

# Über Median, arithmetisches Mittel und Ungleichungen

FRANK MAROHN, WÜRZBURG

**Zusammenfassung:** In diesem Artikel geht es um Ungleichungen, die sich auf Daten beziehen. Es werden einfache Schranken für den Abstand zwischen empirischem Mittelwert und Median vorgestellt. Ferner wird die Samuelson-Ungleichung diskutiert, die eine Aussage darüber macht, wie weit ein einzelner Datenwert vom empirischen Mittelwert entfernt sein kann.

## 1 Einleitung

In der Wahrscheinlichkeitsrechnung spielen Ungleichungen eine wichtige Rolle. Man denke etwa an die Tschebyschow-Ungleichung, Grundlage für den Beweis des schwachen Gesetzes der großen Zahlen (Henze (2021), Irle (2001), Kütting und Sauer (2011)). In der deskriptiven Statistik hat man es mit Daten zu tun, stochastische Modelle spielen hier keine Rolle. Ungleichungen im Rahmen der deskriptiven Statistik können sich daher nur auf Daten beziehen. Im Folgenden wollen wir Ungleichungen im Zusammenhang mit dem empirischen Mittelwert und Median vorstellen.

Median und arithmetisches Mittel sind Maßzahlen für die zentrale Tendenz in den Daten. Der Median ist bekanntlich eine robuste Alternative zum arithmetischen Mittel. Ausgangspunkt bilden *quantitative* Daten  $x_1, \dots, x_n$ ,  $n \geq 2$ . Quantitativ bedeutet, dass die Daten „in natürlicher Weise“ als Zahlen vorliegen (hinter den Zahlen steckt somit keine Kodierung). Mit anderen Worten: Das zugrunde liegende Merkmal ist *metrisch skaliert*. Um den Median zu definieren (genauer müssten wir im Folgenden vom *empirischen* Median sprechen, da dieser sich auf Daten bezieht und nicht auf ein stochastisches Modell), ordnen wir die Daten der Größe nach und schreiben für die geordneten Daten

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

In unserer Notation ist demnach  $x_{(1)}$  das Minimum und  $x_{(n)}$  das Maximum. Der Median, auch Zentralwert genannt, ist der Datenwert, der sich in der mittleren Position der *geordneten* Daten befindet (bei einer *geraden* Anzahl von Daten gibt es zwei Beobachtungen in einer mittleren Position und es wird gemittelt). Formal: Der *Median* von  $x_1, \dots, x_n$  ist die Zahl

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ ungerade} \\ (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})/2, & n \text{ gerade.} \end{cases}$$

Die folgende Eigenschaft ist leicht nachzuweisen: *Mindestens* 50% der Daten sind kleiner oder gleich  $\tilde{x}$ , *mindestens* 50% der Daten sind größer oder gleich  $\tilde{x}$ . Ohne den Zusatz „mindestens“ wird dies Aussage falsch, siehe etwa nachfolgendes Beispiel. Im Fall  $n$  ungerade gilt

$$|\{i : x_i \leq \tilde{x}\}| \geq \frac{n+1}{2}$$

und

$$|\{i : x_i \geq \tilde{x}\}| \geq n - \frac{n+1}{2} + 1 = \frac{n+1}{2}.$$

Ist  $n$  gerade, so gilt im Fall  $x_{(\frac{n}{2})} < x_{(\frac{n}{2}+1)}$

$$|\{i : x_i \leq \tilde{x}\}| = \frac{n}{2} = |\{i : x_i \geq \tilde{x}\}|,$$

da  $x_{(\frac{n}{2})} < \tilde{x} < x_{(\frac{n}{2}+1)}$ . Ist  $x_{(\frac{n}{2})} = x_{(\frac{n}{2}+1)}$ , so erhält man

$$|\{i : x_i \leq \tilde{x}\}| \geq \frac{n}{2} + 1$$

und

$$|\{i : x_i \geq \tilde{x}\}| \geq \frac{n}{2} + 1.$$

Ist  $n$  gerade und  $x_{(\frac{n}{2})} < x_{(\frac{n}{2}+1)}$ , so ist  $\tilde{x}$  kein Datenwert. Möchte man sicher stellen, dass der Median mit einem Datenwert übereinstimmt, so muss dieser anders definiert werden. Es bieten sich zwei Möglichkeiten an. Man wählt als „unteren“ bzw. „oberen“ Median

$$\tilde{x}_u := x_{(\frac{n}{2})} \quad \text{bzw.} \quad \tilde{x}_o := x_{(\frac{n}{2}+1)}.$$

Für  $n$  gerade gilt stets

$$\tilde{x}_u \leq \tilde{x} \leq \tilde{x}_o.$$

Auch für  $\tilde{x}_o$  und  $\tilde{x}_u$  gilt: *Mindestens* 50% der Daten sind kleiner oder gleich  $\tilde{x}$ , *mindestens* 50% der Daten sind größer oder gleich  $\tilde{x}$ .

**Beispiel 1:** Zunächst ein historisches Beispiel über eine Studie zur Wirksamkeit zweier Medikamente, das in der Statistik berühmt geworden ist. Bei 10 Versuchspersonen wurde die durchschnittliche *Schlafverlängerung* durch Medikament A gegenüber Medikament B gemessen (in Stunden: Quelle Cushny und Peebles 1905, „Student“ 1908). Die Ergebnisse waren

$$1.2, 2.4, 1.3, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4.$$

Die geordneten Daten sind

0, 0.8, 1.0, 1.2, 1.3, 1.3, 1.4, 1.8, 2.4, 4.6.

