Konfidensintervall

Låt $X_1, \ldots, X_n$ vara oberoende och normalfördelade,

$$X_i \text{ är } \mathrm{N}(m, \sigma).$$

Skattningsvariabeln av m är $\overline{X}$ och vi vet att

$$\overline{X} \text{ är } \mathrm{N}(m, \sigma/\sqrt{n}) \qquad \text{alt.} \qquad \frac{\overline{X} - m}{\sigma/\sqrt{n}} \text{ är } \mathrm{N}(0,1)$$

Vi kan ur normalfördelningstabeller för varje sannolikhet α bestämma kvantilen $\lambda_{\alpha/2}$ så att

$$\mathsf{P}\left(\left|\frac{\overline{X} - m}{\sigma/\sqrt{n}}\right| \leq \lambda_{\alpha/2}\right) = 1 - \alpha.$$

Alltså, med sannolikhet $1 - \alpha$ är

$$-\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X} - m \leq \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

eller

$$\overline{X} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \leq \quad m \quad \leq \quad \overline{X} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Detta är ett [symmetriskt] konfidensintervall för väntevärdet, med konfidensgrad $1 - \alpha$.

OBS! Konfidensintervallet rör väntevärdet m och *inte* de enskilda observationerna.

Definition: Ett konfidensintervall för θ är ett stokastiskt intervall $I(X_1, \ldots, X_n)$ som täcker θ med en given sannolikhet:

$$\mathsf{P}\left(\theta \in I(X_1, \ldots, X_n)\right) = 1 - \alpha = \text{konfidensgraden}.$$

I vårt exempel har konfidensintervallet dynamiken:

- Större/mindre σ ger längre/kortare intervall.
- Fler/färre observationer n ger kortare/längre intervall.
- Större/mindre konfidensgrad $1 - \alpha$ ger längre/kortare intervall.

Generellt: Om θ^* är $\mathrm{N}(\theta, D)$ så med sannolikhet $1 - \alpha$ är

$$-\lambda_{\alpha/2} \leq \frac{\theta^* - \theta}{D} \leq \lambda_{\alpha/2}$$

eller

$$\theta \in \theta^* \pm \lambda_{\alpha/2} D$$

är ett (symmetriskt) konfidensintervall för θ med konfidensgrad $1 - \alpha$. ”Punktskattning plus/minus normal-kvantil$*$standardavvikelse för skattningen.”

För att räkna ut intervallet måste vi känna D, standardavvikelsen. Om vi inte måste den skattas.

Generellt: Om θ^* är approximativt $\mathrm{N}(\theta, D)$ så är

$$\theta \in \theta^* \pm \lambda_{\alpha/2} D^*$$

är ett (symmetriskt) konfidensintervall för θ med konfidensgrad approximativt $1-\alpha$. Approximativa intervall: "Punktskattning plus/minus normal-kvantil$*$medelfel för skattningen."

Exempel: Opinionsundersökning. Låt X räkna antalet (s)-sympatisörer vid $n = 2854$ intervjuer. Modell: X är $\mathrm{Bin}(n, p)$ där p är andelen sympatisörer i populationen. Skattningen $p^*(x) = x/n$ modelleras med skattningsvariabeln $p^*(X) = X/n$ där $\mathsf{E}\left[p^*(X)\right] = p$ och

$$\mathrm{D}\left(p^*(X)\right) = \sqrt{\frac{1}{n^2}\mathrm{V}\left(X\right)} = \sqrt{\frac{p(1-p)}{n}}.$$

Med utfallet $x = 1085$ fås skattningen $p^* = x/n = 1085/2854 = 0.380$ och medelfelet

$$D^* = \sqrt{\frac{p^*(1-p^*)}{n}} = 0.00909.$$

Om X är approximativt normalfördelad är

$$p^* \text{ approximativt } \mathrm{N}\left(p, D(p^*)\right)$$

och

$$p \in p^* \pm \lambda_{\alpha/2} D^*(p^*) = p^* \pm \lambda_{\alpha/2}\sqrt{\frac{p^*(1-p^*)}{n}}$$

ett konfindensintervall för p med konfidensgrad approximativt $1-\alpha$. För $1-\alpha = 0.95$ fås $\lambda_{\alpha/2} = 1.96$ och

$$p \in 0.380 \pm 0.0178, \qquad 0.362 \le p \le 0.398 \qquad (\approx 95\%)$$

Normalapproximationen motiveras av att skattningen av

$$\mathrm{V}\left(X\right) = np(1-p)$$

är $n\,p^*(1-p^*) = 673 > 10$.

