VOM MITTELWERT ZUR VARIANZ: EINE BETRACHTUNG ZUR FEHLERFORTPFLANZUNG

von ROBERT M. LYNCH

Originaltitel in "Teaching Statistics" Vol. 10 (1988), No. 2:
The Variance with Computational Error in the Mean

R. M. Lynch lehrt am Western Australian College of Advanced Education
Übertragung: I. Strauß, Kronberg im Taunus

Zusammenfassung: Formeln zur Berechnung der Varianz reagieren bei fehlerbehafteten Daten unterschiedlich sensibel. Im folgenden wird an zwei speziellen Formeln gezeigt, wie sich Abweichungen beim Mittelwert auf die Größe der Varianz auswirken.

ZDM-Klassifikation: K40

Ich bereitete mich auf eine Unterrichtsstunde vor, die sich mit der Varianz bei der Normalverteilung beschäftigen sollte. Dabei entdeckte ich eine Diskrepanz zwischen den beiden meistgebrauchten Formeln für die Berechnung dieser Varianz. Zunächst bespricht man aus pädagogisch-methodischen Gründen

$$S^2 = \frac{\sum (X-\bar{X})^2}{N}$$

Später geht man, der einfacheren Berechenbarkeit wegen, zu der Formel

$$S^2 = \frac{\sum X^2}{N} - \bar{X}^2$$

über. Algebraisch sind beide Ausdrücke äquivalent. Sie werden auch in vielen einführenden Statistik-Büchern unproblematisiert nebeneinander gestellt.

Da die zweite Formel aus der ersten abgeleitet ist, erwartet man bei der Anwendung irrigerweise immer identische Resultate. Ist der Mittelwert frei von zufallsbedingten Fehlern, liefern selbstverständlich beide Formeln gleiche und richtige Werte. Diese algebraische Äquivalenz geht jedoch verloren, wenn der

Mittelwert fehlerbehaftet ist; die Ergebnisse differieren, beide sind inkorrekt.

Die Diskrepanz muß aufgeklärt werden. Wir diskutieren sie zunächst aus algebraischer Sicht. Dann folgt ein numerisches Beispiel.

ALGEBRAISCHE ABWEICHUNG

Den Fehler bei der Berechnung des Mittelwertes nennen wir E:

$$\bar{X}_{berechnet} = \bar{X}_{wahr} + E.$$

Leicht kann gezeigt werden, wie dieser Fehler sich in der 'pädagogischen' Formel auswirkt:

$$S^2 = \frac{\sum (X - (\bar{X}_{wahr} + E))^2}{N};$$

er erhöht die geschätzte Varianz um E^2, d. h. der berechnete Wert wird den wahren um E^2 übersteigen. Damit läßt sich schreiben:

$$S^2 = \frac{\sum (X - \bar{X}_{wahr})^2}{N} + E^2 .$$

Die Einführung desselben Fehlers in die zweite Formel erbringt jedoch ein anderes Resultat. Aus

$$S^2 = \frac{\sum X^2}{N} - (\bar{X}_{wahr} + E)^2$$

folgt

$$S^2 = \frac{\sum X^2}{N} - \bar{X}^2_{wahr} - 2\bar{X}_{wahr}E - E^2 .$$

In letzterem Fall beträgt also die Differenz zwischen korrekter und geschätzter Stichprobenvarianz $-2\bar{X}_{wahr}E - E^2$.

Verschiedene Beobachtungen schließen sich an. Gleichgültig, ob der Fehler beim Mittelwert groß oder klein ist, ob er vom Runden, Abschneiden oder aus arithmetischen Falschberechnungen

stammt, er führt zu unterschiedlichen Resultaten. Die Differenz zwischen der korrekten Berechnung und der aus der 'pädagogischen' Formel ist E^2, was bedeutet, daß immer eine positive Abweichung (ein Überschätzen) von S^2 vorliegt. Bei der berechnungsfreundlichen Formel wird die Differenz nicht nur von E beeinflußt, sondern auch von der absoluten Größe des Stichproben-Mittelwertes. Sie kann daher durchaus erheblich sein, auch wenn der Fehler selbst klein ist. Bei einem großen Stichproben-Mittelwert ist demnach die Möglichkeit einer beachtlichen Abweichung immer präsent. Das Fazit: Die 'pädagogische' Formel muß stets dann unsere Wahl sein, wenn Genauigkeit gefordert ist.

Obgleich wir hier nur eine von mehreren möglichen Formeln zur handlichen Berechnung der Varianz diskutiert haben, gelten unsere Erörterungen grundsätzlich auch für andere Umformungen.

BEISPIEL

In Tabelle 1 ist eine Datenmenge mit verschiedenen relevanten Berechnungsschritten abgedruckt. Zunächst gehen wir davon aus, daß E = 0 ist; die Ergebnisse sind korrekt und identisch. Dann setzen wir E = -2. Wie der untere Teil der Tabelle zeigt, wird die Varianz in beiden Fällen richtig zu S^2 = 33,33 berechnet. Sodann wird erneut, jetzt mit E = -2 und einem Mittelwert von 18, gerechnet. Die 'pädagogische' Formel liefert eine geschätzte Varianz von 37,33, also eine Abweichung von E^2 = 4. Die 'schnellere' Formel schätzt dagegen 109.33, was eine Diskrepanz von $-2\overline{X} - E^2$, d. h. 76 ergibt. Im letzten Teil sind die algebraischen Äquivalente aufgeführt. Sie verifizieren die vorigen Resultate bei Benutzung eines Mittelwertes von 20 und eines Fehlers von -2.

Tabelle 1: BERECHNUNGEN FÜR DIE STICHPROBEN-VARIANZ
(Die Tabelle wird unübersetzt, weil im Text erläutert, abgedruckt. Bei * und ** haben sich im Original Druckfehler eingeschlichen: Es muß 225 resp. ΣfX^2 heißen. I. Strauß)

				$\bar{X}+E=20$ $\bar{X}=20,\ E=0$		$\bar{X}+E=18$ $\bar{X}=20,\ E=-2$	
X	f	X^2	fX^2	$(X-\bar{X})^2$	$f(X-\bar{X})^2$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
30	5	900	4 500	100	500	144	720
25	10	625	6 250	25	250	49	490
20	15	400	6 000	0	0	4	60
15	10	235*	2 250	25	250	9	90
10	5	100	500	100	500	64	320
$N=45$		$\Sigma X^2=19\,500$**		$\Sigma(X-\bar{X})^2=1\,500$		$\Sigma(X-\bar{X})^2=1\,680$	

Correct estimates

Pedagogic

$$S^2=\frac{\Sigma(X-\bar{X})^2}{N}=33{\cdot}33$$

Computational

$$S^2=\frac{\Sigma X^2}{N}-\bar{X}^2=33{\cdot}33$$

Incorrect estimates—($E=-2$ and $\bar{X}_{\text{computed}}=18$)

Pedagogic

$$S^2=\frac{\Sigma(X-\bar{X})^2}{N}=37{\cdot}33$$

Computational

$$S^2=\frac{\Sigma X^2}{N}-\bar{X}^2=109{\cdot}33$$

Algebraic equivalents—($E=-2,\ \bar{X}_{\text{true}}=20$)

Pedagogic

$$S^2=\frac{\Sigma(X-\bar{X})^2}{N}+E^2=37{\cdot}33$$

Computational

$$S^2=\frac{\Sigma X^2}{N}-\bar{X}^2-2\bar{X}E-E^2=109{\cdot}33$$