Da  $n = 10$  gerade ist, gilt

$$\tilde{x} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{1.3 + 1.3}{2} = 1.3.$$

60% der Daten sind kleiner oder gleich  $\tilde{x}$ , 60% der Daten sind größer oder gleich  $\tilde{x}$ . Ferner ist

$$\tilde{x}_u = x_{(5)} = 1.3 = x_{(6)} = \tilde{x}_o.$$

Das wohl populärste Lagemaß für die zentrale Tendenz der Daten ist das *arithmetische Mittel*

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i.$$

Gegenüber dem Median, der die „Mitte“ der Daten repräsentiert, ist das arithmetische Mittel ein Durchschnittswert. Alle Datenwerte werden berücksichtigt und gleich gewichtet (mit  $1/n$ ). Das *arithmetische Mittel* besitzt nicht die oben angesprochene Eigenschaft des Medians. Im obigen Beispiel erhalten wir

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_{(i)} \\ &= \frac{1}{10} (0.0 + \dots + 4.6) = \frac{15.8}{10} = 1.58. \end{aligned}$$

Welches Lagemaß charakterisiert die Daten besser? Die Antwort hängt zum einen von der Problemstellung ab, zum anderen auch von Interessen, die man verfolgt. Anders als das arithmetische Mittel ist der Median robust. Ein Lagemaß heißt *robust*, wenn dass Verändern, Hinzufügen oder Weglassen eines einzelnen (extremen) Datenwertes (sog. Ausreißer) seinen Wert nicht stark beeinflusst. Angenommen, im obigen Beispiel wäre versehentlich 46 statt 4.6 notiert worden. Der Median bleibt unverändert, während der arithmetische Mittel von 1.58 auf 5.72 steigt.

Was lässt sich über den Abstand  $|\bar{x} - \tilde{x}|$  sagen? Wie weit kann ein einzelner Datenwert  $x_j$  vom empirischen Mittelwert  $\bar{x}$  entfernt sein? In den Abschnitten 4 und 5 werden wir auf Abschätzungen zu sprechen kommen. Zuvor ein Exkurs über aktuelle Publikationen, die zeigen, dass der Median gegenüber dem arithmetischen Mittel kein Schattendasein führt.

<sup>1</sup> Ct = Cycle threshold

RT-PCR = reverse transcription-polymerase chain reaction

## 2 Der Median in Publikationen zu COVID-19: Ein kurzer Streifzug

Das folgende Beispiel bezüglich des Merkmals *Alter* findet sich im Deutschen Ärzteblatt (Busse und Nimptsch 2021):

„Die im Krankenhaus behandelten COVID-19-Patienten waren im Median 71 Jahre alt.“

Auch beim *Titer* (Maß für die Menge von Antikörpern) wird der Median verwendet (Gallais et al. (2021).

Ausführlicher wollen wir auf den vielzitierten *Ct-Wert* bei einem PCR-Test auf eine SARS-CoV-2-Infektion eingehen. So ist in einer Studie von Brown et al. (2021) folgendes zu lesen:

„Real-time RT-PCR Ct values <sup>1</sup> in specimens from 127 fully vaccinated patients (median = 22.77) were similar to those among 84 patients who were unvaccinated, not fully vaccinated, or whose vaccination status was unknown (median = 21.54).“

Dazu eine Erläuterung: Der Ct-Wert, so ist häufig zu lesen, gibt die Anzahl der Vermehrungszyklen an, die innerhalb einer PCR-Probe durchgeführt werden, um das Corona-Virus nachweisen zu können. Ist der Prozess der Vermehrung (=Verdoppelung) abgeschlossen, beginnt der nächste Zyklus. Je höher der Ct-Wert ist, desto niedriger ist die ursprüngliche Viruskonzentration in der untersuchten Probe. Der aufmerksame Leser wird sich jetzt die Frage stellen, wie die Medianwerte in der obigen Studie zustande gekommen sind, wo doch das Merkmal *Anzahl* ganzzahlig ist (die Werte des Medians können daher nur ganzzahlig sein oder ganzzahlig + 0.5). Der Grund ist folgender. Beim Ct-Wert handelt es sich genauer um den Zeitpunkt, bei dem zuerst ein exponentieller Anstieg des Fluoreszenz-Signals beobachtet werden kann (siehe Abb. 1).

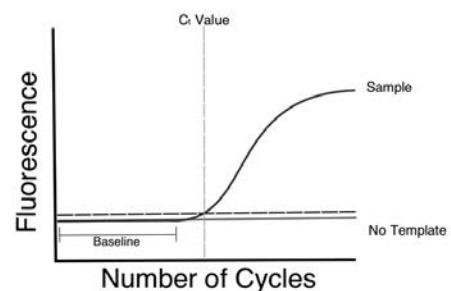


Abbildung 1: Kurvenverlauf eines RT-PCR-Prozesses

Dem Prinzip nach ist der Ct-Wert daher eine kontinuierliche Größe. Daher können Messwerte wie Ct = 22.34 (im 22sten Zyklus) auftreten. Bei Ct-Werten wird häufig der Median dem arithmetischen Mittel vorgezogen. Warum? Ein Grund dürfte folgender sein: Testergebnisse mit hohen Ct-Werten sind als negativ anzusehen. Aber was bedeutet hoch? Es gibt (weltweit) keinen einheitlichen, standardisierten Ct-Wert, ab dem eine positiv getestete Person nicht als infektiös gilt (etwaige Grenzwerte sind Richtwerte, z.B. Ct > 30 laut RKI). Daher bietet sich der Median hier an, wenn es um die Angabe eines mittleren Ct-Wertes von infektiös geltenden Personen geht. Natürlich kann man hier auch an das *getrimmte Mittel* denken, eine Robustifizierung des arithmetischen Mittels. Hier werden extreme Ct-Werte (z. B. ct > 30) weggelassen und über die restlichen Werte wird das arithmetische Mittel berechnet. Der Median umgeht die Verwendung eines Richtwertes.

