Punktskattning

En funktion av en samling stok. var., $\eta = f(\xi_1, \dots, \xi_n)$, är en ny stok. var. som kallas för en statistika. Ex: $f(\xi_1, \dots, \xi_n) = \min\{\xi_1, \dots, \xi_n\}$ och $f(\xi_1, \dots, \xi_n) = \frac{1}{n}\sum_{i=1}^{n} \xi_i$.

- En statistika $\hat{\theta} = \hat{\theta}(\xi_1, \dots, \xi_n)$ som används för att skatta en parameter θ kallas för
- en skattare/estimator. När erhållna realiseringar/utfall $x_1 = \xi_1(\omega), \dots, x_n = \xi_n(\omega)$ (data) av $\xi_1, \dots, \xi_n$ sätts in i skattaren får vi en skattning/ett estimat $\hat{\theta}(x_1, \dots, x_n)$ av θ.

En bra skattare $\hat{\theta}$ för en parameter θ uppfyller att medelkvadratfelet/mean squared

- error $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$

- är litet. Vi vill alltså använda $\hat{\theta}$ s.a. $MSE(\hat{\theta})$ är litet.

Två komponenter påverkar $MSE(\hat{\theta})$:

$$MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = Var(\hat{\theta}-\theta) + E[\hat{\theta}-\theta]^2 =$$

$E[\xi^2] = Var(\xi) + E[\xi]^2$, $\xi = \hat{\theta} - \theta$

$$= Var(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2 = SE(\hat{\theta})^2 + bias(\theta)^2$$ där

standardfelet är $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$ [standardavv. för $\hat{\theta}$] och bias/systematiska felet är

$$bias(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

Ofta börjar man med att försäkra sig om att $\hat{\theta}$ är väntevärdesriktig, dvs

$\text{bias}(\hat{\theta}) = 0 \iff \mathbb{E}[\hat{\theta}] = \theta$ (så att $MSE(\hat{\theta}) = \text{Var}(\hat{\theta})$)

och har man två konkurrerande väntevärdesriktiga skattare $\hat{\theta}_1$ och $\hat{\theta}_2$ så väljer man den effektivaste, dvs den med lägst standardfel/varians:

$\hat{\theta}_1$ mer effektiv än $\hat{\theta}_2 \iff \text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$

$\iff$

den relativa effektiviteten uppfyller

$$\frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)} \geq 0.$$

En väntevärdesriktig skattare genererar skattningar som, i medeltal, givet upprepade försök, ger den sanna parametern. Dock kan det svänga otroligt mycket från försök till försök såvida inte $SE(\hat{\theta})$ är litet (varvid $MSE(\hat{\theta})$ blir litet).

Ex ξ_1, ξ_2, ξ_3 oberoende och likafördelade med $\mathbb{E}[\xi_i] = \mu$, $\text{Var}[\xi_i] = \sigma^2$

Två skattare för μ: $\hat{\mu}_1 = \frac{\xi_1 + 2\xi_2 + 3\xi_3}{6}$, $\hat{\mu}_2 = \frac{\xi_1 + 4\xi_2 + \xi_3}{6}$

$$\Rightarrow \text{bias}(\hat{\mu}_1) = \mathbb{E}\left[\frac{\xi_1 + 2\xi_2 + 3\xi_3}{6}\right] - \mu = \frac{\mathbb{E}[\xi_1] + 2\mathbb{E}[\xi_2] + 3\mathbb{E}[\xi_3]}{6} - \mu = \frac{6\mu}{6} - \mu = 0$$

$$\text{bias}(\hat{\mu}_2) = \frac{\mathbb{E}[\xi_1] + 4\mathbb{E}[\xi_2] + \mathbb{E}[\xi_3]}{6} - \mu = \frac{6\mu}{6} - \mu = 0$$

