Robuste Statistik

Musterlösung zu Blatt 3

Aufgabe 3.1:

Gegeben ist der Datensatz 2, 3, 5, 6, 9.

a)

Eine R-Funktion, die die Lokations-Tiefe implementiert, sieht beispielsweise so aus:

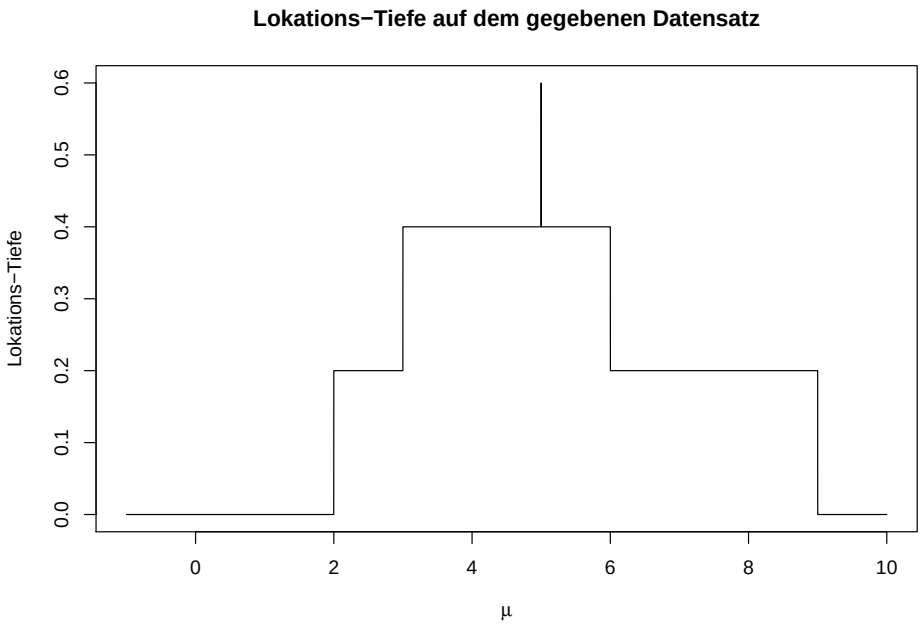
```
# lokationstiefe: Funktion zur Berechnung der Lokations-Tiefe
# Eingabe: mu (numeric): Numerischer Vektor der Parameter mu, fuer die die
#                        Lokations-Tiefe berechnet werden soll
#          y (numeric): Numerischer Vektor mit den Daten
# Ausgabe: (numeric): Numerischer Vektor der gleiche Laenge wie mu, der die
#                     Lokations-Tiefen angibt
lokationstiefe <- function(mu, y){
  sapply(mu, function(x) 1/length(y) * min(sum(y <= x), sum(y >= x)))
}
```

Für $\mu \in \{-1, 2.5, 5, 7, 10\}$ bekommt man dann die jeweiligen Lokations-Tiefen mit:

```
data1 <- c(2, 3, 5, 6, 9)
mu <- c(-1, 2.5, 5, 7, 10)
lokationstiefe(mu, data1)
[1] 0.0 0.2 0.6 0.2 0.0
```

b)

Grafische Darstellung der Lokations-Tiefe auf dem Intervall $[-1, 10]$



Die maximale Lokations-Tiefe ist bei $\mu = 5$ gegeben. Bekannterweise ist hier der Median der Daten.

c)

Nach Satz 3.2.6a) ist μ genau dann ein Lokations-Nonfit, wenn alle Datenpunkte entweder kleiner oder größer als μ sind. In diesem Fall sind also die -1 und die 10 Lokation-Nonfits, da die -1 kleiner als alle Datenpunkte ist und die 10 größer als alle Datenpunkte ist.

