



Johan Koskinen, Statistiska institutionen, Stockholms universitet

Finansiell statistik, vt-05

F14 hypotesprövning, regressionsanalys

Hypotesprövning

Am. Air Transport Association vill utreda hur mycket genomsnittspassageraren har som handbagage. Flygplan anpassade för i genomsnitt $\mu = 5,4$ kilo och flygsäkerhetsverket oroade för säkerhet.



Låt X vara vikten på en slumpvis vald kabinväska i kilo

Stickprov ger ober. stokastiska variabler X_1, X_2, \dots, X_{144} ,

testa $H_0: \mu = 5,4$

mot $H_1: \mu \neq 5,4$ på signifikansnivån $\alpha = 0,05$



Johan Koskinen, Department of Statistics

2005-05-01

2

Hypotesprövning

Vi vet inget om fördelningen men teststatistikan

$$\frac{\bar{X} - \mu_0}{\sqrt{\text{Var}(X)/n}} = \frac{\bar{X} - 5,4}{\sqrt{\text{Var}(X)/144}} \stackrel{\text{approx.}}{\in} N(0,1) \quad \text{då } H_0: \mu = 5,4 \text{ är sann enligt CGS ty } n \text{ stort}$$

Vi vet ej heller $\text{Var}(X)$ utan skattar denna med stickprovsvariansen

$$\frac{\bar{X} - 5,4}{\sqrt{s^2/144}} \stackrel{\text{approx.}}{\in} N(0,1)$$

Vi förkastar $H_0: \mu = 5,4$ på signifikansnivån $\alpha = 0,05$ då det observerade värdet på teststatistikan är större än $z_{\alpha/2}$ eller mindre än $-z_{\alpha/2}$



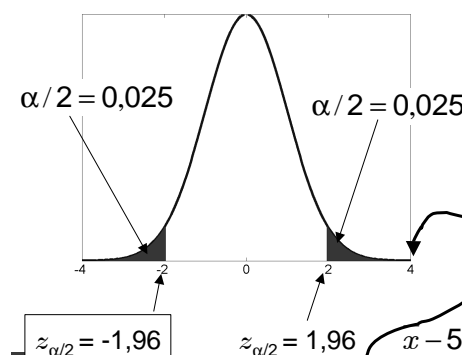
Johan Koskinen, Department of Statistics

2005-05-01

3

Hypotesprövning

De kritiska värdena $z_{\alpha/2}$ och $-z_{\alpha/2}$ ges av tabell som



Stickprov x_1, x_2, \dots, x_{144}

med medelvärde

$$\bar{x} = 6,62$$

och varians

$$s^2 = 12,52$$

det obs. värdet på teststatistikan

$$\frac{\bar{x} - 5,4}{\sqrt{s^2/144}} = \frac{6,62 - 5,4}{\sqrt{12,52/144}} = 4,14$$



Johan Koskinen, Department of Statistics

2005-05-01

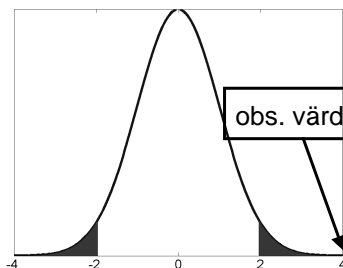
4

Hypotesprövning

Eftersom det observerade värdet på teststatistikan är större än den kritiska gränsen förkastar vi nollhypotesen att $\mu = 5,4$

Det har skett en förändring!

Alternativ: p -värdes ansats



testa $H_0: \mu = 5,4$

mot $H_1: \mu \neq 5,4$

$$\frac{\bar{X} - 5,4}{\sqrt{s^2 / 144}} \in N(0,1)$$

Förkasta H_0 : om sannolikheten att få ett obs. värde på teststatistikan lika långt ifrån $\mu = 5,4$ som det vi observerar är liten.



