2.4 (Fortsetzung)

- Weitere Möglichkeit: Klassierung der Daten in Intervalle (jetzt $H(a_i)$ bzw. $h(a_i)$ absolute bzw. relative **Klassen**häufigkeit)

Klasse Nr.	von ... bis unter ... Stunden	$H(a_j)$	$h(a_i)$	$\frac{h(a_i)}{\text{Klassenbreite}}$
1	0 - 200	7	7/30	7/6000
2	200 - 300	6	6/30	6/3000
3	300 - 400	8	8/30	8/3000
4	400 - 600	9	9/30	9/6000

Beispiel 2.4 (Fortsetzung)

- **Histogramm:** Betrachte aneinander angrenzende Rechtecke in Klassenbreite; Höhe der Rechtecke: $h(a_i)$/Klassenbreite

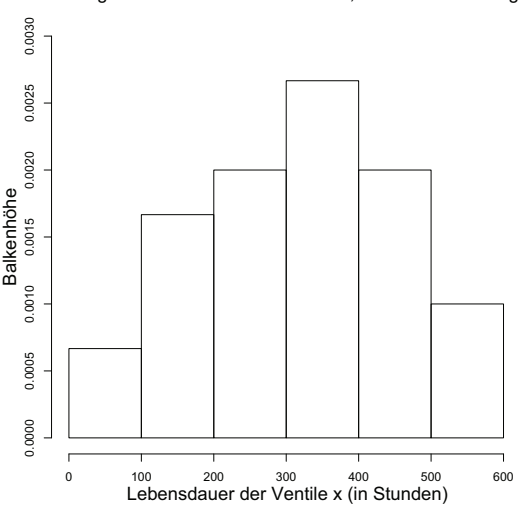


Bemerkung

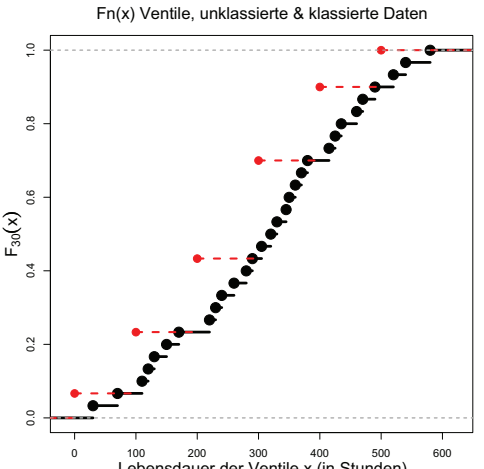
- Die Fläche der einzelnen Balken im Histogramm ist proportional zur relativen Häufigkeit im entsprechenden Intervall:
  Balkenhöhe=$h(a_i)$/Klassenbreite
  $\rightarrow$ $h(a_i)$ = Balkenhöhe $\cdot$ Klassenbreite = Balkenfläche
- Probleme bei zu grober Klasseneinteilung: Zu viel Informationsverlust
- Probleme bei zu feiner Klasseneinteilung: Unübersichtlichkeit, da viele Klassen gering/gar nicht besetzt sind
- Bei großer Variation der Daten können unterschiedliche Klassenbreiten sinnvoll sein, wenn möglich sind jedoch Klassen mit gleicher Breite wünschenswert

Beispiel 2.5

(Lebensdauer Ventile, vgl. Beispiel 2.4)



Beispiel 2.5 (Fortsetzung)



Sprungstelle hier: Klassenuntergrenze; weitere Möglichkeiten: Klassenobergrenze, Klassenmitte,...

Bemerkung

- Säulen/Stab-, Balken- und Kreisdiagramm für nominal, ordinal und kardinal skalierte Merkmale geeignet
- Empirische Verteilungsfunktion für ordinal und kardinal skalierte Merkmale geeignet
- Histogramm nur für kardinal skalierte Merkmale geeignet

Kapitel 3: Lagemaße

Ziel

Komprimierung der Daten zu einer Kenngröße, welche die Lage, das Zentrum der Daten beschreibt

Definition 3.1

Seien $x_1, ..., x_n$ Ausprägungen eines kardinal skalierten Merkmals X, dann heißt

$$\bar{x}^a = \frac{1}{n} \sum_{i=1}^{n} x_i$$

arithmetisches Mittel von X.

