Hans HUMENBERGER, Wien

Was bewirkt eine Veränderung eines einzelnen Wertes bei der Varianz der zugehörigen Datenliste, und warum?

Der Auslöser für die im Titel genannte Fragestellung war eine Aufgabe in einem Schulbuchentwurf (8. Schulstufe) im Kapitel Beschreibende Statistik (hier nur sinngemäß wiedergegeben):

Aufgabe: Thomas und Carina haben 20-mal dasselbe Computerspiel gespielt und ihre Ergebnisse in einer Tabelle festgehalten, von beiden weiß man also, wie oft sie jeweils die möglichen Punktezahlen (100, 200, 300, 400, 500) erreicht hatten.

- Berechne das arithmetische Mittel $\overline{x}$ und die Varianz der Punktezahlen von Thomas und Carina!
- Carina hat sich geirrt und ein Spiel mit 200 statt mit 300 Punkten eingetragen. Wie wirkt sich dieser Irrtum bei der Reparatur aus: Wird der wirkliche Mittelwert dadurch größer oder kleiner als der bisher berechnete? Wird die wirkliche Varianz dadurch größer oder kleiner? Stelle eine Vermutung auf bevor du rechnest!
- Begründe deine Vermutung!

Während die Begründung im Falle des Mittelwertes leicht machbar ist, schien uns das im Fall der Varianz genau genommen nicht mehr so einfach zu sein. Angenommen Carinas Mittelwert lag mit dem falschen Wert (200 Punkte) bei 320 Punkten. Dann ist zunächst natürlich sofort klar, dass der neue (richtige Wert) 300 näher beim bisherigen Mittelwert liegt, so dass es intuitiv nahe liegt, dass dadurch auch die Varianz kleiner wird, weil ja ein entscheidender quadratischer Abstand kleiner wird. So oder so ähnlich war wohl eine mögliche Begründung im Schulbuch auch gemeint. Wenn man nicht tiefer über die Sache nachdenkt, scheint die Angelegenheit damit erledigt zu sein. Aber ist das wirklich so einfach? Ist es wirklich immer so (unabhängig von der Lage der anderen Werte): Wann immer ein Wert näher an den momentanen Mittelwert heranrückt, wird die Varianz dadurch immer kleiner? Oder umgekehrt formuliert: Wann immer ein Wert vom momentanen Mittelwert wegrückt, wird die Varianz dadurch immer größer? Immerhin ändert sich bei der Verschiebung eines Wertes ja auch der Mittelwert selbst (und damit alle Abstände zu ihm), und man weiß ja i. A. nicht, wie viele der Werte kleiner bzw. größer als $\overline{x}$ sind. Wenn man das alles bedenkt, ist es gar nicht mehr so leicht die Auswirkungen auf alle anderen quadratischen Abstände zum neuen Mittelwert, und insbesondere auf deren Summe begründet abzuschätzen.

Es wird sich herausstellen, dass die obigen Intuitionen normalerweise richtig sind, aber leider nicht immer. Im obigen Beispiel mit dem ursprünglichen Mittelwert bei 320 und der Veränderung eines Wertes 200 → 300 nimmt die

In: IDMI-Primar Goethe-Universität Frankfurt (Hrsg.),
Beiträge zum Mathematikunterricht 2022.
*56. Jahrestagung der Gesellschaft für Didaktik der Mathematik.*WTM.
https://doi.org/10.37626/GA9783959872089.0

Varianz tatsächlich ab. Aber wenn der ursprüngliche Mittelwert z. B. 248 ist, so dass der neue Wert weiter entfernt (52) vom ursprünglichen Mittelwert ist als der alte (48, aber auf der anderen Seite), dann verringert sich die Varianz ebenfalls. Zu sehen, wann und warum das passiert, hilft die zugehörigen intuitiven und nichtintuitiven Aspekte zu verstehen.

Qualitative und quantitative Aspekte

Zunächst ist klar, dass man die Gesamtheit aller Werte auf der Zahlengeraden beliebig verschieben kann, ohne dass das einen Einfluss auf die Varianz hat. D. h. man kann o. B. d. A. den Mittelwert $\overline{x}$ in den Nullpunkt legen: $\overline{x} := \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = 0$. Den n-ten Wert x_n betrachten wir als variabel, die anderen x_i bleiben fest (unverändert). Wir interessieren uns für die zugehörige Änderung Δv der Varianz v (wir verzichten dabei der Einfachheit halber auf die Division durch n - 1 bzw. durch n) bei einer Veränderung von x_n.

