

Kapitel 8 Urvalsmetoder och CGS

8.1 Väntevärdesriktighet och effektivitet

DEFINITION 1: Man säger att en skattning är väntevärdesriktig (eng: unbiased estimator) om väntevärdet för skattningen är lika med parametern man försöker skatta.

Exempel 1: Vi har tidigare visat att $E(\bar{X}) = \mu$ d.v.s. att väntevärdet för ett stickprovsmedelvärde är lika med populationsmedelvärdet μ . Med andra ord är $\bar{X}$ en väntevärdesriktig skattning av μ .

DEFINITION 2: Av två väntevärdesriktiga skattningar säger man att den är effektivast (eng: more efficient) som har minst standardavvikelse.

Exempel 2: I exempel 1 ovan såg vi att $\bar{X}$ är en väntevärdesriktig skattning av μ . Vi har tidigare visat att $\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}}$. Om vi har två olika stora stickprov från samma population och beräknar stickprovsmedelvärdet i respektive stickprov, så är alltså stickprovsmedelvärdet som baseras på det större stickprovet en effektivare skattning av μ än det medelvärde som baseras på det mindre stickprovet.

Eftersom $\sigma_{\bar{X}}$ är (ungefär) ”genomsnittsavståndet till populationsmedelvärdet μ ” så bör det vara troligast att skattningen $\bar{X}$ från det större stickprovet hamnar närmast μ . Den bör alltså vara den bättre skattningen och den statistiska termen är alltså att säga att skattningen från det större stickprovet är den effektivaste av de två.

8.2 Centrala gränsvärdessatsen

Vi har tidigare i avsnitt 6.5 träffat på den så kallade Stora Talens Lag (STL). Denna sats säger ungefär att ett stickprovsmedelvärde $\bar{X}$, troligen hamnar nära väntevärdet $\mu = E(X)$ om vi tagit ett ”stort” stickprov.

I detta avsnitt ska vi stifta bekantskap med den så kallade Centrala gränsvärdessatsen (CGS, på engelska Central Limit Theorem, CLT, se sid. 263 i boken).

Denna sats kan man säga är en slags fördjupning av STL. Detta eftersom CGS inte bara säger att $\bar{X}$ antagligen hamnar nära μ utan med hjälp av CGS kan vi även räkna fram hur nära μ som $\bar{X}$ troligen hamnar. I exempel 1 nedan visas hur man kan gå till väga.

Det som CGS säger är att om vi tar ett tillräckligt stort stickprov så är stickprovsmedelvärdet ungefär normalfördelat.

Anmärkning 1: För att använda CGS krävs inte att populationen är normalfördelat.

Anmärkning 2: Skulle populationen faktiskt vara exakt normalfördelning så är stickprovsmedelvärdet också exakt normalfördelat oavsett hur litet stickprovet är.

Detta beror på egenskaperna hos normalfördelningen och har inget med CGS att göra. Man kan motiveras det med hjälp av Sats 1 och Sats 2 i avsnitt 7.5. (Först skriver man om medelvärdet $\bar{X}$ som $\frac{1}{n} \cdot (X_1 + X_2 + \dots + X_n)$, därefter använder man Sats 2 respektive Sats 1.)

Anmärkning 3: Vi sa ovan att stickprovsmedelvärdet är ungefär normalfördelat om stickprovet är ”tillräckligt stort”. Då är en naturlig följdfråga: Hur stort måste n vara för att det ska vara ”tillräckligt stort”?

Svaret på denna fråga är att det beror på hur populationen ser ut.

Skulle pop vara exakt normalfördelat så konstaterade vi ovan (i anmärkning 2) att även $\bar{X}$ är exakt normalfördelat oavsett stickprovets storlek.

Om populationen i sig är ”nästan normalfördelat” så befinner vi oss ”nästan” i samma situation som i anmärkning 2. Därför borde $\bar{X}$ vara ”nästan exakt” normalfördelningen även för små stickprov.

Om populationen i stället är väldigt olik en normalfördelning så kan man säga att CGS inte har någon ”draghjälp” av anmärkning 2. Det kan då behövas ett mycket stort n för att $\bar{X}$ ska vara ungefär normalfördelat. Tumregeln som nämns i boken är att $n \geq 30$ för att $\bar{X}$ ska vara ungefär normalfördelat, men detta inte är någon ”helig” gräns. Det kan behövas ännu större n om populationen i sig är väldigt olik en normalfördelning. Speciellt om populationen är skevt fördelat kan det behövas stora stickprov för att denna skevhet inte ska slå igenom hos $\bar{X}$. För det mesta räcker dock tumregelns ”minst 30” och i många fall, speciellt om populationen är hyfsat symmetrisk, så är $\bar{X}$ nästan normalfördelat även för betydligt mindre värde på n . (Studera figuren på sidan 265 i boken.)