Aufgabe 3.2:

Zu zeigen ist

$$\arg\min_{\mu\in\mathbb{R}} \sum_{n=1}^{N} \rho(z_n - \mu) = \arg\max_{\mu\in\mathbb{R}} f_\mu(z_1, \ldots, z_N),$$

wobei f_μ die Dichtefunktion einer Verteilungsklasse mit Parameter μ sein soll. Mit folgendem Ansatz können wir die Darstellung der Dichtefunktion f_μ gewinnen (siehe Bemerkung 3.2.2):

$$\begin{aligned} \rho(z_n - \mu) &\overset{!}{=} -\log(f_\mu(z_n)) \\ |z_n - \mu| &\overset{!}{=} -\log(f_\mu(z_n)) \\ \exp\left(-|z_n - \mu|\right) &\overset{!}{=} f_\mu(z_n) \end{aligned}$$

Nun ist zu untersuchen, ob f_μ tatsächlich die Dichtfunktion einer Verteilungsklasse zum Parameter μ ist. Man beachte, dass $f_\mu(z) \geq 0$ für alle $z \in \mathbb{R}$ und $\mu \in \mathbb{R}$ gilt, aber

$$\int_{-\infty}^{\infty} \exp\left(-|z - \mu|\right)\, dz = 2$$

und nicht 1 ist. Deswegen setzen wir stattdessen

$$f_\mu(z) := \frac{1}{2} \exp\left(-|z_n - \mu|\right),$$

was gerade der Dichtefunktion der Doppelexponentialverteilung entspricht (dieses Wissen wurde nicht explizit bei der Lösung der Aufgabe erwartet). Nun können wir rechnen

$$\begin{aligned} \arg\min_{\mu\in\mathbb{R}} \sum_{n=1}^{N} \rho(z_n - \mu) &= \arg\min_{\mu\in\mathbb{R}} \sum_{n=1}^{N} |z_n - \mu| \\ &= \arg\min_{\mu\in\mathbb{R}} \left(-\log\left(\exp\left(-\sum_{n=1}^{N} |z_n - \mu| \right) \right) \right) \\ &= \arg\max_{\mu\in\mathbb{R}} \left(\log\left(2^N \prod_{n=1}^{N} \frac{1}{2} \exp\left(-|z_n - \mu|\right) \right) \right) \\ &= \arg\max_{\mu\in\mathbb{R}} \left(\prod_{n=1}^{N} f_\mu(z_n) \right) \\ &\overset{\text{i.i.d.}}{=} \arg\max_{\mu\in\mathbb{R}} f_\mu(z_1, \ldots, z_N). \end{aligned}$$

Man beachte hierbei, die Rechenregeln für arg min und arg max nach streng monoton wachsenden und fallenden Transformationen.

Aufgabe 3.3:

Gegeben sind die Daten 0, 0, 0, 0, 10 und die beiden Schätzer

$$\hat{\theta}_1^p(\boldsymbol{y}) = p \cdot \arg\min_{\theta \in \mathbb{R}} \sum_{n=1}^{N} (y_n - \theta)^2 + (1-p) \cdot \arg\min_{\theta \in \mathbb{R}} \sum_{n=1}^{N} |y_n - \theta|, \ p \in [0,\, 1],$$

$$\hat{\theta}_2^p(\boldsymbol{y}) = \arg\min_{\theta \in \mathbb{R}} \sum_{n=1}^{N} (p \cdot (y_n - \theta)^2 + (1-p) \cdot |y_n - \theta|), \ p \in [0,\, 1].$$

Die Idee hinter den beiden Schätzern ist es das arithmetische Mittel und den Median mittels einer Konvexkombination zu verbinden. Beim ersten Schätzer werden die Optimierungsprobleme, die als Lösung das arithmetische Mittel bzw. den Median haben, mit einer Konvexkombination verbunden. Beim zweiten Schätzer wird zunächst eine Konvexkombination gebildet, welche anschließend optimiert wird.

a) 1. Möglichkeit: Analytische Optimierung

$$\begin{aligned}
\hat{\theta}_1^p(\boldsymbol{y}) &= p \cdot \arg\min_{\theta \in \mathbb{R}} \sum_{n=1}^{N} (y_n - \theta)^2 + (1-p) \cdot \arg\min_{\theta \in \mathbb{R}} \sum_{n=1}^{N} |y_n - \theta| \\
&= p \cdot \bar{y} + (1-p) \cdot med(\boldsymbol{y}) \\
&= p \cdot 2 + (1-p) \cdot 0 \\
&= 2 \cdot p
\end{aligned}$$

Die zweite Zeile gilt, da (Beweis siehe Vorlesung) $\arg\min\limits_{\theta \in \mathbb{R}} \sum\limits_{n=1}^{N} (y_n - \theta)^2 = \bar{y}$ und $\arg\min\limits_{\theta \in \mathbb{R}} \sum\limits_{n=1}^{N} |y_n - \theta| = med(\boldsymbol{y})$. Es ist danach leicht zu sehen, dass für diesen Datensatz $\bar{y} = 2$ und $med(\boldsymbol{y}) = 0$ gilt.

