Was bewirkt eine Veränderung eines einzelnen Wertes bei der Varianz der zugehörigen Datenliste, und warum?[1]

HANS HUMENBERGER, WIEN

[1] Original in Englisch unter CC BY 4.0 Lizenz:
Teaching Statistics 42(3), 2020, S. 87–90.
https://doi.org/10.1111/test.12230.
Übersetzung: HANS HUMENBERGER

Zusammenfassung: *Dieser Beitrag geht einer naheliegenden und interessanten Frage nach, die in einem Schulbuchentwurf zu lesen war. Was passiert mit der Varianz einer Datenliste, wenn einer ihrer Werte verändert wird, und warum? Einige Teilantworten sind wenig überraschend, aber im Folgenden geht es um die „ganze" Antwort und um ein zugehöriges wichtiges Verständnis von Lernenden.*

1 Einleitung

Zunächst betrachten wir die analoge Frage für den (arithmetischen) Mittelwert $\bar{x}$ eines Datensatzes. Wir gehen davon aus, dass es n Datenwerte $x_1, \ldots, x_{n-1}, x_n$ gibt, wobei $x_1, \ldots, x_{n-1}$ festbleiben und $x = x_n$ als variabel gedacht werden sollen. Hier ist alles viel leichter. Es ist unmittelbar klar, wie sich $\bar{x}$ verändert, wenn genau ein Wert der x_i vergrößert wird und die anderen gleich bleiben: Er wird größer (*qualitativ*). Dies ist ohne Formeln klar, wenn man z. B. die Schwerpunktvorstellung von $\bar{x}$ hat (an den Stellen x_i sind gleiche, punktförmige Gewichte platziert, $\bar{x}$ entspricht dann jener Stelle, an der man die Zahlengerade unterstützen muss, so dass diese – als Balken mit Gewichten gedacht – im Gleichgewicht bleibt). Denn wenn der Balken im Gleichgewicht ist und ein Wert auf diesem Balken nach rechts (links) verschoben wird, ist klar, dass der ganze Balken nach rechts (links) kippt. Es ist sogar ganz leicht auszurechnen, um wieviel (*quantitativ*) sich der Mittelwert ändert bei $x_n \to x_n + h$, nämlich um h/n. Es ist eigentlich unerheblich, welcher Wert verändert wird, aber zur Vereinfachung nehmen wir an, dass es der *letzte beobachtete* Wert x_n ist.

Im Folgenden betrachten wir die analoge Situation mit der *Varianz* statt des Mittelwerts. Es ist für das qualitative Verhalten völlig unerheblich, ob wir dabei durch n oder durch $n - 1$ dividieren, oder die Wurzel ziehen (Standardabweichung), nur bei der quantitativen Sichtweise kommt es darauf an; der Einfachheit halber verzichten wir bei unseren Überlegungen auf einen Nenner (n bzw. $n - 1$), dann werden die entsprechenden Formeln etwas einfacher und kürzer. Wir betrachten also nur die Summe der quadrierten Abstände vom Mittelwert $v := \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Der Auslöser für alle diese Überlegungen war eine Aufgabe in einem Schulbuchentwurf (8. Schulstufe) im Kapitel *Beschreibende Statistik* (hier nur sinngemäß wiedergegeben):

> 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.
>
> 1) Berechne das arithmetische Mittel $\bar{x}$ und die Varianz der Punktezahlen von Thomas und Carina!
>
> 2) 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!
>
> 3) *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, sodass es intuitiv naheliegt, dass dadurch auch die Varianz kleiner wird, weil ja *ein* entscheidender quadratischer Abstand kleiner wird. So ähnlich war eine mögliche Begründung im Schulbuch wohl 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 Mittelwert heranrückt, wird die Varianz dadurch *immer* kleiner?

Oder umgekehrt formuliert: Wenn ein Wert vom 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 i. A. nicht, wie viele der Datenwerte kleiner bzw. größer als $\bar{x}$ 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.