Kehren wir zur deskriptiven Statistik zurück.

### 3 Die Minimierungseigenschaft

Für den Median  $\tilde{x}$  gilt

$$\min_{a \in \mathbb{R}} \sum_{i=1}^n |x_i - a| = \sum_{i=1}^n |x_i - \tilde{x}|.$$

Anschaulich und intuitiv ist dies auch leicht einzusehen (siehe z. B. die Diskussion in Henze (2021), Abschnitt 5.6). Um diese Minimierungseigenschaft zu beweisen, ist es aufgrund der Definition des Medians naheliegend, die Fälle  $n$  gerade und  $n$  ungerade zu betrachten, siehe Kütting und Sauer (2011), Kap. 11, Seite 375f. Wir wollen einen alternativen Beweis vorstellen, der auf diese Fallunterscheidung verzichtet.

Sei  $a > \tilde{x}$ . Zum Nachweis spalten wir den Indexbereich  $\{1, \dots, n\}$  in drei disjunkte Mengen auf:

$$I_1 = \{i : x_i \geq a\}, I_2 = \{i : x_i \leq \tilde{x}\}, I_3 = \{i : \tilde{x} < x_i < a\}.$$

Wir wissen aufgrund der Überlegungen aus Abschnitt 1

$$|I_2| \geq \frac{n}{2}$$

und somit, wegen  $I_1 \cup I_3 = \{1, \dots, n\} \setminus I_2$ ,

$$|I_1 \cup I_3| = n - |I_2| \leq \frac{n}{2}.$$

Wir erhalten

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

$$\begin{aligned} &= \sum_{i \in I_1} (x_i - a - (x_i - \tilde{x})) \\ &\quad + \sum_{i \in I_2} (a - x_i - (\tilde{x} - x_i)) \\ &\quad + \sum_{i \in I_3} \underbrace{(a - x_i - (x_i - \tilde{x}))}_{> -(a - x_i) - (x_i - \tilde{x}) = \tilde{x} - a} \\ &> \sum_{i \in I_1} (\tilde{x} - a) + \sum_{i \in I_2} (a - \tilde{x}) + \sum_{i \in I_3} (\tilde{x} - a) \\ &= \sum_{i \in I_1 \cup I_3} \underbrace{(\tilde{x} - a)}_{< 0} + \sum_{i \in I_2} \underbrace{(a - \tilde{x})}_{> 0} \\ &\geq \frac{n}{2}(\tilde{x} - a) + \frac{n}{2}(a - \tilde{x}) = 0. \end{aligned}$$

Analog wird der Fall  $a < \tilde{x}$  gezeigt.

Das arithmetische Mittel hingegen minimiert nicht die Summe der Abstände, sondern die Summe der quadratischen Abweichungen

$$\min_{a \in \mathbb{R}} \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Diese bekannte Eigenschaft kann man über die Ableitung der Funktion  $f(a) := \sum_{i=1}^n (x_i - a)^2$ ,  $a \in \mathbb{R}$ , nachweisen. Es geht auch ohne die Differentialrechnung: Nutzt man die *Schwerpunkteigenschaft* des arithmetischen Mittels aus,

$$\sum_{i=1}^n (x_i - \bar{x}) = 0,$$

so ergibt sich die Abschätzung

$$\begin{aligned} \sum_{i=1}^n (x_i - a)^2 &= \sum_{i=1}^n ((x_i - \bar{x}) - (a - \bar{x}))^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + \underbrace{n(a - \bar{x})^2}_{\geq 0} \\ &\quad - 2(a - \bar{x}) \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} \\ &\geq \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned}$$

### 4 Der Abstand zwischen Median und arithmetischem Mittel

Zunächst wollen wir auf den Fall  $\tilde{x} = \bar{x}$  eingehen, der eintritt, wenn die Daten symmetrisch verteilt sind. Dies ist eine hinreichende, aber keineswegs notwendige Bedingung, wie das Beispiel der 5 (ganzzahligen) Ct-Werte 16, 19, 20, 22, 23 zeigt. Betrachten wir dazu den Fall  $n$  ungerade. Dann ist (anschaulich)

klar, dass der Median  $\tilde{x} = x_{(\frac{n+1}{2})}$  das Symmetriezentrum ist und es gilt  $\tilde{x} = \bar{x}$ . Denn: (Exakte) Symmetrie meint Folgendes:

$$\begin{aligned}\tilde{x} - x_{(1)} &= x_{(n)} - \tilde{x} \\ \tilde{x} - x_{(2)} &= x_{(n-1)} - \tilde{x} \\ &\vdots \\ \tilde{x} - x_{(\frac{n-1}{2})} &= x_{(\frac{n+3}{2})} - \tilde{x}.\end{aligned}$$

Dies sind  $(n-1)/2$  Gleichungen. Summiert man beide Seiten auf und bringt die geordneten Werte sowie den Median jeweils auf eine Seite der Gleichung, so ergibt sich

$$\sum_{1 \leq i \leq n, i \neq \frac{n+1}{2}} x_{(i)} = \frac{n-1}{2} \cdot 2 \cdot \tilde{x}.$$

Addiert man jetzt noch auf beiden Seiten den Median  $\tilde{x} = x_{(\frac{n+1}{2})}$  dazu, so erhält man

$$\bar{x} = \sum_{i=1}^n x_{(i)} = \tilde{x}.$$

Es stellt sich nun die Frage, wie groß der Abstand zwischen Median und arithmetischem Mittel sein kann. (Obere) Schranken für den Abstand hängen zwangsläufig von den Daten ab. Bevor wir zu einer ersten und sehr groben Abschätzung kommen, sozusagen als Einstieg, sei an die folgende Ungleichung erinnert:

$$\min\{a_1, \dots, a_n\} \leq \frac{1}{n} \sum_{i=1}^n a_i \leq \max\{a_1, \dots, a_n\}$$

für reelle Zahlen  $a_1, \dots, a_n$ .

