Eine geometrische Sicht auf das arithmetische Mittel[1]

JYOTIRMOY SARKAR, INDIANAPOLIS, INDIANA UND MAMUNUR RASHID, GREENCASTLE, INDIANA

[1] Original: A geometric view of the mean of a set of numbers.
In *Teaching Statistics* 38 (2016) 3, 77–82.
Kürzung, Bearbeitung und Übersetzung: JÖRG MEYER

Zusammenfassung: *Häufig wird das arithmetische Mittel als Hebelpunkt unter einem Punkteplot visualisiert. In diesem Aufsatz werden statt dessen die kumulative Häufigkeitsverteilung und die kumulativen Histogrammdaten verwendet.*

1 Einleitung

Das arithmetische Mittel als weit verbreiteter Lageparameter einer Menge von n quantitativen Daten $\{x_1, x_2, x_3, \ldots, x_n\}$ ist definiert durch

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

Bei der Verwendung ist Obacht geboten, da das arithmetische Mittel nicht immer ein aussagekräftiger Lageparameter ist, wie etwa bei Savage (2009) erläutert wird.

Der hauptsächliche Anlass für diesen Artikel ist eine gut verständliche Visualisierung des arithmetischen Mittels. Wir entwickeln sie im Kontext eines Beispiels, das mit Daten beginnt und danach in Histogrammdaten umgewandelt wird.

2 Das arithmetische Mittel von Daten

Die in einem Test erreichten Punkte von 25 Schülerinnen und Schülern sind die folgenden:

96, 94, 93, 90, 88, 88, 88, 85, 85, 85, 82, 80, 77, 75, 75, 75, 72, 72, 71, 66, 65, 65, 55, 50, 48.

Das arithmetische Mittel ist $\bar{x} = \frac{1920}{25} = 76{,}8$.

Einer Möglichkeit zur Visualisierung besteht darin, die Daten als kleine Bälle mit einheitlicher Masse auf einer masselosen Zahlengeraden darzustellen; man erhält so ein Punkteplot (Abb. 1).

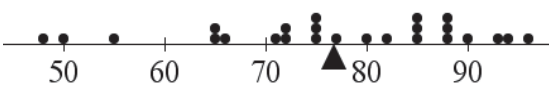

Abb. 1: Daten als Bälle

Andere Punktplots wie Cleveland-Punktplots oder Histodot-Plots werden in Wilkinson (1999) erläutert. Nach dem Hebelgesetz muss der Hebelpunkt genau im Schwerpunkt angebracht werden, dessen Stelle mit dem arithmetischen Mittel übereinstimmt (Devore 2015).

Diese Visualisierung hilft zu erkennen, dass das arithmetische Mittel empfindlich ist gegenüber der Änderung von einem Datum oder von mehreren Daten. Jedoch ist die genaue Lage des Hebelpunkts unmöglich festzustellen, ohne das arithmetische Mittel tatsächlich auszurechnen. Somit können leicht Fehler gemacht werden, insbesondere dann, wenn viele Daten vorliegen.

In diesem Artikel erläutern wir eine andere geometrische Visualisierung des arithmetischen Mittels. Man betrachte den Graphen (Abb. 2) zur empirischen kumulativen Häufigkeitsverteilung $y = F(x)$. Es ist $F(x) = N(x)/n$, wobei $N(x)$ die Anzahl der Werte ist, die nicht größer als x sind. F ist eine Treppenfunktion, die bei 0 startet und bei 1 endet (Rice 2007). Die in Abb. 2 eingezeichnete senkrechte Gerade führt dazu, dass die Flächeninhalte A und B gleich groß sind. Dass das arithmetische Mittel diese Eigenschaft hat, wird im Anhang bewiesen.

