F11, Konfidensintervall

Christian Tallberg

Avdelningen för Nationalekonomi/Statistik

Karlstads universitet

Punktskattning av en populationsparameter

Vi tänker oss att vi vill skatta en viss populationsparameter θ (som kan vara t ex ett populationsmedelvärde, en populationsvarians, eller vad som helst) som antar värden i ett parameterrum Ω_Θ. Säg att $X_1, X_2, ..., X_n$ är ett stickprov av n oberoende observationer från populationen. På grundval av dessa observerade stickprovsvärden beräknar vi en *punktskattning*, $\theta^* = \theta^*(X_1, X_2, ...X_n)$, av den okända populationsparametern θ.

Eftersom slumpen bestämmer vilka våra observationer kommer att bli, så är även *skattningen* θ^* en slumpvariabel. Vår förhoppning är naturligtvis att vi skall få ett värde på θ^* som ligger så nära det sanna (okända) värdet på θ som möjligt. Vi kan dock aldrig komma ifrån att skattningen medför en viss *osäkerhet*. En sannolikhetsfördelning (*samplingfördelningen*) beskriver denna osäkerhet, dvs hur värdet på θ^* kan variera från stickprov till stickprov. Samplingfördelningen kan (precis som vilken sannolikhetsfördelning som helst) beskrivas med hjälp av väntevärde, $E(\theta^*)$, och varians, $V(\theta^*)$.

Egenskaper hos skattningen som brukar betraktas som önskvärda (eller åtminstone intressanta) är:

- att den är *väntevärdesriktig*, dvs att $E(\theta^*) = \theta$ för varje $\theta \in \Omega_\Theta$.

- att den är *konsistent*. Om för varje fixt $\theta \in \Omega_\Theta$ och för varje givet $\varepsilon > 0$

  $$P(|\theta_n^* - \theta| > \epsilon) \to 0$$

  då stickprovsstorleken $n \to \infty$, sägs θ^* vara konsistent. Det innebär att då stickprovsstorleken ökar, minskar variansen för skattningen. Dvs, vår osäkerhet om θ minskar.

Vi tänker oss hypotetiskt en lång serie upprepade stickprov från samma population. Då kan vi tolka egenskaperna *väntevärdesriktig* och *liten varians* på följande sätt:

- En väntevärdesriktig skattning kommer i det långa loppet att i genomsnitt träffa rätt. (OBS, inte varje gång, men *i genomsnitt*).

- Variansen är ett mått på skattningens osäkerhet. Ju mindre varians en skattning har, desto oftare kommer den att träffa i närheten av det sanna parametervärdet.

En skattning som inte är väntevärdesriktig har en *bias* (systematiskt fel):

$$Bias(\theta^*) = E(\theta^*) - \theta$$

Om θ^* är väntevärdesriktig, så är $Bias(\theta^*) = 0$.

(Mer eller mindre kuriosa). Av bekvämlighetsskäl använder vi termen "skattning" för två begrepp, nämligen *estimator* och *estimat*. Egentligen är:

- Estimator = skattningen (eller skattningsfunktionen) betraktad som slumpvariabel, alltså innan vi observerat några data.

- Estimat = det observerade värde som skattningen antar efter att data erhållits.

Exempel: Säg att $X_1, X_2, ..., X_n$ är ett slumpmässigt stickprov av n oberoende observationer från en population. Låt θ vara en okänd populationsparameter, och θ^* en väntevärdesriktig skattning av θ.

- Om $X_i \sim N(\mu; 1), i = 1, 2, ..., n$ då är

$$\begin{aligned} \theta &= \mu \\ \theta^* &= \bar{X} \\ E\left(\bar{X}\right) &= \mu \\ V\left(\bar{X}\right) &= \frac{\sigma^2}{n}. \end{aligned}$$

- Om $X_i \sim N\left(0; \sigma^2\right), i = 1, 2, ..., n$ då är

$$\begin{aligned} \theta &= \sigma^2 \\ \theta^* &= S^2 \\ E\left(S^2\right) &= \sigma^2. \end{aligned}$$

Intervallskattning för ett populationsmedelvärde

I praktiken har vi bara *ett* stickprov (inte hela *samplingfördelningen*). Vad kan vi säga om osäkerheten i den punktskattning vi erhållit från stickprovet?