Die erste Abschätzung verwendet die *Spannweite* (= Maximum minus Minimum). Da Median und arithmetisches Mittel zwischen dem Minimum und dem Maximum liegen,  $x_{(1)} \leq \tilde{x} \leq x_{(n)}$  bzw.  $x_{(1)} \leq \bar{x} \leq x_{(n)}$ , gilt

$$|\bar{x} - \tilde{x}| \leq x_{(n)} - x_{(1)}. \quad (1)$$

Anschaulich ist diese Ungleichung völlig klar. Als nächstes verwenden wir die Aussage

$$a \leq b \leq c \Rightarrow |b| \leq \max\{|a|, |c|\}, \quad a, b, c \in \mathbb{R}.$$

Wenden wir diese Implikation auf  $x_{(1)} - \tilde{x} \leq \bar{x} - \tilde{x} \leq x_{(n)} - \tilde{x}$ , so erhalten wir die (etwas) bessere Abschätzung

$$|\bar{x} - \tilde{x}| \leq \max\{\tilde{x} - x_{(1)}, x_{(n)} - \tilde{x}\}. \quad (2)$$

Sei  $m := (x_{(1)} + x_{(n)})/2$  die Mitte des Intervalls  $[x_{(1)}, x_{(n)}]$ . Liegt der Median *links* von  $m$ , so ist  $\max\{\tilde{x} - x_{(1)}, x_{(n)} - \tilde{x}\} = x_{(n)} - \tilde{x}$ . Liegt der Median *rechts* von  $m$ , so ist  $\max\{\tilde{x} - x_{(1)}, x_{(n)} - \tilde{x}\} = \tilde{x} - x_{(1)}$ .

Der nächste Satz zeigt, dass die Schranke in (2) halbiert werden kann.

**Satz 1:** Es gilt

$$|\bar{x} - \tilde{x}| \leq \frac{1}{2} \max\{\tilde{x} - x_{(1)}, x_{(n)} - \tilde{x}\}. \quad (3)$$

**BEWEIS :** Setze  $y_i := x_i - \tilde{x}$ ,  $i = 1, \dots, n$ . Dann gilt

$$\begin{aligned}|\bar{x} - \tilde{x}| &= \left| \frac{1}{n} \sum_{i=1}^n y_i \right| \\ &= \left| \frac{1}{n} \sum_{i: y_i < 0} y_i + \frac{1}{n} \sum_{i: y_i > 0} y_i \right| \\ &\leq \max \left\{ -\frac{1}{n} \sum_{i: y_i < 0} y_i, \frac{1}{n} \sum_{i: y_i > 0} y_i \right\},\end{aligned}$$

wobei wir im letzten Schritt die Ungleichung

$$|b - a| \leq \max\{a, b\}, \quad a, b \geq 0$$

verwendet haben. Nun gilt  $|\{i : y_i < 0\}| \leq n/2$  und  $|\{i : y_i > 0\}| \leq n/2$ , da der Median von  $y_1, \dots, y_n$  gleich Null ist:  $\tilde{y} = 0$ . Daher gilt

$$-\frac{1}{n} \sum_{i: y_i < 0} y_i \leq -\frac{1}{n} \frac{n}{2} y_{(1)} = -\frac{y_{(1)}}{2}$$

und

$$\frac{1}{n} \sum_{i: y_i > 0} y_i \leq \frac{1}{n} \frac{n}{2} y_{(n)} = \frac{y_{(n)}}{2}.$$

Die Behauptung folgt nun aus  $y_{(1)} = x_{(1)} - \tilde{x}$  und  $y_{(n)} = x_{(n)} - \tilde{x}$ .

In Beispiel 1 ist  $|\bar{x} - \tilde{x}| = 1.58 - 1.3 = 0.28$  und (3) liefert die Schranke  $\max\{1.3 - 0, 4.6 - 1.3\}/2 = 3.3/2 = 1.65$ . Somit zeigt sich, dass die Abschätzung in konkreten Situationen sehr schlecht sein kann. Dies ist auch zu erwarten, da die Ungleichung (3) allgemein gültig ist, d. h. für alle Daten (vom Umfang  $n$ ) gilt.

**Diskussion von (3):** In (3) gilt Gleichheit nur im Fall  $x_1 = x_2 = \dots = x_n$ . Um dies einzusehen, müssen wir uns den Beweis von Satz 1 anschauen. Ohne Einschränkung sei  $\tilde{x} = 0$  (sonst Übergang zu  $x_i - \tilde{x}$ ). Im Beweis der Abschätzung (3) wurde die Ungleichung  $|b - a| \leq \max\{a, b\}$ ,  $a, b \geq 0$ , angewendet auf  $a := -\frac{1}{n} \sum_{i: x_i < 0} x_i$  und  $b := \frac{1}{n} \sum_{i: x_i > 0} x_i$ . Gleichheit kann nur gelten, wenn  $a = 0$  und/oder  $b = 0$  ist. Betrachten wir den Fall  $\frac{1}{n} \sum_{i: x_i < 0} x_i = 0$  (der Fall  $\frac{1}{n} \sum_{i: x_i > 0} x_i = 0$

wird ähnlich behandelt). Wenn in (3) Gleichheit gelten soll, muss  $\sum_{i: x_i > 0} x_i = \frac{n}{2} x_{(n)}$  sein. Sei  $n$  ungerade. Da  $|\{i : x_i > 0\}| < n/2$ , kann diese Gleichung nur gelten, wenn  $\{i : x_i > 0\} = \emptyset$  und somit  $x_{(n)} = 0$  ist. Entsprechend argumentiert man im Fall  $n$  gerade.