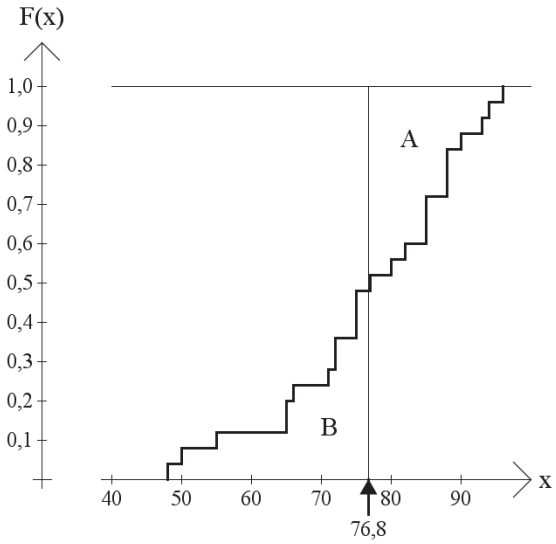

Abb. 2: Die empirische kumulative Verteilungsfunktion

Eine heuristische Begründung geht so: Vertauscht man die beiden Achsen in Abb. 2, d. h. betrachtet man die inverse Funktion $y = F^{-1}(x)$, so ist das arithmetische Mittel aufgrund seiner Definition die Gesamtfläche unter der inversen Funktion über dem Intervall

[0; 1] , es ist aber auch so groß wie die Fläche des Rechtecks [0; 1] × [0; $\bar{x}$]. Daher muss $A = B$ sein.

Diese Visulisierung mit der senkrechten Linie schlägt eine Brücke zwischen dem arithmetischen Mittel und der kumulativen Häufigkeitsverteilung. Man wird bei vielen Daten die senkrechte Linie leichter einzeichnen können als den Hebelpunkt.

3 Das arithmetische Mittel von Histogrammdaten

Oftmals hat man eine so große Datenmenge, dass die genaue Konstruktion der kumulativen Häufigkeitsverteilung schwierig ist. Auch wenn sie mit Software erstellt wurde, mag es immer noch schwierig sein, die Lage der senkrechten Linie gut einzuschätzen. Häufig werden jedoch die Daten in Form eines Histogramms klassifiziert. Dadurch verliert man zwar an Präzision, aber das Verständnis wird besser.

Beim Histogramm werden die Daten in – normalerweise gleich breite – Klassen eingeteilt; zur Ermittlung der Anzahl der Klassen kann man sich an der Regel von Sturges orientieren (Sturges 1926). Abb. 3 zeigt ein solches Histogramm. Das arithmetische Mittel beträgt nun 76,9.

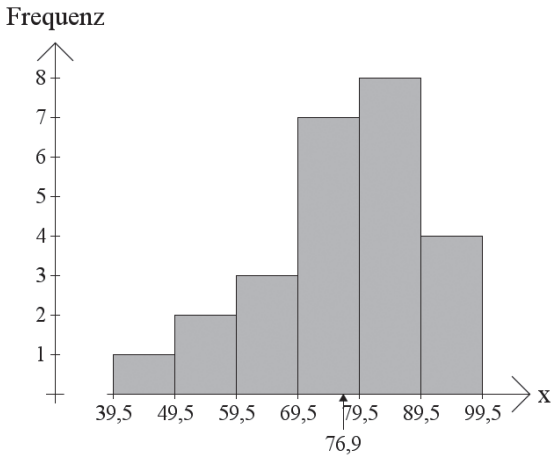

Abb. 3: Histogramm der Daten

Die gewöhnliche Art, das arithmetische Mittel durch den Hebelpunkt zu visualisieren und sich dabei das Histogramm als aus Blechstreifen bestehend vorzustellen, macht insbesondere Schwierigkeiten, wenn man die Lage des Hebelpunktes schätzen soll.

Statt dessen wandeln wir das Histogramm in ein kumulatives Histogramm wie in Abb. 4 um. Dabei handelt es sich um den Graphen zu $y = H(x)$, wo $H(x)$ die Fläche unter dem Histogramm links von x ist, wenn man sie als Anteil der Gesamtfläche unter dem Histogramm ausdrückt. Da das Histogramm eine Treppenfunktion ist, ist das kumulative Histogramm stückweise linear.