Das ist ein klassisches Beispiel dafür, dass durch genaueres Nachdenken der Grad des Zweifels erhöht werden kann. Manchmal sind solche Zweifel ja auch höchstangebracht und entlarven Fallen bzw. Fehlvorstellungen bei intuitiven Herangehensweisen. Wir werden sehen, dass die obigen Intuitionen normalerweise richtig sind, aber leider nicht immer. Im obigen Beispiel mit dem ursprünglichen Mittelwert bei 320 und der Veränderung eines Datenwertes $200 \to 300$ nimmt die Varianz tatsächlich ab. Aber wenn der ursprüngliche Mittelwert z. B. 248 ist, so dass der neue Datenwert weiter entfernt vom ursprünglichen Mittelwert ist als der alte (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.

2 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 den Mittelwert $\overline{x}$ o. B. d. A. 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 unverändert. Wir interessieren uns für die zugehörige Änderung Δv bei Veränderung von x_n.

Bei der ursprünglichen Summe der quadrierten Abstände 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 quadratischer Abweichungen vom Mittelwert) bezeichnen wir mit $\overline{x}_{\text{neu}}$ und v_{neu}. Es gelten

$$\overline{x}_{\text{neu}} = \frac{h}{n} \text{ und}$$

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

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

$$\Delta v = (x_n + h)^2 - h^2/n - x_n^2$$

bzw. vereinfacht

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

Daraus erkennt man unmittelbar: 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 Datenwert näher an den ursprünglichen Mittelwert heranrückt. Wenn der neue Datenwert $x_n + h$ immer noch auf derselben Seite des ursprünglichen Mittelwertes, d. h. negativ ist, dann gelten

$$h < -x_n \text{ und } 2x_n + \frac{n-1}{n} \cdot h < 2x_n - \frac{n-1}{n} \cdot x_n < 0.$$

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) erhielte man statt (1):

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

Auch in dieser allgemeinen Formel bestätigt sich die Intuition: Wenn x_n über dem Mittelwert liegt (d. h. $x_n - \overline{x} > 0$) und weiter weg vom Mittelwert bewegt wird (d. h. $h > 0$), dann vergrößert sich die Varianz. Analog ist die Situation, wenn x_n unter dem Mittelwert liegt und weiter weg vom Mittelwert bewegt wird. Die Beziehungen (1) und (2) beschreiben die Varianzänderung nicht nur qualitativ, sondern auch quantitativ.

Ab nun benutzen wir die vereinfachende Annahme $\overline{x} = 0$ nicht mehr, so dass Leser*innen nicht selbst die Frage beantworten müssen: Wie ist die Situation im allgemeinen Fall $\overline{x} \neq 0$?

Man beachte, dass die obigen Überlegungen den Spezialfall $x_n = \overline{x}$ schon enthalten: Wenn ein Datenwert 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.

Der dafür nötige algebraische Aufwand ist zwar nicht hoch, aber vorhanden. Schöner und anschaulicher wäre es, wenn es dafür auch eine (korrekte) Argumentation ohne Rechnung gäbe (ähnlich zum Mittelwert), aber da haben wir leider nichts gefunden, vielleicht gibt es so etwas auch nicht?

Ab nun wollen wir auch Veränderungen von x_n zulassen, die über den Mittelwert hinwegführen, und sehen was dabei passiert. Dabei nehmen wir an, dass $x_n < \overline{x}$ und $h > 0$ ist, d. h., die Bewegung $x_n \to x_n + h$ geht zunächst in Richtung $\overline{x}$ und dann darüber hinaus (wir vernachlässigen die Behandlung der analogen Situation, in der $x_n > \overline{x}$ ist und schließlich auf die andere Seite kommt; die Details sind fast gleich). 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

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

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

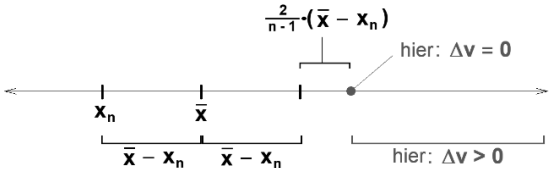

Abb. 1: Der Punkt auf der anderen Seite von $\overline{x}$ mit der Eigenschaft $\Delta v = 0$ liegt ein wenig weiter weg als der symmetrische Punkt.