Låt $X_1, X_2, ..., X_n$ vara ett stickprov av oberoende normalfördelade variabler med okänt väntevärde μ och känd varians σ^2. Vi vill skatta μ med $\bar{X}$ och uttala oss om osäkerheten i $\bar{X}$. Vi vet att $\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$. Därför gäller följande sannolikhet:

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{sqrt(n)}} < 1.96\right) = 0.95$$

$$P\left(-1.96\frac{\sigma}{sqrt\,(n)} < \bar{X} - \mu < 1.96\frac{\sigma}{sqrt\,(n)}\right)$$
$$= 0.95$$
$$P\left(\mu - 1.96\frac{\sigma}{sqrt\,(n)} < \bar{X} < \mu + 1.96\frac{\sigma}{sqrt\,(n)}\right)$$
$$= 0.95$$

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

$$P\left(\bar{X} - 1.96\frac{\sigma}{sqrt\,(n)} < \mu < \bar{X} + 1.96\frac{\sigma}{sqrt\,(n)}\right)$$
$$= 0.95,$$

där *ändpunkterna* alltså är *slumpmässiga*. Kan variera från stickprov till stickprov. Men med sannolikheten 0.95 kommer de att ligga på varsin sida av μ, vilket innebär att intervallet mellan dessa två ändpunkter med sannolikheten 0.95 kommer att fånga upp det sanna, okända värdet på μ.

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

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

eller

$$\bar{x} \pm 1.96\frac{\sigma}{sqrt\,(n)}.$$

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

Tolkning av ett uträknat intervall:

- *Innan vi dragit stickprovet*: Med sannolikheten 0.95 kommer vi att få ett intervall som täcker μ.

- *Efter vi dragit stickprovet* och beräknat ett intervall: Antingen täcker intervallet μ (med sannolikheten 1), eller så täcker intervallet *inte* μ (med sannolikheten 1). Poängen är att vi vet inte hur det är. Det vi vet är att intervallet har beräknats enligt en metod som i det långa loppet producerar intervall som i 95% av fallen innehåller μ. Dvs, med 95 % *konfidens* (tillförlitlighet) ligger den okända parametern μ i intervallet
$$\bar{x} \pm 1.96\frac{\sigma}{sqrt\,(n)}.$$

Normalfördelade populationsvariabler med känd standardavviklelse σ

Exempel:

Låt $X_1, ..., X_n$ vara ett stickprov ($n = 16$) av oberoende normalfördelade variabler med okänt medelvärde μ och känd standardavvikelse $\sigma = 12$. Punktskattningen (det observerade estimatet) är $\bar{x} = 62$. Ett 95 %-igt k.i. för μ ges av

$$\begin{aligned} \bar{x} \pm 1.96\frac{\sigma}{sqrt\,(n)} &= 62 \pm 1.96\frac{12}{4} \\ &= 62 \pm 5.88. \end{aligned}$$

Tolkning: Med konfidensgraden 95 % täcker intervallet $(56.12; 67.88)$ populationsparametern μ.

Några vanliga konfidensgrader (ges i tabellsamlingen):

Konfidensgrad	$\lambda_{\alpha/2}$-värde
90 %	1.64
95 %	1.96
99 %	2.58
99.9 %	3.29

där $\lambda_{\alpha/2}$ är ett värde (sannolikhetstal) från $N\,(0; 1)$ som bestämmer konfidensgraden och α är signifikansnivån (risknivån). Notera att $\alpha = 1-$ konfidensgraden.