**Bemerkung:** Völlig analog zu (2) erhält man die Abschätzung

$$|\bar{x} - \tilde{x}| \leq \max\{\bar{x} - x_{(1)}, x_{(n)} - \bar{x}\}.$$

Interessanterweise lässt sich die Aussage von Satz 1 nicht mehr auf diese Abschätzung übertragen. Betrachten wir dazu die folgenden 10 Ct-Werte (Angaben ganzzahlig):

$$9, 10, 10, 10, 10, 10, 26, 26, 26, 26.$$

Dann gilt  $\bar{x} = 16.3$ ,  $\tilde{x} = 10$ ,  $x_{(1)} = 9$  und  $x_{(10)} = 26$ . Also

$$|\bar{x} - \tilde{x}| = 6.3 \geq 4.85 = \frac{1}{2} \max\{16.3 - 10, 26 - 16.3\}.$$

Eine weitere Schranke, die man vielleicht erwarten konnte, findet sich im nachfolgenden Satz.

**Satz 2:** Es gilt

$$|\bar{x} - \tilde{x}| \leq s, \quad (4)$$

wobei

$$s := \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

die (empirische) Standardabweichung bezeichnet.

**Bemerkung:** Eine zu (4) entsprechende Ungleichung lässt sich auch im Rahmen eines stochastischen Modells herleiten (siehe z. B. Irle (2001), Satz 9.16).

In Beispiel 1 ist  $s \approx 1.17$  und diese Schranke ist besser als die durch (3) berechnete Schranke von 1.65. Liefert die Standardabweichung immer ein bessere Schranke als (3)? Keineswegs! Betrachten wir dazu das folgende Beispiel der 7 Ct-Werte

$$18, 20, 22, 22, 28, 28, 30,$$

wobei wir die Daten gleich der Größe nach geordnet haben. Dann gilt  $\bar{x} - \tilde{x} = 24 - 22 = 2$ ,  $s \approx 4.28$  und  $\max\{\bar{x} - x_{(1)}, x_{(n)} - \bar{x}\}/2 = \max\{22 - 18, 30 - 22\}/2 = 4$ .

**BEWEIS VON SATZ 2:** Wir benötigen (in dieser Reihenfolge) die allgemeine *Dreiecks-Ungleichung* ( $\triangle$ -Ungleichung)

$$\left| \sum_{i=1}^n a_i \right| \leq \sum_{i=1}^n |a_i|,$$

die Minimierungseigenschaft des Medians, sowie die *Cauchy-Schwarz-Ungleichung* (CS-Ungleichung)

$$\sum_{i=1}^n |a_i \cdot b_i| \leq \sqrt{\sum_{i=1}^n a_i^2} \cdot \sqrt{\sum_{j=1}^n b_j^2},$$

wobei  $a_i \in \mathbb{R}$ ,  $b_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ . Im Fall  $a_i = x_i - \tilde{x}$  und  $b_i = 1$ ,  $i = 1, \dots, n$ , erhalten wir

$$\begin{aligned} |\bar{x} - \tilde{x}| &= \left| \frac{1}{n} \sum_{i=1}^n (x_i - \tilde{x}) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}| \\ &\leq \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \\ &\leq \frac{1}{n} \left( \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{n} \right) \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \\ &= s. \end{aligned}$$

**Bemerkung:** Alle Schranken sind nicht robust. Insbesondere ist die Standardabweichung  $s$  kein robustes Streuungsmaß. Dass die oberen Schranken nicht robust sind (nicht robust sein können), darf nicht überraschen, da die Differenz  $\bar{x} - \tilde{x}$ , die es abzuschätzen gilt, nicht robust ist. Schauen wir uns dazu die Daten aus Beispiel 1 an. Dann gilt  $|\bar{x} - \tilde{x}| = |1.58 - 1.3| = 0.28 < 1.17 \approx s$ . Wäre 46 statt 4.6 notiert worden, so ist  $|\bar{x} - \tilde{x}| = |5.72 - 1.3| = 4.42 < 13.44 \approx s$ .

**Diskussion von (4):** (i) Der Median  $\tilde{x}$  liegt immer im Intervall  $[\bar{x} - s, \bar{x} + s]$ . Somit liegt  $\tilde{x}$  auch im Intervall  $[\bar{x} - \sqrt{2}s, \bar{x} + \sqrt{2}s]$ , in dem mindestens die Hälfte aller Daten liegen (wird in Abschnitt 6 gezeigt).

(ii) Für schmale ( $s \approx 0$ ) und symmetrische Verteilungen werden die beiden Lagemaße fast gleich sein. Für breite ( $s$  groß) und asymmetrische Verteilungen können sie weit voneinander abweichen, allerdings nicht weiter als die Standardabweichung.