Relativ effektivitet:

$$\frac{SE(\hat{\mu}_2)^2}{SE(\hat{\mu}_1)^2} = \frac{Var(\hat{\mu}_2)}{Var(\hat{\mu}_1)} \overset{\text{bias}=0}{=} \frac{Var\left(\frac{\xi_1+4\xi_2+\xi_3}{6}\right)}{Var\left(\frac{\xi_1+2\xi_2+3\xi_3}{6}\right)} \overset{\text{obs. med samma varians}}{=} \frac{\frac{\sigma^2+4^2\sigma^2+\sigma^2}{6^2}}{\frac{\sigma^2+2^2\sigma^2+3\sigma^2}{6^2}} = \frac{18\sigma^2}{14\sigma^2} \approx 1.29 > 0$$

$\Rightarrow \hat{\mu}_1$ är en bättre skattare.

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$, $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, är en väntevärdesriktig skattare för $\sigma^2 = Var(\xi_i)$

$$(n-1)S^2 = \sum_{i=1}^{n}\left((\xi_i-\mu)+\underbrace{(\mu-\bar{\xi})}_{\text{bevara ej på i}}\right)^2 = \sum_{i=1}^{n}(\xi_i-\mu)^2 - 2(\bar{\xi}-\mu)\underbrace{\sum_{i=1}^{n}(\xi_i-\mu)}_{=n(\bar{\xi}-\mu)} + n(\mu-\bar{\xi})^2$$

$$= \sum_{i=1}^{n}(\xi_i-\mu)^2 - n(\bar{\xi}-\mu)^2$$

$$\Rightarrow (n-1)E[S^2] = \sum_{i=1}^{n}\underbrace{E[(\xi_i-\mu)^2]}_{=Var(\xi_i)=\sigma^2} - n\underbrace{E[(\bar{\xi}-\mu)^2]}_{=Var(\bar{\xi})+\text{ty } E[\bar{\xi}]=\mu} =$$

$$Var(\bar{\xi}) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n}\xi_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$= n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2$$

$\Rightarrow E[S^2] = \sigma^2$

I ett binomialexperiment har vi $\xi \sim Bin(n,p)$

$\xi = \sum_{i=1}^{n}\xi_i$, där $\xi_i \sim Bin(1,p)$, med medelvärdet $\hat{p} = \frac{\xi}{n}$ och $E[\hat{p}] = \frac{E[\xi]}{n} = \frac{np}{n} = p$, $Var(\hat{p}) = \frac{Var(\xi)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$.

M.a.o. är $\hat{p}$ väntevärdesriktig för p där $\xi \sim Bin(n,p)$ med p okänd.

Konfidensintervall / intervallskattning

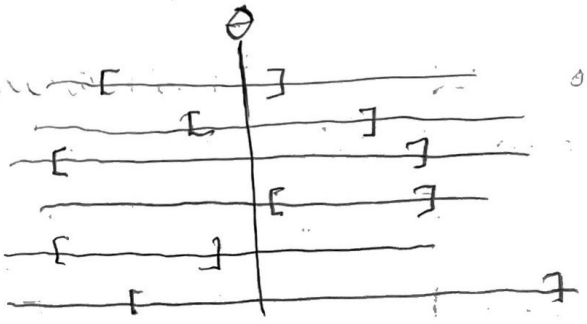
Hittills har vi fokuserat på punktskattningar av en parameter av intresse, θ. Men beroende på det observerade stickprovet kommer skattningen $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ att ligga olika långt ifrån det sanna värdet θ (Om $MSE(\hat{\theta})$ är litet vet vi att detta avstånd tenderar att vara litet). Men kan man på något sätt ge en skattning som inkorporerar osäkerheten som följer av att vi använder oss av ett enda observerat stickprov $x_1, \dots, x_n$ från $\xi_1, \dots, \xi_n$?

Ett konfidensintervall $[L, U] = [L(\xi_1, \dots, \xi_n), U(\xi_1, \dots, \xi_n)]$ för θ är ett stokastiskt intervall som uppfyller

$$P(\theta \in [L, U]) = 1 - \alpha \quad , \alpha \in (0, 1)$$

där $(1-\alpha)100\%$ kallas konfidensgraden.

