



Johan Koskinen, Statistiska institutionen, Stockholms universitet

Finansiell statistik, vt-05

F12 Sampling, samplingfördelning, punktskattningar, intervallskattningar, hypotesprövning

Sampling - grundläggande begrepp

- Population
- element
- stickprov/sample
- sannolikhetsurval
- slumpfel



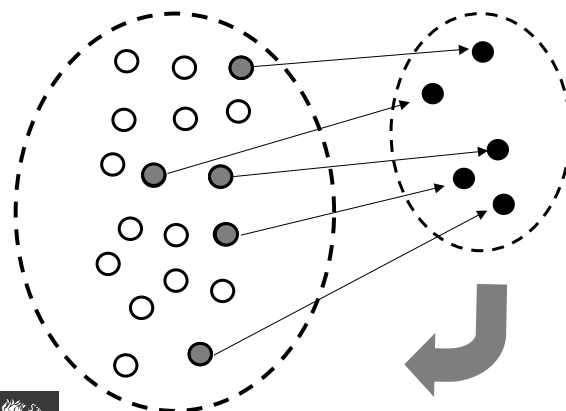
Johan Koskinen, Department of Statistics

2005-04-28

2

Sampling - population, element, stickprov

Population vi vill undersöka dra element



stickprov

inferens:
dra slutsatser från
stickprovet

till populationen



Johan Koskinen, Department of Statistics

2005-04-28

3

Sampling - sannolikhetsurval

För varje element i i populationen $\{1, 2, 3, \dots, N\}$



vet vi vad sannolikheten är att detta element skall komma med i stickprovet

Varför bra?

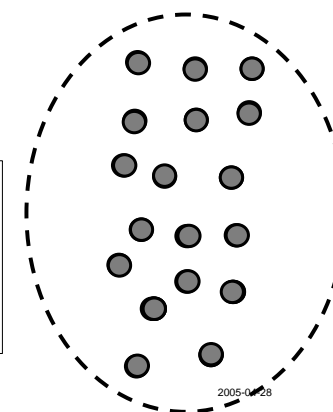
Beräkna slumpfelet...

Exempel sannolikhetsurval:

Obundet slumpmässigt urval (OSU)

stickprovsstorlek n :

alla stickprov lika sannolika



Johan Koskinen, Department of Statistics

2005-04-28

4

Sampling - slumpfel

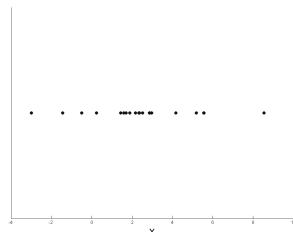
Säg t.ex. vi skall skatta populationstotalen μ .

Vi drar stickprov och beräknar skattningen

$$\bar{x}$$

Eftersom vi endast dragit stickprov kommer slumpen göra så skattningen skiljer sig ifrån μ .

Detta kallas slumpfel



Johan Koskinen, Department of Statistics

2005-04-28

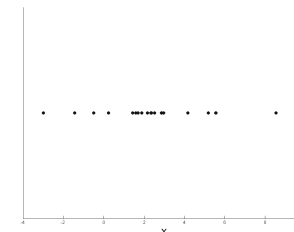
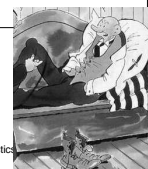
5

Sampling - slumpfel

Om vi drar stickprov med kända sannolikheter kan vi beräkna slumpfelet

Systematiskt fel: om vi systematiskt väljer väljer "fel" kan vi ej beräkna storleken på slumpfelet

vi vill undersöka Sveriges befolkning och drar slumpmässigt telefonnummer...



