Warum nur n-1 und nicht n? Erwartungstreue – leicht gemacht

Raphael Diepgen, Bochum

Zusammenfassung: Es wird der Vorschlag gemacht, im Unterricht über Beschreibende Statistik mit einfachen nichtprobabilistischen Repräsentativitätsüberlegungen den Nenner n-1 für den erwartungstreuen Varianzschätzer plausibel zu machen.

1. Beschreibende Statistik im Stochastikcurriculum

Der neue Lehrplan für die gymnasiale Oberstufe in Nordrhein-Westfalen sieht für die Jahrgangsstufe 11 neben Koordinatengeometrie und Differentialrechnung ganz-rationaler Funktionen einen obligatorischen Kurs über Beschreibende Statistik vor. Die inhaltlichen Beziehungen dieses isolierten Statistikunterrichtes zum übrigen schulischen Stochastikunterricht bleiben freilich im Lehrplan unklar, und dies wohl aus gutem Grund: Denn einerseits ist es jeweils nicht sicher, dass in der Sekundarstufe I die dort „offiziell" vorgesehene Stochastik tatsächlich unterrichtet wurde, auf die sich dann ein Statistikkurs in der Jahrgangsstufe 11 verlässlich beziehen könnte, und andererseits soll wohl auch der einzig abiturrelevante Stochastikunterricht in der Jahrgangsstufe 12/13 unabhängig davon unterrichtet werden können, ob in der Jahrgangstufe 11 tatsächlich Statistik unterrichtet wurde oder nicht. Diese weise Rücksicht auf die üblichen Kluften zwischen geduldigem Lehrplanpapier und widerständiger Schulwirklichkeit hat allerdings Folgen für den Statistikunterricht: Es handelt sich notgedrungen um einen nichtprobabilistischen Unterricht, der nur Häufigkeiten kennt, aber keine Wahrscheinlichkeiten, nur Stichproben, aber keine Populationen. Und dennoch: Auch dieser nichtprobabilistische Statistikunterricht erleichtert – so bestimmt der Lehrplan – „den Zugang zu stochastischen Problemstellungen in den nachfolgenden Jahrgangsstufen". Nur wie? Hier ein kleiner Vorschlag.

2. Motivation der Fragestellung

Der Unterricht über die Varianz führt unschwer zu der verblüffenden Begegnung mit dem seltsamen Nenner n-1. Jeder statistikfähige Taschenrechner fordert die Wahl zwischen den Nennern n und n-1 – ebenso jedes Statistikkochbuch, jede statistische Formelsammlung. Für den Schüler, der gerade die Varianz als mittlere quadratische Abweichung kennengelernt hat, dürfte der plötzlich angebotene Nenner n-1 völlig unnatürlich, kontraintuitiv und verwirrend sein. Was haben sich die verrückten Mathematiker dabei bloß wieder gedacht? Dass man die Welt selbst nicht versteht – damit kann man notfalls leben. Aber dass man noch nicht einmal versteht, was sich andere gedacht haben, das fällt denn

Stochastik in der Schule **19**(1999), Nr. 1, S. 10-13

doch schwer. Kurzum: Motivation für und Neugier auf das Warum der seltsamen Sitte (oder Unsitte?) mit dem Nenner n-1 dürfte leicht zu wecken sein.

Hat der Lehrer diese Neugier einmal geweckt, so hilft ihm die didaktische Literatur kaum. Die Vorschläge dort – man studiere etwa den jüngsten Überblick von Pöppelmann (1997) – sind mathematisch wohl zu komplex, um in der Jahrgangsstufe 11 leicht bewältigt werden zu können. Selbstverständlich lässt sich dort in einem nichtprobabilistischen Statistikunterricht nicht das probabilistische Konzept der Erwartungstreue in seiner ganzen Tragweite einführen; schließlich kann man sich hier ja noch nicht einmal auf einen elaborierten Begriff des Erwartungswertes beziehen. Aber: Es lässt sich ganz elementar mit einem basalen Verständnis von „Repräsentativität" argumentieren – und zwar dann, wenn man den Varianzbegriff nicht über die Abweichungen der Messwerte von ihrem Mittelwert eingeführt hat, sondern über die Abweichungen der Messwerte voneinander.

3. Abweichungen voneinander versus Abweichungen vom Mittelwert

Die Varianz ist nämlich die Hälfte der mittleren Abweichung der Messwerte voneinander, wie auch in Jahrgangstufe 11 mit einfachen algebraischen oder geometrischen Mitteln zu zeigen ist (vgl. Bielig-Schulz u.a. 1999):

Es seien n Messwerte x_i gegeben. Dann gilt für deren mittlere quadratische Abweichung voneinander:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((x_i - \bar{x}) + (\bar{x} - x_j)\right)^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - x_j) + (\bar{x} - x_j)^2\right)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})(\bar{x} - x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{x} - x_j)^2\right]$$

$$= \frac{1}{n^2}\left[2\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}\left((x_i - \bar{x})\sum_{j=1}^{n}(\bar{x} - x_j)\right)\right]$$

$$= \frac{1}{n^2}\left[2n\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})\sum_{j=1}^{n}(\bar{x} - x_j)\right]$$

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

Dies ist aber gerade das Zweifache der mittleren quadratischen Abweichung vom Mittelwert, also das Zweifache der Varianz.

