

Sannolikhet och statistik A3

Johan Jonasson

May 2019

Styrka. Signifikansnivån för ett test är sannolikt felaktigt förkasta H_0 . Man vill heller inte felaktigt acceptera H_0 . Styrkan är sannolikheten att förkasta nollhypotesen som funktion av vad sanningen verkligen är. Konkret:

Testa $H_0 : \theta = \theta_0$ mot $H_A : \theta \neq \theta_0$. Styrkan är en funktion av vad θ verkligen är:

$$g(\theta_1) = \mathbb{P}_{\theta_1}(H_0 \text{ förkastas}).$$

Det gäller förstås att $g(\theta_0) = \alpha$.

Exempel: Slantsingling. Test av $H_0 : p = 1/2$ mot $H_A : p \neq 1/2$. $n = 100$. Låt X vara antal klave. Vi kom fram till att förkasta på 5% signnivå om $|X - 50| \geq 10$. Vad är $g(0.7)$? Om $p = 0.7$ gäller

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{X - 70}{\sqrt{21}} \approx N(0, 1).$$

Alltså

$$\mathbb{P}_{0.7}(|X - 10| \geq 10) = \mathbb{P}_{0.7}(X \geq 59.5) + \mathbb{P}_{0.7}(X \leq 40.5).$$

Den andra termen är försumbar och

$$\begin{aligned} \mathbb{P}_{0.7}(X \geq 59.5) &= \mathbb{P}_{0.7}\left(\frac{X - 70}{\sqrt{21}} \geq \frac{59.5 - 70}{\sqrt{21}}\right) \\ &\approx 1 - \Phi(-2.291) \\ &\approx 0.989. \end{aligned}$$

Alltså: $g(0.7) \approx 0.989$.

Vad är $g(0.6)$? Med samma räkningar blir

$$g(0.6) \approx 1 - \Phi\left(\frac{59.5 - 60}{\sqrt{24}}\right) \approx 1 - \Phi(-0.102) \approx 0.541.$$

Stickprov X_1, \dots, X_n på $N(\mu, \sigma^2)$ med σ^2 känd. Testa $H_0 : \mu = 0$ mot $H_A : \mu > 0$ på 1% signnivå. Hur stort behöver n vara för att få $\beta(1) = 0.9$? (Dvs sannol att förkasta ska vara 90% om det verkliga värdet på μ är 1.)

Testet förkastar H_0 om

$$\frac{\bar{X}}{\sigma/\sqrt{n}} \geq \Phi^{-1}(0.99) \approx 2.33.$$

Vi ska alltså beräkna

$$\mathbb{P}_1 \left(\frac{\bar{X}}{\sigma/\sqrt{n}} \geq 2.33 \right).$$

Om $\mu = 1$ är $(\bar{X} - 1)/(\sigma/\sqrt{n}) \sim N(0, 1)$ och

$$\begin{aligned} \mathbb{P}_1 \left(\frac{\bar{X}}{\sigma/\sqrt{n}} \geq 2.33 \right) &= \mathbb{P}_1 \left(\frac{\bar{X} - 1}{\sigma/\sqrt{n}} \geq 2.33 - \frac{1}{\sigma/\sqrt{n}} \right) \\ &= \Phi \left(\frac{\sqrt{n}}{\sigma} - 2.33 \right). \end{aligned}$$

Eftersom $\Phi^{-1}(0.9) \approx 1.28$ är högerledet ≥ 0.9 då

$$\frac{\sqrt{n}}{\sigma} - 2.33 \geq 1.28,$$

dvs då

$$n \geq (3.61\sigma)^2.$$

Bayesiansk statistik. Istället för att se okända parametrar som fixa tal, se dem som stokastiska variabler. Inte onaturligt, men vad ska vi tro att en parameter har för fördelning?? Med mycket data spelar det inte alltid så stor roll.

Ibland naturligt, ibland inte.

Exempel: Ett mynt singlar tio ggr. Vi vet att myntet antingen ger klave m.s. $2/3$ eller m.s. $1/3$. Om vi får klave sex ggr, vad ska vi tro om sannol att myntet ger klave?