Johan Koskinen, Department of Statistics

2005-04-28

6

Samplingfördelning

Fördelningen för en statistika /stickprovsvariabel (funktion av stickprovet)

Exempel från tidigare

För ober. likaförd. s.v. $X_1, X_2, \dots, X_n, X_i \in N(\mu, \sigma^2)$

$$E(\bar{X}) = E\left[\frac{\sum_{i=1}^n X_i}{n}\right] = E(X) = \mu \quad \text{och}$$

$$\text{Var}(\bar{X}) = \text{Var}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n} \text{Var}(X) = \frac{\sigma^2}{n}$$



Johan Koskinen, Department of Statistics

2005-04-28

7

Samplingfördelning

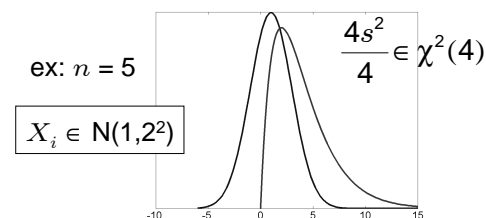
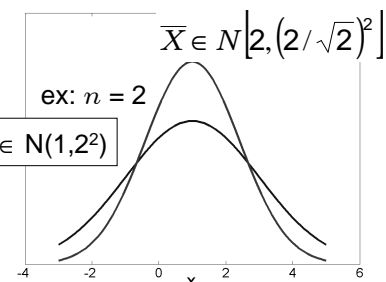
Samt

$$\bar{X} \in N\left[\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right]$$

Vidare, för ober. likaförd. s.v.

$X_1, X_2, \dots, X_n, X_i \in N(\mu, \sigma^2)$

$$\frac{(n-1)s^2}{\sigma^2} \in \chi^2(n-1)$$



Johan Koskinen, Department of Statistics

8

Samplingfördelning

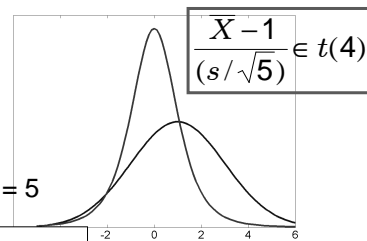
ober. likaförd. s.v.

$$X_1, X_2, \dots, X_n, X_i \in N(\mu, \sigma^2)$$

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

ex: $n = 5$

$$X_i \in N(1, 2^2)$$



Johan Koskinen, Department of Statistics

2005-04-28

9

Inferens

- Vad är en parameter?
 - Anpassa fördelningar
 - Skatta konstanta populationsegenskaper
- Punktskattningar
- Intervalskattningar
- Testa hypoteser



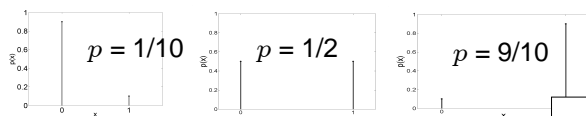
Johan Koskinen, Department of Statistics

2005-04-28

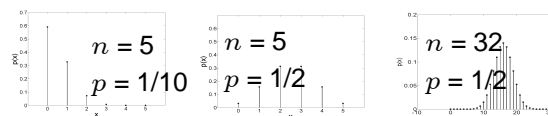
10

Inferens - anpassa fördelningar

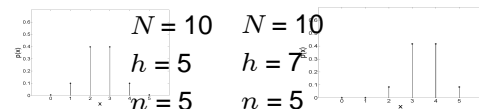
Bernoulli:



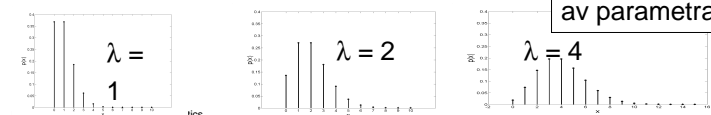
Binomial:



Hyperg.:



Poisson:



modeller som anger hur sannolikt det är att observera saker vi kan observera
- bestäms helt av parametrar

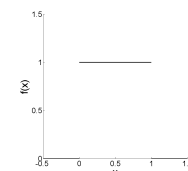


11

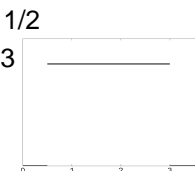
Inferens - anpassa fördelningar

uniform:

$$a = 0 \\ b = 1$$

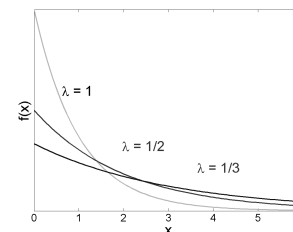


$$a = 1/2 \\ b = 3$$

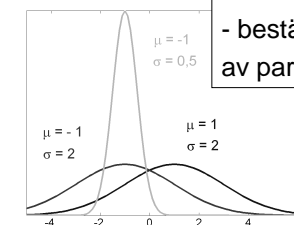


modeller som anger hur sannolikt det är att observera saker vi kan observera

exp:



norm:



