7 Parameterskattning

7.1 Punktskattning

Ofta antar man att populationen man är intresserad av har en viss fördelning och man vill skatta t.ex. väntevärdet (eller variansen) av fördelningen. Man kan göra det genom att använda ett stickprov (en samling av oberoende variabler från fördelningen). Väntevärdet kan man skatta genom att använda stickprovsmedelvärdet och variansen genom att använda stickprovsvariansen.

Ett **stickprov** (sample) av storlek n från fördelningen av X är en samling av n oberoende stokastiska variabler som har samma fördelning som X.

Låt $X_1, X_2, \ldots, X_n$ vara ett stickprov från fördelningen av X. Då kallas

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ett **stickprovsmedelvärde** (sample mean).

Stickprovsvariansen (sample variance) är

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

och **stickprovsstandardavvikelsen** (sample standard deviation)

$$S = \sqrt{S^2}.$$

En funktion av ett stickprov $X_1, X_2, \ldots, X_n$ kallas en **statistika**. Till exempel $\bar{X}$ och S^2 är statistikor.

En statistika som används för att skatta en parameter θ (t.ex. μ eller σ^2) kallas en **punktskattning** (point estimator). Den betecknas av $\hat{\theta}$, t.ex. $\hat{\mu} = \bar{X}$ och $\hat{\sigma^2} = S^2$.

$\hat{\theta}$ är en bra skattning av en parameter θ om

1) $\hat{\theta}$ är **väntevärdesriktig**, dvs. att $\mathbf{E}[\hat{\theta}] = \theta$.

2) $\hat{\theta}$ har liten varians då stickprovsstorleken är stor.

Obs! $\bar{X}$ är en väntevärdesriktig skattning för μ.

Ex.7.1.1: Man kastar en tärning. Låt X vara antalet ögon. Då är X en diskret stokastisk variabel med frekvensfunktionen

$$P(X = x) = \frac{1}{6}\,, \quad x = 1, 2, 3, 4, 5, 6$$

dess väntevärde är

$$\mathbf{E}[X] = \mu = 3.5.$$

Man kan kasta en tärning, säg, 30 gånger och räkna $\bar{X}$ för att skatta μ. Om man upprepar 30 tärningskast många gånger, får $\bar{X}$ olika värden. Säg att man har 56 repetitioner (av 30 kast), dvs. att man har 56 (olika) $\bar{X}$-värden. Medelvärdet av de $\bar{X}$-värdena är 3.548, litet högre än det teoretiska värdet 3.5.

Även om skattningen är väntevärdesriktig betyder det inte att man alltid får ett värde som är nära det sanna värdet. Därför räcker det inte för en bra skattning att vara väntevärdesriktig. Den måste också ha liten varians för stora stickprovsstorlekar.

Obs! $Var(\bar{X})$ minskar då n ökar, dvs. att desto större är stickprovsstorleken, desto mindre är variansen av $\bar{X}$.

Obs! Stickprovsmedelvärdet som baserar sig på ett stort stickprov är en bra skattning för väntevärdet.

Låt $\bar{X}$ vara stickprovsmedelvärdet av ett stickprov av storlek n från en fördelning med standardavvikelse σ. Standardavvikelsen av $\bar{X}$, $\sigma/\sqrt{n}$, kallas **standardfelet** (standard error) av medelvärdet.

Obs! S^2 är en väntevärdesriktig skattning för σ^2. $Var(S^2)$ minskar då n ökar. Dvs. att S^2 är en bra skattning för σ^2.

7.4 Intervallskattning

Punktskattning säger ingenting om noggranheten av skattningen. Intervallskattning ger ett intervall, som täcker det sanna parametervärdet med en viss sannolikhet (t.ex. 95%). Intervallet kallas ett **konfidensintervall**.

Ett $100(1-\alpha)\%$ **konfidensintervall** för en parameter θ är ett stokastiskt intervall $[L_1, L_2]$ sådant att

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha$$

Till exempel ett 95% konfidensintervall för θ är ett intervall sådant att

$$P(L_1 \leq \theta \leq L_2) = 0.95$$

För att hitta ett konfidensintervall för μ måste man hitta en statistika som innehåller μ och dess fördelning man känner till. $\bar{X}$ är en väntevärdesriktig skattning av väntevärdet och vi vet dess fördelning när X är normalfördelad:

Låt $X_1, X_2, ..., X_n$ vara ett stickprov från $N(\mu, \sigma)$-fördelningen. Då är $\bar{X}$ normalfördelat med väntevärde μ och varians σ^2/n (standardavvikelse $\sigma/\sqrt{n}$).

Ex.7.4.3: Leukemi är en av de dödligaste cancersorterna. Enligt några studier är tiden som en patient lever efter diagnosen normalfördelad med väntevärde 13 månader och standardavvikelse 3 månader. Man undersöker ett nytt läkemedel, som borde förlänga överlevnadstiden (utan att påverka variansen).

Generellt: Låt $X_1, X_2, ..., X_n$ vara ett stickprov från $N(\mu, \sigma)$-fördelningen. $100(1-\alpha)\%$ konfidensintervall för μ är

$$\bar{X} \pm z_{\frac{\alpha}{2}} \sigma/\sqrt{n},$$

där $z_{\frac{\alpha}{2}}$ är ett värde sådant att $P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ och $Z \sim N(0,1)$.

Konfidensvallet för μ (ovan)

1) är centrerat i $\bar{X}$, symmetriskt m.a.p. $\bar{X}$
2) längden av konfidensintervallet beror på
   - konfidensen man vill ha (desto större konfidens, desto längre intervall)
   - hur mycket $\bar{X}$ varierar (desto större varians, desto längre intervall)
   - stickprovsstorleken n (desto större n, desto kortare intervall)
3) kan användas endast om σ känd. Vanligen känner man inte till den och måste skatta den.
4) X är normalfördelad. Hur gör man om X inte är normalfördelad? Även då är det här konfidensintervallet en bra approximation om n är tillräckligt stort, säg $n \geq 25$.

Centrala gränsvärdessatsen: Låt $X_1, X_2, ..., X_n$ vara ett stickprov (av storlek n) från en fördelning med väntevärdet μ och variansen σ^2. För stora n är $\bar{X}$ approximativt normalfördelat med väntevärdet μ och variansen σ^2/n och

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