Lernzettel

Empirische Verteilungen: Klassenbildung, Histogramme und Verteilungsformen

Universität: Technische Universität Berlin
Kurs/Modul: Statistik I für Wirtschaftswissenschaften
Erstellungsdatum: September 19, 2025

Lernzettel: Statistik I für Wirtschaftswissenschaften

(1) Empirische Verteilungen: Klassenbildung, Histogramme und Verteilungsformen

(1) Häufigkeiten und Klassenbildung.

Gegeben sei eine Stichprobe $\{x_1, \ldots, x_n\}$ mit diskreten oder stetigen Messwerten. Die wichtigsten Größen sind

$$H_k = \#\{i : \; x_i \in K_k\}, \qquad k = 1, \ldots, m,$$

wobei $K_k = [a_k, b_k)$ die k-te Klasse (Klassenintervall) bezeichnet, und

$$p_k = \frac{H_k}{n} \quad (k = 1, \ldots, m)$$

die relative Häufigkeit der Klasse K_k.

Klassenbreite und Klassenmittel.
Die Klassenbreite ist

$$w = b_k - a_k \quad (\text{für alle } k),$$

und das Klassenmittelpunkt ist

$$m_k = \frac{a_k + b_k}{2}.$$

Zahl der Klassen.
Eine gängige Orientierung ist die Anwendung von Sturges' Formel

$$m = \lceil 1 + \log_2 n \rceil \,,$$

wobei $\lceil \cdot \rceil$ die aufgerundete Ganzzahl bezeichnet. Die Klassenbreite folgt dann aus

$$w \approx \frac{x_{(n)} - x_{(1)}}{m},$$

mit $x_{(1)} = \min_i x_i$ und $x_{(n)} = \max_i x_i$.

Offene Klassen.
Bei sehr großen oder offenen Enden kann man die ersten/letzten Klassen offen $[a_1, b_1), [a_m, \infty)$ bzw. $[-\infty, b_m)$ wählen.

Beispiel. Ohne numerisches Beispiel geht es hier nur um das Vorgehen: Bestimme n, wähle m über Sturges', bilde die Klassen $[a_k, b_k)$ und berechne H_k und p_k pro Klasse.

(2) Histogramme.

Ein Histogramm ist eine graphische Darstellung der Häufigkeitsverteilung. Die Höhe der Balken hängt von der gewählten Art der Skalierung ab. Zwei gängige Varianten:

- absolute Häufigkeit H_k pro Klasse, oder
- relative Häufigkeit p_k pro Klasse.

Histogramm-Höhenformel.
Bei der gängigen Diagrammskala mit Fläche pro Klasse gilt

$$h_k = \frac{H_k}{n\,w} = \frac{p_k}{w},$$

so dass die Fläche jeder Bar $h_k w = p_k$ entspricht.

Hinweise zur Gestaltung.

- Gleichmäßige Klassenbreite verwenden, wenn möglich.
- Achsen sinnvoll beschriften (Beschriftung der Klassen, Einheiten, Legende).
- Ausschnitte vermeiden, Balken nicht überlappen.

(3) Verteilungsformen.
Empirische Verteilungen zeigen oft typische Formen. Wichtige Kategorien:

- Symmetrische/unimodale Verteilung (z. B. Normalform).
- Rechtssteilung (positive Schiefe) oder Linkssteilung (negative Schiefe).
- Unimodal oder multimodal (z. B. bimodal).
- Gleichverteilung, schmal oder breit verteilt.

Empirische Verteilungsfunktion.
Die empirische Verteilungsfunktion (EDF) ordnet Werte der Größe nach. Eine gängige Definition ist

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \le x\}},$$

wobei $\mathbf{1}_{\{\cdot\}}$ der Indikatorfunktion ist. $F_n(x)$ steigt stufenweise an und nähert sich der wahren Verteilungsfunktion $F(x)$.

Beispiele für Verteilungsformen.

- Normalverteilung: spiegelbildlich symmetrisch um das Zentrum.
- Rechtssteilung: lange Schwanz nach rechts.
- Linksschief: langer Schwanz nach links.
- Multimodal: mehrere Gipfel in der Verteilung.

(4) Kennzahlen und Form der Verteilung.
Aus empirischen Verteilungen lassen sich zentrale Größen ableiten:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \text{Median } \tilde{x} \text{ (zentres Wert bei geordneter Liste)},$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \mathrm{IQR} = Q_3 - Q_1.$$

(5) Praktische Hinweise.

- Wähle die Anzahl der Klassen (m) sinnvoll, z. B. mit Sturges' Formel.
- Verwende möglichst gleich breite Klassen.
- Prüfe, ob die Histogrammform durch Ausreißer oder Fehlklassierungen beeinflusst wird.
- Nutze die EDF, um quantile-basierte Aussagen zu treffen.