

§ 2 Häufigkeitsverteilungen und grafische Darstellung

Sei Ω eine statistische Grundgesamtheit;

$X: \Omega \rightarrow W$ ein Merkmal mit Wertebereich W ;

$D = \{e_1, \dots, e_N\}$ eine Stichprobe der Größe N .

- Annahme:
- X ist entweder nominal, ordinal oder quantitativ diskret.
 - W ist endlich, $\#W =: M$.
 - $W = \{z_1, \dots, z_M\}$ wurde nummeriert

Definition 2.1 (Häufigkeitsverteilung).

Sei $1 \leq j \leq M$. Wir definieren

(1) $N_j := \#\{i \mid 1 \leq i \leq N, X(e_i) = z_j\} = \#X^{-1}(z_j)$,
ist die absolute Häufigkeit von z_j .

(2) $f_j = \frac{N_j}{N} =$ relative Häufigkeit von z_j

(3) (N_1, \dots, N_M) ist absolute Häufigkeitsverteilung des Merkmals X der Daten D .

(4) (f_1, \dots, f_M) ist relative Häufigkeitsverteilung.

Beispiel: $\omega = \{a, b, c\}$ (nominal Data mit $M=3$)

$D = \{e_1, \dots, e_5\}$ ($N=5$)

$$X(e_1) = a, \quad X(e_2) = b, \quad X(e_3) = a, \quad X(e_4) = a, \quad X(e_5) = a.$$

Sei etwa $z_1 := a, z_2 := b, z_3 := c$.

Dann gilt:

$$\begin{aligned} N_1 &= \#\{i \mid X(e_i) = z_1 = a\} \\ &= \#\{1, 3, 4, 5\} = 4 \\ N_2 &= 1, \quad N_3 = 0 \end{aligned} \quad \left. \begin{array}{l} (N_1, N_2, N_3) \\ = (4, 1, 0) \end{array} \right\}$$

$$\begin{aligned} f_1 &= \frac{N_1}{N} = \frac{4}{5} = 0.8 \\ f_2 &= \frac{1}{5} = 0.2 \\ f_3 &= 0 \end{aligned} \quad \left. \begin{array}{l} (f_1, f_2, f_3) \\ = (0.8, 0.2, 0) \end{array} \right\}$$

→ In R: `table(...)`.

Bemerkung: Es gilt: (a) $\sum_{j=1}^M N_j = N$

$$(b) \sum_{j=1}^M f_j = \sum_{j=1}^M \frac{N_j}{N} = \frac{1}{N} \sum_{j=1}^M N_j = \frac{N}{N} = 1$$

Definition 2.2 (Balkendiagramm)

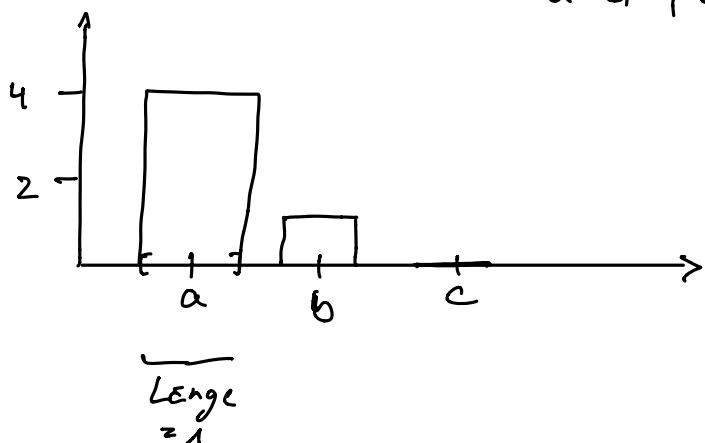
Ein Balkendiagramm stellt die Häufigkeitsverteilung eines nominalen oder ordinalen Merkmals dar.

Auf der x-Achse werden die Werte aus ω aufgetragen.

Über $z_j \in \omega$ wird ein Balken der Länge 1 und Höhe N_j gezeichnet.