- bestäms helt av parametrar



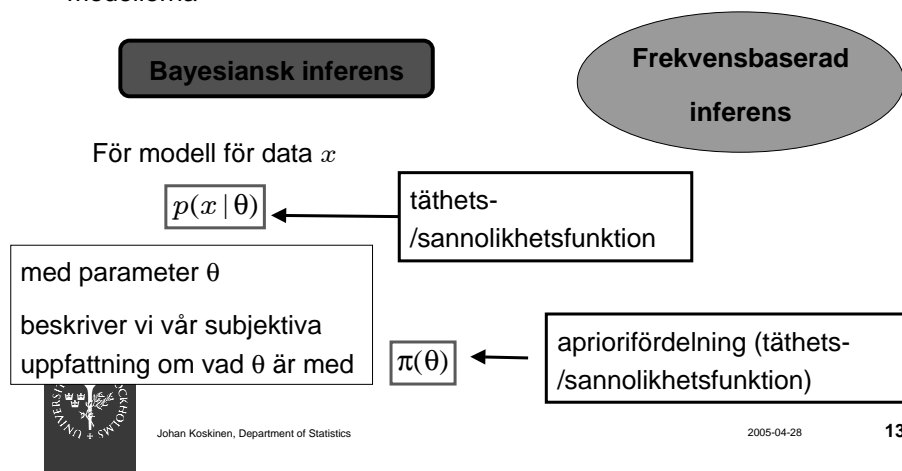
Johan Koskinen, Department of Statistics

2005-04-28

12

Inferens - parameterskattningar

Vi behöver något sätt att skatta parametrarna för att kunna anpassa modellerna

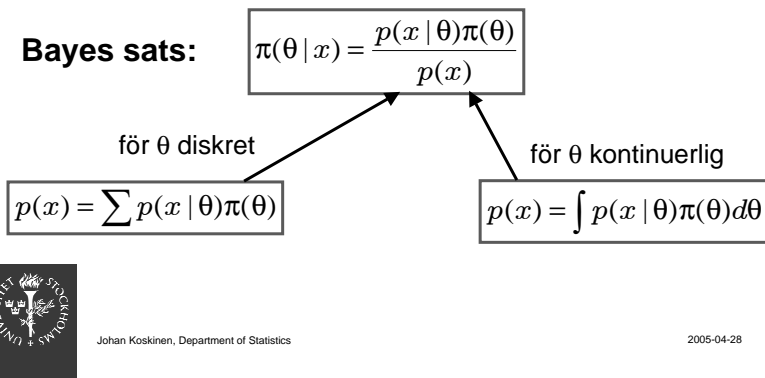


Inferens - parameterskattningar

När vi väl observerat data x

osäkerhet endast angående parameter θ

osäkerheten ges av aposteriorifördelningen för θ givet data



Inferens - parameterskattningar

Ex: Antag att längden för vuxna störfiskar i Donau är $N(\mu, 0,19^2)$

utifrån mina fiskerfarenheter beskriver jag min subjektiva tro angående medellängden μ med apriorifördelningen $N(1,8; 0,25^2)$

m.a.o. tror jag exempelvis att medellängden ligger mellan 1,31 och 2,29 med sannolikheten 0,95.

Vi drar ett stickprov X_1, X_2, \dots, X_n om $n = 10$ störfiskar

eftersom

$$\bar{X} \in N\left[\mu, \left(\sigma/\sqrt{n}\right)^2\right] \text{ så } p(\bar{x} | \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2}}$$