(iii) Wegen  $(x_i - \bar{x})^2 \leq (x_{(n)} - x_{(1)})^2$  ist die Standardabweichung stets kleiner oder gleich der Spannweite:

$$s \leq x_{(n)} - x_{(1)}.$$

Aufgrund der Minimierungseigenschaft des arithmetischen Mittels gilt

$$s \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$



(iv) In der Literatur wird die Standardabweichung auch häufig so definiert:

$$\hat{s} := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Wegen  $s < \hat{s}$  bleibt die Aussage (4) auch mit  $\hat{s}$  gültig. Der Vorteil von  $\hat{s}$  begründet sich darin, dass  $\hat{s}^2$  eine *erwartungstreue* Schätzung für den (unbekannten) Modellparameter *Varianz* ist (siehe z. B. Kütting und Sauer (2011), Abschnitt 9.2).

(v) Jetzt wollen wir der Frage nachgehen, ob und wann die Schranke in (4) angenommen wird. Wir werden zeigen, dass dies nur im Fall  $x_1 = x_2 = \dots = x_n$  möglich ist (und in diesem Fall ist  $s = 0$ ). Warum? Im Beweis von (4) wurden die  $\triangle$ -Ungleichung und die CS-Ungleichung verwendet. Eine gute Gelegenheit zu fragen, wann in diesen Ungleichungen das Gleichheitszeichen gilt. In der  $\triangle$ -Ungleichung gilt Gleichheit, wenn  $a_i \leq 0$  für  $i = 1, \dots, n$  oder  $a_i \geq 0$  für  $i = 1, \dots, n$ . In der vorliegenden Situation ist  $a_i = (x_i - \bar{x})$  und wegen  $x_{(1)} - \bar{x} \leq 0$  und  $x_{(n)} - \bar{x} \geq 0$  muss  $x_{(1)} = x_{(n)}$  gelten, d.h.  $x_1 = x_2 = \dots = x_n$  (damit gilt  $\bar{x} = \tilde{x}$  und  $s = 0$ ). Gleichheit tritt in der CS-Ungleichung genau dann ein, wenn  $a_i \cdot b_j = a_j \cdot b_i$  für  $1 \leq i, j \leq n$  gilt, wenn also die Vektoren  $a = (a_1, \dots, a_n)$  und  $b = (b_1, \dots, b_n)$  linear abhängig sind (siehe z. B. Fischer (2000), Abschnitt 5.1, Königsberger (1992), Abschnitt 9.8). In der vorliegenden Situation ist  $a_i = (x_i - \bar{x})$  und  $b_j = 1$ . Also  $(x_i - \bar{x}) \cdot 1 = (x_j - \bar{x}) \cdot 1$ , d. h.  $x_i = x_j$ . Fassen wir zusammen:  $|\bar{x} - \tilde{x}| = s$  impliziert Gleichheit in der  $\triangle$ -Ungleichung und in der CS-Ungleichung. Sowohl aus der Gleichheit in der  $\triangle$ -Ungleichung als auch aus der Gleichheit in der CS-Ungleichung folgt  $x_1 = \dots = x_n$ . Für die Implikation  $|\bar{x} - \tilde{x}| = s \Rightarrow x_1 = \dots = x_n$  genügt es, wenn in einer der beiden Ungleichungen das Gleichheitszeichen gilt. Umformulierung dieser Implikation (Kontraposition): Sind mindestens zwei Datenwerte verschieden, so gilt  $|\bar{x} - \tilde{x}| < s$ .

## 5 Die Samuelson-Ungleichung

Eine Abschätzung darüber, wie weit ein einzelner Datenwert vom arithmetischen Mittel entfernt sein kann, ist schnell gefunden. Da das Weglassen nicht-negativer Summanden eine Summe immer kleiner macht (jedenfalls nicht größer), gilt

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

und somit

$$|x_j - \bar{x}| \leq \sqrt{n} \cdot s. \quad (5)$$

Zentrale Erkenntnis: Kein Datenwert kann weiter als  $\sqrt{n} \cdot$  Standardabweichung vom arithmetischem Mittel entfernt sein. Abschätzung (5) zeigt deutlich, dass mit zunehmender Anzahl an Daten die Abschätzung sehr grob ausfallen kann.

Eine etwas bessere Schranke als (5) ist

$$|x_j - \bar{x}| \leq \sqrt{n-1} \cdot s, \quad j = 1, \dots, n \quad (6)$$

(Samuelson (1968)). Insbesondere gilt

$$|x_{(1)} - \bar{x}| \leq \sqrt{n-1} \cdot s \quad \text{und} \quad |x_{(n)} - \bar{x}| \leq \sqrt{n-1} \cdot s.$$

Somit ist die trivialerweise geltende Schranke

$$|x_j - \bar{x}| \leq \max\{\bar{x} - x_{(1)}, x_{(n)} - \bar{x}\}. \quad (7)$$

sogar besser. In Beispiel 1 ist diese  $4.6 - 1.58 = 3.02$ , die Samuelson-Ungleichung (6) liefert die Schranke  $\sqrt{9} \cdot 1.17 = 3.51$ .

**Bemerkung:** Geht man über zu den *standardisierten* Daten

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, \dots, n,$$

so impliziert die Samuelson-Ungleichung  $|z_j| \leq \sqrt{n-1}$ ,  $j = 1, \dots, n$ . Beachte: Die Daten  $z_1, \dots, z_n$  haben den Mittelwert null,  $\bar{z} = 0$ , und die Standardabweichung 1,  $s_z = 1$ .

Am Ende dieses Abschnittes wollen wir der Vollständigkeit halber den Beweis von (6) vorstellen, der auf Arnold (1974) zurückgeht.

