

§73: ROBUSTE STATISTIK

73.1 Motivation:

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

Bsp.:

9 Studierende benötigen 10 Semester für ihr Studium. 1 benötigt 40 Semester. Das arithmetische Mittel liefert eine mittlere 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 .

Für die Verteilungsfunktion F gilt:

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

73.3 Bemerkungen

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

73.4 Empirischer Median

Hat man $2k + 1$ der Größe nach geordnete Messwerte $x_1 \leq x_2 \leq \dots \leq x_{2k+1}$, dann nennt man

$$\hat{\mu}_m := x_{k+1}$$

den (empirischen) Median.

Es sind 50% der Daten $\geq \hat{\mu}_m$ und 50% sind $\leq \hat{\mu}_m$.

Bei einer geraden Anzahl von Messungen $x_1 \leq x_2 \leq \dots \leq x_{2k}$ 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 aus 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 einer $(2k + 1) \times (2k + 1)$ -Maske:

32	17	24
35	251	21
12	24	25

Ordnen der Grauwerte:

$$12 \leq 17 \leq 21 \leq 24 \leq \underbrace{24}_{\text{Median}} \leq 25 \leq 32 \leq 35 \leq 251$$

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

73.6 M-Schätzer

Sei $x_1, \dots, x_n$ Messwerte und $\Psi : [0, \infty) \rightarrow \mathbb{R}$ eine monoton wachsende Straffunktion (engl.: penaliser). Dann nennt man dasjenige μ , das

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

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

73.4 Beispiele

- (a) Beliebige 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 die quadratische Abstandssumme

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

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

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

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

- (iv) $p \rightarrow \infty$ ergibt den Midrange $\frac{\max\{x_i\} + \min\{x_i\}}{2}$

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

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

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

§74: FEHLERFORTPFLANZUNG

74.1

Es werden 2 physikalische Größen x und y mehrmals gemessen. Man erhält als Mittelwert $\bar{x}, \bar{y}$ und als (empirische) Standardabweichung s_x, s_y (vgl. 70.4). Nun will man hieraus eine neue Größe $z = f(x, y)$ berechnen. Als Schätzer für den Erwartungswert verwendet man

$$\bar{z} := f(\bar{x}, \bar{y})$$

Wie gehen jedoch die Meßfehler s_x, s_y in die Standardabweichung s_z ein? Man kann zeigen:

74.2 Satz: (Fehlerfortpflanzungsgesetz von Gauß)

Schätzt man aus 2 Fehler behafteten Größen x, y mit Mittelwert $\bar{x}, \bar{y}$ und Standardabweichung s_x, s_y eine neue Größe $z = f(x, y)$, so ist der Schätzer für ihren Erwartungswert gegeben durch $\bar{z} = f(\bar{x}, \bar{y})$. Für die Varianz von z gilt:

$$s_z^2 = \left(\frac{\partial f}{\partial x} \right)^2 \bigg|_{\bar{z}} s_x^2 + \left(\frac{\partial f}{\partial y} \right)^2 \bigg|_{\bar{z}} s_y^2$$

Beweisidee: Taylorentwicklung.

74.3 Beispiel:

Zwei Widerstände R_1, R_2 werden mehrmals gemessen:

$$\begin{aligned} \bar{R}_1 &= 100 \text{ (Ohm)} \\ \bar{R}_2 &= 200 \\ s_{R_1} &= 0.8 \quad (\text{Schreibweise: } R_1 = 100 \pm 0.8) \\ s_{R_2} &= 1 \end{aligned}$$

Wie groß sind Gesamt Widerstand und Fehler in Parallelschaltung?

$$\begin{aligned} \frac{1}{R} &= \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_2 + R_1}{R_1 R_2} \Rightarrow R = \frac{R_1 R_2}{R_1 + R_2} \\ \bar{R} &= \frac{\bar{R}_1 \cdot \bar{R}_2}{\bar{R}_1 + \bar{R}_2} = \frac{100 \cdot 200}{100 + 200} \approx 66.67 \end{aligned}$$

$$\frac{\partial R}{\partial R_1} = \frac{(R_1 + R_2)R_2 - R_1R_2}{(R_1 + R_2)^2} = \frac{R_2^2}{(R_1 + R_2)^2}$$

$$\left. \frac{\partial R}{\partial R_1} \right|_{\bar{R}} \approx 0.44$$

$$\frac{\partial R}{\partial R_2} = \frac{R_1^2}{(R_1 + R_2)^2}$$

$$\left. \frac{\partial R}{\partial R_2} \right|_{\bar{R}} \approx 0.11$$

$$\begin{aligned} s_R &= \sqrt{\left(\left. \frac{\partial R}{\partial R_1} \right|_{\bar{R}}^2 s_{R_1}^2 + \left(\left. \frac{\partial R}{\partial R_2} \right|_{\bar{R}}^2 s_{R_2}^2 \right)} \\ &= \sqrt{0.44^2 \cdot 0.8^2 + 0.11^2 \cdot 1^2} \\ &\approx 0.37 \end{aligned}$$

$$R = \bar{R} \pm s_R \approx 66.67 \pm 0.37$$