Inferens - parameterskattningar

Och enligt Bayes sats

$$\pi(\mu | \bar{x}) = \frac{p(\bar{x} | \mu)\pi(\mu)}{p(\bar{x})}$$

där

$$p(\bar{x} | \mu) = \frac{1}{\sqrt{2\pi \frac{0,19^2}{10}}} e^{-\frac{(\bar{x}-\mu)^2}{2 \times 0,19^2 / 10}}$$

och

$$\pi(\mu) = \frac{1}{\sqrt{2\pi 0,25^2}} e^{-\frac{(\mu-1,8)^2}{2 \times 0,25^2}}$$

aposteriorifördelningen för μ givet data

$\mu | \bar{x} \in N[m_1, h_1]$ där

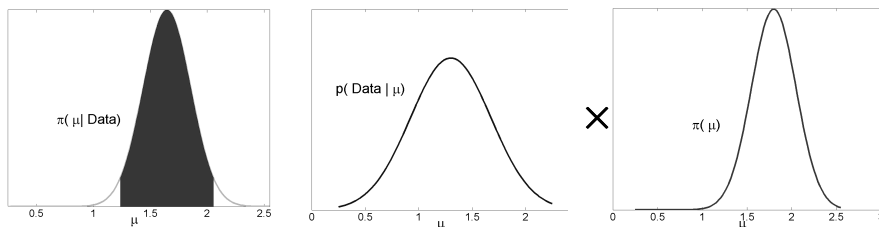
$$m_1 = \frac{1/0,25^2}{1/0,25^2 + 10/0,19^2} 1,18 + \frac{1/0,19^2}{1/0,25^2 + 10/0,19^2} \bar{x}$$

$$h_1 = \frac{1}{1/0,25^2 + 10/0,19^2}$$



Inferens - parameterskattningar

Så om för våra $n = 10$ störor X_1, X_2, \dots, X_n
ger $\bar{x} = 1,3$



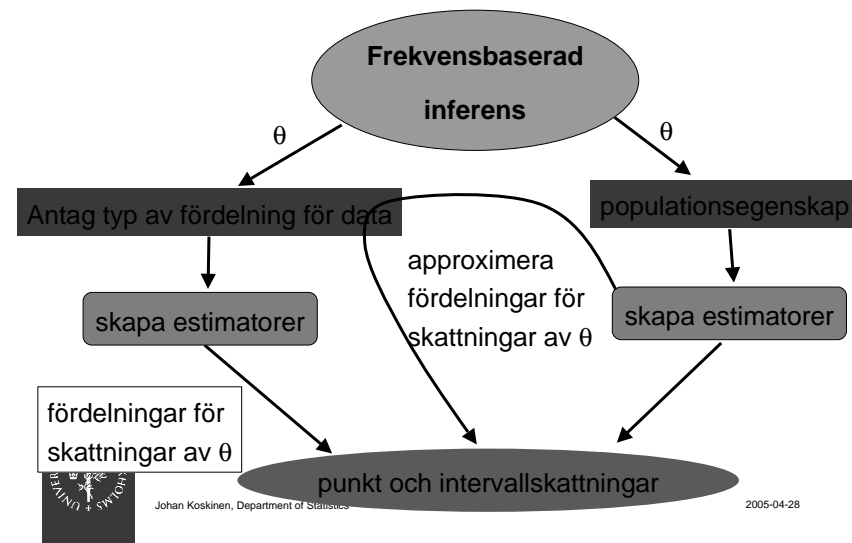
$$m_1 = \frac{1/0,25^2}{1/0,25^2 + 10/0,19^2} 1,18 + \frac{10/0,19^2}{1/0,25^2 + 10/0,19^2} 1,3 = 1,65$$

Johan Koskinen, Department of Statistics

2005-04-28

17

Inferens - parameterskattningar



Johan Koskinen, Department of Statistics

2005-04-28

18

Inferens - parameterskattningar

Antag att parametern θ är en pop.egenskap eller
fördelningsparameter

Ex: Antag att längden för vuxna störfiskar i Donau är $N(\theta, 0,19^2)$

Låt θ vara medelinkomsten (på kommunnivå) i Sverige

En estimator

$\hat{\theta}$ är en funktion av stickprov som
skall skatta θ

är en stokastisk variabel med en
fördelning

Ex för stickprov: Stickprovsmedelvärde är en estimator för
medellängden för störfiskar i Donau

Stickprovsmedelvärde är en estimator för medelinkomsten i Sverige

9

Inferens - parameterskattningar

En realisation av en estimator $\hat{\theta}$

det vill säga ett observerat värde, är en skattning av θ

Så om för våra $n = 10$ störor X_1, X_2, \dots, X_n ger

$$\bar{x} = 1,3$$

är 1,3 en skattning av medellängden



Johan Koskinen, Department of Statistics

2005-04-28

20

Inferens - punktskattning

När estimatoren är ett tal kallas detta en punktskattning

Eftersom estimatoren är en stokastisk variabel (vi får olika värden för varje stickprov vi drar) har den ett väntevärde

$$E(\hat{\theta})$$

Skall tolkas som:

