

F8, Punktskattningar och intervallskattningar

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Statistisk slutledning (statistisk inferens)

Totalundersökning

Urvalsundersökning; stickprov

OSU (Obundet slumpmässigt urval)

Stratifierat urval

Klusterurval

Snöbollsurval

Systematiska fel vid urvalsundersökning:

- Ramproblem
- Svarsbortfall
- Mätfel
- Oberoende
- Bearbetningsfel

Slumpfel vid urvalsundersökning pga att vi drar stickprov

Punktskattning

θ : Okänd populationsparameter.

Exempel: X och Y är två stokastiska variabler.

$$X \sim N(\mu; 1), \theta = \mu$$

$$X \sim N(0; \sigma^2), \theta = \sigma^2$$

$$(X - Y) \sim N(\mu_1 - \mu_2; 1 + 1), \theta = \mu_1 - \mu_2$$

$$X \sim \text{Bin}(10; p), \theta = p$$

$$\rho(X, Y), \theta = \rho$$

Skatta θ med en funktion av stickprovet (estimator), $\hat{\theta}$.

	Populationsparameter (θ)	Skattnings- funktion Estimator ($\hat{\theta}$)
Medelvärde	μ	$\bar{X}$
Varians	σ^2	S^2
Proportion	p	$\hat{P}$
Skillnad mellan medelvärden	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$
Skillnad mellan proportioner	$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2$

	Estimat (observerat värde)
Medelvärde	$\bar{x}$
Varians	s^2
Proportion	$\hat{p}$
Skillnad mellan medelvärden	$\bar{x}_1 - \bar{x}_2$
Skillnad mellan proportioner	$\hat{p}_1 - \hat{p}_2$

Hur skall man skatta (estimera) de okända parametrarna i populationen, t ex μ och σ^2 ? Eftersom μ och σ^2 är populationsmedelvärdet och populationsvariansen, kan det tyckas naturligt att använda stickprovsmedelvärdet $\bar{X}$ och stickprovsvariansen S^2 som skattningar. Men det är inte helt självklart. Det finns alternativa skattningar, t ex medianen $\tilde{X}$, minsta värdet $X_{\min}$ eller största värdet $X_{\max}$ istället för $\bar{X}$.

Egenskaper hos estimatorer

Väntevärdesriktighet

För estimatorer (eller skattningar) gäller alltid att

$$E(\hat{\theta}) = \theta + b$$

där b är skattningsfunktionens systematiska fel (bias). För väntevärdesriktiga skattningar är $b = 0$, dvs en estimator (eller skattning) är väntevärdesriktig om väntevärdet för estimatoren är lika med populationssparametern, $E(\hat{\theta}) = \theta$.

Exempel:

$$E(\bar{X}) = \mu$$

$$E(\hat{P}) = p$$

$$E(S^2) = \sigma^2$$

där

$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

är stickprovsvariansen, och där nämnaren $n-1$ brukar kallas antalet frihetsgrader. I Körner sid 142 visas att $E(S^2) = \sigma^2$.

Konsistens

När stickprovsstorleken ökar minskar osäkerheten om den okända populationsparametern θ , dvs variansen minskar i samplingfördelning. T ex då $\theta = \mu$ och $\hat{\theta} = \bar{X}$ blir $V(\bar{X}) = \frac{\sigma^2}{n}$.

Effektiv

Den väntevärdesriktiga estimator som har den minsta variansen är effektivast. T ex är $\bar{X}$ en effektivare skattning än $\tilde{X}$, $X_{\min}$ och $X_{\max}$.

