Musterlösung zur Übung 1

1. a) Zur Erinnerung: Das α-gestutzte Mittel von $(X_1, \ldots, X_n)$ ist

$$\hat{\mu}_n = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)} \,.$$

Es seien $X_{(1)} = x_1 \,, X_{(2)} = x_2 \ \ldots, X_{(n)} = x_n$ und x beobachtet worden. Von diesen $n+1$ (geordneten) Beobachtungen werden sowohl die kleinsten $k = [(n+1)\alpha]$ als auch die grössten k Beobachtungen weggestutzt. Falls $x \leq x_k$, dann wird x selbst auch gestutzt, und es gilt

$$\hat{\mu}_{n+1}(x_1, \ldots, x_n, x) = \frac{1}{n+1-2k} \sum_{i=k}^{n-k} x_i \quad (\text{falls } x \leq x_k) \,.$$

Wenn x gross ist, wird es ebenfalls weggestutzt:

$$\hat{\mu}_{n+1}(x_1, \ldots, x_n, x) = \frac{1}{n+1-2k} \sum_{i=k+1}^{n-k+1} x_i \quad (\text{falls } x \geq x_{n-k+1}) \,,$$

und in allen übrigen Fällen gilt

$$\hat{\mu}_{n+1}(x_1, \ldots, x_n, x) = \frac{1}{n+1-2k} \Big(\sum_{i=k+1}^{n-k} x_i + x \Big) \quad (\text{sonst}) \,.$$

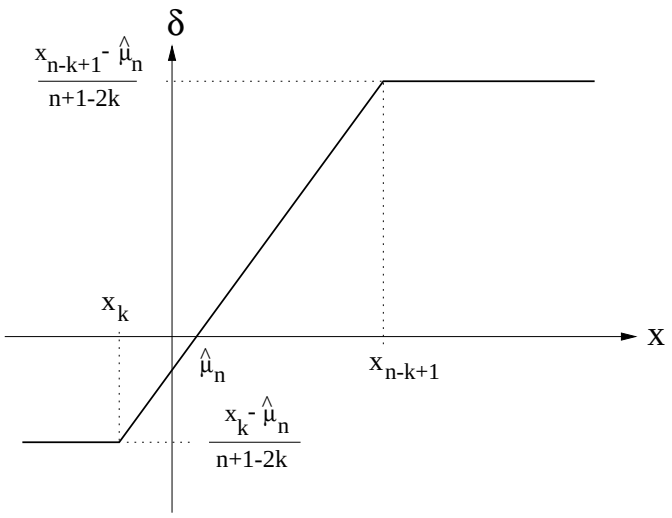
Ohne Fallunterscheidung ausgedrückt:

$$\hat{\mu}_{n+1}(x_1, \ldots, x_n, x) = \frac{1}{n+1-2k} \Big(\sum_{i=k+1}^{n-k} x_i + \max\{x_k, \min\{x, x_{n-k+1}\}\} \Big)$$

Als nächstes betrachten wir $\hat{\mu}_n$. Wir nehmen an, dass von der kleineren Stichprobe gleich viele Beobachtungen gestutzt werden wie von der grösseren, d. h. $[n\alpha] = [(n+1)\alpha] = k$. (Möglich wäre ja auch $[n\alpha] = [(n+1)\alpha] - 1$. Die Ergebnisse sind jedoch dieselben (o. Bew.).)

$$\begin{aligned}
\delta &:= \hat{\mu}_{n+1}(x_1, \ldots, x_n, x) - \hat{\mu}_n(x_1, \ldots, x_n) \\
&= \Big(\frac{1}{n+1-2k} - \frac{1}{n-2k} \Big) \sum_{i=k+1}^{n-k} x_i + \frac{1}{n+1-2k} \max\{x_k, \min\{x, x_{n-k+1}\}\} \\
&= \frac{\max\{x_k, \min\{x, x_{n-k+1}\}\} - \hat{\mu}_n(x_1, \ldots, x_n)}{n+1-2k}
\end{aligned}$$

Die folgende Figur zeigt die gesuchte Veränderung δ als Funktion der zusätzlichen Beobachtung x.