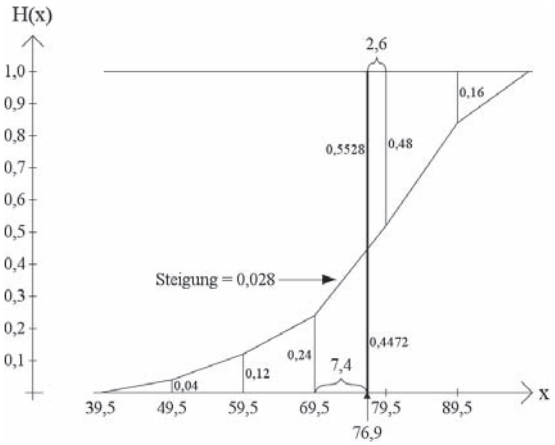

Abb. 4: Kumulatives Histogramm

Die zu 76,9 gehörige senkrechte Linie hat die Eigenschaft, dass die Fläche zwischen Graph und x-Achse links mit der Fläche zwischen Graph und der Geraden zu $H(x) = 1$ rechts übereinstimmt.

Natürlich könnte man die in Rede stehenden Flächen für jede mögliche Wahl der senkrechten Linie ausrechnen. Aber durch eine geschickte Wahl geht es einfacher. Wir nehmen das gewichtete Mittel der Abszissen-Mittelpunkte der Histogramm-Streifen, wobei die Gewichte die Frequenzen der Streifen sind. Sind $m_1, m_2, \ldots, m_K$ die Streifenmitten und $f_1, f_2, \ldots, f_K$ die entsprechenden Frequenzen der K Streifen, so ist das gewichtete Mittel nach Ramachandran/Tsokos (2015) gegeben durch

$$\bar{x}_M = \frac{\sum_{i=1}^{K} m_i \cdot f_i}{\sum_{i=1}^{K} f_i}; \qquad (2)$$

der Index M gibt an, dass die Streifenmitten gemeint sind.

Die Berechnung von $\bar{x}_M$ nach (2) geht schneller als die Berechnung von $\bar{x}$ nach (1), da K i. a. viel kleiner ist als n. Für das Histogramm von Abb. 2 bekommt man mit Formel (2) den Wert 76,9.

Für jedes Histogramm gilt, dass die zu $\bar{x}_M$ gehörige senkrechte Gerade die Eigenschaft hat, dass die Fläche zwischen Graph und x-Achse links mit der Fläche zwischen Graph und der Geraden zu $H(x) = 1$ rechts übereinstimmt. Man braucht für die Ermittlung der senkrechten Geraden nicht zu probieren, da man (2) verwenden kann.

4 Anmerkung

Bei einem kumulativen Histogramm ist die Aufgabe, eine Gerade zu finden, die die erwähnte Flächenteil-Eigenschaft hat, genauso leicht wie die Ermittlung des Medians bei einem Histogramm. Denn bei der Median-Aufgabe braucht man nur vom kumulativen Histogramm auszugehen und den Graphen mit $H(x) = 1/2$ zu schneiden.

5 Anhang

Ohne Beschränkung der Allgemeinheit seien alle x_i positiv (falls nicht, kann man eine hinreichend große Zahl C hinzuaddieren; nach Berechnung des arithmetischen Mittels wird diese Zahl C wieder subtrahiert).

Egal, wie man die Gerade g: $x = t$ wählt, stets sind die in Rede stehenden (von g abhängigen) Flächeninhalte A_t und B_t endlich.

In Abb. 5 wird am Beispiel {3, 9, 12, 12, 14} gezeigt, dass

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = A_0$$

gilt, und in Abb. 6, dass $[A_0 - A_t] + B_t = t \cdot 1 = t$ gilt. Kombiniert man beide Gleichungen, bekommt man $\bar{x} = A_0 = t + A_t - B_t$ für jede Wahl von t.