Intervallet innehåller de av data föreslagna rimliga värdena på p. Ett påstående (en utsaga) om att p är 0.40 skulle vi inte tro på.

Vi skall nu betrakta specialfallet med konfidensintervall för väntevärdet i normalfördelningen.

Minns:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Jämför:

$$\frac{\overline{X} - m}{\sigma/\sqrt{n}} \qquad \text{och} \qquad \frac{\overline{X} - m}{S/\sqrt{n}}$$

Storheten till höger har en fördelning som inte beror på m eller σ utan bara på antalet observationer n. Vi kan ställa upp tabeller över dess fördelning precis på samma sätt som vi har gjort för normalfördelningen. Vi säger att

$$\frac{\overline{X} - m}{S/\sqrt{n}} \text{ är } t(n-1),$$

utläst: "t-fördelad med $n-1$ frihetsgrader". Utnyttjar vi tabellerna för t-fördelningen kan vi för varje sannolikhet α bestämma kvantiler $t_{\alpha/2}$ så att

$$\mathsf{P}\left(\left|\frac{\overline{X} - m}{S/\sqrt{n}}\right| \le t_{\alpha/2}\right) = 1 - \alpha.$$

Då är med sannolikhet $1 - \alpha$

$$-t_{\alpha/2} \le \frac{\overline{X} - m}{S/\sqrt{n}} \le t_{\alpha/2}$$

eller

$$-t_{\alpha/2}\frac{S}{\sqrt{n}} \le \overline{X} - m \le t_{\alpha/2}\frac{S}{\sqrt{n}}$$

eller

$$\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \le m \le \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}.$$

Detta är ett [symmetriskt] konfidensintervall för m med konfidensgrad $1-\alpha$.

Exempel: Från följande observationer på $n = 10$ normalfördelade stokastiska variabler

$$x_i : 5 \;\; 4.1 \;\; 5.4 \;\; 5.5 \;\; 4.5 \;\; 6.1 \;\; 6.1 \;\; 5.3 \;\; 5.5 \;\; 5.4$$

fick vi punktskattningar

$$\overline{x} = 5.29, \qquad s^2 = 0.39433 \qquad \text{och} \qquad s = 0.62796.$$

Med konfidensgrad $1-\alpha = 0.95$ får man $t_{\alpha/2} = t_{0.025} = 2.2622$ ur tabellen för $t(n-1) = t(9)$-fördelningen. Det observerade intervallet blir

$$4.84 = \overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le m \le \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} = 5.74 \qquad m \in \overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 5.29 \pm 0.45 \qquad (95\%).$$

"Punktskattning plus/minus t-kvantil$*$medelfel för skattningen."

Konfidensintervall för σ^2

Fördelningen för

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 = \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2$$

är fördelningen för en summa av kvadrater av oberoende $\mathrm{N}(0,1)$-variabler, och beror inte på m eller σ. Fördelningen kallas χ^2-fördelningen och parametern, antalet termer i summan, kallas dess frihetsgrader. Fördelningen finns tabulerad i formelsamlingen.

Man kan visa att

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2.$$

är $\chi^2(n-1)$-fördelad. Vi kan således för varje sannolikhet α bestämma kvantilerna $\chi^2_{1-\alpha/2}$ och $\chi^2_{\alpha/2}$ så att

$$\mathsf{P}\left(\chi^2_{1-\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2}\right) = 1-\alpha.$$

Så med sannolikhet $1-\alpha$ är

$$\chi^2_{1-\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2}$$

eller

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}.$$

Detta är ett [symmetriskt] konfidensintervall för σ^2 med konfidensgrad $1-\alpha$. Vi kan få ett intervall för σ genom

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}} \le \sigma \le \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}.$$

Exempel: Med $n = 10$ observationer fick vi punktskattningen $s^2 = 0.39433$ av σ^2. Ur $\chi^2(9)$-tabeller får vi att med $1-\alpha = 0.95$ är tabellvärdena

$$\chi^2_{1-\alpha/2} = \chi^2_{0.975} = 2.7004 \qquad \chi^2_{\alpha/2} = \chi^2_{0.025} = 19.023.$$

Alltså, ett observerat intervall för σ^2 är

$$0.18657 = \frac{(10-1)0.39433}{19.023} = \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} = \frac{(10-1)0.39433}{2.7004} = 1.3143 \qquad (95\%).$$

Motsvarande intervall för standardavvikelsen blir

$$0.43 = \sqrt{0.18657} \leq \sigma \leq \sqrt{1.3143} = 1.16 \qquad (95\%).$$