Beispiel 3.1

(Ventillebensdauern, vgl. Kapitel 2)

$$\bar{x}^a = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{30} \cdot (110 + 520 + ... + 580) = 313,17$$

Beispiel 3.1 (Fortsetzung)

- Vorgehen, wenn Daten lediglich in klassierter Form vorliegen?
- Betrachte etwa Klassierung wie in Beispiel 2.5

Klasse (von ... bis unter ... Stunden)	$h(a_j)$	Klassenmittel	Klassenmitte
0 - 100	2/30	50	50
100 - 200	5/30	136	150
200 - 300	6/30	253,33	250
300 - 400	8/30	345	350
400 - 500	6/30	449,17	450
500 - 600	3/30	546,67	550

Definition 3.2

Gegeben sei ein kardinal skaliertes Merkmal X mit Ausprägungen $x_1, ..., x_n$ und zugehörigen Gewichten $g_1, ..., g_n$, für die

$$g_i \geq 0 \text{ für alle } i = 1, ..., n \text{ und } \sum_{i=1}^{n} g_i = 1$$

gelte. Dann heißt

$$\bar{x}^{ga} = \sum_{i=1}^{n} g_i x_i = g_1 x_1 + ... + g_n x_n$$

gewichtetes arithmetisches Mittel von X.

Beispiel 3.2

(Ventillebensdauern, Klassierung wie in Beispiel 3.1)

- Verwende relative Häufigkeiten $h(a_i)$ als Gewichte g_i

a) Annahme: Klassenmittel bekannt

$$\bar{x}^{ga} = \frac{2}{30} \cdot 50 + \frac{5}{30} \cdot 136 + ... + \frac{3}{30} \cdot 546,67 = 313,17 = \bar{x}^{a}$$

klar, da

$$\begin{aligned}\bar{x}^{ga} &= \tfrac{2}{30} \cdot [\tfrac{1}{2}(30+70)] + \tfrac{5}{30} \cdot [\tfrac{1}{5}(110+...+170)] + ... \\ &\quad + \tfrac{3}{30} \cdot [\tfrac{1}{3}(520+540+580)] = \tfrac{1}{30} \cdot (30+70+...+580) \\ &= \bar{x}^{a}\end{aligned}$$

Beispiel 3.2 (Fortsetzung)

b) Annahme: Klassenmittel unbekannt

$$\bar{x}^{ga} = \frac{2}{30} \cdot 50 + \frac{5}{30} \cdot 150 + ... + \frac{3}{30} \cdot 550 = 316,67$$

bei unbekanntem Klassenmittel stimmen $\bar{x}^{a}$ und $\bar{x}^{ga}$ in der Regel nicht überein

Beispiel 3.3

a) Betrachte für die letzten 15 Jahre die Platzierungen des BVB in der Bundesliga-Abschlusstabelle: 5, 6, 13, 9, 7, 7, 6, 3, 1, 3, 11, 4, 10, 3, 1 $\rightarrow$ Durchschnittlicher Tabellenplatz (gemäß des arithmetischen Mittels): $\bar{x}^a = 5,9\bar{3} \rightarrow$???

- Derartige Angabe nicht sinnvoll interpretierbar, da Tabellenplätze normalerweise ganzzahlig
- Tabellenplätze außerdem ordinal skaliert $\rightarrow$ die möglichen Platzierungen (1-18) sind nicht naturgegeben, könnten daher (unter Beibehaltung der Reihenfolge) auch willkürlich in andere Zahlen transformiert werden (z.B. 1; 2,5; 3; 5; 7,7; ... ; 99); $\bar{x}^a$ und $\bar{x}^{ga}$ gegenüber derlei Umskalierungen nicht robust