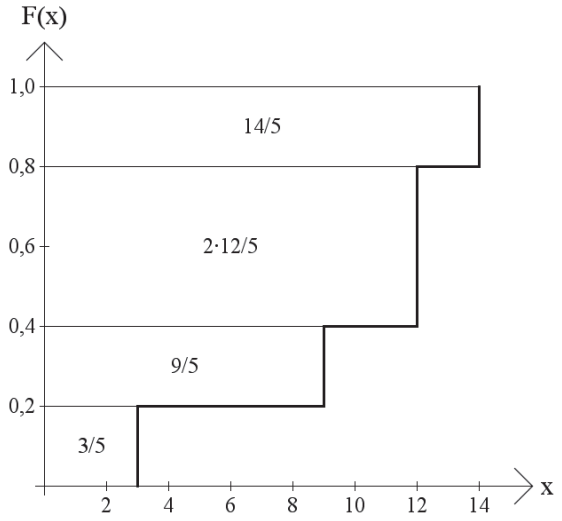

Abb. 5: Das arithmetische Mittel

Der Flächeninhalt A_t nimmt monoton für $t \in (-\infty; x_{max})$ ab und nähert sich 0, und der Flächeninhalt B_t nimmt für $t \in (x_{min}; \infty)$ ausgehend von 0 monoton zu. Nach dem Satz von Rolle (Anton et al. 2012) gibt es einen eindeutigen Wert τ mit $A_\tau = B_\tau$.

Für diesen speziellen Wert τ gilt $\bar{x} = t + (A_\tau - B_\tau) = \tau$. Daher ist $\tau = \bar{x}$ die eindeutige Lösung von $A_\tau = B_\tau$, wie in Abb. 2 dargestellt wird.

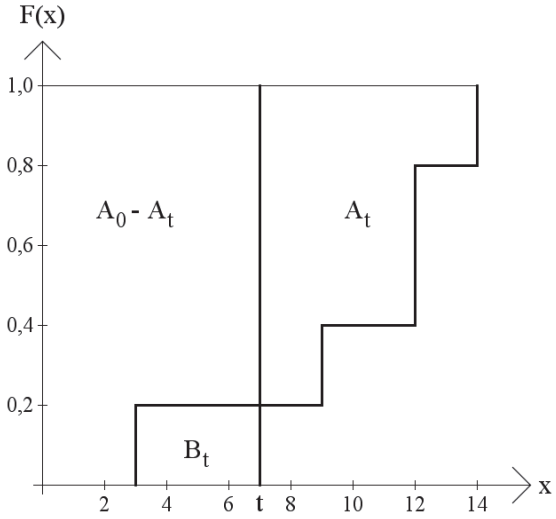

Abb. 6: Eine weitere Beziehung

Literatur

Anton, H.; Bivens, I.; Davis, S. (2012): Calculus: Early Transcendentals, 10th edn. New York, NY: Wiley.

Devore, J. (2015): Probability and Statistics for Engineering and Sciences, 9th edn. Boston, MA: Brooks/Cole, Cengage Learning.

Ramachandran, M. K.; Tsokos, C. P. (2015): Mathematical Statistics with Applications in R, 2nd edn. San Diego, CA: Elsevier Inc.

Rice, J. (2007): Mathematical Statistics and Data Analysis, 3rd edn. Boston, MA: Brooks/Cole, Cengage Learning.

Savage, S. (2009): The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty, Hoboken, NJ: John Wiley & Sons, Inc.

Sturges, H. (1926): The choice of a class interval. In: *Journal of the American Statistical Association*, 21, S. 65–66.

Wilkinson, L. (1999): Dot plot. In: *The American Statistician*, 53(3), S. 276–281.

Anschrift der Verfasser

Jyotirmoy Sarkar
Indiana University – Purdue University Indianapolis, Indiana; USA
jsarkar@iupui.edu

Mamunur Rashid
DePauw University, Greencastle, Indiana; USA
mrashid@depauw.edu