Zur Berechnung des zweiten Schätzers definiere zunächst:

$$\begin{aligned}
f(\theta) :&= \sum_{n=1}^{N} (p \cdot (y_n - \theta)^2 + (1-p) \cdot |y_n - \theta|) \\
&= 4 \cdot (p \cdot (0-\theta)^2 + (1-p) \cdot |0-\theta|) + p \cdot (10-\theta)^2 + (1-p) \cdot |10-\theta| \\
&= 4p\theta^2 + 4 \cdot (1-p) \cdot |\theta| + p \cdot (100 - 20\theta + \theta^2) + |10-\theta| - p \cdot |10-\theta| \\
&= 4p\theta^2 + 4|\theta| - 4p|\theta| + 100p - 20\theta p + p\theta^2 + |10-\theta| - p|10-\theta| \\
&= 5p\theta^2 + 4|\theta| - 4p|\theta| + 100p - 20\theta p + |10-\theta| - p|10-\theta|
\end{aligned}$$

Nun kann man sich überlegen, dass ein Lageschätzer bei dieser Datensituation Werte im Intervall [0, 10] annehmen sollte/muss. Unter der Annahme, dass $\theta \in [0,\, 10]$ ist, sind alle Beträge in obiger Funktion positiv. Dies führt dann zu:

$$\begin{aligned}
\tilde{f}(\theta) &= 5p\theta^2 + 4\theta - 4p\theta + 100p - 20\theta p + 10 - \theta - p \cdot (10-\theta) \\
&= 5p\theta^2 + 4\theta - 4p\theta + 100p - 20\theta p + 10 - \theta - 10p + p\theta \\
&= 5p\theta^2 - 23p\theta + 3\theta + 90p + 10
\end{aligned}$$

Zur Optimierung muss nun bekanntermaßen die erste Ableitung gleich Null gesetzt werden und zur Überprüfung, ob ein Minimum vorliegt, die zweite Ableitung an der in Frage kommenden Stelle positiv sein.

$$\tilde{f}'(\theta) = 10p\theta - 23p + 3$$

$\Rightarrow$

$$\begin{aligned} 10p\theta - 23p + 3 &= 0 \\ 10p\theta &= 23p - 3 \\ \theta &= \frac{23p-3}{10p} \end{aligned}$$

$$\tilde{f}''(\theta) = 10p > 0 \forall \theta \Rightarrow \text{Lokales Minimum}$$

Somit ist $\hat{\theta} = \frac{23p-3}{10p}$ der gesuchte Schätzer. Allerdings nur für $\theta \in [0, 10]$. Für kleines p wird der Bruch jedoch negativ. Genauer gesagt:

$$\begin{aligned} \frac{23p-3}{10p} &< 0 \\ 23p - 3 &< 0 \\ 23p &< 3 \\ p &< \frac{3}{23} \end{aligned}$$

Somit ist obiger Schätzwert nur für alle $p \geq \frac{3}{23}$ korrekt. Für $p < \frac{3}{23}$ hat obige Funktion auf dem kompakten Intervall $[0, 10]$ kein lokales Minimum. Eine Randwertbetrachtung führt zu dem Ergebnis, dass $\hat{\theta} = 0$ für alle $p < \frac{3}{23}$.