Det värde vi får i genomsnitt om vi drar "många" stickprov och beräknar skattningen för varje stickprov



Inferens - punktskattning

osäkerheten för skattningen

≈ hur säkra vi är på att skattningen är riktig

mäts med variansen (standardavvikelsen) för estimatoren

$$\text{Var}(\hat{\theta}) \quad \text{standardfel: } SD(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$$

Ex: om estimatoren är stickprovsmedelvärdet för ober. likaförd. s.v.

X_1, X_2, \dots, X_n alla med $\text{Var}(X)$

$$\text{Var}(\bar{X}) = \text{Var}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n} \text{Var}(X)$$



Egenskaper för estimatorer - vvr

Om en estimator skall skatta θ

är det önskvärt att vi gissar rätt i långa loppet

d.v.s. att det förväntade felet vi gör är litet

$$\text{Systematiskt fel} = E(\hat{\theta}) - \theta$$



def En estimator för θ är väntevärdesriktig om det förväntade felet är 0:

$$E(\hat{\theta}) - \theta = 0$$



Egenskaper för estimatorer - vvr

Ex: antag att vi har ober. likaförd. s.v. X_1, X_2, \dots, X_n fördelade som $X \in \text{Exp}(1/\theta)$

För estimatoren för θ

$$\hat{\theta} = \bar{X}$$

har vi

$$E(\hat{\theta}) = E(\bar{X})$$

$$= E\left[\frac{\sum_{i=1}^n X_i}{n}\right]$$

$$= E(X) = \theta$$

m.a.o. är stickprovsmedelvärdet en väntevärdesriktig estimator av θ



Egenskaper för estimatorer -vvr

Ex: För ober. likaförd. s.v. X_1, X_2, \dots, X_n alla med $\text{Var}(X) = \theta$

låt estimatorn vara stickprovsvariansen

$$\begin{aligned}
 E(\hat{\theta}) &= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}\right] \\
 &= E\left[\frac{\sum_{i=1}^n X_i^2}{n-1}\right] - E\left[\frac{n\bar{X}^2}{n-1}\right] \\
 &= \frac{1}{n-1} \sum_{i=1}^n E(X_i^2) \\
 &\quad - \frac{n}{n-1} (\text{Var}(\bar{X}) + E[\bar{X}]^2)
 \end{aligned}$$

$$\begin{aligned}
 &\sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \\
 &= \sum_{i=1}^n X_i^2 - 2 \frac{\sum_{i=1}^n X_i}{n} \sum_{i=1}^n X_i + n\bar{X}^2 \\
 &= \sum_{i=1}^n X_i^2 - n\bar{X}^2
 \end{aligned}$$

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2$$

2005-04-28
25



Egenskaper för estimatorer -vvr

$$\begin{aligned}
 E(\hat{\theta}) &= \frac{1}{n-1} \sum_{i=1}^n E(X_i^2) - \frac{n}{n-1} (\text{Var}(\bar{X}) + [E(\bar{X})]^2) \\
 &= \frac{1}{n-1} \sum_{i=1}^n (\text{Var}(X_i) + [E(X_i)]^2) - \frac{n}{n-1} (\text{Var}(\bar{X}) + [E(\bar{X})]^2) \\
 &= \frac{n}{n-1} (\text{Var}(X) + [E(X)]^2) - \frac{n}{n-1} \left(\frac{\text{Var}(X)}{n} + [E(X)]^2 \right) \\
 &= \frac{n}{n-1} \text{Var}(X) - \frac{n}{n-1} \frac{\text{Var}(X)}{n} \quad \boxed{= \text{Var}(X)}
 \end{aligned}$$

