

Motivation

- Visuellen Überblick über Daten zu gewinnen
- Informationen gehen womöglich verloren

Empirische Verteilungsfunktion

Sei X_1, \dots, X_n eine Zufallsstichprobe mit Verteilung F . Die empirische Verteilungsfunktion ist

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}},$$

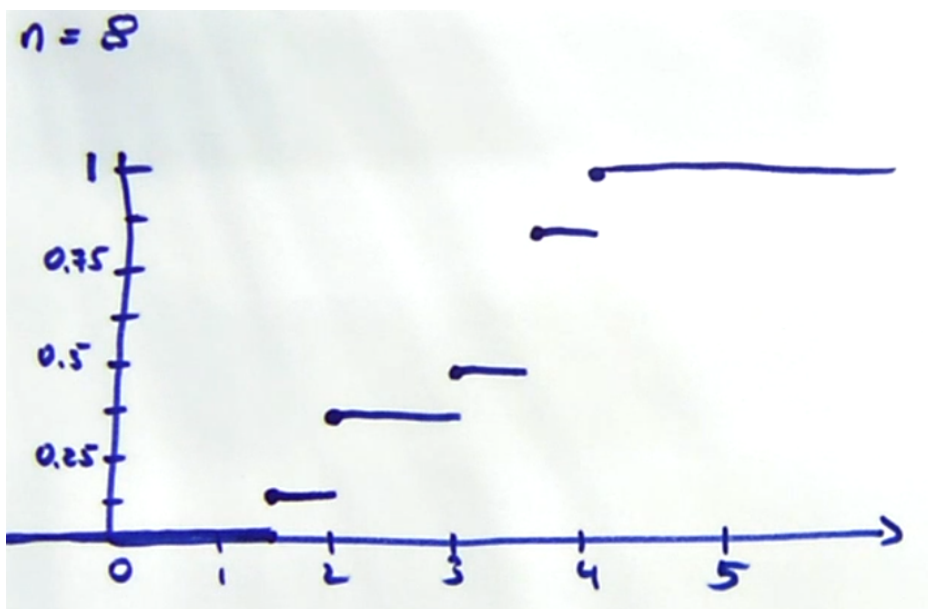
wobei $1_{\{X_i \leq x\}} = 1$, wenn $X_i \leq x$, und $1_{\{X_i \leq x\}} = 0$ sonst.

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0,$$

- Fundamentalsatz der [[Statistik]]
 - Approximation unbekannter Verteilungen möglich
 - wenn [[Stichprobe]] groß genug
- Beispiel

Skizzieren Sie die empirische Verteilungsfunktion der Daten:

1.5 2.0 2.0 3.0 3.5 3.5 3.5 4.0.



Histogramm

Die Konstruktion geschieht folgendermaßen:

1. Wähle $a \leq x_{(1)}$ und $b \geq x_{(n)}$.
2. Zerlege $[a, b]$ in k äquidistante Intervalle

$$[a, a + \Delta] \quad (a + \Delta, a + 2\Delta] \quad \dots \quad (b - \Delta, b],$$

wobei $\Delta = \frac{b-a}{k}$.

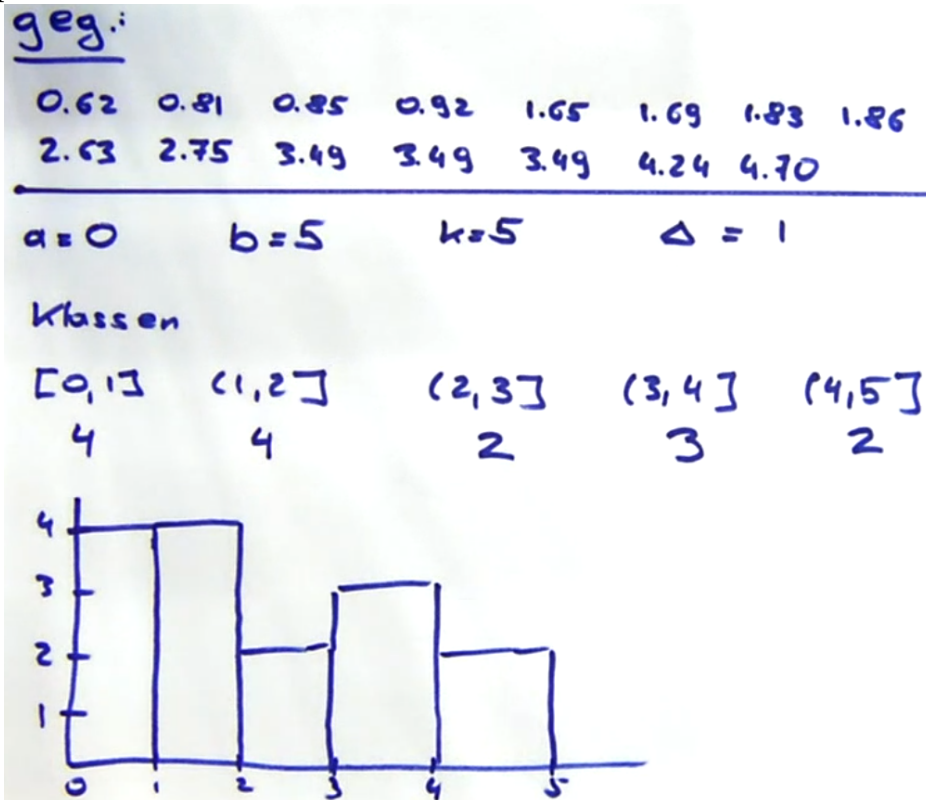
3. Sei n_i die Anzahl der Beobachtungen im i -ten Intervall.
Konstruiere über dem i -ten Intervall ein Rechteck der Höhe
 - $n_i \dots$ für absolute Häufigkeiten oder
 - $n_i/(n\Delta) \dots$ für relative Häufigkeiten.