Diese Argumentation benutzt – deshalb wurde sie hier ausführlich wiedergegeben – lediglich einfache algebraische Argumente wie Binomische Formeln, Vertauschen, Zusammenfassen und Ausklammern – und natürlich die Ausgleichseigenschaft des Mittelwerts $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Auf dieser Basis lässt sich nun der Nenner n-1 für den Schätzer der Populationsvarianz, also der halben mittleren quadratischen Abweichung der Messwerte voneinander, ganz einfach begründen:

4. Repräsentativitätsüberlegungen

Bei n Werten gibt es n^2 Abweichungen der Werte voneinander. Von diesen Abweichungen sind aber n „von vornherein" gleich Null, nämlich die Abweichungen eines Wertes jeweils von sich selbst. Der relative Anteil dieser „Von-vornherein-Nulldifferenzen" beträgt bei n Werten $n/n^2=1/n$. In der Population mit „riesigem" n ist dieser Anteil $1/n$ verschwindend gering; die „Von-vornherein-Nulldifferenzen" spielen also für die mittlere quadratische Abweichung der Werte voneinander und damit für deren Hälfte, also die Populationsvarianz, so gut wie keine Rolle. In einer kleinen Stichprobe jedoch mit dem Umfang n=10 beispielsweise ist der relative Anteil der „Von-vornherein-Nulldifferenzen" mit 1/10 recht groß; die relativ vielen „Von-vornherein-Nulldifferenzen" machen sich hier bei der Berechnung der mittleren quadratischen Abweichung der Messwerte voneinander und deren Hälfte, also der Stichprobenvarianz mit Nenner n, deutlich bemerkbar. Wie? Sie machen oder lassen klein. Kurzum: Die in der Stichprobe berechnete mittlere quadratische Abweichung der Messwerte voneinander bzw. Varianz ist „systematisch" kleiner als – und unterschätzt daher – die mittlere Abweichung bzw. Varianz in der Population, weil in der Stichprobe die „Von-vornherein-Nulldifferenzen" überrepräsentiert sind.

Will man diese Unterschätzung der Populationsvarianz vermeiden, muss man die Überrepräsentanz der „Von-vornherein-Nulldifferenzen" in der Stichprobe neutralisieren, also dafür sorgen, dass in der Stichprobe der relative Anteil der „Von-vornherein-Nulldifferenzen" genauso groß ist wie in der Population, nämlich praktisch Null. Das aber heißt: Man muss in der Stichprobe für die Berechnung der mittleren quadratischen Abweichung der Messwerte voneinander die n „Von-vornherein-Nulldifferenzen" von vornherein unberücksichtigt lassen; man hat es also nur noch mit n^2-n=n(n-1) quadratischen Differenzen zu tun, deren Mittel es zu finden gilt. Der Nenner n^2 der mittleren quadratischen Abweichung der Messwerte voneinander reduziert sich also auf n(n-1). Beginnt man mit dieser Veränderung die Ableitung oben, so ergibt sich für den Nenner der Stichprobenvarianz nur noch n-1.

Diese einfache Argumentation löst das Problem mit der „nichtprobabilistischen" Überlegung, dass ein Mittelwert in einer Population durch einen Mittelwert in einer Stichprobe systematisch „verfehlt" wird, wenn bestimmte Gruppen von Werten – hier die Nullen – in der Stichprobe gegenüber der Population über- oder unterrepräsentiert werden. Ein solches Konzept von Repräsentativität dürfte Schülern in der Jahrgangsstufe 11 allemal geläufig sein.

5. Fazit

Dieses Beispiel zeigt, dass man auch in einem nichtprobabilistischen Unterricht über Beschreibende Statistik die nichttriviale inferenzstatistische Thematik der Erwartungstreue mit ganz einfachen Überlegungen bearbeiten kann, ausgehend von einer für die Schüler überaus verblüffenden Begegnung mit einer seltsamen Praxis.

6. Ergänzung

Selbstverständlich kann man diese theoretischen Überlegungen empirisch bestätigen, indem man etwa von den Schülern hinreichend viele kleine Versuchsserien etwa mit n=3 durchführen, die Varianzen jeweils mit Nenner n und n-1 berechnen und deren Mittelwerte mit der theoretisch berechneten „wahren" Varianz vergleichen lässt. Diese Simulation repräsentiert das Konzept der Erwartungstreue, insofern hier nach dem Durchschnitt eines Schätzers auf lange Sicht gefragt wird. Indessen: Wenn man das Ganze bei normalverteilten Variablen dann auch noch mit dem Nenner n+1 durchführen lässt, zeigt sich plötzlich, dass der Varianzschätzer mit diesem seltsamen Nenner in einem gewissen Sinne noch besser ist, weil er nämlich auf lange Sicht im Schnitt die geringste quadratische Abweichung von der „wahren" Varianz hat. Auch wenn sich dies im Unterricht über Beschreibende Statistik nicht weiter begründen lässt, so macht es doch darauf aufmerksam, dass es unter inferenzstatistischer Perspektive ganz verschiedene Wünsche an einen „guten" Schätzer gibt – von denen die Erwartungstreue sicherlich nicht der wichtigste ist.

Literatur:

Bielig-Schulz, G.; Diepgen, R.; Jahnke, T.; Lapport, G.; Wuttke, H. (1999): Mathematik 11. Schuljahr. Berlin: Cornelsen

Pöppelmann, T. (1997): Bemerkungen zur Division durch n-1 bei der empirischen Varianz. – In: Der Mathematikunterricht 43 (H. 4), S. 26-35

Dipl.-Psych. Dr. Raphael Diepgen
Ruhr-Universität Bochum / Fakultät für Psychologie
D-44780 Bochum
e-Mail: raphael.diepgen@ruhr-uni-bochum.de