Schluss von der Stichprobe auf die Grundgesamtheit

Martin Wabnik

Jeder Mensch, der schon mal den „Schluss von der Stichprobe auf die Grundgesamtheit“ oder den „Hypothesentest“ unterrichtet hat, weiß wohl ein Lied davon zu singen, wie schwierig es ist, Schüler davon zu überzeugen, dass mit diesen Methoden nichts über die Wahrscheinlichkeit ausgesagt wird, mit der eine vorliegende Stichprobe aus einer bestimmten Grundgesamtheit gezogen wurde.

Dieses Problem ist auch deshalb so hartnäckig, weil wir im Alltag ohnehin „falschherum“ schließen und damit normalerweise gut durch's Leben kommen.

Um es mal ganz einfach zu formulieren: Wir sind davon überzeugt, dass, wenn wir eine Stichprobe mit wenigen roten Kugeln haben, diese höchstwahrscheinlich aus einer Grundgesamtheit mit wenigen roten Kugeln gezogen wurde und nicht aus einer Grundgesamtheit mit vielen roten Kugeln.

Haben wir z.B. eine Stichprobe mit 30 % roter Kugeln und 70 % gelber Kugeln, schließen wir intuitiv, diese Stichprobe wahrscheinlich aus einer Grundgesamtheit gezogen zu haben, die ca. 30 % rote Kugel enthält.

Das ist sinnvoll, weil wir aus der Kombinatorik wissen: Eine Grundgesamtheit mit 30 % roter Kugeln hat viel mehr mögliche Stichproben mit 30 % roter Kugeln als eine Grundgesamtheit mit 40 % oder mit 50 % roter Kugeln.

Gefühlsmäßig stellen wir uns vor, unsere Stichprobe aus einem Topf gezogen zu haben, in dem sich alle möglichen Stichproben mit 30 % roter Kugeln aus allen möglichen Grundgesamtheiten befinden. Und da der Anteil der Stichproben in diesem Topf, die aus Grundgesamtheiten mit ungefähr 30 % roter Kugeln stammen, viel größer als der Anteil der Stichproben aus anderen Grundgesamtheiten, ist die Wahrscheinlichkeit groß, eine Stichprobe aus einer ungefähr-30-%-Grundgesamtheit zu ziehen.

Mathematisch gesehen stellen wir uns folgendes vor: Es wird ein zweistufiger Zufallsversuch durchgeführt: Auf der ersten Stufe wird zufällig eine der möglichen Grundgesamtheiten gezogen, auf der zweiten Stufe wird aus dieser Grundgesamtheit zufällig eine Stichprobe entnommen.

Mit der bedingten Wahrscheinlichkeit lassen sich dann den einzelnen Grundgesamtheiten die Wahrscheinlichkeiten zuordnen, mit denen die vorliegende Stichprobe aus genau dieser Grundgesamtheit gezogen wurde.

Konkret sieht das Prinzip so aus: Es ist eine geordnete Stichprobe $S_{n;k}$ vom Umfang n mit Zurücklegen gezogen worden[1]. In der Stichprobe $S_{n;k}$ sollen sich k Elemente (mit $0 < k < n$)[2] mit der Eigenschaft E befinden.

Die Stichprobe $S_{n;k}$ sei aus einer Grundgesamtheit vom Umfang N gezogen worden. Sei nun K die Anzahl der Elemente der Grundgesamtheit, die die Eigenschaft E hat. Dann ist $0 < K < N$ (denn wäre z.B. $K = 0$, so wäre auch $k = 0$).

Es gibt

[1] Hier wird von einer *geordneten* Stichprobe und von einem Ziehen *mit Zurücklegen* ausgegangen, weil man so auf eine Formel kommt, die der Formel für die Berechnung von Wahrscheinlichkeiten einzelner Werte binomialverteilter Zufallsgrößen sehr ähnlich ist.

[2] um wenig lehrreichen Spezialfällen wie $k = 0$ und $k = n$ aus dem Weg zu gehen

$$\binom{n}{k} \cdot K^k \cdot (N-K)^{n-k} \tag{1}$$

mögliche Stichproben $S_{n;k}$ aus einer Grundgesamtheit mit N Elementen, von denen K die Eigenschaft E haben.

Insgesamt gibt es

$$\sum_{K=1}^{N-1} \binom{n}{k} \cdot K^k \cdot (N-K)^{n-k} \tag{2}$$

mögliche geordnete Stichproben $S_{n;k}$ aus allen möglichen Grundgesamtheiten.

