Robuste Statistik

Musterlösung zu Blatt 1

Aufgabe 1.1:

Wir betrachten den Datensatz 2, 3, 5, 6, 9.

Zunächst: Die Standardabweichung des Originaldatensatzes beträgt $\sqrt{7.5} \approx 2.739$ und der MAD beträgt 2.

a)

x_0	sd	mad
-100 000	44723.148	2
-10 000	4473.925	2
-1 000	449.005	2
-100	46.537	2
-10	6.458	2
2	1.817	1
3	1.643	1
5	1.643	1
6	1.817	1
9	2.739	2
10	3.114	2
100	42.962	2
1 000	445.428	2
10 000	4470.347	2
100 000	44719.571	2

Tabelle 1: Ersetzen der 9 durch x_0

b)

x_0	sd	mad
-100 000	54774.082	3
-10 000	5479.051	3
-1 000	549.549	3
-100	56.608	3
-10	7.382	3
2	1.304	0
3	1.095	0
5	1.414	0
6	1.817	1
9	3.286	3
10	3.808	3
100	52.958	3
1 000	545.898	3
10 000	5475.400	3
100 000	54770.430	3

Tabelle 2: Ersetzen der 6 und 9 durch x_0

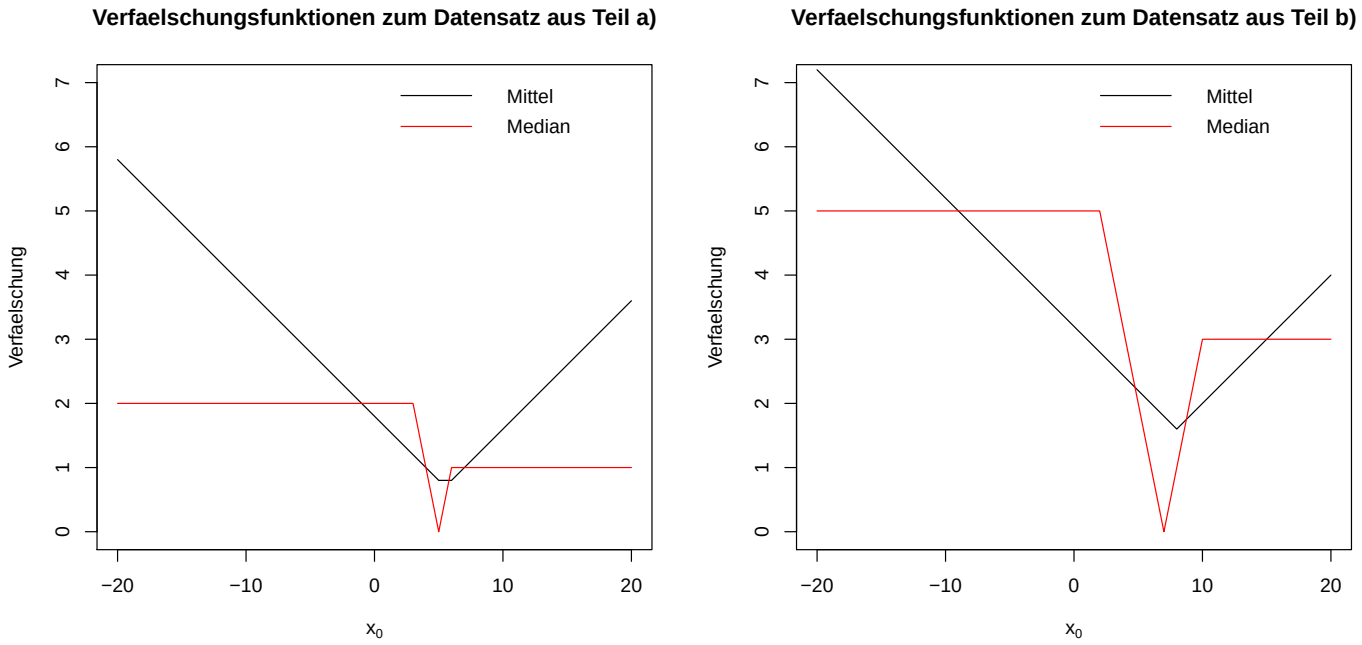
c)

x_0	sd	mad
-100 000	54773.625	0
-10 000	5478.595	0
-1 000	549.092	0
-100	56.143	0
-10	6.856	0
2	0.447	0
3	0.447	0
5	1.414	0
6	1.949	0
9	3.578	0
10	4.123	0
100	53.404	0
1 000	546.353	0
10 000	5475.856	0
100 000	54770.886	0

Tabelle 3: Ersetzen der 5, 6 und 9 durch x_0

Es ist zu erkennen, dass die Standardabweichung größer wird bei betragsmäßig großen Ausreißern. Dabei genügt schon ein einziger Ausreißer. Ein zweiter oder dritter Ausreißer ändert nicht mehr viel am Gesamtbild. Der MAD hingegen zeigt ein anderes Bild: Während ein Ausreißer den MAD um maximal Eins verändert, verändert sich bei zwei bzw. drei Ausreißern der MAD um bis zu Zwei. Dies klingt zwar nach nicht viel und vor allem deutlich weniger als bei der Standardabweichung, dennoch ist hier ein fatales Verhalten für einen Streuungsschätzer zu erkennen: Er wird Null. Man kann sich leicht überlegen, dass der MAD immer Null wird, wenn mehr als die Hälfte der Daten konstant ist. Dieses Verhalten ist unerwünscht, da ein Streuungsschätzer nur dann den Wert Null annehmen sollte, wenn die Daten komplett konstant sind.

Aufgabe 1.2:



Für den Datensatz aus Teil a) (2, 3, 5, 6, 9) ergibt sich eine maximale Verfälschung des arithmetischen Mittels von ∞, da die Verfälschungsfunktion nach oben unbeschränkt ist. Für den Median ist die maximale Verfälschung 2. Diese wird für alle Werte ≤ 3 angenommen.

Auch bei dem Datensatz aus Teil b) (0, 2, 7, 10, 16) ist die maximale Verfälschung des arithmetischen Mittels unendlich. Der Median hat eine maximale Verfälschung von 5 für alle Werte ≤ 2.

Allgemein gilt: Der Median verschiebt sich durch das Ersetzen einer Beobachtung durch eine beliebige Zahl in einem Datensatz mit einer ungeraden Anzahl an Beobachtungen maximal um eine Stelle des geordneten Datensatzes. Daher ist die maximale Verfälschung gegeben durch das Maximum des Abstands des ursprünglichen Medians zu seiner nächstkleineren Beobachtung und des Abstands des ursprünglichen Medians zu seiner nächstgrößeren Beobachtung.