Johan Koskinen, U

$z_{\alpha/2} = 1,96$

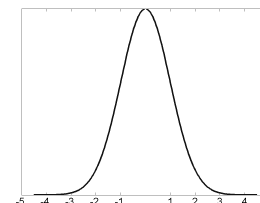
2005-05-01

5

Hypotesprövning - p -värde

$$\frac{\bar{X} - 5,4}{\sqrt{s^2 / 144}} \in N(0,1)$$

då $H_0: \mu = 5,4$ är sann enligt CGS ty n stort



Förkasta H_0 : om sannolikheten att få ett obs. värde på teststatistikan lika långt ifrån $\mu = 5,4$ som det vi observerar är liten.

$$\frac{\bar{x} - 5,4}{\sqrt{s^2 / 144}} = \frac{6,62 - 5,4}{\sqrt{12,52 / 144}} = 4,14$$

sannolikheten att få något lika långt ifrån $\mu = 5,4$ när H_0 sann = p -värde

$$p\text{-värde: } P\left(\left\{\frac{\bar{x} - 5,4}{\sqrt{s^2 / 144}} \geq 4,14\right\} \cup \left\{\frac{\bar{x} - 5,4}{\sqrt{s^2 / 144}} \leq -4,14\right\}\right)$$



Johan Koskinen, Department of Statistics

2005-05-01

6

Hypotesprövning - p -värde

$$P\left(\left\{\frac{\bar{x} - 5,4}{\sqrt{s^2 / 144}} \geq 4,14\right\} \cup \left\{\frac{\bar{x} - 5,4}{\sqrt{s^2 / 144}} \leq -4,14\right\}\right)$$

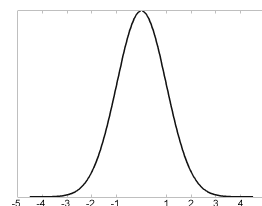
$$= P(\{Z \geq 4,14\} \cup \{Z \leq -4,14\})$$

$$= P(Z \geq 4,14) + P(Z \leq -4,14)$$

$$= P(Z \geq 4,14) + P(Z \geq 4,14)$$

$$= 2[1 - P(Z < 4,14)]$$

$$= 2[1 - \Phi(4,14)] \approx 0$$



om H_0 vore sann skulle sannolikheten att få ett obs. värde på teststatistikan lika långt ifrån $\mu = 5,4$ som 4,14 vara nästan 0: vi förkastar H_0 .



Johan Koskinen, Department of Statistics

2005-05-01

7

Hypotesprövning - p -värde

Använder man p -värden brukar man jämföra det observerade p -värdet med de vanliga signifikansnivåerna

Man efterkonstruerar på sätt och vis signifikansnivån!

Observera att man inte kan/skall tolka p -värdet som en sannolikhet.



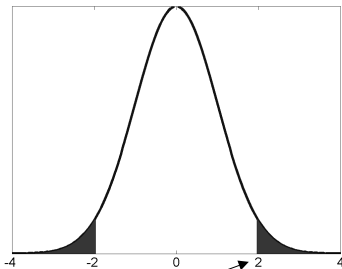
Johan Koskinen, Department of Statistics

2005-05-01

8

Hypotesprövning - k.i.

Alternativ: konfidenstervall ansats



$$z_{\alpha/2} = 1,96$$

Förkasta H_0 på signifikansnivån $\alpha = 0,05$: om nollhypotesens värde $\mu = 5,4$ inte täcks av k.i.

testa $H_0: \mu = 5,4$
mot $H_1: \mu \neq 5,4$ på signifikansnivån $\alpha = 0,05$

$$\bar{X} \overset{\text{approx.}}{\in} N(\mu, s^2 / 144)$$

Ett approximativt $(1 - \alpha) \times 100\% = 95\%$ -igt konfidenstervall för μ ges av

$$\bar{X} \pm z_{\alpha/2} \sqrt{\frac{s^2}{n}}$$

9

Hypotesprövning - k.i.

Observerat medelvärde

$$\bar{x} = 6,62 \quad \text{och varians} \quad s^2 = 12,52$$

Ett approximativt $(1 - \alpha) \times 100\% = 95\%$ -igt konfidenstervall för μ ges av

$$\bar{x} \pm z_{\alpha/2} \sqrt{\frac{s^2}{n}}$$

$$6,62 \pm 1,96 \sqrt{\frac{12,52}{144}} = 5,578$$

$$I_\mu = (6,04; 7,20)$$



Tolkning: medelvikten ligger med 95% konfidens i intervallet eftersom 5,4 inte täcks av itervalltet förkastar vi H_0 på 5%-nivån

