

Sannolikhet och statistik XII

Johan Jonasson

April 2019

Centrala gränsvärdesatsen: Låt X_1, X_2, \dots vara oberoende och likafördelade med $\mathbb{E}[X_1] = \mu$ och $\text{Var}(X_1) = \sigma^2 < \infty$. Låt

$$S_n = \sum_{k=1}^n X_k.$$

Då gäller för alla $x \in \mathbb{R}$ att

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \rightarrow \Phi(x)$$

då $n \rightarrow \infty$. M.a.o.: S_n är approximativt $N(n\mu, n\sigma^2)$ -fördelad när n är stort. (Oavsett vilken fördelning X_k :na har, men hur stort n ska vara för att approximationen ska vara god beror förstås på X_k :nas fördelning.)

Exempel: I världen inträffar i genomsnitt 100 större jordbävningar per år. Vad är sannolikheten att det ett givet år inträffar minst 110 större jordbävningar?

Rimligt att tro att jb kommer som en Poi-process med int 100. Antal jb $\sim Poi(100)$, lite jobbigt att räkna på. Alt: tiderna T_1, T_2, \dots mellan jb är ober och $exp(100)$ -förd. Låt

$$S_n = \sum_{k=1}^n T_k.$$

Vi söker $\mathbb{P}(S_{110} \leq 1)$. Vi har $\mu = \mathbb{E}[T_k] = 1/100 = 0.01$,
 $\sigma^2 = \mathbb{V}\text{ar}(T_k) = 1/100^2 = 0.0001$. Alltså, enl CGS,

$$\begin{aligned} \mathbb{P}(S_{110} \leq 1) &= \mathbb{P}\left(\frac{S_{110} - 110 \cdot 0.01}{0.01 \cdot \sqrt{110}} \leq \frac{1 - 110 \cdot 0.01}{0.01 \cdot \sqrt{110}}\right) \\ &\approx \Phi\left(-\frac{0.1}{\sqrt{0.011}}\right) \approx 0.17. \end{aligned}$$

Exempel: Låt $X \sim \text{Bin}(n, p)$ med n stort och p ickextremt. ($1/n \ll p \ll 1 - 1/n$; vi tänker att $n \rightarrow \infty$ och p fixt.) Skriv

$$X = \sum_{k=1}^n I_k.$$

Vi har $\mu = \mathbb{E}[I_k] = p$ och $\sigma^2 = \text{Var}(I_k) = p(1-p)$. Enl CGS

$$\mathbb{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \leq x\right) \rightarrow \Phi(x).$$

Ex.vis slå tärning 700 ggr och låt X vara antal sexor. Vad är $\mathbb{P}(X \geq 100)$?

(Halvkorrektion: X diskret så $\mathbb{P}(X \geq 100) = \mathbb{P}(X > 99) = \mathbb{P}(X > 99.5)$.)

$$\begin{aligned} \mathbb{P}(X > 99.5) &= 1 - \mathbb{P}\left(\frac{X - 700 \cdot \frac{1}{6}}{\sqrt{700 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \leq \frac{99.5 - 700 \cdot \frac{1}{6}}{\sqrt{700 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) \\ &\approx 1 - \Phi(-1.74) = \Phi(1.74) \approx 0.959. \end{aligned}$$

□

CGS fungerar även utan likafördelning under vissa enkla förutsättningar. Även utan oberoende under lite tuffare krav. I båda fallen måste variansen förstås justeras. Se upp med extrema sannolikheter.

STATISTIK

Somliga parametrar okända. Skatta parametrar utifrån data.

Skrivsätt: iid = independent and identically distributed = oberoende och likafördelade.

Definition: Låt X_1, X_2, \dots, X_n vara iid och förd som X . Då kallas X_1, X_2, \dots, X_n för ett *stickprov* (sample) på X (eller på F_X).

Antag att F_X beror på en parameter, θ . Ex.vis $X \sim Poi(\theta)$, $X \sim exp(\theta)$, $X \sim N(\theta, 1)$ eller $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = \theta$. Parametern kan vara flerdimensionell, t.ex. $\theta = (\mu, \sigma^2)$ och $X \sim N(\mu, \sigma^2)$.

En fkn av X_1, \dots, X_n som anv till att skatta θ kallas för en *punktskattning* (estimator) av θ . Standardbeteckningar $\hat{\theta}$, $\hat{\theta}_n$, $\hat{\theta}(X_1, \dots, X_n)$.

Obs att $\hat{\theta}$ är en stokastisk variabel. Efter data $X_1 = x_1, \dots, X_n = x_n$ observerats får man ett givet värde $\hat{\theta}(x_1, \dots, x_n)$, ett estimat av θ .

Definition: Om det, oavsett det korrekta värdet på θ , gäller att

$$\mathbb{E}[\hat{\theta}] = \theta$$

kallas $\hat{\theta}$ för en *väntevärdesriktig* skattning sv θ . (Unbiased) (vvr).

Definition: Om

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

kallas $\hat{\theta}$ för en *konsistent* skattning av θ .

Proposition: Om $\hat{\theta}_n$ är vvr och $\text{Var}(\hat{\theta}_n) \rightarrow 0$, så är $\hat{\theta}_n$ konsistent.