Aus der Figur kann man sehr schön die Sensitivität γ^* des α-gestutzten Mittels ablesen:

$$\gamma^* = \frac{1}{n+1-2k} \max\{X_{(n-k+1)} - \hat{\mu}_n, \hat{\mu}_n - X_{(k)}\}$$

b)

$$\begin{aligned}
\hat{\beta} &= \operatorname{argmin}_\beta \sum_{i=1}^n |y_i - \beta x_i| = \operatorname{argmin}_\beta \sum_{i=1}^n \frac{1}{\sum_{j=1}^n x_j} |z_i x_i - \beta x_i| \\
&\overset{x_i>0}{=} \operatorname{argmin}_\beta \sum_{i=1}^n \omega_i |z_i - \beta| = \operatorname{argmin}_\beta \sum_{i=1}^n P[Z = z_i] \, |z_i - \beta| \\
&= \operatorname{argmin}_\beta \mathbf{E}[|Z - \beta|] \overset{\text{A.8}}{=} \operatorname{median}(Z)
\end{aligned}$$

Wenn in einem einzigen Pärchen (y_i, x_i) x_i so gross gemacht wird, dass $\omega_i = x_i / \sum_{j=1}^n x_j > \frac{1}{2}$ (möglich, da $\omega_i \to 1 \, (x_i \to \infty)$), dann gilt also $P[Z = z_i] > \frac{1}{2}$. Also ist z_i der Median von Z, und $\hat{\beta} = z_i$. Lässt man $y_i \to \infty$, dann $\hat{\beta} = z_i \to \infty$. Ein einziger Ausreisser (y_i, x_i) kann also zum Zusammenbruch von $\hat{\beta}$ führen:

$$\varepsilon^*(\hat{\beta}) = \frac{1}{n}$$

c) Sei $I = \{i_1, \ldots, i_k\}$ mit $\sum_{j \in I} \omega_j > \frac{1}{2}$. Dann folgt mit derselben Argumentation wie in b), dass falls $y_j \to \infty \quad \forall j \in I$, dann auch $\hat{\beta} \to \infty$. Gesucht ist also das kleinste m, für das $\exists \; i_1, \ldots, i_m$:

$$\sum_{j=1}^m \omega_{i_j} > \frac{1}{2}$$

Wegen $x_i = \frac{i}{n}$ ist dies offensichtlich für $i_j = n-m+1, \ldots, n$ der Fall. Mit

$$\begin{aligned}\sum_{j=n-m+1}^{n} \omega_j &= \sum_{j=n-m+1}^{n} \frac{\frac{j}{n}}{\frac{n(n+1)}{2n}} = \frac{2}{n(n+1)} \cdot \Big(\frac{n(n+1)}{2} - \frac{(n-m)(n-m+1)}{2}\Big) \\ &= \frac{2mn - m^2 + m}{n^2 + n}\end{aligned}$$

folgt für den Bruchpunkt

$$\varepsilon^*(\hat{\beta}) = \frac{\min\{m; \frac{2mn-m^2+m}{n^2+n} > \frac{1}{2}\}}{n} \overset{\alpha=\frac{m}{n}}{=} \min\{\alpha; \frac{2\alpha n^2 - \alpha^2 n^2 + \alpha n}{n^2+n} > \frac{1}{2}\}.$$

Für $n \to \infty$

$$\varepsilon^*_\infty(\hat{\beta}) = \min\{\alpha; 2\alpha - \alpha^2 > \frac{1}{2}\} = 1 - \frac{\sqrt{2}}{2}$$

2. a)

$$\begin{aligned}f_{\varepsilon_1+\varepsilon_2}(z) &= \int_{-\infty}^{\infty} f_{\varepsilon_1}(y) f_{\varepsilon_2}(z-y)\, dy = \int_{-\infty}^{\infty} \frac{1}{\pi}\frac{1}{1^2+y^2}\frac{1}{\pi}\frac{1}{1^2+(z-y)^2}\, dy \\ &= \frac{1}{\pi^2}\frac{1}{z(4+z^2)}\int_{-\infty}^{\infty}\Big(\frac{2y+z}{1+y^2} + \frac{2(z-y)+z}{1+(z-y)^2}\Big)\, dy \\ &= \frac{1}{\pi^2}\frac{1}{z(4+z^2)}\big[0 + z\,\arctan y\big|_{-\infty}^{\infty} + 2\cdot 0 + z(-\arctan(z-y))\big|_{-\infty}^{\infty}\big] \\ &= \frac{1}{\pi^2 z(4+z^2)} 2\pi z = \frac{2}{\pi(2^2+z^2)}\end{aligned}$$

Also ist $\varepsilon_1 + \varepsilon_2$ Cauchy(2)-verteilt.