**Diskussion von (6):** (i) Da die (triviale und von (6) unabhängige) Schranke (7) besser als die Schranke aus der Samuelson-Ungleichung ist, stellt sich die Frage nach dem Nutzen der Samuelson-Ungleichung. Der Punkt ist der, dass die Schranke aus der Samuelson-Ungleichung nur von  $n$  und  $s$  abhängt, so dass etwa der größte und kleinste Datenwert gar nicht vorliegen müssen, um diese anzuwenden. Angenommen, der Durchschnitt von 10 Daten beträgt 1.58 und für die Standardabweichung gilt  $s = 1.17$ . Die Daten sind leider verloren gegangen. Wie groß kann der maximale bzw. minimale Datenwert sein? Lösung:

$$|x_j - 1.58| \leq \sqrt{10-1} \cdot 1.17 = 3.51$$

liefert

$$-1.93 \leq x_j \leq 5.09$$

(siehe Beispiel 1).

(ii) Die Samuelson-Ungleichung kann nicht verbessert werden. D. h.,  $\sqrt{n-1}$  ist die kleinste Konstante  $c$  mit  $|x_j - \bar{x}| \leq c \cdot s$ ,  $j = 1, \dots, n$ . Dazu betrachte man

das Beispiel  $n = 2$  mit  $x_1 = a, x_2 = -a, a > 0$  (z. B.  $a = 1$ ). Dann gilt  $\bar{x} = 0, s = a$  und folglich

$$|x_j - 0| = a \leq \sqrt{2-1} \cdot a = a, \quad j = 1, 2.$$

Im Unterschied zu (5) ist die Schranke aus der Samuelson-Ungleichung scharf.

(iii) Ersetzt man auf der linken Seite von (6) das arithmetische Mittel durch den Median, so gilt die Samuelson-Ungleichung i. A. nicht mehr. So zeigt das Beispiel der 7 Ct-Werte 22, 22, 22, 22, 24, 26, 30:  $|x_{(n)} - \tilde{x}| = 30 - 22 = 8$  und  $\sqrt{n-1} \cdot s = \sqrt{6} \cdot \sqrt{8} \approx 6.93$ . Mittels der  $\triangle$ -Ungleichung, (4) und (6) lässt sich sofort eine Samuelson-Ungleichung für den Median formulieren:

$$\begin{aligned} |x_j - \tilde{x}| &\leq |x_j - \bar{x}| + |\bar{x} - \tilde{x}| \\ &\leq (\sqrt{n-1} + 1) \cdot s, \quad j = 1, \dots, n. \end{aligned}$$

BEWEIS VON (6) NACH ARNOLD (1974): Zunächst gilt

$$|x_j - \bar{x}| = \left| \sum_{i \neq j} (x_i - \bar{x}) \right|,$$

denn

$$\begin{aligned} \bar{x} - x_j &= n\bar{x} - (n-1)\bar{x} - x_j \\ &= \sum_{i \neq j} x_i - (n-1)\bar{x} \\ &= \sum_{i \neq j} (x_i - \bar{x}). \end{aligned}$$

Die  $\triangle$ -Ungleichung und die CS-Ungleichung, angewendet auf  $a_i = x_i - \bar{x}, b_i = 1, 1 \leq i \leq n, i \neq j$ , ergeben

$$|x_j - \bar{x}| \leq \sum_{i \neq j} |x_i - \bar{x}| \leq \sqrt{(n-1) \sum_{i \neq j} (x_i - \bar{x})^2}.$$

Quadrieren führt zur Aussage

$$\begin{aligned} (x_j - \bar{x})^2 &\leq (n-1) \sum_{i \neq j} (x_i - \bar{x})^2 \\ &= (n-1) \sum_{i=1}^n (x_i - \bar{x})^2 - (n-1)(x_j - \bar{x})^2. \end{aligned}$$

Dies ist äquivalent zu

$$n(x_j - \bar{x})^2 \leq (n-1) \sum_{i=1}^n (x_i - \bar{x})^2.$$

Division durch  $n$  und ziehen der Quadratwurzel liefert die Behauptung.

## 6 Vergleich der Samuelson-Ungleichung mit der Tschebyschow-Ungleichung

Im letzten Abschnitt wollen wir die Samuelson-Ungleichung vergleichen mit der eingangs erwähnten Tschebyschow-Ungleichung. Sie lautet

$$P(|X - E(X)| > \varepsilon) \leq \frac{V(X)}{\varepsilon^2}, \quad \varepsilon > 0 \quad (8)$$

bzw. durch Übergang zur komplementären Wahrscheinlichkeit

$$P(|X - E(X)| \leq \varepsilon) \geq 1 - \frac{V(X)}{\varepsilon^2}, \quad \varepsilon > 0.$$

Dabei bezeichnet  $E(X)$  den Erwartungswert und  $V(X)$  die Varianz einer Zufallsvariablen  $X$ . Um diese Ungleichung mit der Samuelson-Ungleichung zu vergleichen, müssen wir die Aussage der Tschebyschow-Ungleichung (8), die sich auf ein stochastisches Modell bezieht, auf die „Daten-Ebene“ übertragen:

$$\frac{|\{i : |x_i - \bar{x}| > \varepsilon\}|}{n} \leq \frac{s^2}{\varepsilon^2}, \quad \varepsilon > 0. \quad (9)$$

Auf der linken Seite steht der Anteil der Daten, die einen Abstand vom arithmetischem Mittel größer als  $\varepsilon$  haben. Nachweis von (9):

$$\begin{aligned} s^2 &\geq \frac{1}{n} \sum_{i: |x_i - \bar{x}| > \varepsilon} (x_i - \bar{x})^2 \\ &\geq \frac{1}{n} \sum_{i: |x_i - \bar{x}| > \varepsilon} \varepsilon^2 \\ &= \frac{\varepsilon^2}{n} \sum_{i: |x_i - \bar{x}| > \varepsilon} 1 \\ &= \varepsilon^2 \frac{|\{i : |x_i - \bar{x}| > \varepsilon\}|}{n}. \end{aligned}$$

**Bemerkung:** Wählt man in (8) als Verteilung von  $X$  die relativen Häufigkeiten, mit der die Daten auftreten, so gilt  $E(X) = \bar{x}$  und  $V(X) = s^2$  und (9) ergibt sich als Spezialfall von (8).