Bevis: Enligt Chebyshev's olikhet:

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \leq \frac{\text{Var}(\hat{\theta}_n)}{\epsilon^2} \rightarrow 0.$$



Låt X_1, \dots, X_n vara ett stickprov på en sv X med $\mathbb{E}[X] = \mu$ och $\text{Var}(X) = \sigma^2 < \infty$. Kom ihåg att

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k.$$

Vi vet att $\mathbb{E}[\bar{X}] = \mu$ och $\text{Var}(\bar{X}) = \sigma^2/n$, så enl prop är \bar{X} en konsistent skattning av μ . □

Man ska se konsistens som ett mycket önskvärt krav och som i sig själv inte räcker för att säga att en skattning är bra.

Definition: Om $\hat{\theta}$ och $\tilde{\theta}$ är två vvr skattningar av θ och

$$\text{Var}(\hat{\theta}) < \text{Var}(\tilde{\theta})$$

för alla θ , säger man att $\hat{\theta}$ är mer effektiv än $\tilde{\theta}$.

Skattning av varians: Låt X_1, \dots, X_n vara ett stickprov på en sv X med $\mathbb{E}[X] = \mu$ och $\text{Var}(X) = \sigma^2 < \infty$. Man brukar skatta σ^2 med

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2.$$

Proposition: s^2 är vvr. Om $\mathbb{E}[X^4] < \infty$ gäller också att s^2 är konsistent.

Delbevis: Det är lätt att se att $s^2 = \frac{1}{n-1} \left(\sum_{k=1}^n X_k^2 - n\bar{X}^2 \right)$. Det gäller att

$$\mathbb{E}[\bar{X}^2] = \text{Var}(\bar{X}) + \mathbb{E}[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Vi har också att $\mathbb{E}[X_k^2] = \text{Var}(X_k) + \mathbb{E}[X_k]^2 = \sigma^2 + \mu^2$. Detta ger

$$\mathbb{E}[s^2] = \frac{1}{n-1} \left(n(\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) \right) = \sigma^2.$$

□

Konfidensintervall. Låt X_1, \dots, X_n vara ett sp på en sv X vars förd beror av en okänd parameter θ . Om T_1 och T_2 är två funktioner av X_1, \dots, X_n och

$$\mathbb{P}(T_1 \leq \theta \leq T_2) = q$$

så kallas $[T_1, T_2]$ för ett konfidensintervall för θ av konfidensgrad q .
Man skriver

$$T_1 \leq \theta \leq T_2 (q).$$

Obs: Sannolikheten q gäller innan T_1 och T_2 observerats. När vi obs $T_1 = t_1$ och $T_2 = t_2$ och skriver $t_1 \leq \theta \leq t_2 (q)$, är detta ett påstående som antingen är sant eller falskt, men som, innan data observerats, hade sannol q att få värden t_1 och t_2 som skulle gjort det sant.

Låt $X \sim \text{likf}(0, \theta)$, θ okänd och X_1, \dots, X_n ett sp på X . Vi vet att $\mathbb{E}[\bar{X}] = \theta/2$, så

$$\hat{\theta} = 2\bar{X}$$

är en vvr skattning av θ .

Alt: skriv $M = \max\{X_1, \dots, X_n\}$. Skulle vara dumt att skatta θ som mindre än M . Låt oss göra en skattning baserad på M .

$$F_M(x) = \mathbb{P}(M \leq x) = \mathbb{P}(X \leq x)^n = \left(\frac{x}{\theta}\right)^n.$$

$$f_M(x) = \frac{1}{\theta^n} nx^{n-1}, \quad x \in (0, \theta).$$

$$\mathbb{E}[M] = \int_0^\theta x \frac{1}{\theta^n} nx^{n-1} dx = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1} \theta.$$

Alltså: $\tilde{\theta} = \frac{n+1}{n}M$ är en annan vvr skattning av θ . Vilken är mest effektiv?

$$\text{Var}(\hat{\theta}) = \frac{4}{n} \text{Var}(X) = \frac{\theta^2}{3n}.$$

$$\mathbb{E}[M^2] = \frac{n}{\theta^n} \int_0^\theta x^{n+1} dx = \frac{n}{n+2} \theta^2.$$

$$\text{Var}(M) = \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} \right) \theta^2$$

$$\text{Var}(\tilde{\theta}) = \left(\left(\frac{n+1}{n} \right)^2 \frac{n}{n+2} - 1 \right) \theta^2 = \frac{\theta^2}{n(n+2)}.$$

Vi har alltså att $\tilde{\theta}$ är mycket mer effektiv än $\hat{\theta}$.

Konfidensintervall: Verkar bra att låta T_1 och T_2 vara fkner av M . Vi gör ett symmetriskt konfintervall med konfidensgrad 95%, dvs tar $\mathbb{P}(T_1 \geq \theta) = \mathbb{P}(T_2 \leq \theta) = 0.025$. Vi har

$$\mathbb{P}(M \leq x) = \frac{x^n}{\theta^n} = 0.025$$

då $x = 0.025^{1/n}\theta$. Alltså

$$0.025 = \mathbb{P}(M \leq 0.025^{1/n}\theta) = \mathbb{P}\left(\theta \geq \frac{M}{0.025^{1/n}}\right).$$

På samma sätt

$$\mathbb{P}\left(\theta \leq \frac{M}{0.975^{1/n}}\right) = 0.025.$$

Alltså

$$\frac{M}{0.975^{1/n}} \leq \theta \leq \frac{M}{0.025^{1/n}} \quad (95\%).$$

I boken finns ett ex med $n = 10$ och $M = 8.69$. Vi får

$$8.71 \leq \theta \leq 12.57$$

med konfgrad 95%. (Ej sannolikhet!)