Petra HAUER-TYPPELT, Wien

Die Normalverteilung als Fehlerverteilung

1. Warum dieser Fokus

Zufällige, nicht systematische Messfehler werden im Stochastikunterricht, sowohl im Schul- als auch im Hochschulbereich, häufig als Beispiele für normalverteilte Größen herangezogen. Hat man die Normalverteilung mit den Lernenden erarbeitet, lässt sich bereits anhand der Glockenkurve und ihren Eigenschaften gut plausibel machen, warum Messfehler normalverteilt sind. Die Begründung mittels zentralen Grenzwertsatzes geht von der Interpretation aus, dass Messfehler aus einer Summe von sehr vielen, voneinander unabhängigen Fehlerkomponenten aufgebaut sind, wobei jede der einzelnen Komponenten nur einen geringfügigen Beitrag liefert.

Im Folgenden wird die Normalverteilung nicht wie üblich im Nachhinein als Messfehlerverteilung interpretiert, sondern direkt aus dem praxisnahen Umgang mit Messfehlern entwickelt. Dazu regt auch ein historischer Aspekt an. So war C.F. Gauß durch seine Tätigkeit als Geodät unweigerlich mit der Analyse zufälliger Messfehler konfrontiert.

2. Ausgangspunkte – Setzen des Ankers

Eine bestimmte Größe *t* sei *n*-mal mit gleich bleibender Sorgfalt gemessen. Dabei sind Abweichungen der Messwerte vom wahren Wert zufällig und voneinander unabhängig. Es liegen keinerlei systematische Messfehler vor. Der Fehler der *i*-ten Messung wird mit $u_i = t - t_i$ bezeichnet.

Wie lässt sich nun der unbekannte Wert *t* aus den *n* Messwerten $t_1, t_2, \ldots, t_n$ schätzen? In der Praxis ist es üblich, als besten Schätzer für *t* das arithmetische Mittel $\bar{t} = \frac{t_1 + t_2 + \ldots + t_n}{n}$ zu verwenden.

Lernende haben meiner Erfahrung nach kein Problem sich dieser Vorgangsweise anzuschließen, im Gegenteil sie schlagen sie meist selbst vor. Dennoch oder gerade deswegen ist es essentiell, hier auch andere mögliche Schätzer zu thematisieren. Denn es wird sich zeigen, dass die Entscheidung $\bar{t}$ als besten Schätzer für *t* zu verwenden, bereits die Verteilung der Fehler u_i festlegt. Die Wahl eines anderen Schätzers führt zu einer anderen Fehlerverteilung, so gelangt man beispielsweise durch die Wahl des Medians zur Laplace-Verteilung (siehe z. B. [3], S.78).

Die Forderung $\bar{t}$ als besten Schätzer für *t* zu verwenden, ist ident mit der

Forderung $\sum_{i=1}^{n} u_i = 0$. Diesen entscheidenden Ansatzpunkt der Herleitung bezeichnen wir als **Anker**:

$$\sum_{i=1}^{n} u_i = 0 \qquad \text{Forderung 1 = Anker}$$

Die Werte u_i dürfen als Realisierungen von Zufallsvariablen $U_1, U_2, ..., U_n$ betrachtet werden. Da alle u_i aus derselben Messreihe stammen, haben alle U_i dieselbe Wahrscheinlichkeitsverteilung bzw. dieselbe Dichtefunktion *f*.

3. Anwenden der Maximum-Likelihood-Methode

Zur Bestimmung der Dichtefunktion f wird die Maximum-Likelihood-Methode in umgekehrter Argumentationsrichtung zur Anwendung gebracht: Der beste Schätzer ist nicht gesucht, sondern (als Wunsch) vorgegeben und jene Funktion gesucht, die an dieser Stelle ein Maximum annimmt.
Wird also $\bar{t}$ als bester Maximum-Likelihood-Schätzer für t angesehen, muss die gemeinsame Dichtefunktion $g(U_1, U_2, ..., U_n)$ der n-dimensionalen Zufallsvariablen $(U_1, U_2, ..., U_n)$ für $\bar{t}$ ein Maximum annehmen.
Da die Messungen voneinander unabhängig sind und daher $U_1, U_2, ..., U_n$ unabhängige Zufallsvariablen sind, lässt sich die gemeinsame Dichtefunktion g als Produkt schreiben: $g(u_1, u_2, ..., u_n) = f(u_1) \cdot f(u_2) \cdot ... \cdot f(u_n)$

Um beim anstehenden Differenzieren anstelle des Produkts mit einer Summe arbeiten zu können, betrachten wir ln g. Da g eine nicht negative und ln x eine streng monoton wachsende Funktion ist, nimmt ln g an derselben Stelle wie g ihre Maxima an.

$$\begin{aligned} \ln g(u_1, u_2, ..., u_n) &= \ln f(u_1) + \ln f(u_2) + ... + \ln f(u_n) \\ &= \ln f(t - t_1) + \ln f(t - t_2) + ... + \ln f(t - t_n) := h(t) \end{aligned}$$

Differenzieren und Nullsetzen liefert

$$h'(t) = \frac{f'(t - t_1)}{f(t - t_1)} + \frac{f'(t - t_2)}{f(t - t_2)} + ... + \frac{f'(t - t_n)}{f(t - t_n)} = 0 \text{ für } t = \bar{t}\,.$$