Beispiel $\omega = \{a, b, c\}$, $(N_1, N_2, N_3) = (4, 1, 0)$

$$a = z_1, b = z_2, c = z_3$$



In R: `barplot(--)`

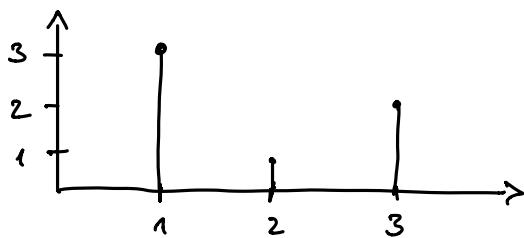
Definition 2.3 (Stabdiagramme)

Ein Stabdiagramm stellt die Häufigkeitsverteilung eines quantitativ diskreten Merkmals dar.

Auf der x-Achse werden die Werte in ω aufgetragen.

Über $z_j \in \omega$ wird ein Stab der Länge N_j gezeichnet.

Beispiel $\omega = \{1, 2, 3\}$, $(N_1, N_2, N_3) = (3, 1, 2)$.



In R: `plot(...)`.

Bisher: ordinale, nominale Merkmale (\rightarrow Balkendiagramm), oder quantitativ diskrete Merkmale (\rightarrow Stabdiagramm).

Jetzt Für quantitativ kontinuierliche Daten benutzen wir Histogramme.

Idee Einen kontinuierlichen Wertebereich $\omega \subseteq \mathbb{R}$.

Definition 2.4 (Diskretisierung)

Sei $\omega \subseteq \mathbb{R}$ ein kontinuierlicher Wertebereich.

Es seien $I_j = (v_{j-1}, v_j]$, $1 \leq j \leq k$, halboffene Intervalle, so dass,

$$(1) \quad \omega \subseteq \bigcup_{j=1}^k I_j$$

$$(2) \quad v_j \leq v_{j+1} \text{ für alle } 1 \leq j \leq k-1, \text{ (d.h. } I_j \cap I_\ell = \emptyset, \text{ falls } j \neq \ell\text{).}$$

(d.h. jeder Punkt in ω liegt in genau einem Intervall I_j).

Dann nennen wir (I_1, \dots, I_k) eine Discretisierung von ω .

Definition 2.5 (Häufigkeitsverteilung quantitativ stetiger Merkmale).

Sei X ein quantitativ stetiges Merkmal mit Wertebereich ω .
Und sei $I = (I_1, \dots, I_k)$ eine Discretisierung von ω .

$$(1) N_j := \#\{i \mid 1 \leq i \leq N : X(e_i) \in I_j\} = \# X^{-1}(I_j)$$

$$(2) f_j := \frac{N_j}{N}$$

(3) (N_1, \dots, N_k) ist die absolute Häufigkeitsverteilung von X bzgl der Discretisierung I .

(4) (f_1, \dots, f_k) ist die relative Häufigkeitsverteilung.

Bemerkung Eine Discretisierung transformiert ein quantitativ stetiges Merkmal in ein quantitativ diskretes mit Wertebereich $\{I_1, \dots, I_k\}$.

Definition 2.6.

Ein Histogramm stellt ein quantitativ stetiges Merkmal als Balkendiagramm nach Transformation in ein quantitativ diskontes Merkmal dar.

In R: `hist(...)`

Eine weitere Möglichkeit die Häufigkeitsverteilung eines quantitativen Merkmals zu beschreiben ist die empirische Verteilungsfunktion

Definition 2.7

Sei X ein Merkmal mit Wertebereich $W \subseteq \mathbb{R}$.

Sei $D = \{e_1, \dots, e_N\}$ und $x_i := X(e_i)$.

Die zugehörige empirische Verteilungsfunktion ist:

$$F_N(x) = \frac{1}{N} : \# \{ i \mid 1 \leq i \leq N : x_i \leq x \}$$

Bemerkung: $\lim_{x \rightarrow -\infty} F_N(x) = 0$, $\lim_{x \rightarrow +\infty} F_N(x) = 1$.

Beispiel $N = 4$, $x_1 = 0$, $x_2 = 2$, $x_3 = 2$, $x_4 = 5$