Data $X \sim \text{Bin}(n, \theta)$. Vi fick $X = 6$. Är θ lika med $1/3$ eller $2/3$? Vi kan tänka oss att θ är en sv där $\mathbb{P}(\theta = 1/3) = \mathbb{P}(\theta = 2/3) = 1/2$ (ex.vis om draget ur en hatt med två mynt).

Isf:

$$\begin{aligned}
 \mathbb{P}(\theta = 2/3|X = 6) &= \frac{\mathbb{P}(X = 6|\theta = 2/3)\mathbb{P}(\theta = 2/3)}{\mathbb{P}(X = 6|\theta = 2/3)\mathbb{P}(\theta = 2/3) + \mathbb{P}(X = 6|\theta = 1/3)\mathbb{P}(\theta = 1/3)} \\
 &= \frac{\binom{10}{6}(2/3)^6(1/3)^4 \cdot (1/2)}{\binom{10}{6}(2/3)^6(1/3)^4 \cdot (1/2) + \binom{10}{6}(1/3)^6(2/3)^4 \cdot (1/2)} \\
 &= \frac{2^2}{2^2 + 1^2} \\
 &= \frac{4}{5}.
 \end{aligned}$$

Den fördelning av parametern vi tror på innan experimentet (i ex. $\mathbb{P}(\theta = 2/3) = 1/2$) kallas för θ :s *prior* och den betingade fördelningen givet data (i ex. $\mathbb{P}(\theta = 2/3|X = 6) = 4/5$) kallas för θ :s *posterior*.

(På svenska egentligen *à-priorifördelning* och *à-posteriorifördelning*.)

Bayes formel för tätheter:

$$\begin{aligned}
 f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \\
 &= \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|t)f_X(t)dt} \\
 &= Cf(x, y) = Cf_{Y|X}(y|x)f_X(x).
 \end{aligned}$$

De sista likheterna pga av att nämnaren inte har med x att göra.

I Bayesianisk statistik har vi alltid en parameter θ vars prior $f_{\theta}(t)$ är känd, liksom den betingade fördelningen för data givet θ : $f_{X|\theta}(x|t)$. Vi vill veta posterior $f_{\theta|X}(t|x)$.

Att ge en objektiv prior är oftast omöjligt, så därför är Bayesiansk statistik ibland uteslutet (tänkt till ex ett läkemedelsbolag som rapporterar till regulatoriska myndigheter).

Däremot mycket vanligt inom AI: en AI-algoritm har en uppfattning om världen i termer av sannolikheter. Den får data och uppdaterar sin uppfattning. Eller ett spelföretag som sätter odds.

Exempel: Om du vet att $X \sim \exp(\theta)$, att $0 < \theta < 1$ och obs $X = x$, vad tror du om θ ?

Man kan tänka sig en prior som är likformig: $f_{\theta}(t) = 1$, $0 < t < 1$. Då får vi

$$f_{\theta|X}(t|x) = C f_{X|\theta}(x|t) f_{\theta}(t) = C t e^{-tx}, 0 < t < 1$$

Obs att detta ska ses som en funktion av t . Vi känner igen den som täthet för $\Gamma(2, x)$ -förd, men bara för $0 < t < 1$, så posterior är en *trunkerad* gammafördelning. □

Exempel: Error correction. En dataström av 0/1:or kommer så att den startar på måfå och sedan följs en etta av en etta m.s. 0.8 och en nolla av en nolla m.s. 0.8. Varje siffra avläses fel m.s. 0.1. Om man ser 0100, vad är den mest sannolika korrekta dataströmmen, 0100 eller 0000?

Kalla obs för Y och det korrekta för X .

$$\begin{aligned}\mathbb{P}(X = 0100|Y = 0100) &\propto \mathbb{P}(Y = 0100|X = 0100)\mathbb{P}(X = 0100) \\ &= 0.9^4 \cdot 0.5 \cdot 0.2 \cdot 0.2 \cdot 0.8 \\ &\approx 0.0105.\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X = 0000|Y = 0100) &\propto \mathbb{P}(Y = 0100|X = 0000)\mathbb{P}(X = 0000) \\ &= 0.9^3 \cdot 0.1 \cdot 0.5 \cdot 0.8^3 \\ &\approx 0.0187.\end{aligned}$$