§73: ROBUSTE STATISTIK

73.1. Motivation

Will man aus realen Daten statistische Parameter schätzen (z.B. arithmetisches Mittel als Schätzer für den Erwartungswert; Parameter einer Regressionskurve), kann es sein, dass das Ergebnis auf Grund von Ausreißern stark verfälscht wird.

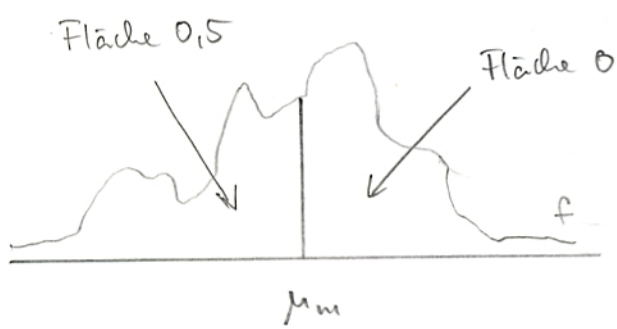
Beispiel: 9 Studierende benötigen 10 Semester für ihr Studium, 1 benötigt 40 Semester. Das arithmetische Mittel ergibt eine mittl. Studiendauer von 13 Semestern. Sie ist jedoch nicht repräsentativ für die Mehrzahl der Studierenden.

Gibt es statistische Verfahren, die robuster gegenüber Ausreißern sind?

73.2. Median

Sei X eine Zufallsvariable. Dann nennt man jede Zahl μ_m mit $P(X \geq \mu_m) \geq \frac{1}{2}$ und $P(X \leq \mu_m) \geq \frac{1}{2}$ einen Median von X.

Veranschaulichung für kontinuierliche Zufallsvariable mit Dichte f:

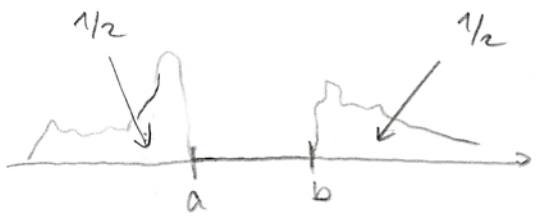


Für die Verteilungsfunktion F gilt:

$$F(\mu_m) = \frac{1}{2}$$

73.3. Bemerkung

a) Nicht immer gibt es einen eindeutigen Median.
Gibt es ein Intervall $[a,b]$ mit $P(X \leq a) = \frac{1}{2}$ und $P(X \geq b) = \frac{1}{2}$, so ist jede Zahl aus $[a,b]$ ein Median:



b) I.A. stimmen Erwartungswert und Median nicht überein.

73.4. Empirischer Median

Hat man $2k+1$ der Größe nach geordnete Messwerte

$$x_1 \leq x_2 \leq \ldots \leq x_{2k+1}$$

dann nennt man $\hat{\mu}_m := x_{k+1}$ den (empirischen) Median dieser Daten. Es sind 50% der Daten $\geq \hat{\mu}_m$ und 50% der Daten $\leq \hat{\mu}_m$.

Bei einer geraden Anzahl von Messungen

$$x_1 \leq x_2 \leq \ldots \leq x_{2k}$$

gilt für jedes $\hat{\mu} \in [x_k, x_{k+1}]$: $\geq 50\%$ der Daten sind $\geq \hat{\mu}$, und $\geq 50\%$ sind $\leq \hat{\mu}$. Man definiert in diesem Fall

$$\hat{\mu}_m := \frac{1}{2}(x_{k+1} + x_k)$$

als „den" (empirischen) Median.

73.5. Beispiele

a) Der Median der Studiendauer in 73.1. beträgt 10 Semester. Der Ausreißer mit 40 Semestern hat somit keinen Einfluss auf den Median.

b) In der Bildverarbeitung ersetzt der Medianfilter einen Grauwert durch seinen Median innerhalb eines $(2k+1) \times (2k+1)$-Fensters:

32	17	24
35	251	21
12	24	25

Ordnen der Grauwerte:

$$12 \leq 17 \leq 21 \leq 24 \leq 24 \leq 25 \leq 32 \leq 35 \leq 251$$

↑ Median

Der Grauwert 251 (Ausreißer) wird durch den Median 24 ersetzt. Medianfilter sind robust gegenüber Impulsrauschen (Ausreißer nach oben oder unten) und erhalten Kanten.

73.6. M-Schätzer

Seien $x_1, \dots, x_n$ Messwerte und $\psi : [0, \infty) \to \mathbb{R}$ eine mon. wachsende Straffunktion (engl.: penalises).

Dann nennt man dasjenige μ, das

$$E(x) = \sum_{i=1}^{n} \psi(|x - x_i|)$$

minimiert, den M-Schätzer von $x_1, \dots, x_n$.

73.7. Beispiele

a) Beliebt ist die Familie $\Psi(s) = s^p$ mit $p \geq 0$.

Man kann zeigen:

i) $p = 2$ liefert das arithmetische Mittel $\bar{x}$. Es minimiert den quadratischen Abstand

$$E(x) = \sum_{i=1}^{n} (x_i - x)^2$$

ii) $p = 1$ liefert den Median $\hat{\mu}$. Er minimiert die Abstandssumme

$$E(x) = \sum_{i=1}^{n} |x_i - x|$$

iii) $p \to 0$ liefert als Minimierer die Modalwerte (Maxima des Histogramms)

iv) $p \to \infty$ ergibt den Midrange $\frac{\max\{x_i\} + \min\{x_i\}}{2}$.

Kleinere Werte für p liefern robustere M-Schätzer, da sie Ausreißer x_i, für die $\Psi(|x_i - x|) = |x_i - x|^p$ groß wird, weniger stark bestrafen.

b) Eine andere Straffunktion, die robuster als die übliche quadr. Straffunktion $\Psi(s) = s^2$ ist, ist z.B. die Lorentz-Straffunktion

$$\Psi(s) = \ln\left(1 + \tfrac{1}{2} s^2\right)$$