När ett k.i. med konf.grad $(1 - \alpha)\,100$ % skrivs som
$$\bar{x} \pm \lambda_{\alpha/2}\frac{\sigma}{sqrt\,(n)},$$

betyder det att intervallets längd är
$$2 \cdot \lambda_{\alpha/2}\frac{\sigma}{sqrt\,(n)}.$$

Intervallets längd bestäms av

- standardavvikelsen σ. Större $\sigma \rightarrow$ *längre* intervall (givet n och konfidensgrad).
- konfidensgraden (som bestäms genom värdet på $\lambda_{\alpha/2}$). Högre konfidensgrad $\rightarrow$ *längre* intervall (givet n och σ).
- stickprovets storlek n. Större stickprov $\rightarrow$ *kortare* intervall (givet σ och konfidensgrad).

Forts. exemplet: Ett 90 % igt k.i. för μ ges av

$$\begin{aligned}\bar{x} \pm 1.64 \frac{\sigma}{sqrt\,(n)} &= 62 \pm 1.64 \frac{12}{4} \\ &= 62 \pm 4.92 \\ & \quad (57.08; 66.92)\end{aligned}$$

Ett 99 % igt k.i. för μ ges av

$$\begin{aligned}\bar{x} \pm 2.58 \frac{\sigma}{sqrt\,(n)} &= 62 \pm 2.58 \frac{12}{4} \\ &= 62 \pm 7.74 \\ & \quad (54.26; 69.74)\end{aligned}$$

Normalfördelade populationsvariabler med okänd standardavvikelse

Sedan tidigare vet vi att om $X_1, ..., X_n$ är ett stickprov av oberoende normalfördelade variabler med okänt medelvärde μ och känd standardavvikelse σ så gäller att

$$\frac{\bar{X} - \mu}{\frac{\sigma}{sqrt(n)}} \in N\,(0; 1)\,.$$

Om vi har samma förutsättningar men σ är okänd och skattas i stickprovet med s, då gäller att

$$\frac{\bar{X} - \mu}{\frac{s}{sqrt(n)}} \in t\,(n-1)\,,$$

dvs den standardiserade stokastiska variabeln är t-fördelad med $(n-1)$ frihetsgrader. Sannolikhetstalet, $t_{\alpha/2}\,(n-1)$, är ett värde från t-fördelningen som bestämmer konfidensgraden. Detta värde ges i tabellsamlingen. Ett $(1-\alpha)\,100$ %-igt k.i. för μ ges då av

$$\bar{x} \pm t_{\alpha/2}\,(n-1)\,\frac{s}{sqrt\,(n)}.$$

Exempel:

Ett telefonbolag vill skatta den genomsnittliga längden på långdistanssamtal under helger. Låt $X =$ "samtalslängd" vara en normalfördelad variabel med okänt medelvärde μ och *okänd* standardavvikelse σ. I ett slumpmässigt stickprov av 16 samtal är $\bar{x} = 62$ minuter och $s = 12$ minuter. Ett 95 %-igt k.i. för μ ges av

$$\begin{aligned} \bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{sqrt(n)} &= 62 \pm 2.13\frac{12}{4} \\ &= 62 \pm 6.39 \end{aligned}$$

Tolkning: Med 95 % tillförlitlighet ligger den genomsnittliga samtalslängden (i populationen) i intervallet $(55.61; 68.39)$.

- Notera att konfidensintervall beräknade med t-fördelningen blir något längre än om man använt $N(0;1)$. Återspeglar ökad osäkerhet pga att populationsvariansen är okänd.

