

Statistisk inferens teori, punktskattningar

I sannolikhets teori har vi arbetat med s.v. vars fördelningar och parametrar varit kända. Utgående från en given slumpmodell har vi bestämt sannolikheter för olika händelser. I statistik teorin är förhållandet det omvända. Vi har då en mängd data från någon fördelning vars parametrar i regel är okända och vill använda data för att uppskatta parametrarna så bra som möjligt. Detta kallas statistisk inferens (=slutledning). Nufortiden skapas kontinuerligt mycket stora datamängder (finansiella transaktioner, konsumtionsmönster sociala nätverk). Hur ska dessa omsättas i relevant kunskap? Låt $x_1, x_2, \dots, x_n$ beteckna ett stickprov, dvs en samling observationer av de oberoende och likafördelade s.v. $X_1, X_2, \dots, X_n$. Den gemensamma fördelningen för X_i beror av en en- eller flerdimensionell parameter θ . De möjliga värden på θ bildar ett parameterrum $\mathcal{R}_\theta$ (Om $X_i \in N(\mu, \sigma)$, så är $\theta = (\mu, \sigma)$ och $\mathcal{R}_\theta = \mathbb{R} \times (0, \infty)$.)

Def: En punktskattning av en parameter θ är en funktion av stickprovet $x_1, \dots, x_n$ som för varje val av data ger ett värde $\hat{\theta}_{obs}^*$ på parametern θ , dvs $\hat{\theta}_{obs}^* = \theta^*(x_1, \dots, x_n)$

Stickprovet $x_1, \dots, x_n$ kan ses som ett utfall av de s.v. $X_1, \dots, X_n$. Därmed blir $\hat{\theta}_{obs}^*$ ett utfall av den s.v. $\theta^*(X_1, \dots, X_n)$ som kallas stickprovsvariabeln. Stickprovsvariabelns fördelning beror på fördelningarna för $X_1, \dots, X_n$ som i sin tur beror av θ . Vi skriver ofta θ^* istället för $\theta^*(X_1, \dots, X_n)$.

Ex: Låt θ beteckna sannolikheten för minst 1 mm nederbörd i Stockholm en godtyckligt vald dag. Under ett års tid registrerar vi antalet nederbördsdagar och betecknar detta med X . Stickprovet x är ett utfall av en s.v. $X \in \text{Bin}(365, \theta)$, vars fördelning beror av den okända parametern θ . En punktskattning $\hat{\theta}_{obs}^*$ ska ge ett värde på θ som är troligt givet stickprovet x . Eftersom $E(X) = 365\theta$, så är en rimlig punktskattning $\hat{\theta}_{obs}^* = \frac{x}{365}$.

Vi ska nu undersöka tre kriterier för vad som är en god punktskattning.

Def: En punktskattning $\hat{\theta}_{obs}^*$ är väntevärdesriktig om $E(\theta^*) = \theta$ för alla $\theta \in \mathcal{R}_\theta$.

En väntevärdesriktig punktskattning har inget systematiskt fel. Punktskattningen i exemplet är väntevärdesriktig eftersom

$$E(\theta^*) = E(X/365) = E(X)/365 = 365\theta/365 = \theta.$$

Vi önskar också att skattningarna ska bli bättre ju mer data som vi har tillgängligt. För att kvantifiera detta låter vi nedan $\hat{\theta}_n^*$ beteckna en stickprovsvariabel baserad på ett stickprov med n element.

Def: En punktskattning $\hat{\theta}_{obs}^*$ är konsistent om det för varje fixt $\theta \in \mathcal{R}_\theta$ och varje givet $\varepsilon > 0$ gäller att $\mathbb{P}(|\hat{\theta}_n^* - \theta| > \varepsilon) \rightarrow 0$ då $n \rightarrow \infty$.

Det är också önskvärt att välja en punktskattning så att stickprovsvariabeln har så liten spridning (varians) som möjligt.

Def: Om $\hat{\theta}_{obs}^*$ och $\hat{\theta}_{obs}^{\circ}$ är två väntevärdesriktiga punktskattningar och $V(\theta^*) \leq V(\theta^{\circ})$ för alla $\theta \in \mathcal{R}_\theta$ (med strikt olikhet för något θ), så sägs θ^* vara mer effektiv än θ° .

Den väntevärdesriktiga punktskattningen i exemplet uppfyller

$$V(\theta^*) = V(X/365) = \frac{1}{365^2} V(X) = \frac{1}{365^2} 365\theta(1-\theta) = \theta(1-\theta)/365$$

Notera att variansen beror på det sanna parametervärdet.