10

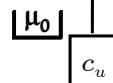
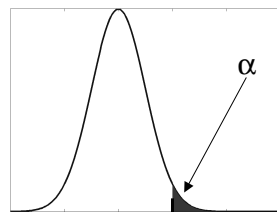
Mer om typ II fel

vi utgår ifrån att $H_0: \mu = \mu_0$ är sann

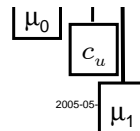
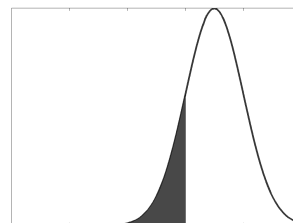
men om $\mu = \mu_1$ så att H_0 är falsk

$$\bar{X} \in N(\mu_0, \sigma^2 / \sqrt{n})$$

$$\bar{X} \in N(\mu_1, \sigma^2 / \sqrt{n})$$



vad är slh ej
förkasta H_0 ?
slh ej förkasta |
 \Leftrightarrow
slh medel minc
än kritisk gräns



$$P(\bar{X} \leq c_u | \bar{X} \in N(\mu_1, \frac{\sigma^2}{n}))$$

vi bestämmer $\alpha \rightarrow$ ger kritisk gräns c_u

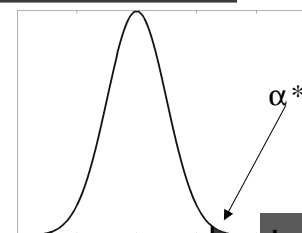
11

Mer om typ II fel

Hade vi satt α lägre, till α^* \rightarrow ger annan kritisk gräns c_u^*

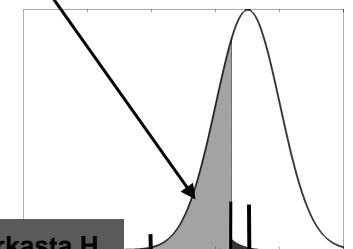
vi utgår ifrån att $H_0: \mu = \mu_0$ är sann

$$\bar{X} \in N(\mu_0, \sigma^2 / \sqrt{n})$$



men $\mu = \mu_1$

$$\beta = P(\bar{X} \leq c_u^* | \bar{X} \in N(\mu_1, \frac{\sigma^2}{n}))$$



ju mindre risk att förkasta H_0
vi tar desto mindre chans att
vi upptäcker att H_1 gäller