2. Möglichkeit: Numerische Optimierung

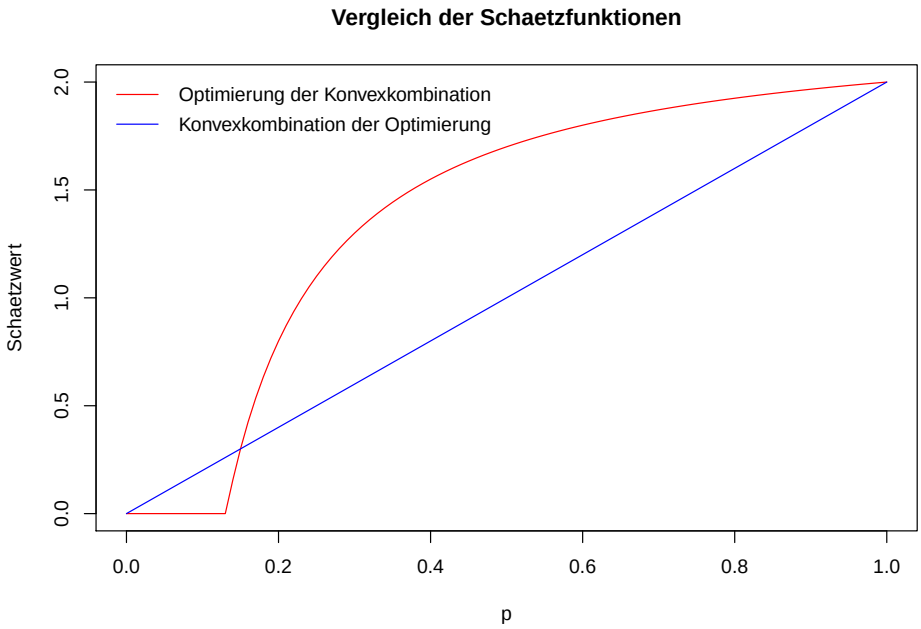
Funktionen für die beiden Schätzer:

```
# estimator1: Funktion, die den Schaetzer theta_1 implementiert
# Eingabe: p (numeric): Numerischer Vektor mit Werten fuer p, fuer die der
#                       Schaetzer berechnet werden soll
#          y (numeric): Datensatz, fuer den der Schaetzer berechnet werden soll
# Ausgabe: (numeric): numerischer Vektor der gleichen Laenge wie p, der die
#                     Schaetzwerte angibt
estimator1 <- function(p, y){
  p * optimize(function(theta) sum(y - theta)^2, lower = 0, upper = 10)$minimum +
    (1 - p) * optimize(function(theta) sum(abs(y - theta)), lower = 0, upper = 10)$minimum
}


# estimator2: Funktion, die den Schaetzer theta_2 implementiert
# Eingabe: p (numeric): Numerischer Vektor mit Werten fuer p, fuer die der
#                       Schaetzer berechnet werden soll
#          y (numeric): Datensatz, fuer den der Schaetzer berechnet werden soll
# Ausgabe: (numeric): numerischer Vektor der gleichen Laenge wie p, der die
#                     Schaetzwerte angibt
estimator2 <- function(p, y){
  sapply(p, function(x){
    optimize(function(theta) sum(x * (y - theta)^2 + (1 - x) * abs(y - theta)),
      lower = 0, upper = 10)$minimum
```

```
  })
}
```

Grafische Darstellung:



Wie zu sehen ist, sind die beiden Schätzer nicht identisch, obwohl sie den gleichen Grundgedanken haben. Bei $p = 0$ kommt bei beiden Schätzern Null heraus, was dem Median entspricht. Für $p = 1$ liefern beide Schätzer den Wert des arithmetischen Mittels.

b)

Während der erste Schätzer von jedem Anteil $p > 0$, den das arithmetische Mittel zum Schätzwert beiträgt, „verfälscht“ wird, bleibt der zweite Schätzer beim Schätzwert von 0 bis zu einem p von $\frac{3}{23}$. Dafür liefert der zweite Schätzer für große p stets größere (also mehr verfälschte) Schätzwerte als der erste Schätzer. Somit hat jeder der beiden Schätzer seine Vor- und Nachteile. Es fällt daher schwer einen klaren „Sieger“ zwischen den beiden Schätzern zu küren. Rechnerisch und von der Interpretation her ist der erste Schätzer deutlich einfacher. Dennoch sollten die Vorteile des zweiten Schätzers nicht unterschlagen werden. Generell haben beide Schätzer den Nachteil, dass die Wahl von p nicht klar ist und ein gewisser willkürlicher Spielraum vorzuliegen scheint.