Stickprovsvariansen är en vvr skattning av variansen för X



Egenskaper för estimatorer - effektivitet

Om en estimatorn skall skatta θ

är det önskvärt att vi får ungefär samma skattning om vi upprepar skattningsproceduren många gånger

d.v.s. variansen för skattningen är liten \approx estimatorn effektiv

Relativ effektivitet: En estimatorn

$\hat{\theta}_1$ är mer effektiv än $\hat{\theta}_2$ $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$

m.a.o. ett sätt att säga vilken av två vvr estimatorer som är "bäst"



Egenskaper för estimatorer - effektivitet

Ex: Ober. likaförd. s.v. X_1, X_2, \dots, X_n alla med $E(X) = \mu$

Estimatorer

$$-\hat{\theta}_1 = \overline{X} \quad \text{---} \quad \hat{\theta}_2 = \overline{X} + Y$$

Där oberoende av X , $Y = 1$ med sannolikheten $1/2$, -1 med slh $1/2$

$$E(\hat{\theta}_1) = E(\bar{X}) = E(X) = \mu$$

$$E(\hat{\theta}_2) = E(\bar{X} + Y) = E(\bar{X}) + E(Y) = \mu + 0$$

men:

$$\text{Var}(\hat{\theta}_1) = \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} \quad \text{och}$$

$$\text{Var}(\hat{\theta}_2) = \text{Var}(\bar{X} + Y) = \text{Var}(\bar{X}) + \text{Var}(Y) = \frac{\text{Var}(X)}{n} + 1$$

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$$



Egenskaper för estimatorer - konsistens

Om en estimator skall skatta θ

är det önskvärt att vi ju mer information vi får desto bättre skattar vi θ

en estimator är konsistent om vi skattar rätt när antalet observationer i stickprovet blir "många"

Ex: Ober. likaförd. s.v. X_1, X_2, \dots, X_n från fördelning med parameter θ
en estimator av θ är konsistent om
Sannolikheten att vi gör ett fel större än ε

$$P(|\hat{\theta} - \theta| > \varepsilon) \text{ går mot } 0 \text{ när } n \rightarrow \infty$$



Johan Koskinen, Department of Statistics

2005-04-28

29

Inferens - intervallskattning

Ett sätt att redovisa skattning och osäkerhet samtidigt är:

intervallskattning

isf estimator som ger ett tal \rightarrow estimator ger två tal

$$\hat{\theta}_1 \text{ och } \hat{\theta}_2 \text{ med } \hat{\theta}_1 < \hat{\theta}_2$$

och tolkning t.ex. "med 95% konfidens ligger θ mellan

$$\hat{\theta}_1 \text{ och } \hat{\theta}_2 "$$

observera: from nu antar vi hela tiden Normalfördelning el. approximativ normalfördelning genom CGS, för data



Johan Koskinen, Department of Statistics

2005-04-28

30

Inferens - intervallskattning

En intervallestimator för θ

är en funktion av ett stickprov som ger en

$$\hat{\theta}_1 \text{ och } \hat{\theta}_2 \text{ med } \hat{\theta}_1 < \hat{\theta}_2$$

så att om tar många stickprov: kommer θ ligga mellan "ett visst antal gånger"

allt, om tar många stickprov: med föreskriven sannolikhet $1 - \alpha$ kommer

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$$



Johan Koskinen, Department of Statistics

2005-04-28

31

Inferens - intervallskattning

En intervallestimator för θ

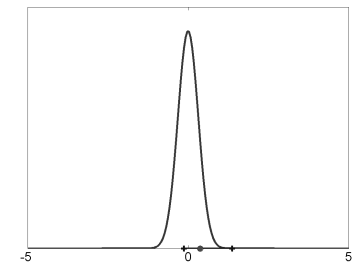
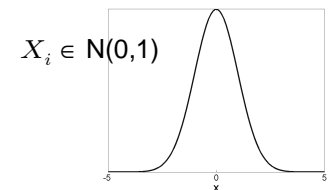
är alltså ett slumpmässigt intervall

ex på slumpmässigt intervall:

dra ober. likaförd. s.v. $X_1, X_2, \dots, X_{10}, X_i \in N(0,1)$

beräkna stickprovsmedelvärdet, lägg till 1 och dra ifrån .5

$$\bar{X} \in N(0, \sqrt{.10})$$



Johan Koskinen, Department of Statistics

2005-04-28

32

Inferens - intervallskattning

Om en punktestimator för θ

$$\hat{\theta} \in N(\theta, SD(\hat{\theta})^2)$$

kan vi konstruera en intervallestimator för θ genom att lägga till och dra ifrån B

$$\hat{\theta}_1 = \hat{\theta} - B; \hat{\theta}_2 = \hat{\theta} + B$$

för att bestämma B så att vi får rätt täckningssannolikhet $1 - \alpha$ lös ut B

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = P(\hat{\theta} - B < \theta < \hat{\theta} + B) = 1 - \alpha$$