Johan Koski

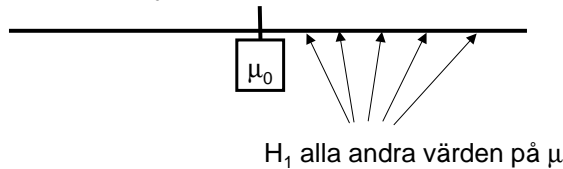
200

12

Mer om typ II fel - styrkefunktion

$$H_0: \mu \leq \mu_0$$

$$H_1: \mu > \mu_0$$



vi utgår ifrån att $H_0: \mu = \mu_0$ är sann

vi bestämmer $\alpha \rightarrow$ ger kritisk gräns c_u

För varje värde μ^* i H_1 blir sannolikheten för typ II fel olika

$$\beta = P(\bar{X} \leq c_u \mid \bar{X} \in N(\mu^*, \frac{\sigma^2}{n}))$$

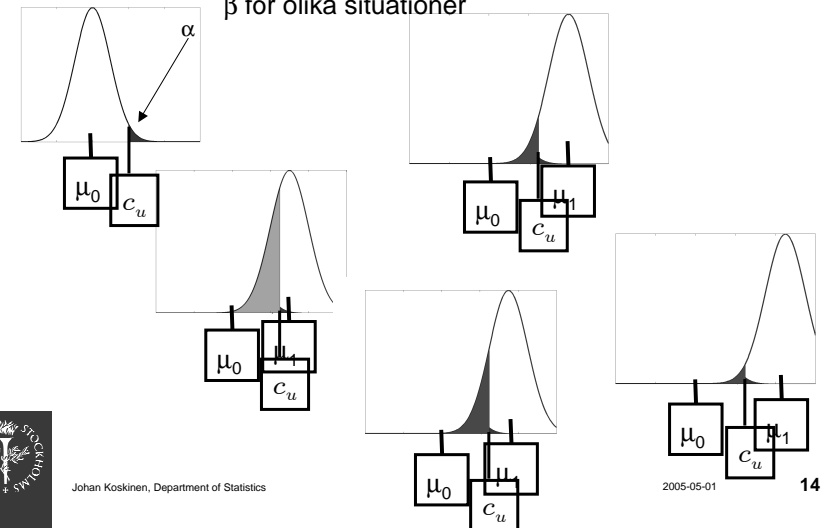


Johan Koskinen, Department of Statistics

13

Mer om typ II fel - styrkefunktion

β för olika situationer



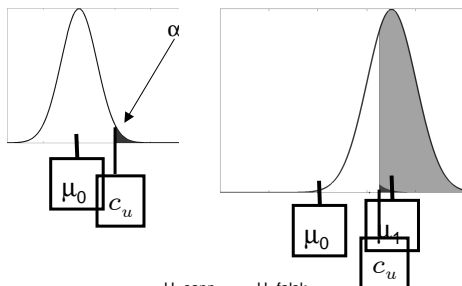
14

Mer om typ II fel - styrkefunktion

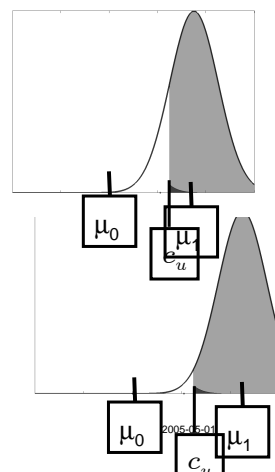
Styrkan för ett test:

sannolikheten att göra korrekt beslut givet ett parametervärde i H_1

$= 1 - \beta$ för olika situationer



	H_0 sann	H_0 falsk
handling	förkasta ej H_0	förkasta H_0
	rätt beslut	fel av typ I
	fel av typ II	rätt beslut



15

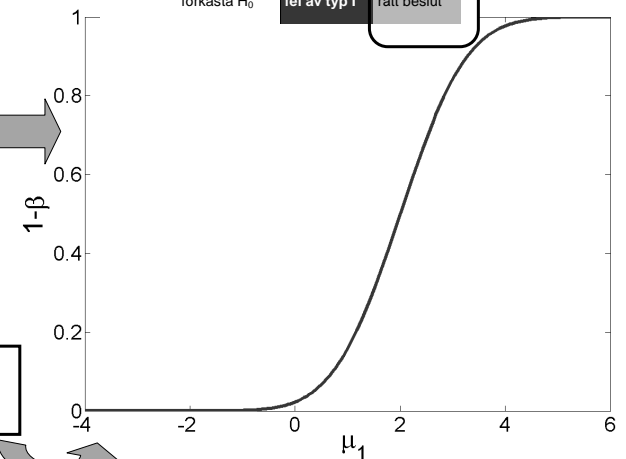
Mer om typ II fel - styrkefunktion

Styrkefunktionen

	H_0 sann	H_0 falsk
handling	förkasta ej H_0	förkasta H_0
	rätt beslut	fel av typ I
	fel av typ II	rätt beslut

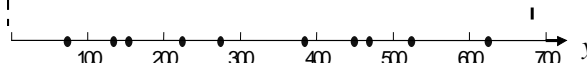
Skulle sannolikheten att göra korrekt beslut

Om det här var det sanna värdet



Regressionsanalys - gissa med medelvärden

Ex: Oljeförbrukning i liter

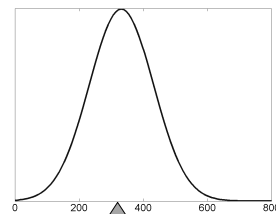


$$\bar{y} = 332$$

Vår bästa gissning av ett värde på y



Johan Koskinen, Department of Statistics

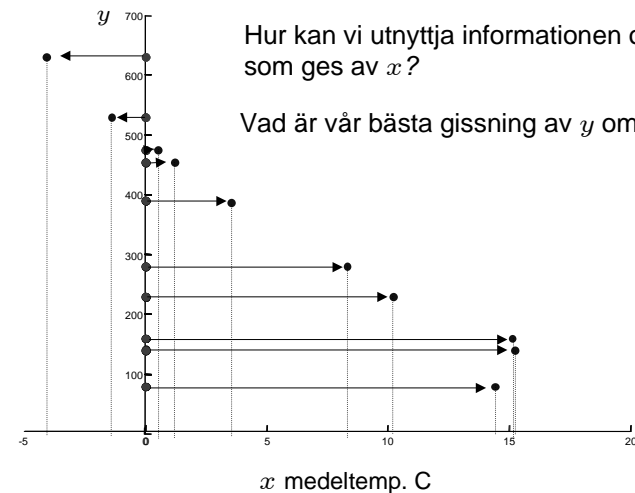


$$E(Y) = 332$$

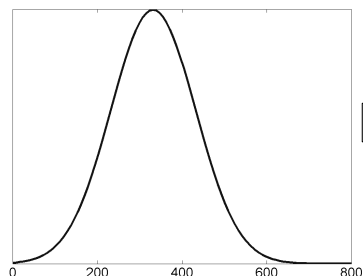
Vår bästa gissning
av ett värde på y

17

Regressionsanalys - gissa med medelvärden



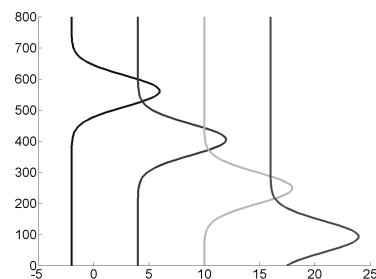
Regressionsanalys - gissa med medelvärden