Bei der ursprünglichen Summe der quadrierten Abstände (vom Mittelwert $\overline{x} = 0$) spalten wir den Beitrag von x_n absichtlich ab: $v = \sum_{i=1}^{n-1} x_i^2 + x_n^2$. Nun wird x_n verändert ($x_n \to x_n + h$), die dadurch entstehenden neuen Parameter (Mittelwert, Summe der quadratischen Abweichungen vom Mittelwert) bezeichnen wir mit $\overline{x}_{\text{neu}}$ bzw. v_{neu}. Es gelten $\overline{x}_{\text{neu}} = h / n$ und

$$v_{\text{neu}} = \sum_{i=1}^{n-1} \underbrace{\left(x_i - \frac{h}{n} \right)^2}_{x_i^2 - 2 \cdot \frac{h}{n} \cdot x_i + \frac{h^2}{n^2}} + \underbrace{\left(\left(x_n + h \right) - \frac{h}{n} \right)^2}_{\left(x_n + h \right)^2 - 2 \cdot \frac{h}{n} \cdot \left(x_n + h \right) + \frac{h^2}{n^2}} = \sum_{i=1}^{n-1} x_i^2 + \left(x_n + h \right)^2 -$$

$$- 2 \cdot \frac{h}{n} \cdot \underbrace{\left(\sum_{i=1}^{n-1} x_i + \left(x_n + h \right) \right)}_{= h \text{ wegen } \sum_{i=1}^{n} x_i = 0} + n \cdot \frac{h^2}{n^2} = \sum_{i=1}^{n-1} x_i^2 + \left(x_n + h \right)^2 - \frac{h^2}{n}.$$

Wir sind interessiert an der Änderung $\Delta v := v_{\text{neu}} - v$, wenn diese positiv ist, findet eine Varianzvergrößerung statt, bei $\Delta v < 0$ eine Verkleinerung.

Wir erhalten $\Delta v = \left(x_n + h \right)^2 - h^2 / n - x_n^2$ bzw. vereinfacht

$$\Delta v = h \cdot \left(2 \cdot x_n + \frac{n-1}{n} \cdot h \right). \tag{1}$$

Daraus erkennt man: Sowohl für $h, x_n > 0$ als auch für $h, x_n < 0$ ist $\Delta v > 0$.

Nun nehmen wir an, dass h und x_n verschiedenes Vorzeichen haben, z. B. $x_n < 0$ und $h > 0$. Für relativ kleine Werte von h bedeutet das, dass der Wert näher an den ursprünglichen Mittelwert heranrückt. Wenn der neue Wert $x_n + h$ immer noch auf derselben Seite des ursprünglichen Mittelwertes, d. h. negativ ist, dann gelten $h < -x_n$ und $2 \cdot x_n + \frac{n-1}{n} \cdot h < 2 \cdot x_n - \frac{n-1}{n} \cdot x_n < 0$. Das bedeutet: Wenn x_n näher an den ursprünglichen Mittelwert heranrückt, aber auf derselben Seite bleibt, dann gilt $\Delta v < 0$.

Ohne unsere Annahme $\overline{x} = 0$ (die den algebraischen Aufwand verkleinert, die wir aber ab nun nicht mehr benutzen, so dass auch der allgemeine Fall $\overline{x} \neq 0$ abgedeckt ist) erhielte man statt (1):

$$\Delta v = h \cdot \left(2 \cdot (x_n - \overline{x}) + \frac{n-1}{n} \cdot h \right) \qquad (2)$$

Die Beziehungen (1) und (2) beschreiben die Varianzänderung nicht nur qualitativ, sondern auch quantitativ.

Man beachte, dass die obigen Überlegungen den Spezialfall $x_n = \overline{x}$ schon enthalten: Wenn ein Wert genau bei $\overline{x}$ liegt und geändert wird, nimmt die Varianz zu: $\Delta v = \frac{n-1}{n} \cdot h^2 > 0$. Dieser Spezialfall $x_n = \overline{x}$ könnte im Unterricht sogar noch vor dem obigen schon etwas allgemeineren Fall behandelt werden.

Ab nun wollen wir auch Veränderungen von x_n zulassen, die über den Mittelwert hinwegführen, und untersuchen, was passiert. Dabei nehmen wir $x_n < \overline{x}$ und $h > 0$ an, d. h. die Bewegung $x_n \to x_n + h$ geht zunächst in Richtung $\overline{x}$ und dann sogar darüber hinaus (analog die andere Richtung $x_n > \overline{x}$, $h < 0$). Anfänglich wird v tatsächlich abnehmen (wir wissen bereits: zumindest bis der neue Wert $x_n + h$ bei $\overline{x}$ ist), aber was passiert danach?