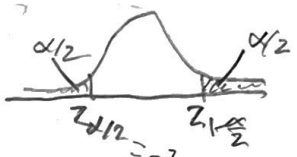
Om vi gör om ett experiment 1000 ggr får vi 1000 st observerade stickprov $x_j = \{x_{1j}, \dots, x_{nj}\}$, $j = 1, \dots, 1000$, från $\xi_1, \dots, \xi_n \Rightarrow$ intervallen $I_j = [L(x_{1j}, \dots, x_{nj}), U(x_{1j}, \dots, x_{nj})]$, $j = 1, \dots, 1000$, där ungefär $(1-\alpha) \cdot 1000$ av dem kommer att innehålla den sanna men okända parametern θ.



Vi vill så klart ha α så litet som möjligt men $[L, U] \to (-\infty, \infty)$ när $\alpha \to 0$; $P(\theta \in (-\infty, \infty)) = 1$ är inte så användbart.

Vad som är "tillräcklig säkerhet" varierar men oftast låter man $1 - \alpha = 0.95$

Men hur kan man skapa rättriga konfidens-intervall?



Om $\xi_1, \dots, \xi_n \sim N(\mu, \sigma^2)$ vet vi att

$$1-\alpha = P\left(z_{\alpha/2} \le \underbrace{\frac{\bar{\xi}-\mu}{\sigma/\sqrt{n}}}_{=Z} \le z_{1-\alpha/2} \right) = P\left(-z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \le \bar{\xi} - \mu \le z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

$$= P\left(\bar{\xi} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{\xi} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

$\Rightarrow L = L(\xi_1, \dots, \xi_n) = \bar{\xi} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$, $U = U(\xi_1, \dots, \xi_n) = \bar{\xi} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$

(funktion av $\xi_1, \dots, \xi_n$)

- Vilket man givet data $x_1, \dots, x_n$, med $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, uttrycker som:

Ett $(1-\alpha)\cdot 100\%$ konfidensintervall för väntevärdet $\mu = E[\xi_i]$ ges av $\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$

- Notera att ett observerat intervall antingen innehåller den sanna okända parametern θ eller inte. Tolkningen är att vi med sannolikhet α har fått ett så "ovanligt" observerat stickprov att det erhållna intervallet ej innehåller θ.

- Ex Observationer av temperaturen i 49 komponenter antas normalfördelad med $\sigma = 14\,C^\circ$. Medeltemp. uppmätt till $\bar{x} = 68\,C^\circ$ $\Rightarrow$ 99%-igt konfidensintervall för μ ges av

$$\bar{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} = 68 \pm 2.58 \frac{14}{\sqrt{49}} \Leftrightarrow [62.84, 73.16]$$

eftersom $\alpha = 0.01 \Rightarrow z_{1-\alpha/2} = z_{0.995} = 2.58$ ($0.9951 = \Phi(2.58)$ från tabellen) ☐

Allmänt gäller för en skattare $\hat{\theta}$ som har en $N(\theta, \mathrm{Var}(\hat{\theta}))$-fördelning att ett $(1-\alpha)\cdot 100\%$-igt konfidensintervall ges av $\hat{\theta} \pm z_{1-\alpha/2} \underbrace{\sqrt{\mathrm{Var}(\hat{\theta})}}_{=SE(\hat{\theta})}$

I praktiken är det dock (nästan) aldrig fallet att σ är känt och ofta är antagandet att $x_1,\dots,x_n$ kommer från en normalfördelning inte sann.