I stället för t.ex. $Y_i \in N(\mu, \sigma^2)$



Johan Koskinen, Department of Statistics



En fördelning för varje värde
på x_i :

$$Y_i \in N(\mu_{x_i}, \sigma_{x_i}^2)$$

2005-05-01

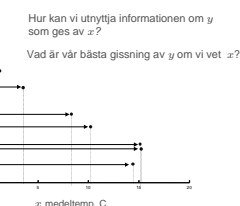
19

Regressionsanalys - mer info mindre osäkerhet

Data



Vår bästa gissning av ett värde på y



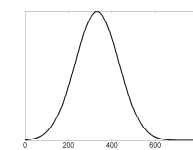
gissar med
medelvärden

tillför
information

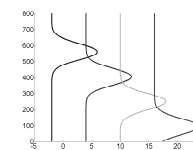
minskad
osäkerhet

vi gissar mindre
fel

Model



I stället för t.ex. $Y_i \in N(\mu, \sigma^2)$



En fördelning för varje värde
på x_i :

$$Y_i | x_i \in N(\mu_{x_i}, \sigma_{x_i}^2)$$

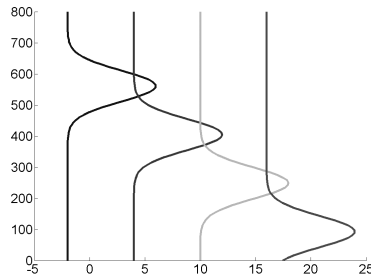
2005-05-01

20

Regressionsanalys

Om vi nu har en modell för data

$$Y_i \in N(\mu_{x_i}, \sigma_{x_i}^2)$$



borde vi kunna skatta
medelvärdena

$$\mu_{x_i}$$

med den vanliga estimatorn
stickprovsmedelvärdet

$$\bar{Y}_{x_i}$$

$$\bar{Y}_x = \frac{\sum y_j}{\# j \text{ som har } x_j = x}$$

kan vi göra när flera av
variablerna har samma x-
värde



Johan

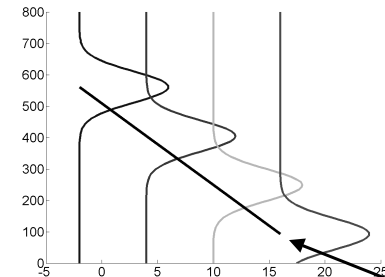
2005-05-01

21

Regressionsanalys

Modell för data

$$Y_i \in N(\mu_{x_i}, \sigma_{x_i}^2)$$



problem:

kanske endast en observation
per värde på x

vi får "många" skattningar - hur
tolka?

Lösning 1:

$$\mu_{x_i} = \beta_0 + \beta_1 x_i$$



Johan Koskinen, Department of Statistics

2005-05-01

22

Regressionsanalys

med den vanliga estimatorn: stickprovsmedelvärdet

"behöver" vi dock ober. likaförd. s.v. Y_1, Y_2, \dots, Y_n ,

Lösning 2: antag spridningen dvs variansen är lika för alla
 $i = 1, 2, \dots, n$.

Tolkning: fördelningen för Y givet värdet på β_0 , β_1 och x_i är

$$Y_i \in N(\mu_{x_i}, \sigma_{x_i}^2) = N(\beta_0 + \beta_1 x_i, \sigma^2)$$

För varje värde x_i är Y normalfördelad med värdvärde $\beta_0 + \beta_1 x_i$



Johan Koskinen, Department of Statistics

2005-05-01

23

Regressionsanalys

M.a.o., ökar man x_i en enhet ökar värdet för Y
med β_1 enheter om $\beta_1 > 0$

och minskar med β_1 enheter om $\beta_1 < 0$

när $x_i = 0$ är det förväntade värdet på Y , β_0

Modellen kan skrivas:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

där $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ är oberoende likafördelade $N(0, \sigma^2)$



Johan Koskinen, Department of Statistics

2005-05-01

24