Beispiel 3.3 (Fortsetzung)

b) Betrachte 10 Personen, 9 davon haben ein Jahreseinkommen von 40.000 Euro; Person 10: Jahreseinkommen von 500.000 Euro (fiktive Zahlen) $\rightarrow$ $\bar{x}^a = 86.000$ Euro $\rightarrow$ $\bar{x}^a$ (und auch $\bar{x}^{ga}$) sehr anfällig gegenüber „Ausreissern“

Definition 3.3

Sei X ein mindestens ordinal skaliertes Merkmal mit beobachteten Ausprägungen $x_1, x_2, \ldots, x_n$. Mit $x_{(i)}$ ist der i-te Wert der aufsteigend geordneten Daten bezeichnet. Dann heißt

$$\bar{x}^m = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \quad \text{ungerade} \\ \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right), & n \quad \text{gerade} \end{cases}$$

Median von X.

Beispiel 3.4

(vgl. Beispiel 3.3)

a) Im Durchschnitt hat der BVB in der Bundesliga-Abschlusstabelle auf Basis der letzten 15 Jahre den **6.** Platz belegt, denn

$$n = 15 = \text{ungerade} \;\rightarrow\; \bar{x}^m = x_{(8)}$$

und

$$x_{(1)} = x_{(2)} = 1,\; x_{(3)} = ... = x_{(5)} = 3,\; x_{(6)} = 4,\; x_{(7)} = 5,$$
$$x_{(8)} = x_{(9)} = 6,\; x_{(10)} = x_{(11)} = 7,\; x_{(12)} = 9,\; x_{(13)} = 10,$$
$$x_{(14)} = 11,\; x_{(15)} = 13$$

Beispiel 3.4 (Fortsetzung)

b) Das Durchschnittseinkommen der 10 Personen im fiktiven Beispiel aus Beispiel 3.3 b) beträgt (gemäß des Medians) 40.000 Euro, denn

$$n = 10 = \text{gerade} \;\rightarrow\; \bar{x}^m = \frac{1}{2} \cdot (x_{(5)} + x_{(6)})$$

und

$$x_{(1)} = ... = x_{(9)} = 40.000, \;\; x_{(10)} = 500.000$$
$$\rightarrow \;\; \bar{x}^m = \frac{80.000}{2} = 40.000$$

Bemerkung 1

- Der Median stimmt oft mit einer beobachteten Ausprägung überein
- Der Median ist robuster gegenüber Ausreissern als $\bar{x}^a$ und $\bar{x}^{ga}$
- Nachteil des Medians: Häufig großer Informationsverlust, da nur die mittleren Beobachtungen relevant sind

Bemerkung 2 (Eigenschaften von arithm. Mittel und Median)

- Bei linearen Datentransformationen der Form

$$y_i = a \cdot x_i + b \quad \text{mit} \quad a \neq 0 \quad (i = 1, \ldots, n)$$

gilt:

$$\bar{y}^a = a \cdot \bar{x}^a + b \quad \text{und} \quad \bar{y}^m = a \cdot \bar{x}^m + b.$$

- Beide Lagemaße minimieren jeweils eine Zielfunktion:

$$\bar{x}^a = \underset{z \in \mathbb{R}}{\text{argmin}} \left(\sum_{i=1}^{n} (x_i - z)^2 \right) \text{ und } \bar{x}^m = \underset{z \in \mathbb{R}}{\text{argmin}} \left(\sum_{i=1}^{n} |x_i - z| \right)$$

Beispiel 3.5

- Kardinal skaliertes Merkmal: Arithmetisches Mittel; Ordinal skaliertes Merkmal: Median; Nominale Skalierung: ???
- Notiere etwa Farbe der Fahrzeuge auf dem Uniparkplatz: rot, grün, grün, blau, blau, rot, schwarz, weiss, rot, schwarz (vergleiche Beispiel 1.1) $\rightarrow$ sinnvolles Lagemaß?