Die Wahrscheinlichkeit, dass $S_{n;k}$ aus einer Grundgesamtheit mit K_0 Elementen mit der Eigenschaft E gezogen wurde, ist dann

$$P(S_{n;k} \ \ aus \ \ G_{K_0}) = \frac{\binom{n}{k} \cdot {K_0}^k \cdot (N-K_0)^{n-k}}{\sum_{K=1}^{N-1} \binom{n}{k} \cdot K^k \cdot (N-K)^{n-k}}. \tag{3}$$

Besteht die Grundgesamtheit aus allen Bürgern der Bundesrepublik Deutschland, gibt es bezüglich K ca. 83000000 unterschiedliche mögliche Grundgesamtheiten. Die Wahrscheinlichkeit, dass eine bestimmte Stichprobe aus einer Grundgesamtheit mit einem bestimmten K gezogen wurde, ist demnach ziemlich gering. Vermutlich will man das meist auch gar nicht wissen, sondern es ist eine viel gröbere Abschätzung sinnvoll. Dazu kann z.B. $N = 10$ gesetzt werden. (Wenn N viel kleiner als n ist, ist das kein Problem.) Ist $K_0 = 3$ sieht Gleichung (**??**) dann so aus:

$$\frac{\binom{n}{k} \cdot 3^k \cdot (10-3)^{n-k}}{\sum_{K=1}^{9} \binom{n}{k} \cdot K^k \cdot (10-K)^{n-k}}. \tag{4}$$

Das ist die Wahrscheinlichkeit, dass die Stichprobe $S_{n;k}$ aus einer Grundgesamtheit mit 30% Anteil von Elementen mit der Eigenschaft E gezogen wurde, verglichen mit den Möglichkeiten 10%, 20%, ...,90%.

Es könnte aber auch interessant sein, wie wahrscheinlich es ist, dass eine Stichprobe aus einem bestimmten Bereich von Grundgesamtheiten kommt.

Dazu addiert man alle möglichen Stichproben aus diesem Bereich - z.B. von K_i bis K_s mit $0 < K_i \leq K_0 \leq K_s < n$ - und teilt diese Summe durch die Summe aller möglichen Stichproben aus allen möglichen Grundgesamtheiten.

$$\frac{\sum_{K=K_i}^{K_s} \binom{n}{k} \cdot K^k \cdot (N-K)^{n-k}}{\sum_{K=1}^{N-1} \binom{n}{k} \cdot K^k \cdot (N-K)^{n-k}} \tag{5}$$

Nun sind die verwendeten Zahlen wegen ihrer Größe doch recht unhandlich und so können wir mit $\frac{1}{K^n}$ erweitern, um so direkt auf die diskrete Likelihood-Funktion zu kommen.

$$\frac{\sum_{K=K_i}^{K_s} \binom{n}{k} \cdot \left(\frac{K}{N}\right)^k \cdot \left(\frac{N-K}{N}\right)^{n-k}}{\sum_{K=1}^{N-1} \binom{n}{k} \cdot \left(\frac{K}{N}\right)^k \cdot \left(\frac{N-K}{N}\right)^{n-k}} \tag{6}$$

Was kann man tun, wenn man die Anzahl der Elemente der Grundgesamtheit nicht einschränken möchte, um vielleicht ein beliebig genaues Ergebnis zu haben? Ersetzen wir $\frac{K}{N}$ durch $p \in [0;1] \subset \mathbb{R}$ und ersetzen wir die diskreten Summen durch Integrale,

können wir mit der üblichen Likelihood-Funktion rechnen (p_i entspricht dabei $\frac{K_i}{N}$ und p_s entspricht $\frac{K_s}{N}$).

$$\frac{\int\limits_{p=p_i}^{p_s} \binom{n}{k} \cdot (p)^k \cdot (1-p)^{n-k}\, dp}{\int\limits_{p=0}^{1} \binom{n}{k} \cdot (p)^k \cdot (1-p)^{n-k}\, dp} \tag{7}$$

Für $n = 50$ und $k = 15$ ergibt sich z.B.:

$$\frac{\int\limits_{0,25}^{0,35} \binom{50}{15} \cdot (p)^{15} \cdot (1-p)^{50-15}\, dp}{\int\limits_{0}^{1} \binom{50}{15} \cdot (p)^{15} \cdot (1-p)^{50-15}\, dp} \approx 0,5669 \tag{8}$$

Für $n = 100$ und $k = 30$ ist

$$\frac{\int\limits_{0,25}^{0,35} \binom{100}{30} \cdot (p)^{30} \cdot (1-p)^{100-30}\, dp}{\int\limits_{0}^{1} \binom{100}{30} \cdot (p)^{30} \cdot (1-p)^{100-30}\, dp} \approx 0,8790 \tag{9}$$

und für $n = 1000$ und $k = 300$ ist

$$\frac{\int\limits_{0,25}^{0,35} \binom{1000}{300} \cdot (p)^{300} \cdot (1-p)^{1000-300}\, dp}{\int\limits_{0}^{1} \binom{1000}{300} \cdot (p)^{300} \cdot (1-p)^{1000-300}\, dp} \approx 0,9994 \tag{10}$$