Schätzen mit Modus, Mittel, Median

H. R. Schneebeli, Wettingen

Zusammenfassung: *Angenommen, wir beobachten eine Bernoullikette mit der Länge n und es zeigen sich k Erfolge. Wie lässt sich aus dieser Beobachtung die Erfolgswahrscheinlichkeit p schätzen? Die Maximum Likelihood Schätzung findet $p = \frac{k}{n}$. Diese Methode lässt sich mit dem* Modus *einer Wahrscheinlichkeitsverteilung rechtfertigen. Sinngemäss entspricht der* Mittelwert *oder der* Median *derselben stetigen Verteilung je einer Schätzmethode. In Beispielen werden einige Eigenschaften der drei Schätzungen verglichen, die p ausgehend von den Daten n und k bestimmen. Jede der drei Schätzmethoden zeichnet sich durch eine* Extremaleigenschaft *aus.*

1 Das Vorgehen

Für jedes p im Intervall $[0,1]$ tritt die beobachtete Folge mit der Wahrscheinlichkeit $p^k(1-p)^{n-k}$ unter allen möglichen Folgen von n Ergebnissen auf. Unser Vorgehen variiert den Ansatz der Maximum Likelihood Schätzung. Dazu betrachten wir eine Zufallsvariable X auf $[0,1]$ mit einer Wahrscheinlichkeitsdichte $\rho : x \mapsto C \cdot x^k \cdot (1-x)^{n-k}$, wobei die Konstante C die Normierung $\int_0^1 \rho(x)\mathrm{d}x = 1$ besorgt. Die Funktion ρ ist im Intervall $[0,1]$ unimodal. Ihre Maximalstelle führt zur Maximum Likelihood Schätzung für p. Diese Schätzung entspricht dem Modus für die stetige Zufallsvariable X mit der Dichte ρ.

Würden wir ‘alle’ Bernoulliketten mit beliebigen Wahrscheinlichkeiten x und dem beobachteten Ergebnis herausgreifen können, so spielte in unseren Gedanken die Dichtefunktion ρ die Rolle der Werteverteilung bei den experimentellen Daten in der Statistik. Die Begriffe *arithmetisches Mittel* und *Median* sollten nun auf diese Situation angepasst werden. Es entsteht eine Aufgabe mit einem fast offensichtlichen Lösungsansatz. Aber was sind die Konsequenzen? Welchen Schätzwert für p gewinnen wir mit den verschiedenen Methoden aus den angenommenen Daten n und k? Wie gehen wir damit um, dass drei verschiedene Schätzungen für eine Grösse auftreten können? Was zeichnet die einzelnen Verfahren aus?

2 Erwartungswert und Median

1. Der Erwartungswert von X ist gegeben durch
$$E(X) = \int_0^1 x \cdot \rho(x)\mathrm{d}x$$
Er lässt sich salopp ausgedrückt als Verallgemeinerung des arithmetischen Mittels ‘aller’ Daten interpretieren.

2. Dem Median ‘aller’ Ergebnisse entspricht jene Zahl m, für welche
$$\int_0^m \rho(x)\mathrm{d}x = \frac{1}{2} = \int_m^1 \rho(x)\mathrm{d}x$$
gilt.

2.1 Elementare Beispiele: lauter Erfolge!

1. Es ist fast absurd, aber doch aufschlussreich, eine Bernoullikette der Länge 0 zu betrachten, bei der wir 0 Erfolge beobachten. Die Dichte der Gleichverteilung $\rho : x \mapsto 1$ gehört zu dieser Annahme. Mit der Maximum Likelihood Schätzung ist jede Zahl $0 \le p \le 1$ verträglich, alle Chancen bleiben offen. Erwartungswert und Median legen sich auf $p = \frac{1}{2}$ fest.

2. Angenommen, wir haben in *einem* Bernoulliversuch *einen* Erfolg beobachtet. Das heisst, $k = n = 1$. Dann ist $\rho : x \mapsto 2x$.
   a) Die Maximum Likelihood Methode schätzt $p_{ML} = 1$. Aufgrund eines einzigen Erfolgs wird eine Erfolgsgarantie ausgegeben.
   b) Mit dem Erwartungswert $E(X)$ wird $p_E = 2/3$.
   c) Der Median bestimmt die Schätzung $p_m = 1/\sqrt{2}$.

3. Allgemein ist der Fall $k = n$ leicht zu handhaben.
   a) Der Maximum Likelihood Schätzer hat keinen Grund, die ‘gute Erfahrung’ $p_{ML} = 1$ zu korrigieren.
   b) Der Erwartungswert im Fall der Dichte $\rho : x \mapsto (n+1)x^n$ schätzt $p_E = \frac{n+1}{n+2}$, ein Wert, der mit wachsendem n die stets guten Erfahrungen vorsichtiger berücksichtigt als p_{ML}.

c) Wer p mit dem Median von $\rho : x \mapsto (n+1)x^n$ schätzt, findet $p_m = 2^{-\frac{1}{n+1}}$. Diese Schätzung hängt ebenfalls von n ab. Obwohl lauter Erfolge beobachtet wurden, schätzt der Median vorsichtiger als der Maximum Likelihood Schätzer, aber weniger zurückhaltend als der Erwartungswert.

3 Drei Schätzungen im Vergleich

3.1 Erwartungswert als Schätzung

Im allgemeinen Fall ist $\rho : x \mapsto C \cdot x^k \cdot (1-x)^{n-k}$ die Dichte, mit der der Erwartungswert $E(X)$ zu bestimmen ist. Die Antwort lautet $E(X) = \frac{k+1}{n+2}$.