Okänd populationsfördelning (stora stickprov)

Om fördelningen för X är *okänd*, utnyttjar vi att fördelningen för $\bar{X}$ är *approximativt normalfördelad* om n är stort ($n \geq 30$). Dvs

$$\bar{X} \sim approx\ N\left(\mu; \frac{\sigma}{sqrt(n)}\right) \text{ enligt CGS.}$$

Om populationsvariansen σ^2 också är okänd skattas den med stickprovsvariansen s^2. Ett $(1-\alpha)\,100$ %-igt k.i. för μ ges då av

$$\bar{x} \pm \lambda_{\alpha/2}\frac{s}{sqrt(n)}$$

Exempel: I en effektiviseringsprocess vill ett stort bilföretag skatta genomsnittliga antalet arbetsoperationer per arbetare och dag vid en viss typ av monteringsarbete. Man drar därför ett slumpmässigt stickprov bestående av 100 monteringsarbetare (från en stor population av monteringsarbetare), och studerar antalet arbetsoperationer per arbetare under en dag vardera. Man fann att de igenomsnitt hann med 140 arbetsoperationer per dag. Då den skattade standardavvikelsen i stickprovet blev 15, ges ett 90 %-igt k.i. för genomsnittliga antalet arbetsoperationer per arbetare och dag i populationen av

$$\begin{aligned} \bar{x} \pm 1.64\frac{s}{sqrt(n)} &= 140 \pm 1.64\frac{15}{10} \\ &= 140 \pm 2.46 \\ & \quad (137.54; 142.46) \end{aligned}$$

Tolkning: Med 90 % tillförlitlighet ligger genomsnittliga antalet arbetsoperationer per arbetare och dag i populationen i intervallet.

Sammanfattning: k.i. för μ

1. Undersökningsvariabeln är normfördelad (stickprovsstorleken utan betydelse)

   - Om σ är känd ges k.i. av:
   $$\bar{x} \pm \lambda_{\alpha/2} \frac{\sigma}{sqrt\,(n)}.$$
   - Om σ är okänd ges k.i. av:
   $$\bar{x} \pm t_{\alpha/2}\,(n-1)\,\frac{s}{sqrt\,(n)}.$$

2. Undersökningsvariabelns fördelning är okänd (stora stickprov)

   - Om σ är känd ges k.i. av:
   $$\bar{x} \pm \lambda_{\alpha/2} \frac{\sigma}{sqrt\,(n)}.$$
   - Om σ är okänd ges k.i. av:
   $$\bar{x} \pm \lambda_{\alpha/2} \frac{s}{sqrt\,(n)}.$$

Konfidensintervallet för populationsmedelvärdet μ skrivs allmänt på följande sätt:

- Om skattningen $\bar{X}$ är normalfördelad (eller approximativt normalfördelad), där standardavvikelsen för skattningen är
$$\frac{\sigma}{sqrt\,(n)},$$
eller skattade standardavvikelsen för skattningen (medelfelet) är
$$\frac{s}{sqrt\,(n)},$$
gäller att ett k.i.med konfidensgrad $(1-\alpha)\,100$ %, för populationsparametern μ ges av:
$$\bar{x} \pm \text{sanlhetstal } (\lambda_{\alpha/2} \text{ eller } t_{\alpha/2}\,(n-1)) \\ \times \frac{\sigma}{sqrt\,(n)} \left(\text{eller } \frac{s}{sqrt\,(n)} \right).$$

Konfidensintervallet för en godtycklig parameter θ (t ex μ) skrivs ännu mer allmänt på följande sätt:

- Om skattningen θ^* är normalfördelad (eller approximativt normalfördelad), där standardavvikelsen för skattningen är
$$\sigma_{\theta^*},$$
eller skattade standardavvikelsen för skattningen är
$$s_{\theta^*},$$
gäller att ett k.i.med konfidensgrad $(1-\alpha)\,100$ %, för populationsparametern θ ges av:
$$\theta^* \pm \text{sannolikhetstalet } (\lambda_{\alpha/2} \text{ eller } t_{\alpha/2}\,(n-1)) \\ \times \sigma_{\theta^*} \ (\text{eller } s_{\theta^*})\,.$$

Notera att i fallet $\theta = \mu$, är $\theta^* = \bar{x}$,
$$\sigma_{\theta^*} = \frac{\sigma}{sqrt\,(n)}$$
och
$$s_{\theta^*} = \frac{s}{sqrt\,(n)}.$$