Vi kan kvantifiera de systematiska och statistiska felen samtidigt genom att undersöka medeltvadradsfelet (MSE) som ges av

$$MSE = E((\theta^* - \theta)^2) = \dots = V(\theta^*) + (E(\theta^*) - \theta)^2$$

Vi ska undersöka slumpmässigt fel systematiskt fel, som är noll om θ^* är en väntevärdesriktig punktskattning.

trä mycket vanliga punktskattningar och se om de är väntevärdesriktiga, konsistenta och effektiva. Låt $x_1, \dots, x_n$ vara ett stickprov av de oberoende s.v. $X_1, \dots, X_n$ som alla har (okänt) väntevärde μ och (okänt) varians σ^2 . Som punktskattningar av μ och σ kan vi använda stickprovsmedelvärdet och stickprovsvariansen, dvs

$$\mu_{obs} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{respektive} \quad (\sigma^2)_{obs} = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Båda dessa punktskattningar är väntevärdesriktiga. Allt $E(\bar{x}) = E(\frac{1}{n} \sum_{i=1}^n X_i) = \mu$ inses direkt och beviset av motsvarande egenskap för s^2 finns i boken. Kravet på väntevärdesriktighet är en anledning till att vi dividerar med $n-1$ istället för n i definitionen av s^2 . Båda punktskattningarna är också konsistenta. Konsistens för $\bar{x}$ visas på samma sätt som stora talens lag. Ingen av de båda punktskattningarna är dock den mest effektiva i alla situationer, även om de ofta ger goda skattningar. Det finns analytiska metoder för att härleda den bästa punktskattningen i en given situation. Två sådana är minstakvadratmetoden (MK) och maximum likelihoopsmetoden (ML).

Def: MK-skattningen θ_{MK}^* av parametern θ är det värde i $\mathbb{R}_0$ som minimerar kvadratsumman.

$$Q(\theta) = \sum_{i=1}^n (x_i - E(X_i))^2$$

der $E(X_i)$ beror på den okända parametern θ . (Om ex. vis $X_i \in \text{Exp}(\theta)$, så är $E(X_i) = \frac{1}{\theta}$)

Def: ML-skattningen θ_{ML}^* av parametern θ är det värde i $\mathbb{R}_0$ som maximerar den så kallade likelihoodfunktionen.

$$L(\theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta) \quad (\text{om } X_i \text{ är kontinuerlig, byt annars } f_{X_i} \text{ mot } p_{X_i}).$$

Likelihoodfunktionen ger ett mått på sannolikheten att stickprovet $x_1, \dots, x_n$ uppträder för ett visst parametervärde. ML-skattningen θ_{ML}^* är det parametervärde för vilket det är mest troligt att vi får det givna stickprovet. Produkten i definitionen av L gör räkningarna komplicerade, så ofta räknar vi istället med $\ln L(\theta) = \sum_{i=1}^n \ln f_{X_i}(x_i; \theta)$, vilket gör lika bra eftersom L och $\ln L$ antar maximum för samma värde på θ .

Ex: Vi har stickprovet $x_1 = 0,77$, $x_2 = 0,82$, $x_3 = 0,92$, $x_4 = 0,94$ och $x_5 = 0,98$ som antas komma från fördelningen $f_X(x) = \theta x^{\theta-1}$ för $x \in [0, 1]$ och någon okänd parameter $\theta > 0$. Vi vill bestämma ML-skattningen av θ och bestämma likelihoodfunktionen och dess logaritm.

$$L(\theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n \theta x_i^{\theta-1} = \theta^n (x_1 \cdots x_n)^{\theta-1}$$

$$\ln L(\theta) = \ln(\theta^n (x_1 \cdots x_n)^{\theta-1}) = n \ln \theta + (\theta-1) \ln(x_1 \cdots x_n) = n \ln \theta + (\theta-1) \sum_{i=1}^n \ln x_i$$

Derivering med avseende på den okända parametern θ ger nu:

$$\frac{d}{d\theta} \ln L(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \ln x_i = 0 \Rightarrow \frac{n}{\theta_{ML}^*} = - \sum_{i=1}^n \ln x_i \Rightarrow \theta_{ML}^* = - \frac{n}{\sum_{i=1}^n \ln x_i}$$

så för det givna stickprovet är $\theta_{ML}^* = - \frac{5}{\ln(0,77 \cdot 0,82 \cdot 0,92 \cdot 0,94 \cdot 0,98)} = 7,997$