-
- simpler Schätzer für Dichtefunktion
- einfach verständlich
- abhängig von Auswahl der Parameter a, b, k

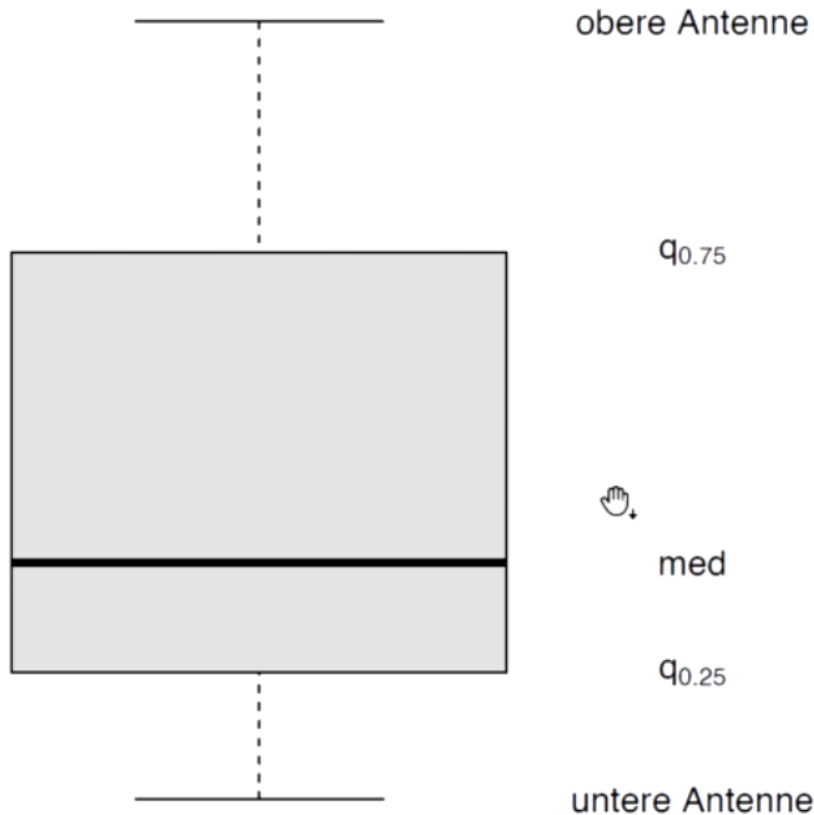
Breite Intervalle verursachen einen höheren Informationsverlust.

Schmale Intervalle ergeben oftmals ein unregelmäßiges Erscheinungsbild.

-
- Beispiel



Boxplot



- Die **Antennen** (Whisker) reichen bis zum kleinsten (größten) Wert, der nicht weiter als $1.5 \cdot iqr$ unter (über) der Boxgrenze liegt.
-
- Ausreißer
 - Punkte außerhalb der Antennen
 - viele Ausreißer → womöglich nicht normalverteilt
- ideal für Vergleich unterschiedener Stichproben
- liefert schnellen Überblick

Q-Q-Plot

- Quantil-Quantil-Plot
- vergleicht Daten mit Referenzverteilung
 - beurteilt Anpassung der Daten an theoretische Verteilung

Dabei plottet man die empirischen Quantile gegen die theoretischen Quantile der Referenzverteilung F .

Sei $0 \leq p_1 < p_2 < \dots < p_n \leq 1$. Man plottet

$$(z_{p_1}, q_{p_1}) \quad (z_{p_2}, q_{p_2}) \quad \dots \quad (z_{p_n}, q_{p_n}),$$

wobei z_p dem theoretischen p -Quantil von F entspricht.

-
- Annahme Normalverteilung
 - Q-Q-Plot weist auf Gegenteil hin