Damit ergibt sich als Anforderung an die gesuchte Dichtefunktion *f*, die wir als Forderung 2 bezeichnen:

$$\frac{f'(u_1)}{f(u_1)} + \frac{f'(u_2)}{f(u_2)} + ... + \frac{f'(u_n)}{f(u_n)} = 0 \qquad \text{Forderung 2}$$

4. Erfüllen der beiden Forderungen

Mit der Substitution $\frac{f'(u_i)}{f(u_i)} := F(u_i)$ lauten nun die beiden Forderungen

$$u_1 + u_2 + ... + u_n = 0 \qquad \text{Forderung 1 = Anker}$$
$$F(u_1) + F(u_2) + ... + F(u_n) = 0 \quad \text{Forderung 2}$$

Beide Gleichungen müssen für alle Werte von *n* gelten, eine Betrachtung für $n = 2$ liefert: $-u_1 = u_2$ und $-F(u_1) = F(u_2)$ und damit $F(-u_1) = -F(u_1)$, *F* ist also eine ungerade Funktion.

Anker und Forderung 2 liefern für ein beliebiges *n* betrachtet:

aus Forderung 1: $F(-u_1) = F(u_2 + u_3 + ... + u_n)$ und
aus Forderung 2: $-F(u_1) = F(u_2) + F(u_3) + ... + F(u_n)$

und damit $F(u_2 + u_3 + ... + u_n) = F(u_2) + F(u_3) + ... + F(u_n)$.

F muss also die Cauchy'sche Funktionalgleichung $F(u_i + u_j) = F(u_i) + F(u_j)$ erfüllen. Für stetige Funktionen *F* bilden genau die linearen Funktionen $F(u_i) = k \cdot u_i$ die Lösungen dieser Gleichung.

5. Lösen der Differentialgleichung und Bestimmen der Konstante

Damit erhalten wir für die Summanden der Forderung 2: $\frac{f'(u_i)}{f(u_i)} = k \cdot u_i, \quad k \in R, \forall i$

Die Lösung dieser einfachen Differentialgleichung kann auch ohne vorangegangene Auseinandersetzung mit Differentialgleichungen leicht erfolgen und führt über $\ln|f(u)| = k \cdot \frac{u^2}{2} + C_1$ zu $f(u) = C \cdot e^{k \cdot \frac{u^2}{2}}$.

Zu bestimmen sind noch *k* und *C*, beginnen wir mit *k*. Entsprechend der Problemstellung treten sehr große Fehler wesentlich seltener als kleine Fehler auf, *k* muss daher negativ sein. Klarerweise muss *k* auch deshalb negativ sein, weil *f* Dichtefunktion ist und daher die Normierungsbedingung $\int_{-\infty}^{\infty} f(u)du = 1$ erfüllen muss. Wir setzen daher $k = -h^2$ und erhalten $f(u) = C \cdot e^{-h^2 \frac{u^2}{2}}$.

Die Normierungsbedingung legt auch die Konstante *C* fest. Die notwendige Berechnung des Integrals $\int_{-\infty}^{\infty} e^{-h^2\frac{u^2}{2}}\,du$ ist per Hand, Bleistift auf Papier, mit einigem Aufwand verbunden[1]. Um an dieser Stelle einen Exkurs zu vermeiden und den roten Faden nicht zu verlieren, kann ein CAS zum Einsatz kommen, das rasch zum Ergebnis $\frac{\sqrt{2\pi}}{|h|}$ führt. Damit hat *f* nun die Gestalt $f(u)=\frac{|h|}{\sqrt{2\pi}}\cdot e^{-h^2\frac{u^2}{2}}$ und das ist gerade die Dichtefunktion einer $N(0,\frac{1}{h^2})$-verteilten Zufallsvariable.

6. Didaktische Bemerkungen

Bei der vorliegenden Darstellung handelt es sich um eine heuristische Annäherung an die Normalverteilung, die auf Betonung eines sehr wichtigen Teilaspekts in der Bedeutung der Normalverteilung abzielt. In Anlehnung an den Wortursprung – heurisko, „ich finde“ – soll das Finden des Weges, begonnen bei der Wahl des Ankers, über die mathematischen Methoden hin zur gesuchten Dichtefunktion, im Zentrum der Überlegungen stehen. Das Zeigen der Äquivalenz von „Wahl des arithmetischen Mittels als besten Schätzer“ und „die Fehler sind normalverteilt“ dient nicht zur Einführung der Normalverteilung, sondern einem fundierten Verständnis der Normalverteilung. Gleichzeitig wird damit exemplarisch Einsicht in das Zusammenspiel von Stochastik und Analysis gegeben. Deshalb eignet sich dieser Zugang zum Einsatz in der Lehramtskandidatenausbildung. Im Schulbereich bleibt dieser Weg zur Normalverteilung wohl eher Spezialgruppen vorbehalten, begründet sowohl durch die Ansprüche an das stochastische Vorwissen (Maximum-Likelihood-Methode, mehrdimensionale Zufallsvariablen, …) als auch an das mathematische Rüstzeug der Lernenden.

7. Literatur

[1] Petra Hauer-Typpelt: Zugänge zur Normalverteilung und ihre fachdidaktische Analyse. Dissertation an der Universität Wien, Wien 1998

[2] Josef Heinhold, Karl Gaede: Ingenieur Statistik. R. Oldenburg Verlag, München, Wien 1968

[3] Jörg Meyer: Schulnahe Beweise zum zentralen Grenzwertsatz. Verlag Franzbecker, Hildesheim, Berlin 2004

[1] Zwei Möglichkeiten dazu werden in [1], S.41 ff. vorgestellt