Jörg MEYER, Hameln

12. 4. 2021

Zur Abhängigkeit der Varianz von einem Einzelwert[1]

Man habe die Daten a, b, x. Der Mittelwert $\frac{x}{3}+\frac{a+b}{3}$ hängt linear von x ab, die Varianz quadratisch, so dass kein monotoner Zusammenhang zu erwarten ist. Sehen wir uns das an: Die Varianz

$$V(a, b, x)=\frac{\left(x-\frac{a+b+x}{3}\right)^2+\left(a-\frac{a+b+x}{3}\right)^2+\left(b-\frac{a+b+x}{3}\right)^2}{3}=\frac{2}{9}\cdot\left(\left(x-\frac{a+b}{2}\right)^2+\frac{3}{4}\cdot(a-b)^2\right)$$

beschreibt in Abhängigkeit von x eine nach oben geöffnete Parabel und nimmt den kleinsten Wert für $x=\frac{a+b}{2}$ an; dieser beträgt $\frac{1}{6}\cdot(a-b)^2$ und ist *kleiner* als die Varianz

$$V(a, b)=\frac{\left(a-\frac{a+b}{2}\right)^2+\left(b-\frac{a+b}{2}\right)^2}{2}=\frac{(a-b)^2}{4}$$ zu a und b.

Für $x=\frac{a+b}{2}\pm\frac{\sqrt{6}}{4}\cdot(a-b)$ ist $V(a, b, x)=V(a, b)$.

Allgemeiner habe man die Daten a_1, a_2, ..., a_n und x, und es sei $S:=\sum_{i=1}^{n} a_i$ und $Q:=\sum_{i=1}^{n} a_i^2$. Der Mittelwert $\frac{S}{n+1}+\frac{x}{n+1}$ hängt wieder linear von x ab.

Mit $b_j:=(n+1)\cdot a_j-S$ ist wegen $\sum_{j=1}^{n} b_j=(n+1)\cdot S-n\cdot S=S$ und wegen

$$\sum_{j=1}^{n} b_j^2=(n+1)^2\cdot Q-2\cdot S^2\cdot(n+1)+n\cdot S^2=(n+1)^2\cdot Q-S^2\cdot(n+2)$$

ist die Varianz gegeben durch

[1] Zu diesen Ausführungen wurde ich angeregt von Hans Humenberger: How does the change of a single data point affect the variance, and why? In: Teaching Statistics 2020; 42: 87-90. Eine deutsche Übersetzung erscheint demnächst in Stochastik in der Schule.

$$V\left(a_1,\ldots,a_n,x\right)=\frac{\left(x-\frac{S}{n+1}-\frac{x}{n+1}\right)^2+\sum_{j=1}^{n}\left(a_j-\frac{S}{n+1}-\frac{x}{n+1}\right)^2}{n+1}$$

$$=\frac{\left(n\cdot x-S\right)^2+\sum_{j=1}^{n}\left(\left(n+1\right)\cdot a_j-S-x\right)^2}{\left(n+1\right)^3}$$

$$=\frac{\left(n\cdot x-S\right)^2+\sum_{j=1}^{n}\left(b_j-x\right)^2}{\left(n+1\right)^3}$$

$$=\frac{x^2\cdot\left(n^2+n\right)-2\cdot x\cdot\left(n\cdot S+\sum_{j=1}^{n}b_j\right)+S^2+\sum_{j=1}^{n}b_j^2}{\left(n+1\right)^3}$$

$$=\frac{x^2\cdot\left(n^2+n\right)-2\cdot x\cdot\left(n\cdot S+S\right)+S^2+\left(n+1\right)^2\cdot Q-S^2\cdot\left(n+2\right)}{\left(n+1\right)^3}$$

$$=\frac{x^2\cdot n\cdot\left(n+1\right)-2\cdot x\cdot S\cdot\left(n+1\right)-S^2\cdot\left(n+1\right)+\left(n+1\right)^2\cdot Q}{\left(n+1\right)^3}$$

$$=\frac{x^2\cdot n-2\cdot x\cdot S-S^2+\left(n+1\right)\cdot Q}{\left(n+1\right)^2}$$

Dieser Term beschreibt wieder eine nach oben geöffnete Parabel. Für $x=\frac{S}{n}=\frac{\sum_{i=1}^{n}a_i}{n}$ wird (analog zum Fall $n=2$) der kleinste Wert angenommen. Insbesondere steigt die Varianz, wenn sich x von $\frac{\sum_{i=1}^{n}a_i}{n}$ entfernt. Der kleinste Wert beträgt

$$V_{min}=\frac{\frac{S^2}{n}-2\cdot\frac{S}{n}\cdot S-S^2+\left(n+1\right)\cdot Q}{\left(n+1\right)^2}=\frac{-S^2\cdot\frac{n+1}{n}+\left(n+1\right)\cdot Q}{\left(n+1\right)^2}=\frac{n\cdot Q-S^2}{n\cdot\left(n+1\right)}$$

und ist (wie im Fall $n=2$) kleiner als die Varianz $V\left(a_1,\ldots,a_n\right)=\frac{\sum_{j=1}^{n}\left(a_j-\frac{S}{n}\right)^2}{n}=\frac{n\cdot Q-S^2}{n^2}$ für n Daten.

Es gibt also wieder zwei x-Werte, für die $V\left(a_1,\ldots,a_n,x\right)=V\left(a_1,\ldots,a_n\right)$ ist.