Wählen wir  $\varepsilon = c \cdot s, c > 0$ , so folgt aus (9)

$$\frac{|\{i : |x_i - \bar{x}| \leq c \cdot s\}|}{n} \geq 1 - \frac{1}{c^2}.$$

Dies ist eine Abschätzung über den Anteil der Daten, die im Intervall  $[\bar{x} - c \cdot s, \bar{x} + c \cdot s]$  liegen. Speziell gilt: Mindestens 50% der Daten liegen im Intervall

$$[\bar{x} - \sqrt{2} \cdot s, \bar{x} + \sqrt{2} \cdot s],$$

mindestens 95% der Daten liegen im Intervall

$$[\bar{x} - \sqrt{20} \cdot s, \bar{x} + \sqrt{20} \cdot s].$$

Die Samuelson-Ungleichung besagt hingegen, dass alle  $n$  Daten im Intervall

$$[\bar{x} - \sqrt{n-1} \cdot s, \bar{x} + \sqrt{n-1} \cdot s]$$

liegen. Speziell für  $c = \sqrt{n-1}$  liefert die Tschebyschow-Ungleichung die Abschätzung

$$\frac{|\{i : |x_i - \bar{x}| \leq \sqrt{n-1} \cdot s\}|}{n} \geq \frac{n-2}{n-1}$$

(die linke Seite ist 1 nach der Samuelson-Ungleichung).

Entsprechende Aussagen über Intervalle kennt man auf Modellebene. Man denke an die  $\sigma$ -Umgebungen („ $\sigma$ -Regeln“) aus der Wahrscheinlichkeitsrechnung. Sie spielen im Schulunterricht eine große Rolle und sind zusammen mit den sogenannten Prognoseintervallen zentrale Bausteine, um Konfidenzintervalle einzuführen und zu verstehen.

## 7 Schlussbemerkungen

Schüler erlernen im Laufe ihrer Mathematik-Ausbildung den Umgang mit Ungleichungen. Im vorliegenden Artikel haben wir Ungleichungen aus der Analysis auf datenabhängige Kenngrößen (Median, Mittelwert) „angewendet“. Die einfachen Abschätzungen können im Rahmen des Stochastik-Curriculums thematisiert und angewendet werden, in Ergänzung zur üblichen Deskription univariater Datensätze.

Die Untersuchung von Schranken bietet eine Möglichkeit, Ungleichungen im Schulcurriculum angemessen(er) zu berücksichtigen, sind sie doch ein unverzichtbares Werkzeug in der Mathematik.

**Danksagung:** Ausdrücklich möchte ich mich bei den Gutachtern für ihre wertvollen Anregungen und Verbesserungsvorschläge bedanken.

## Literatur

Arnold, B.C. (1974): Schwarz, Regression, and Extreme Deviance. In: *The American Statistician* 28, Nr. 1, S. 22-23.

Brown, C.M, Vostok, J., Johnson, H. et al.(2021): Outbreak of SARS-CoV-2-infections, including COVID-19 vaccine breakthrough infections, associated with large public gatherings - Barnstable County, Massachusetts, July 2021. In: *Morbidity and Mortality Weekly Report (MMWR)* 2021. 70 (31), 1059-1062.

Busse, R. und Nimptsch, U. (2021): COVID-19-Pandemie: Historisch niedrige Bettenauslastung In: *Deutsches Ärzteblatt*, Jg. 118, Heft 10, A 504.

Cushny und Peebles (1905): The Action of Optical Isomers. II. Hyoscines. In: *J. Physiology* 32, 501-510.

Fischer, G. (2000): Lineare Algebra. 12. verbesserte Auflage. Vieweg, Wiesbaden.

Gallais, F., Ganter, P., Bruel, T. et al. (2021): Evolution of antibody responses up to 13 months after SARS-CoV-2 infection and risk of reinfection. In: *EBioMedicine* 71, <https://doi.org/10.1016/j.ebiom.2021.103561>

Henze, N. (2021): Stochastik für Einsteiger. 13. Auflage. Springer Spektrum, Wiesbaden.

Irl, A. (2001): Wahrscheinlichkeitstheorie und Statistik. Teubner, Stuttgart-Leipzig-Wiesbaden.

Königsberger, K (1992): Analysis 1, 2. Auflage. Springer Lehrbuch, Berlin-Heidelberg.

Kütting H. und Sauer, M. J: (2011): Elementare Stochastik, 3. Auflage. Spektrum Akademischer Verlag, Heidelberg.

Samuelson, P.A. (1968): How Deviant Can You Be? In: *Journal of the American Statistical Association* 63, 1522-1525.

Student (1908): The Probable Error of a Mean. In: *Biometrika* 6, Nr. 1, S. 1-25.

Anschrift des Verfassers

Frank Marohn

Institut für Mathematik

Universität Würzburg

Emil-Fischer-Str. 30

97074 Würzburg

[marohn@mathematik.uni-wuerzburg.de](mailto:marohn@mathematik.uni-wuerzburg.de)