b)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu + \frac{1}{n} \underbrace{\sum_{i=1}^{n} \varepsilon_i}_{\sim \text{Cauchy}_{(n\lambda)}}$$

Sei $Z \sim \text{Cauchy}(n\lambda)$. Dann ist $Y := \frac{1}{n}Z \sim \text{Cauchy}(\lambda)$ wegen

$$\begin{aligned}f_Y(y) &= \frac{1}{|\det(J(g^{-1}(y)))|} f_Z(g^{-1}(y))\,, \text{ wobei } y = g(z) = \frac{z}{n}\,,\ J(z) = \frac{\partial g}{\partial z}(z) \equiv \frac{1}{n} \\ &= \frac{1}{\frac{1}{n}} f_Z(ny) = \frac{1}{\pi}\frac{n\lambda}{(n\lambda)^2 + (ny)^2}\, n = \frac{\lambda}{\pi(\lambda^2+y^2)}\,.\end{aligned}$$

Also hat $\overline{X} = \mu + Z$, $Z \sim$ Cauchy(λ), die gleiche Verteilung wie X_1. Das ist kein Widerspruch zum GGZ, da die dort nötige Voraussetzung $\mathbf{E}[|X_i|] < \infty$ für $X_i \sim$ Cauchy(λ) nicht erfüllt ist. (Bei der Cauchyverteilung existiert der Erwartungswert nicht einmal!)

c) Aus Symmetriegründen gilt $P[|\hat{\mu}_{2n+1} - \mu| > c] = 2P[\hat{\mu}_{2n+1} > \mu + c]$. Weiter gilt

$$\begin{aligned} 2P[\hat{\mu}_{2n+1} > \mu + c] &= 2P[\text{mindestens } n+1 \text{ Beob.} > \mu + c] \\ &= 2P\big[\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} > n\big] \\ &= 2P\big[\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} > n + \frac{1}{2}\big] \\ &= 2P\big[\frac{1}{2n+1}\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} > \frac{1}{2}\big]\,. \end{aligned}$$

Nach dem schwachen GGZ gilt

$$\frac{1}{2n+1}\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} \xrightarrow{P} \mathbf{E}[\mathbf{I}_{\{X_i > \mu + c\}}] = P[X_i > \mu + c] \overset{c>0}{<} \frac{1}{2}\,,$$

das heisst $\forall \varepsilon > 0$

$$\begin{aligned} & P\big[\frac{1}{2n+1}\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} - P[X_i > \mu + c] > \varepsilon\big] \\ \leq\ & P\big[\,\big|\frac{1}{2n+1}\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} - P[X_i > \mu + c]\,\big| > \varepsilon\big] \to 0\ (n \to \infty)\,. \end{aligned}$$

Insbesondere gilt das für $\varepsilon = \frac{1}{2} - P[X_i > \mu + c] > 0$, also $\forall c > 0$

$$\begin{aligned} P[|\hat{\mu}_{2n+1} - \mu| > c] &= 2P\big[\frac{1}{2n+1}\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} > \frac{1}{2}\big] \\ &= 2P\big[\frac{1}{2n+1}\sum_{i=1}^{2n+1} \mathbf{I}_{\{X_i > \mu + c\}} - P[X_i > \mu + c] > \frac{1}{2} - P[X_i > \mu + c]\big] \\ &\to 0\ (n \to \infty)\,. \end{aligned}$$

d) Für jedes $c > 0$ gilt $P[|\overline{X} - \mu| \leq c] = 1 - 2P[\overline{X} \leq \mu - c] = 1 - 2P[X_1 \leq \mu - c] < 1$ unabhängig von n. Für den Median jedoch folgt aus c) $P[|\hat{\mu}_{2n+1} - \mu| \leq c] \to 1\,(n \to \infty)$. In diesem Sinn ist der Median der bessere Schätzer.