Exempel:

En slumpvariabel X har väntevärdet μ och variansen σ^2 . I syfte att skatta μ görs tre oberoende observationer på variabeln, varefter följande skattningar bildas:

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3 \\ \hat{\mu}_2 &= \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3 \\ \hat{\mu}_3 &= \frac{1}{2}X_1 + \frac{1}{2}X_2\end{aligned}$$

a) Vilka av skattningarna är VVR?

$$\begin{aligned} E(\hat{\mu}_1) &= \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(X_3) \\ &= \frac{1}{3}\mu + \frac{1}{3}\mu + \frac{1}{3}\mu = \mu \\ E(\hat{\mu}_2) &= \frac{1}{4}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{4}E(X_3) \\ &= \frac{1}{4}\mu + \frac{1}{2}\mu + \frac{1}{4}\mu = \mu \\ E(\hat{\mu}_3) &= \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) \\ &= \frac{1}{2}\mu + \frac{1}{2}\mu = \mu \end{aligned}$$

b) Vilken av skattningarna har största precisionen (minsta variansen)?

$$V(\hat{\mu}_1) = \frac{1}{9}V(X_1) + \frac{1}{9}V(X_2) + \frac{1}{9}V(X_3)$$

$$= \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 = \frac{\sigma^2}{3}$$

$$V(\hat{\mu}_2) = \frac{1}{16}V(X_1) + \frac{1}{4}V(X_2) + \frac{1}{16}V(X_3)$$

$$= \frac{1}{16}\sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{16}\sigma^2 = \frac{3\sigma^2}{8}$$

$$V(\hat{\mu}_3) = \frac{1}{4}V(X_1) + \frac{1}{4}V(X_2)$$

$$= \frac{1}{4}\sigma^2 + \frac{1}{4}\sigma^2 = \frac{\sigma^2}{2}$$

c) Vilken av skattningarna tror du är "bäst"?

Tillräcklig

En estimator är tillräcklig om den utnyttjar all information i stickprovet om den okända populationssparametern.

Medelfel

Vi vet att

$$\begin{aligned}V(\bar{X}) &= \frac{\sigma^2}{n} \text{ och} \\V(\hat{P}) &= \frac{p(1-p)}{n},\end{aligned}$$

men σ^2 och p oftast okända. Använd skattningarna S^2 och $\hat{P}$ istället. Då blir den skattade variansen för punktskattningarna

$$\begin{aligned}\hat{V}(\bar{X}) &= S_{\bar{X}}^2 = \frac{S^2}{n} \text{ och} \\ \hat{V}(\hat{P}) &= S_{\hat{P}}^2 = \frac{\hat{P}(1-\hat{P})}{n},\end{aligned}$$

och den skattade standardavvikelsen för skattningarna

$$\begin{aligned}S_{\bar{X}} &= \frac{S}{\text{sqrt}(n)} \text{ och} \\ S_{\hat{P}} &= \text{sqrt}\left(\frac{\hat{P}(1-\hat{P})}{n}\right),\end{aligned}$$

som brukar kallas skattningens *medelfel*.

Intervallskattning

Hur bra är punktskattningen? Vi vill ha en precision av punktskattningen.

Låt $X_1, X_2, \dots, X_n$ vara oberoende $N(\mu; \sigma^2)$ -fördelade variabler. Då har vi att $\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$ och följande sannolikhet gäller

$$\begin{aligned} & \Pr\left(-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) \\ = & 0.95 \\ & \Pr\left(-1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \frac{\sigma}{\sqrt{n}}\right) \\ = & 0.95 \\ & \Pr\left(\mu - 1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96 \frac{\sigma}{\sqrt{n}}\right) \\ = & 0.95 \end{aligned}$$

Omforma vänstra delen:

$$\begin{aligned} -1.96 \frac{\sigma}{\sqrt{n}} &< \bar{X} - \mu \\ \mu &< \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Omforma högra delen:

$$\begin{aligned} \bar{X} - \mu &< 1.96 \frac{\sigma}{\sqrt{n}} \\ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} &< \mu \end{aligned}$$

Alltså

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Sannolikheten kan nu skrivas som

$$\begin{aligned} &\Pr \left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \\ &= 0.95 \end{aligned}$$

Ett 95 %-igt konfidensintervall erhålls genom att byta ut *slumpvariabeln* $\bar{X}$ mot en *observation* $\bar{x}$, dvs

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Om man beräknar ett stort antal sådana intervall kan man vänta sig att 95% av dem täcker μ .

Tolkning av ett uträknat intervall:

Med 95 % konfidens (tillförlitlighet) ligger den okända parametern μ i intervallet $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$.