Um die Länge dieses zusätzlichen Stücks genau zu bestimmen, brauchten wir Formel (2). Wenn es aber gar nicht um die genaue Länge geht, sondern nur die Existenz eines solchen zusätzlichen Stücks intuitiv erklärt werden soll, könnte man folgendermaßen argumentieren: Wenn man x_n vergrößert, so vergrößert sich auch der Mittelwert. Daher kann der bezüglich $\overline{x}$ symmetrische Punkt $x_n + 2 \cdot (\overline{x} - x_n)$ nicht denselben Wert von v haben, man muss zu diesem Zweck noch ein bisschen weiter gehen.

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

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

Daher ist die intuitive Vermutung, dass eine Verschiebung eines Datenwertes näher zum ursprünglichen Mittelwert immer eine Verringerung der Varianz nach sich zieht, korrekt, und zwar unabhängig davon, auf welcher Seite des ursprünglichen Mittelwertes der neue Wert liegt. Wenn ein Datenwert weiter weg vom ursprünglichen Mittelwert bewegt wird (auf derselben Seite von diesem bleibend), dann erhöht sich die Varianz dabei immer. Wenn aber der Datenpunkt auf die andere Seite des ursprünglichen Mittelwertes bewegt wird, dann gibt es ein „Anfangsintervall“ von Werten, die weiter weg vom ursprünglichen Mittelwert sind, in dem die Varianz kleiner als die ursprüngliche ist, aber wenn der Datenwert dann darüber hinaus zu liegen kommt, dann ist die neue Varianz größer als die ursprüngliche.

3 Weitere interessante Einsichten und Zusammenhänge

Die Änderung Δv ist eine quadratische Funktion (bzw. Parabel) in h mit den Nullstellen bei

$$h_1 = 0 \text{ und } h_2 = \frac{2n}{n-1} \cdot (\overline{x} - x_n),$$

sie hat daher ihr Minimum an der Stelle

$$h^* = \frac{n}{n-1} \cdot (\overline{x} - x_n).$$

Daher hat v_{neu} ein Minimum in jener Situation, in der der veränderte Datenwert an folgender Stelle ist:

$$x_n + h^* = x_n + \frac{n}{n-1} \cdot (\overline{x} - x_n) = \frac{n \cdot \overline{x} - x_n}{n-1} \qquad (3)$$

Das ist der Mittelwert der *anderen* $n - 1$ Datenwerte $x_1, \ldots, x_{n-1}$ und kann auch durch

$$\overline{x} + \frac{\overline{x} - x_n}{n-1} = \overline{x} + \frac{h^*}{n}$$

ausgedrückt werden, was dem *neuen Mittelwert* $\overline{x}_{\text{neu}}$ in dieser Situation entspricht.

Das heißt: Genau dann, wenn der neue n-te Datenwert $x_n + h$ dem *neuen Mittelwert* entspricht, dann nimmt die Varianz ihr Minimum an. Der einzige Weg, wie das passieren kann, ist in Abb. 1 nachzuvollziehen (oder umgekehrt, wenn x_n ursprünglich größer als der Mittelwert ist und über diesen hinweg verkleinert wird): Der n-te Datenwert bewegt sich nach „rechts“ in Richtung des ursprünglichen Mittelwerts und „passiert“ ihn; dabei bewegt sich auch der neue Mittelwert nach „rechts“, aber der n-te Datenwert ist „schneller“, holt den neuen Mittelwert ein und zieht schließlich an diesem vorbei. In Abb. 1 erfolgt dies beim Mittelpunkt zwischen x_n und dem „Punkt rechts“ mit $\Delta v = 0$, anders ausgedrückt beim Punkt

$$\bar{x} + \frac{n}{n-1}(\bar{x} - x_n).$$

Dieses Minimum von v_{neu} kann auf mehrere Arten erhalten werden. Erstens, wenn ein Wert genau beim Mittelwert (einer beliebigen Datenliste) liegt, dann trägt dieser ja nichts zur Summe der quadrierten Abstände vom Mittelwert bei. In unserem Fall besagt (3), dass dies der Mittelwert der *anderen* Datenwerte ist. Daher ist das Minimum von v_{neu} durch die *Summe der quadrierten Abstände zum Mittelwert* der anderen $n - 1$ Datenwerte, nämlich

