Sannolikhet och statistik XIII

Johan Jonasson

May 2019

Konfidensintervall för μ i normalfördelning. Vi har $X_1, \dots, X_n$ sp på $N(\mu, \sigma^2)$. Börja med att anta att σ^2 är känd.

Låt oss basera oss på $\overline{X}$, (man kan visa att detta är den mest effektiva av alla skattningar av μ). Det gäller att $\overline{X} \sim N(\mu, \sigma^2/n)$, dvs

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Skriv $z_\alpha = \Phi^{-1}(1-\alpha)$. Då gäller

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Lös ut μ och få det symmetriska konfintervallet

$$\mu \in \overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad (1-\alpha).$$

Uppåt begr konfintervall: Utnyttja att

$$\mathbb{P}\left(-z_\alpha \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right) = 1-\alpha.$$

Lös ut μ och få

$$\mu \leq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}} \qquad (1-\alpha)$$

Om σ^2 är okänd, ersätt σ^2 med s^2. Det gäller att

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Härma nu rakt av:

$$\mu \in \overline{X} \pm F^{-1}_{t_{n-1}}(1-\alpha/2)\frac{s}{\sqrt{n}}$$

med konfgrad $1-\alpha$.

Exempel: I ett eluttag mäts spänningen en gång om dagen i en vecka. Resultat:

$$230.7, 226.9, 228.8, 232.2, 227.3, 227.0, 229.1$$

Rimligt att anta att data är $N(\mu, \sigma^2)$-fördelade. Vi har $\overline{X} = 228.9$ och $s^2 = 2.014^2$, så $s/\sqrt{n} = 2.014/\sqrt{7} = 0.761$. För t_6-förd gäller t.ex. $F_{t_6}^{-1}(0.975) = 2.46$, $F_{t_6}^{-1}(0.995) = 3.71$, $F_{t_6}^{-1}(0.9) = 1.44$. Alltså

$$\mu = 228.9 \pm 2.46 \cdot 0.761 = 228.9 \pm 1.9 \ (95\%).$$
$$\mu = 228.9 \pm 3.71 \cdot 0.761 = 228.9 \pm 2.8 \ (99\%).$$
$$\mu \leq 228.9 + 1.44 \cdot 0.761 = 230.0 \ (90\%).$$

Ok, om vi nu skulle veta att $\sigma^2 = 2^2$ och vill ha ett 95% symmetriskt konfintervall utnyttjar vi att $z_{0.975} = 1.96$ och får

$$\mu = 228.9 \pm 1.96 \cdot 2/\sqrt{7} = 228.9 \pm 1.5 \ (95\%).$$

$\square$

Det gäller att $F_{t_n}^{-1}(a) \to \Phi^{-1}(a)$ då $n \to \infty$. Tumregel: $n \geq 100 \to t_n \approx N(0,1)$. (Vi har t.ex. att $F_{t_{100}}^{-1}(0.975) = 1.98$ och $\Phi^{-1}(0.975) = 1.96$.)

Konfidensintervall för σ^2: vi hoppar över. Man baserar sig på att

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Prediktion. (σ^2 känd). Vad kommer spänningen i uttaget att vara nästa dag? Vi har $X_1, \ldots, X_n$ sp på $N(\mu, \sigma^2)$ och undrar vad vi kan säga om en ny observation Y. Vi har

$$Y - \overline{X} \sim N\left(0, \sigma^2 + \frac{\sigma^2}{n}\right)$$

(ty $\mathbb{V}\mathrm{ar}(Y - \overline{X}) = \mathbb{V}\mathrm{ar}(Y) + \mathbb{V}\mathrm{ar}(\overline{X}) = \sigma^2 + \sigma^2/n$).
Alltså

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{Y - \overline{X}}{\sigma\sqrt{1 + 1/n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Vi får ett *prediktionsintervall*:

$$Y = \overline{X} \pm z_{\alpha/2}\sigma\sqrt{1 + \frac{1}{n}} \qquad (1 - \alpha).$$

I vårt ex med $\sigma^2 = 2^2$ och $n = 7$,

$$Y = 228.9 \pm 1.96 \cdot 2 \cdot \sqrt{\frac{8}{7}} = 228.9 \pm 4.2 \ (95\%).$$

Reflektioner:

- Normalfördelningsantagandet alltid lite fel, ffa i svansarna.
- Ju större n desto mer normal blir $\overline{X}$ och därmed kan vi klara mer extrema konfidensgrader.
- Om beroende mellan observationer kan $\overline{X}$ fortfarande vara normal, men konfidensgraden kan bli rejält fel, eftersom variansen hos $\overline{X}$ blir en helt annan.
- Det finns test (goodness-of-fit) och normalfördelningsplot för att kolla normalfördelning.

Konfidensintervall för p**:** I en opinionsundersökning med 10000 deltagare säger sig 8.4 procent stödja parti A. Ge ett konfidensintervall för stödet för A i hela populationen.

Låt p vara andelen väljare som stöder A. Låt $X_1, \ldots, X_n$ vara indikatorerna om resp väljare stöder A. $\sum_k X_k \sim Bin(n,p)$. Enl CGS är

$$\overline{X} \approx N\left(p, \frac{p(1-p)}{n}\right).$$

dvs

$$\frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

så om $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, gäller

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.$$

Man kan lösa ut p ut olikheten och bilda ett konfintervall för p. Lite bökigt. Man kan utnyttja att $\overline{X}/p$ är appr 1 med stor sannolikhet och få att

$$\frac{\overline{X} - p}{\sqrt{\overline{X}(1-\overline{X})/n}} \approx N(0,1).$$

Detta är De-Moivre-Laplace gränsvärdessats.

Här är lätt att lösa ut p:

$$p = \overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \qquad (1-\alpha).$$

I vårt ex hade vi $\overline{X} = 0.084$. För 95% obs vi att $z_{0.025} = 1.96$. Vi får

$$p = 0.084 \pm 1.96\sqrt{\frac{0.084 \cdot 0.916}{10000}} = 0.084 \pm 0.0054$$

dvs $8.4\% \pm 0.54\%$.

Reflektioner:

- Inga problem i svansarna!
- Mycket bra skattning av p bland
  - dem som svarar
  - och talar sanning.

Svarsbortfall, lögner och felsvar är stora problem i opininonsundersökningar.