I det senare fallet kan vi använda CGS-approximationen och utgå från att $\frac{\bar{\xi}-E[\bar{\xi}]}{\sqrt{Var(\bar{\xi})}} \overset{approx}{\sim} N(0,1)$ för att konstruera $[L,U]$. T.ex. om $\xi \sim Bin(n,p)$ där vi vill intervallskatta $p \in (0,1)$, förutsatt att n är stort:

$$1-\alpha \approx P\left(z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{1-\frac{\alpha}{2}}\right) = \dots = P\left(\hat{p}-z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p}+z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$\hat{p} = \frac{\xi}{n} = \frac{1}{n}\sum_{i=1}^{n} \xi_i \overset{approx}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

($\xi_i \sim Bin(1,p)$; $\frac{p(1-p)}{n}$ = var.)

?? + $p \approx \hat{p}$:

$$P\left(\underbrace{\hat{p}-z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{=L} \le p \le \underbrace{\hat{p}+z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{=U}\right)$$

$\Rightarrow$ intervallet ges av $\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Till fallet där σ är okänd och $\xi_1,\dots,\xi_n \sim N(\mu,\sigma^2)$ (och även om $\xi_1,\dots,\xi_n$ inte är normalförd. men n är stort):

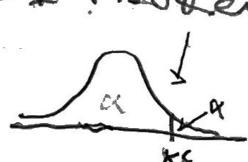
$$\frac{\bar{\xi}-\mu}{S/\sqrt{n}} \sim t(\nu)$$

← t-fördeln. med $\nu = n-1$ "frihetsgrader" (skrivs "df"), $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2$

$t(\nu)$-fördelningen är symmetrisk och ser ut ungefär som en $N(0,1)$-fördelning men har mer massa för stora och små värden ("tjocka svansar"); vi har $\lim_{\nu\to\infty} t(\nu) = N(0,1)$ (eftersom S^2 går mot σ^2 när $n\to\infty$).

Tätheten $f_{t(\nu)}(x)$ (och därmed fördeln.funk.) har en ganska komplex form (se sid. 268) men värden för fördelningsfunkt. finns i Appendix 1 i boken.

$\xi \sim t(\nu) \Rightarrow E[\xi] = 0$, $Var(\xi) = \frac{\nu}{\nu-2}$ för $\nu \ge 3$.

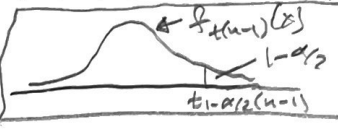


Vi får här

$$1-\alpha = P\left(-x^c_{\alpha/2} \le \frac{\bar{\xi}-\mu}{s/\sqrt{n}} \le x^c_{\alpha/2}\right) = \ldots = P\left(\underbrace{\bar{\xi}-x^c_{\alpha}\frac{s}{\sqrt{n}}}_{=L} \le \mu \le \underbrace{\bar{\xi}+x^c_{\alpha}\frac{s}{\sqrt{n}}}_{=U}\right)$$

(skrivs ofta $t_{1-\frac{\alpha}{2}}(\nu) = t_{1-\frac{\alpha}{2}}(n-1)$)

=> **Ett $(1-\alpha)\cdot 100\%$-igt konfidensintervall för $\mu = E[\xi_i]$ när $\sigma^2 = Var(\xi_i)$ är okänd och skattas med s^2 ges av $\bar{x} \pm t_{1-\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}$**



← fatheten för $t(n-1)$-fördel.

- **Ex** Längderna på 15 stag till båtar antas normalfördelade och vi vet att
- $\sum_{i=1}^{15} x_i = 199$, $\sum_{i=1}^{n} x_i^2 = 2787$

$$\Rightarrow \bar{x} = \frac{199}{15} \approx 13.267 \text{ och } s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\underbrace{\sum_{i=1}^{n} x_i}_{=n\bar{x}} + n\bar{x}^2}{n-1} =$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} = \frac{2787 - \frac{199^2}{15}}{14} \Rightarrow s \approx 3.24$$

Antal frihetsgrader $\nu = df$ är $n-1 = 14$ och vi vill ha ett 95%igt konfidensintervall