$$\frac{n-2}{n-1} \times \text{Varianz der anderen } n-1 \text{ Datenwerte,}$$

gegeben. („Varianz” ist hier gemeint als *Stichprobenvarianz* = Summe der quadrierten Abstände zum Mittelwert geteilt durch $n - 1$ im Fall von n Datenwerten.)

Eine andere Möglichkeit ergibt sich aus dem rekursiven Ansatz, der für das Programmieren effizienter Algorithmen von Bedeutung ist. Wenn neue Daten dazukommen, muss man nicht „von vorne“ bei der Berechnung von Varianzen anfangen. In Chan (1994) werden rekursive Formeln für Mittelwert und Varianz einer Stichprobe der Größe n beschrieben, wenn ein neuer Datenwert hinzugenommen wird. Wenn der hinzugefügte Datenwert der Mittelwert der Stichprobe der Größe n ist, dann ergibt sich der Zusammenhang

$$n \cdot s_{n+1}^2 = (n-1) \cdot s_n^2.$$

In unserem Fall, wenn wir einen Wert zunächst entfernen, so dass nur mehr $n - 1$ Datenwerte übrigbleiben, und dann anschließend einen Datenwert genau beim Mittelwert dieser $n - 1$ Werte hinzufügen, ergibt sich, wie oben,

$$(n-1) \cdot s_n^2 = (n-2) \cdot s_{n-1}^2.$$

4 Zusammenfassung

Eine Aufgabe in einem Schulbuchentwurf hat uns dazu geführt, über die geschilderten Phänomene (Auswirkung des Verschiebens eines Datenwertes auf die Varianz) erstmals genauer nachzudenken. Die Effekte des Bewegens, Streichens bzw. Hinzufügens von Datenwerten gehören zu einem Teilgebiet der Statistik, das man *robuste Schätzverfahren* nennt.

Es wäre u. E. noch schöner, die besprochenen Phänomene auch ohne Rechnung *einsehen* bzw. *begreifen* zu können, aber vielleicht geht das gar nicht, denn Summen von Abweichungsquadraten sind eben ein wenig sperrig. Wieder einmal bestätigt sich: Auch intuitiv naheliegende Phänomene bedürfen zu ihrer Erklärung oft eines gewissen mathematischen Aufwandes.

Die involvierte Mathematik von Abschnitt 2 ist wenig abstrakt und auf Schulniveau zugänglich, so dass dieses Thema auch in der Schule behandelt werden kann. Es ist eine gute Gelegenheit *explorierend* die intuitive Vermutung zu untersuchen:

> Immer wenn man einen Datenpunkt vom Mittelwert wegbewegt (näher zum Mittelwert bringt), dann vergrößert (verkleinert) sich die Varianz.

Das Ergebnis, dass diese Vermutung *immer* stimmt, wenn der neue Datenwert näher beim ursprünglichen Mittelwert ist, aber *nur meistens* stimmt, wenn der neue Datenwert weiter entfernt zum ursprünglichen Mittelwert liegt, bestätigt die Wichtigkeit Intuitionen zu überprüfen.

Beispiele, die das Ergebnis illustrieren, auch solche, in denen die Intuition versagt, sind auch im Schulunterricht leicht zugänglich. Darüber hinaus könnte die Untersuchung für etwas ältere Schüler*innen zu interessanten Zusammenhängen wie in Abschnitt 3 führen.

Literatur

Chan, Y.-M. (1994): Combining means and variances of samples. In: *Teaching Statistics* 16(3), S. 80.

Wiley (*Teaching Statistics*) hat dem Autor die Publikation einer deutschen Fassung in *Stochastik in der Schule* genehmigt.

Anschrift des Autors

Hans Humenberger
Fakultät für Mathematik
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien
hans.humenberger@univie.ac.at