Beim Ermitteln der Normierungskonstante C und des Erwartungswertes kann man die Betafunktion benutzen. Allgemein ist

$$B : (s,t) \mapsto \int_0^1 x^{s-1}(1-x)^{t-1}\,\mathrm{d}x.$$

Wegen $1/C = B(k+1, n-k+1)$ und $E(X) = B(k+2, n-k+1)$ werden nur Werte $B(s,t)$ mit natürlichen Zahlen s und t benötigt. Diese Werte lassen sich wegen einer bekannten Beziehung zwischen der Betafunktion und der Gammafunktion mit Fakultäten ausdrücken.

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} = \frac{(t-1)!(s-1)!}{(s+t-1)!}.$$

Wer eine direkte Berechnung wagt, kommt mit partieller Integration und Rekursion schliesslich zum selben Ziel.

3.2 Median als Schätzung

Wer den Median bei der angenommenen Dichtefunktion $\rho : x \mapsto C \cdot x^k \cdot (1-x)^{n-k}$ berechnen will, muss eine Polynomgleichung vom Grad $n+1$ in $[0,1]$ lösen. Das erfordert fast immer numerische Näherungen. Eine geschlossene Formel ist nur in Ausnahmefällen zu erwarten, etwa für $k = 0$, $k = \frac{n}{2}$, $k = n$. Dieser Umstand mag erklären, weshalb der Median bei diesen Schätzverfahren nicht verbreitet ist.

3.3 Vergleich der drei Methoden, ein Beispiel

Angenommen, $k = 2$ und $n = 5$. Die Schätzungen $p_{ML} = \frac{2}{5}$ und $p_E = \frac{3}{7}$ lassen sich sofort aus k und n berechnen. Aus der Bedingung $\int_0^m \rho(x)\mathrm{d}x = \frac{1}{2}$ folgt die Gleichung für m:

$$-10m^6 + 36m^5 - 45m^4 + 20m^3 = \frac{1}{2}$$

Die einzige Lösung in $[0,1]$ lautet $m \approx 0.421407\ldots$.

Somit sind $p_{ML} = 0.4 < p_m \approx 0.421407 < p_E \approx 0.4286$ die drei verschiedenen Schätzungen für p aufgrund unserer Daten.

Der Erwartungswert der Anzahl Erfolge in einer Bernoullikette mit Erfolgschance p und der Länge n beträgt $n \cdot p$. Die Schätzung p_{ML} bringt die Anzahl der erwarteten Erfolge mit der beobachteten in Übereinstimmung. Für p_m und p_E trifft dies nur ausnahmsweise zu.

4 Extremaleigenschaften

4.1 Erwartungswert

Welche Wahl von μ minimiert das Integral $I(\mu) = \int_0^1 (x-\mu)^2 \cdot \rho(x)\mathrm{d}x$? In Anlehnung an den diskreten Fall könnte man vermuten, dass $\mu = E(X)$ die Minimalstelle ist, während die Varianz

$$V(X) = \int_0^1 (x - E(X))^2 \cdot \rho(x)\mathrm{d}x$$

gerade den Minimalwert der mittleren quadratischen Abweichung von x zu allen möglichen Schätzungen für p angibt.

Das Integral $I(\mu)$ lässt sich in eine Summe von Integralen aufteilen

$$I : \mu \mapsto I(\mu) = \int_0^1 x^2 \cdot \rho(x)\mathrm{d}x - 2\mu \int_0^1 x \cdot \rho(x)\mathrm{d}x + \mu^2$$

Diese Darstellung zeigt, dass $I : \mu \mapsto I(\mu)$ eine quadratische Funktion ist. Die einzige Nullstelle der Ableitung $I'(\mu) = 2\mu - 2\int_0^1 x \cdot \rho(x)\mathrm{d}x = 2\mu - 2E(X)$ liegt bei $\mu = E(X)$ und bezeichnet ein Minimum.

4.2 Median

Gibt es eine Extremaleigenschaft, die den Median der stetigen Verteilung ρ auszeichnet? Im diskreten Fall minimiert der Median die mittlere absolute Abweichung zu den Daten. Stimmt es, dass der Median m das Integral $\int_0^1 |x-m| \cdot \rho(x)\mathrm{d}x$ minimiert? Das trifft tatsächlich zu, wie folgende Überlegung zeigt:

Wir betrachten die Funktion $f : u \mapsto f(u) = \int_0^1 |x-u| \cdot \rho(x)\mathrm{d}x$. Sie ist durch ein Faltungsintegral definiert. Die Funktion f ist differenzierbar, weil ρ glatt ist. Für die Ableitung $f' = D_u f : u \mapsto \int_0^u \rho(x)\mathrm{d}x - \int_u^1 \rho(x)\mathrm{d}x$ gilt dann $f'(0) = -1$, $f'(1) = 1$ und $f'(m) = 0$. Für $x \in [0,1]$ ist $\rho(x) > 0$. Also wächst $g(u) = \int_0^u \rho(x)\mathrm{d}x$ streng monoton, während $h(u) = \int_u^1 \rho(x)\mathrm{d}x$ monoton abnimmt. Daher ist $f'(u) = g(u) - h(u)$ streng monoton wachsend, der Median m ist die einzige Nullstelle von f'

auf $[0,1]$. Für $u < m$ ist $f'(u) < 0$ und $f'(u) > 0$ für $u > m$, daher ist der Median m Minimalstelle von f.

Anschrift des Verfassers

Hans Rudolf Schneebeli
Margelstr. 14
CH-5430 Wettingen
schneebe@othello.ch