Anmärkning 4: CGS skulle lika gärna kunna uttryckas i termer av summan av observationerna i ett stickprov: $X_1 + X_2 + \dots + X_n$. Anledningen till detta är att $X_1 + X_2 + \dots + X_n = n \cdot \bar{X}$, dvs summa erhålls genom att skala om $\bar{X}$ med faktorn n . Med andra ord är summan $X_1 + X_2 + \dots + X_n$ precis lika ”normalfördelningslik” som medelvärdet $\bar{X}$ är.

Exempel 1. Läs igenom hela exempel 2 i avsnitt 6.5. Vi ska nu med hjälp av CGS visa det som påstås där, att $P(3.4 \leq \bar{X} \leq 3.6) \approx 0.22$.

För det första säger ju CGS att $\bar{X}$ är ungefär normalfördelad. När vi väl konstaterat detta gör vi som vanligt när vi har att göra med en normalfördelad slumpvariabel, vi gör om den till $N(0,1)$ och slår sedan upp de sökta sannolikheterna i NF-tabellen.

När vi gör om till en (approximativ) $N(0,1)$ -fördelning bildar vi

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$

För att kunna göra detta måste vi först ta reda på μ och σ :

x	1	2	3	4	5	6	summa
$P(x)$	1/6	1/6	1/6	1/6	1/6	1/6	1
$xP(x)$	1/6	2/6	3/6	4/6	5/6	6/6	$\mu = \sum xP(x) = 21/6 = 7/2 = 3.5$
$x^2P(x)$	1/6	4/6	9/6	16/6	25/6	36/6	$E(X^2) = \sum x^2P(x) = 91/6$

Standardavvikelsen för X blir alltså

$$\sigma = \sqrt{\sigma^2} = \sqrt{E(X^2) - \mu^2} = \sqrt{\frac{91}{6} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{182}{12} - \frac{147}{12}} = \sqrt{\frac{35}{12}} \approx 1.7078.$$

Vi får nu

$$\begin{aligned} P(3.4 \leq \bar{X} \leq 3.6) &= P\left(\frac{3.4 - 3.5}{1.7078/\sqrt{10}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{3.6 - 3.5}{1.7078/\sqrt{10}}\right) \approx P(-0.19 \leq Z \leq 0.19) = \\ &= P(Z \leq 0.19) - P(Z \leq -0.19) = 0.5753 - 0.4247 = 0.1506 \approx 15\% \end{aligned}$$

Vi fick inte 22 % som det påstods i exemplet och detta beror på att vi ovan inte tagit hänsyn till att vi approximerar en diskret, heltalsvärd variabel med en kontinuerlig slumpvariabel. Vi har tidigare konstaterat att det i sådana fall är lämpligt att använda sig av $1/2$ -korrektion.

Det är lättare att få $1/2$ -korrektionen rätt om vi först uttrycker våra olikheter med hjälp av summan $X_1 + X_2 + \dots + X_{10} = 10 \cdot \bar{X}$.

Olikheterna

$3.4 \leq \bar{X} \leq 3.6$ är det samma som

$34 \leq X_1 + X_2 + \dots + X_{10} \leq 36$, vilket är det samma som

$33.5 \leq X_1 + X_2 + \dots + X_{10} \leq 36.5$, vilket i sin tur är det samma som

$3.35 \leq \bar{X} \leq 3.65$.

Med $1/2$ -korrektion får vi alltså

$$\begin{aligned}
 P(3.4 \leq \bar{X} \leq 3.6) &= P(3.35 \leq \bar{X} \leq 3.65) = P\left(\frac{3.4 - 3.35}{1.7078/\sqrt{10}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{3.65 - 3.5}{1.7078/\sqrt{10}}\right) \approx \\
 &\approx P(-0.28 \leq Z \leq 0.28) = 0.6103 - 0.3897 = 0.2206 \approx 22\% \text{ vilket var vad som tidigare} \\
 &\text{påstods.}
 \end{aligned}$$