3. **a)** Mittelwert: 18.2

b) Median: 26.5

c) MAD= $7, S_n = 24.71$. Mittel nach MAD-Verwerfungsregel: 25.11

d) Mittel nach S_n-Verwerfungsregel: 25.11 (Zwei Werte -44 wären als Ausreißer nicht mehr erkannt worden.)

e) 0.1-gestutztes Mittel: 23.25; 0.2-gestutztes Mittel: 25.67

f) Berechnung des Huber-Schätzers: Es gilt MAD= $7, c = 10.5$. Mit

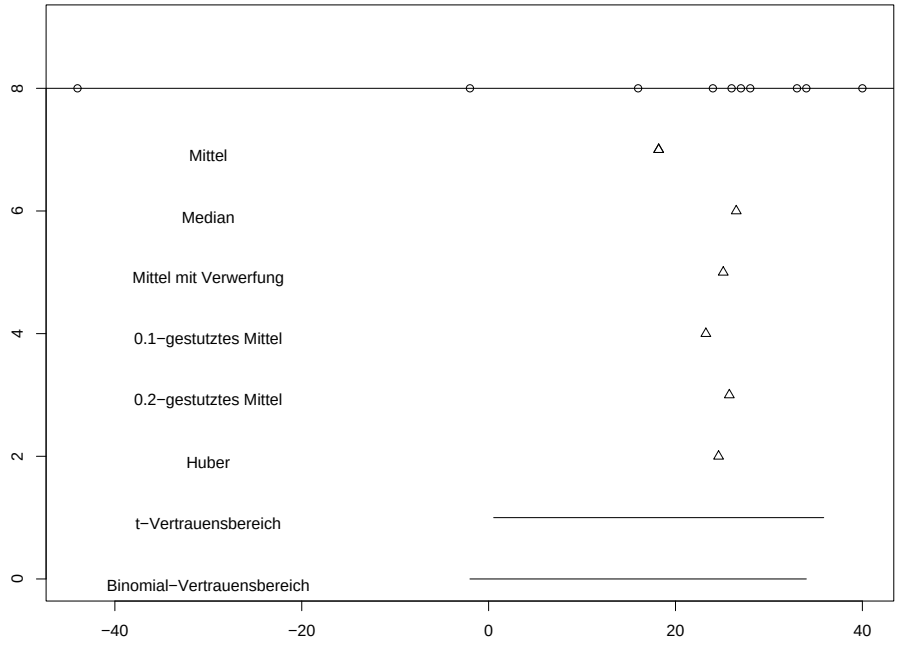
$$f(\mu) = \sum_{i=1}^{n} \rho(X_i - \mu) \Rightarrow f'(\mu) = -\sum_{i=1}^{n} \psi(X_i - \mu),$$
$$\psi(x) = \begin{cases} x & \text{für} \quad |x| \leq c \\ c & \text{für} \quad x > c \\ -c & \text{für} \quad x < -c \end{cases},$$
$$\hat{\mu} = \arg\min_{\mu} f(\mu) \Rightarrow f'(\hat{\mu}) = 0.$$

f' ist offenbar monoton wachsend in μ und konvergiert für $\mu \to \pm\infty$ gegen $\pm nc$. f' hat daher eine (wie sich zeigt, eindeutige) Nullstelle, die eine globale Minimalstelle von f anzeigt und z.B. durch Intervall-Halbierung approximiert werden kann. Es ergibt sich $\hat{\mu} = 24.61$.

g) Vertrauensbereich:

$$\{\mu | \frac{\sqrt{n}(\bar{X} - \mu)}{S_n} \in [-2.26, 2.26]\} = \left[\bar{X} - \frac{2.26 S_n}{\sqrt{n}}, \bar{X} + \frac{2.26 S_n}{\sqrt{n}}\right] = [0.54, 35.86].$$

h) Vertrauensbereich $[-2, 34]$.



Diskussion: Die Schätzer liegen relativ nah beieinander, mit Ausnahme von $\bar{X}$. Die Vertrauenswürdigkeit von $\bar{X}$ muss bezweifelt werden, denn unter den restlichen 56 Beobachtungen (im Skript) kommt nur noch ein einziger Wert vor, der kleiner als unser $\bar{X}$ auf der Basis von 10 Beobachtungen ist. Im Lichte der folgenden Daten kann man von einem extremen Ausreißer (nach unten) und zwei weiteren Ausreißer-verdächtigen Werten (einer nach unten, einer nach oben) ausgehen. Das 0.2-gestutzte Mittel wirft alle diese Werte heraus, das 0.1-gestutzte Mittel scheint dagegen ebenso wie der Huber-Schätzer noch etwas nach unten abgelenkt zu sein. Die Verwerfungsregeln haben hier den Vorteil, nicht oben und unten gleich viele Daten zu verwerfen.

Die beiden Vertrauensbereiche sind ungefähr gleich lang und zeigen, dass die Daten noch keine genaue Angabe von μ erlauben.