Johan Koskinen, Department of Statistics

2005-04-28

33

Inferens - intervallskattning

lös ut B

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = P(\hat{\theta} - B < \theta < \hat{\theta} + B) = 1 - \alpha$$

börja i andra änden

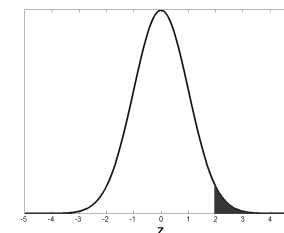
$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{SD(\hat{\theta})} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$-z_{\alpha/2}SD(\hat{\theta}) \leq \hat{\theta} - \theta \leq z_{\alpha/2}SD(\hat{\theta})$$

$$-z_{\alpha/2}SD(\hat{\theta}) - \hat{\theta} \leq -\theta \leq z_{\alpha/2}SD(\hat{\theta}) - \hat{\theta}$$

$$\hat{\theta} + z_{\alpha/2}SD(\hat{\theta}) \geq \theta \geq \hat{\theta} - z_{\alpha/2}SD(\hat{\theta})$$

$$\Phi(z_{\alpha/2}) = 1 - \alpha/2$$



Johan Koskinen, Department of Statistics

2005-04-28

34

Inferens - intervallskattning

m.a.o. eftersom

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{SD(\hat{\theta})} \leq z_{\alpha/2}\right) = 1 - \alpha$$

är ekvivalent med

$$P[\hat{\theta}z_{\alpha/2} + SD(\hat{\theta}) \geq \theta \geq \hat{\theta} - z_{\alpha/2}SD(\hat{\theta})] = 1 - \alpha$$

är sannolikheten att

$$(\hat{\theta} - z_{\alpha/2}SD(\hat{\theta}), \hat{\theta} + z_{\alpha/2}SD(\hat{\theta}))$$

täcker θ precis $1 - \alpha$ (innan man har ett observerat värde...)



Johan Koskinen, Department of Statistics

2005-04-28

35

Inferens - konfidensintervall

För en vvr punktestimator för θ och med B_1 och B_2 sådant att

$$P(\hat{\theta} - B_1 < \theta < \hat{\theta} + B_2) = 1 - \alpha$$

ges ett $(1 - \alpha) \times 100\%$ -igt konfidensintervall för θ av

$$I_{\theta} = (\hat{\theta} - B_1, \hat{\theta} + B_2)$$

Framförallt för

$$\hat{\theta} \in N(\theta, SD(\hat{\theta})^2)$$

ges ett $(1 - \alpha) \times 100\%$ -igt konfidensintervall för θ av

$$\hat{\theta} \pm z_{\alpha/2}SD(\hat{\theta})$$

$$\Phi(z_{\alpha/2}) = 1 - \alpha/2$$



Johan Koskinen, Department of Statistics

2005-04-28

36

Hypotesprövning intro.



Hypotes angående något



Samla in data



Om/när 1 gäller hur troligt är det att observera det vi observerat (i data)?



Om det vi har observerat inte är särskilt troligt när 1 gäller: förkasta 1



Om det vi har observerat är troligt när 1 gäller:
förkasta ej 1



Johan Koskinen, Department of Statistics

2003-04-26

37

Hypotesprövning intro. ex.



Alla får i Derbyshire är vita



Data



Om/när 1 gäller hur troligt är det att observera det vi observerat (i data)? SVAR: omöjligt



Om det vi har observerat inte är särskilt troligt när 1 gäller: förkasta 1. Eftersom vi observerat svart får och svart får enligt 1 inte är möjligt förkastar vi 1



Om det vi har observerat är troligt när 1 gäller:
förkasta ej 1



Johan Koskinen, Department of Statistics

38