- $\Rightarrow t_{1-\alpha/2}(\nu) = t_{1-\frac{0.05}{2}}(14) = 2.15$ enl. tabell
- $\Rightarrow$ konfidensintervallet ges av

$$\bar{x} \pm t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} = 13.267 \pm 2.15\frac{3.24}{\sqrt{15}} = 13.267 \pm 1.799$$

Konfidensintervall för variansen

Eftersom kvadraten av $Z \sim N(0,1)$ har att $E[Z^2] = Var(Z) + E[Z]^2 = \sigma^2 + 0^2 = \sigma^2$ så är $\sum_{i=1}^{n}\left(\frac{\xi_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2$ vettig för att härleda ett konfidensintervall för σ^2 givet $\xi_1, \ldots, \xi_n \sim N(\mu, \sigma^2)$

Konfidensintervall för variansen

Givet $\xi_{1:n} = \xi_1, \dots, \xi_n \sim N(\mu, \sigma^2)$ så gäller att

$$\sum_{i=1}^{n} \left(\frac{\xi_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$$ ← "Chi-två"-fördelad med n frihetsgrader

$\chi^2(\nu)$-fördelningen är skev med en gaska komplex ~~tathetsfunktion~~ (sid. 305).

Vi har att $\xi \sim \chi^2(\nu) \Rightarrow \mathbb{E}[\xi] = \nu$, $\mathrm{Var}(\xi) = 2\nu$.

Om $\nu = 2$ så gäller $\chi^2(\nu) = \chi^2(2) = \mathrm{Exp}(\frac{1}{2})$.

Eftersom $\xi_i - \mu \sim N(0, \sigma^2)$ så skulle vi kunna använda $\frac{1}{\sigma^2}\sum_{i=1}^{n}(\xi_i - \mu) \sim \chi^2(n)$ för att härleda ett konfidensintervall för σ^2. Men, vi känner ej till μ i praktiken. Lösning: Använd $\frac{1}{\sigma^2}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2$, men vad har denna ~~statistika~~ för fördelning?

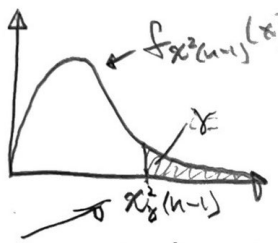
Man kan visa att $\chi^2(n-1) \sim \frac{1}{\sigma^2}\sum_{i=1}^{n}(\xi_i - \bar{\xi}) =$

$$= \frac{n-1}{\sigma^2}\underbrace{\frac{\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2}{n-1}}_{\sim S^2(n-1)} = \frac{S^2(n-1)}{\sigma^2}$$

För $\xi \sim \chi^2(n-1)$, låt $\gamma = \mathbb{P}(\xi > \chi^2_\gamma(n-1))$, $\gamma \in (0,1)$.

$$\Rightarrow 1-\alpha = \mathbb{P}\left(\chi^2_{1-\frac{\alpha}{2}}(n-1) \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2}(n-1)\right) =$$

$$= \mathbb{P}\left(\underbrace{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}}_{=L} \le \sigma^2 \le \underbrace{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}}_{=U}\right), \text{ dvs}$$



Finns i tabell

ett $(1-\alpha)\cdot 100\%$ igt konfidensintervall för $\sigma^2 = \mathrm{Var}(\xi_i)$, där $\xi_{1:n} \sim N(\mu, \sigma^2)$ med både μ och σ^2 okända:

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$ För intervall för σ, ta roten av intervallets gränserna

Ex $n=15$ huvudvärkstabletter där koncentrationen av aktiva ingredienser antas normalfördelad. Givet punktskattningen 0.8 för standardavvikelsen, finn ett 95%igt konfidensintervall för σ.

Lösn: Tabell: $\chi^2_{\alpha/2}(n-1) = \chi^2_{0.025}(14) = 26.119$

$\chi^2_{1-\alpha/2}(n-1) = \chi^2_{0.975}(14) = 5.629$

$n=15$
$s=0.8$

$$\Rightarrow \left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}}\right] = [0,5857, 1.2617]$$