Föreläsning 9-10. Punktskattningar och konfidensintervaller

Väntevärdesriktig punktskattning av μ

Vi antar att vektor $(X_1, X_2, \ldots, X_n)$ består av n oberoende slumpvariabler med samma fördelning som har okänt populations parameter μ och samma populations standardavvikelse σ. Ett stickprov $(x_1, x_2, \ldots, x_n)$ är en given realisering (bland flera möjliga) av slumpvektorn $(X_1, X_2, \ldots, X_n)$.

En naturlig punktskattning av μ är stickprovs medelvärde

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n},$$

som är en realisering (bland flera möjliga) av slumpvariabeln

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

Punktskattning $\bar{x}$ är väntevärdesriktig punktskattning av μ eftersom

$$\mathrm{E}(\bar{X}) = \mathrm{E}(\tfrac{X_1+\ldots+X_n}{n}) = \tfrac{\mathrm{E}(X_1)+\ldots+\mathrm{E}(X_n)}{n} = \tfrac{\mu+\ldots+\mu}{n} = \mu.$$

Väntevärdesriktig betyder att det finns ingen systematiskt fel (unbiased estimate).
Dock som regel finns det ett slumpfel $\bar{x} \neq \mu$ som orsakas av att n stickprovs värden väljs på måfå. Olika stickprovs utfall ger olika värde för $\bar{x}$. Medelfel

$$\sigma_{\bar{X}} = \sqrt{\mathrm{Var}(\bar{X})}$$

mäter slumpfels storlek: ju större $\sigma_{\bar{X}}$ desto större avvikelser $\bar{x} - \mu$ förväntas. Enligt oberoende antagandet

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\tfrac{X_1+\ldots+X_n}{n}) = \tfrac{\mathrm{Var}(X_1)+\ldots+\mathrm{Var}(X_n)}{n^2} = \tfrac{\sigma^2+\ldots+\sigma^2}{n^2} = \tfrac{\sigma^2}{n}.$$

Det innebär att

$$\sigma_{\bar{X}} = \tfrac{\sigma}{\sqrt{n}}$$

ju större är stickprovs storlek n desto mindre medelfel.

Populations andel p och stickprovs andel $\hat{p}$

Population består av N individer. Det finns två slags individer

$$N = N_1 + N_0, \quad p = \tfrac{N_1}{N}, \quad 1 - p = \tfrac{N_0}{N}$$

och vi är intresserade i populations andel p. Parameter p är okänd och i praktiken omöjligt att räkna exakt.

Vi tar ett stickprov av storlek n och för $i = 1, \ldots, n$ skriver $x_i = 1$ om person nummer i tillhör grupp 1 och $x_i = 0$ om personen tillhör grupp 0. Vi tillämpar två olika modeller för att beskriva slumpexperiment för stickprovs samling:

Modell 1: dragning med återläggning (oberoende slumpvariabler $X_1, X_2, \ldots, X_n$),
Modell 2: dragning utan återläggning (beroende slumpvariabler $X_1, X_2, \ldots, X_n$).

Modell 1 tolkas som oändlig population och Modell 2 som ändlig population.

Frågor

1. vilken fördelning har slumvariabel X_i?
2. vilken fördelning har slumvariabel $Y = X_1 + \ldots + X_n$?
3. förklara varför modell 1 kan tolkas som oändlig population?
4. vad är väntevärde och varians av Y?

Vi skattar p med hjälp av stickprovs andel

$$\hat{p} = \bar{x}.$$

Stickprovs andel $\hat{p}$ är en väntevärdesriktig punktskattning av populations andel p eftersom $X_i \sim \text{Ber}(p)$ och därför

$$\mu = p, \qquad \sigma^2 = p(1-p).$$

Medelfelet är

$$\sigma_{\hat{P}} = \sqrt{\tfrac{p(1-p)}{n}}$$

om populationen är oändlig, och

$$\sigma_{\hat{P}} = \sqrt{\tfrac{p(1-p)}{n}\tfrac{N-n}{N-1}}$$

om populationen är ändlig.

Skattad medelfel

Vi antar igen att vektor $(X_1, X_2, \ldots, X_n)$ består av n oberoende slumpvariabler med samma fördelning som har okänt populations parameter μ och samma populations standardavvikelse σ. Om σ är okänd då medelfelet

$$\sigma_{\bar{X}} = \tfrac{\sigma}{\sqrt{n}}$$

kan inte räknas fram. För ett givet stickprov $(x_1, \ldots, x_n)$ beräknar man (skattad) medelfel enligt formel

$$s_{\bar{x}} = \tfrac{s}{\sqrt{n}}, \quad s = \sqrt{\tfrac{(x_1-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}},$$

där s kallas stickprovs standardavvikelse. Man delar med $n-1$ istället för n eftersom

$$\begin{aligned}\text{E}(S^2) &= \text{E}(\tfrac{(X_1-\bar{X})^2+\ldots+(X_n-\bar{X})^2}{n-1}) = \tfrac{n\text{E}(X_1-\bar{X})^2}{n-1} = \tfrac{n\text{Var}(X_1-\bar{X})}{n-1} \\ &= \tfrac{\text{Var}((n-1)X_1-X_2-X_3-\ldots-X_n)}{n(n-1)} = \tfrac{(n-1)^2\sigma^2+\sigma^2+\sigma^2+\ldots+\sigma^2}{n(n-1)} = \sigma^2.\end{aligned}$$

Skattad medelfel för stickprovs andel är

$$s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}}$$

om populationen är oändlig, och

$$s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}\tfrac{N-n}{N}}$$

om populationen är ändlig.

Approximativa konfidensintervaller för p

För dragning med återläggning, om stickprovsstorlek n är stor, då

$$\sigma_{\hat{P}} \approx s_{\hat{p}} \approx \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}},$$
$$\hat{P} \approx \mathrm{N}(p, \sigma_{\hat{P}}) \approx \mathrm{N}(p, \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}}).$$

Det ger formeln

$$I_p = \hat{p} \pm z_{\alpha/2}\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}}$$

för en approximativt $100(1-\alpha)\%$ konfidensinterval för p i oändlig population.

För dragning utan återläggning då stickprovsstorlek n är stor, vi får

$$I_p = \hat{p} \pm z_{\alpha/2}\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}\,\tfrac{N-n}{N}}$$

som en approximativt $100(1-\alpha)\%$ konfidensinterval för p i ändlig population.

Exakta konfidensintervaller för μ

Vi antar igen att vektor $(X_1, X_2, \ldots, X_n)$ består av n oberoende slumpvariabler med samma normalfördelning $\mathrm{N}(\mu, \sigma)$. Om σ är känd då

$$\bar{X} \sim \mathrm{N}(\mu, \tfrac{\sigma}{\sqrt{n}}).$$

Det ger formel

$$I_\mu = \bar{x} \pm z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}$$

för ett exakt konfidensintervall för μ med känd σ.

Om σ är okänd då

$$T = \tfrac{\bar{X}-\mu}{s_{\bar{x}}} \sim \mathrm{t}_{n-1}$$

har så kallade t-fördelning med $n-1$ frihetsgrader. Det innebär

$$\mathrm{P}\Big(-t_{\alpha/2}(n-1) < \tfrac{\bar{X}-\mu}{s_{\bar{x}}} < t_{\alpha/2}(n-1)\Big) = 1-\alpha,$$

som ger formeln

$$I_\mu = \bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

för ett exakt konfidensintervall för μ med okänd σ.