Definition 3.4

Als Modalwert bzw. Modus wird die Ausprägung eines beliebig skalierten Merkmals X bezeichnet, die am häufigsten auftritt; Bezeichnung: $\bar{x}_{mod}$

Beispiel 3.6

(vgl. Beispiel 3.5, Fahrzeugfarben)

- Häufigkeiten der beobachteten Farben: $3\times$rot, $2\times$blau, $2\times$grün, $2\times$schwarz, $1\times$weiss $\rightarrow$ $\bar{x}_{mod}$ =rot

Bemerkung (Nachteile des Modus)

- Modalwert muss nicht eindeutig sein
- Bei quantitativ stetigen Daten sind oft sämtliche Beobachtungen unterschiedlich voneinander; hier liefert der Modus keine Informationen $\rightarrow$ Klassierung der Daten; als Modus kann die Mitte der Klasse mit der größten Klassenhäufigkeit aufgefasst werden (im Rahmen der Klassierung von Beispiel 3.1 gilt also $\bar{x}_{mod} = 350$)

Beispiel 3.7

- Aktienkurse zu drei Zeitpunkten (fiktiv)

Zeitpunkt i	0	1	2
Aktienkurs x_i	100	160	100
Wachstumsrate r_i		0,6	-0,375
Wachstumsfaktor $(1 + r_i)$		1,6	0,625

$$\text{wobei} \quad r_i = \frac{x_i - x_{i-1}}{x_{i-1}}$$

- Durchschnittliche Wachstumsrate?

$$\bar{r}^a = \frac{1}{2} \cdot (0,6 + (-0,375)) = 0,1125$$

$\rightarrow$ Unsinn, da (wegen $x_0 = x_2$) $\bar{r} = 0$ gelten muss ($\bar{r}$ = sinnvolles Lagemaß)

Definition 3.5

- Sei X ein kardinal skaliertes Merkmal mit Ausprägungen $x_1, ..., x_n \geq 0$. Dann heißt

$$\bar{x}^{geo} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

das geometrische Mittel von $x_1, ..., x_n$.

Beispiel 3.8

(vgl. Beispiel 3.7)

- Auch negative Wachstumsraten möglich (hier etwa $r_2 = -0,375$) $\rightarrow$ berechne geometrisches Mittel $\overline{(1+r)}^{geo}$ aus den Wachstumsfaktoren $\rightarrow$ $\bar{r}^{geo} = \overline{(1+r)}^{geo} - 1$

$$\overline{(1+r)}^{geo} = \sqrt{1,6 \cdot 0,625} = 1 \;\rightarrow\; \bar{r}^{geo} = 1 - 1 = 0$$

Bemerkung 1

a) Herleitung des geometrischen Mittels (exemplarisch anhand Situation aus Beispiel 3.7 bzw. 3.8)

- Kurs z. Zeitpkt. 0 : x_0
  Kurs z. Zeitpkt. 1 : $x_0 + r_1 \cdot x_0 = x_0 \cdot (1 + r_1) = x_1$
  Kurs z. Zeitpkt. 2 : $x_2 = x_1 \cdot (1 + r_2) = x_0 \cdot (1 + r_1) \cdot (1 + r_2)$
- Gesucht: Geeigneter Durchschnitt von $r_1, r_2 (= \bar{r})$
- Anforderungen an $\bar{r}$:

$$x_0 \cdot (1 + r_1) \cdot (1 + r_2) \stackrel{!}{=} x_0 \cdot (1 + \bar{r}) \cdot (1 + \bar{r}) = x_0 \cdot (1 + \bar{r})^2$$

$\rightarrow$ Division durch x_0 und Auflösung nach $\bar{r}$:

$$(1 + \bar{r}) = \sqrt[2]{(1 + r_1) \cdot (1 + r_2)} \;\rightarrow\; \bar{r} = \sqrt[2]{(1 + r_1) \cdot (1 + r_2)} - 1$$

Bemerkung 1 (Fortsetzung)

b) Allgemein gilt $\bar{x}^{geo} \leq \bar{x}^{a}$ ($\bar{x}^{geo} = \bar{x}^{a}$ genau dann, wenn $x_1 = ... = x_n$)

c) Verwende $\bar{x}^{geo}$, falls Merkmalsausprägungen relativen Änderungen entsprechen