Aus (2) erhält man, dass $\Delta v \geq 0$ äquivalent ist mit

$$h \geq 2 \cdot \frac{n}{n-1} \cdot \left(\overline{x} - x_n \right) = 2 \cdot \left(\overline{x} - x_n \right) + \frac{2}{n-1} \cdot \left(\overline{x} - x_n \right)$$

Das bedeutet, man muss für $\Delta v \geq 0$ den Wert x_n nicht nur bis zu seinem Symmetriepartner auf der anderen Seite des ursprünglichen Mittelwertes (d. h. zum Wert $x_n + 2 \cdot \left(\overline{x} - x_n \right)$) bewegen, sondern zumindest das Stück $\frac{2}{n-1} \cdot \left(\overline{x} - x_n \right)$ weiter (Abb. 1).

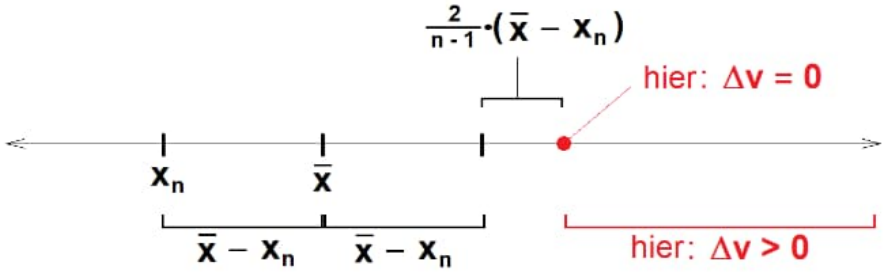

Abb. 1: Der Punkt auf der anderen Seite von $\bar{x}$ mit der Eigenschaft $\Delta v = 0$ liegt ein wenig weiter weg als der symmetrische Punkt

Zurück zur Aufgabe aus dem Schulbuchentwurf, die unser Aufhänger war; angenommen wir wissen $x_n = 200$, $n = 20$ und $h = 100$, aber wir kennen den ursprünglichen Mittelwert $\bar{x}$ nicht. Wir geben hier zwei verschiedene mögliche Werte von $\bar{x}$ an, um das Ergebnis von Abb. 1 zu illustrieren.

- $\bar{x} = 320$: Dann ergibt sich $\Delta v = -14500$ aus (2). Der Abstand des Wertes vom bisherigen Mittelwert hat sich von 120 auf 20 Punkte verkleinert (auf derselben Seite des bisherigen Mittelwertes!), und die Summe der quadrierten Abweichungen vom Mittelwert hat sich um 14500 verringert.
- $\bar{x} = 248$: Dann ergibt sich $\Delta v = -100$ aus (2). Der Abstand des Wertes vom bisherigen Mittelwert hat sich von 48 auf 52 Punkte vergrößert (diesmal liegt der neue Punkt aber auf der anderen Seite des bisherigen Mittelwertes!), und die Summe der quadrierten Abweichungen vom Mittelwert hat sich trotzdem um 100 verringert.

Wenn der Wert auf derselben Seite des ursprünglichen Mittelwertes bleibt, dann stimmt unsere intuitive Vermutung: je näher zum (weiter weg vom) ursprünglichen Mittelwert der neue Wert liegt, desto kleiner (größer) ist die Varianz. Wenn aber bei der Veränderung der ursprüngliche Mittelwert überschritten wird, ist die Lage etwas komplizierter: Nach der symmetrischen Lage zum ursprünglichen Mittelwert gibt es ein Intervall der Länge $2/(n-1)\cdot|\bar{x} - x_n|$, in dem der neue Wert weiter weg vom ursprünglichen Mittelwert, aber trotzdem die neue Varianz kleiner als die ursprüngliche ist.

Ausführlichere Fassungen zu diesem Thema auf Englisch bzw. Deutsch finden sich in Humenberger (2020) bzw. Humenberger (2022).

Literatur

Humenberger, H. (2020). How does the change of a single data point affect the variance, and why? *Teaching Statistics, 42*(3), 87–90.

Humenberger, H. (2022). Was bewirkt eine Veränderung eines einzelnen Wertes bei der Varianz der zugehörigen Datenliste, und warum? *Stochastik in der Schule, 42*(3), 22–25.