EX | 7.2.1

Antag att vi har $X_1 \dots X_5$ från $Bin(\overset{n}{20}, p)$. Vi vill skatta p

$$E(X)^1 = 20p = M_1 = \frac{1}{n}\sum_i X_i = \bar{X} \longrightarrow 20p = \bar{X} \rightarrow p = \frac{\bar{X}}{20}$$

$\longrightarrow$ $\boxed{p = \frac{\bar{X}}{20}}$ för binomialfördelning

EX | $X \sim N(\mu, \sigma^2)$. Vi vill skatta μ och σ

$$\left.\begin{aligned} E(X^1) &= M_1 \\ E(X^2) &= M_2 \end{aligned}\right\} \text{Ekvationssystemet vi ska lösa!}$$

$$\left.\begin{aligned} \mu &= \bar{X} \\ \sigma^2 + \mu^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}\right\} \rightarrow \left.\begin{aligned} \mu &= \bar{X} \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 \end{aligned}\right\} \rightarrow$$

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

MAXIMUM LIKELIHOOD METODEN (ML)

$$f_{X_1 \dots X_n}(x_1, \dots x_n) = \prod_{i=1}^{n} \underbrace{f_{X_i}(x_i)}_{f(x_i)} = \underbrace{\prod_{i=1}^{n} f(x_i)}_{\text{Likelihood}}$$

DEFINITION: $L(\theta) = \prod_{i=1}^{n} f(x_i)$ kallas likelihood

DEFINITION: $\hat{\theta}_{ML} = \arg\max L(\theta)$ är ML skattaren.

Det är ofta lättare att beräkna $\ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i)$

$\boxed{\frac{\partial}{\partial \theta} \ln L(\theta) = 0} \rightarrow$ maximum likelihoodmetoden

EX) Antag $X_1 \dots X_n$ är ett stickprov från $X \sim Po(\mu)$

$$\ln L(\theta) = \sum_{i=1}^{n} \ln \underbrace{\left(\frac{\mu^{x_i}}{x_i!} \cdot e^{-\mu} \right)}_{P_X(x_i)} = \sum_{i=1}^{n} (x_i \ln \mu - (-\mu) - \ln(x_i!)) =$$

$$= \ln \mu \sum_{i=1}^{n} x_i - n\mu - \sum_{i=1}^{n} \ln(x_i!)$$

$$\frac{\partial}{\partial \mu} \ln L(\mu) = \frac{1}{\mu} \sum_{i=1}^{n} x_i - n = 0 \longrightarrow \mu = \underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i}_{} = \bar{X}$$

stickprovsmedelvärdet brukar vara en bra skattare.

EX) Antag $X_1 \dots X_n$ stickprov från $X \sim N(\mu, \sigma^2)$

$$\ln L(\mu, \sigma) = \sum_{i=1}^{n} \ln \left(\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \right) = \dots = -n\sqrt{2\pi} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

(1) $\frac{\partial \ln L(\mu)}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$ (2.) $\frac{\partial}{\partial \sigma} \ln L(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$

Lös ut μ och σ^2

(1.) $\frac{\partial}{\partial \mu} \ln L(\mu) = 0 \rightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{X}$

Stoppa in (1.) i (2.) $\rightarrow \frac{\partial}{\partial \sigma} \ln L(\sigma) = 0 = \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \bar{X})^2 \rightarrow$

$$\rightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2$$

MATEMATISK STATISTIK F8 10/4 2014

TVÅSIDIGT KONFIDENSINTERVALL

DEFINITION: Låt $X_1 \dots X_n$ vara slumpvariabler med fördelningen som har en parameter θ med ett sant okänt värde θ_0.

Ett konfidensintervall för θ med konfidensgrad $1-\alpha$ är:

$[a(X_1 \dots X_n), b(X_1 \dots X_n)]$ så att $P(a < \theta_0 < b) = 1-\alpha$

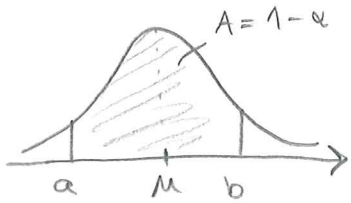
KONFIDENSINTERVALL FÖR μ I $N(\mu, \sigma^2)$ DÅ σ^2 ÄR KÄND

$X_1 \dots X_n \sim N(\mu, \frac{\sigma^2}{n}) \xrightarrow{*} \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ om σ^2 är känd.

kom ihåg: σ^2-regeln

Om $Y \sim N(\mu, \sigma^2) \rightarrow P(\mu - z_{\alpha/2}\sigma < Y < \mu + z_{\alpha/2}\sigma) = 1-\alpha$

$\boxed{\text{Summan av normalfördelade variabler är också normalfördelad.}}$ *



SATS

Om $X_i \sim N(\mu, \sigma^2)$ och σ^2 är känd så är:

$I_\mu = (\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ ett konf-intervall för μ med konf-grad $1-\alpha$

BEVIS

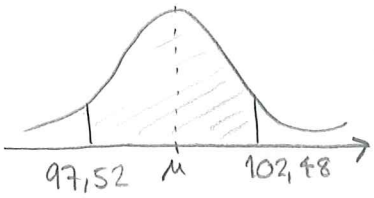
$P(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$ vilket är samma som $\rightarrow$

$\rightarrow P(-\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$ —— || —— $\rightarrow$

$\rightarrow P(\underbrace{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{punkt a}} < \mu < \underbrace{\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{punkt b.}}) = 1-\alpha$

EX Antag $X \sim N(\mu, 16)$ $n=10$ $\bar{X}=100$ $\alpha=0{,}05 \rightarrow z_{\alpha/2}=1{,}96$

$$I_\mu = \left(100 - 1{,}96 \times \frac{4}{\sqrt{100}},\ 100 + 1{,}96 \times \frac{4}{\sqrt{100}}\right) = (97{,}52,\ 102{,}48)$$



* ju mindre α dessto större sln att intervallet täcker det sanna värdet
* ju mindre α dessto bredare intervall

KONFIDENSINTERVALL FÖR μ i $N(\mu, \sigma^2)$ DÅ σ^2 OKÄND

kom ihåg: Om $X \sim N(\mu, \sigma^2) \rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$

SATS: Antag $X_1 \dots X_n \sim N(\mu, \sigma^2)$ Då är $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

T-fördelad med n-1 frihetsgrader

Då $n \rightarrow \infty$ i T-fördelningen går

$t(n-1) \rightarrow N(\mu, \sigma)$

Täthetsfunktionen för T-fördelning

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{2}\right)^{-\frac{n+1}{2}}$$

Γ = gammafunktioner

Vi kommer inte behöva kunna dessa. Men bra att se

SATS Om $X_1 \dots X_n \sim N(\mu, \sigma^2)$ där σ^2 är okänd så är:

$$I_\mu = \left(\bar{X} - \underbrace{t_{\alpha/2}(n-1)}_{\alpha/2 \text{ kvantiler för } t(n-1) \text{ fördelningen finns i tabeller i boken.}} \frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)$$

ett konfidensintervall m. konfgrad $1-\alpha$

T-fördelningskvantilen

$t_{0,025}(\infty) = z_{0,025} = 1{,}96$ alltså $t_{0,025} \rightarrow z_{0,025}$ då $n \rightarrow \infty$

$t_{0,025}(1) = 12{,}71$

extremt dåligt fall ettersom $n-1=1$ ger $n=2$!

CENTRALA GRÄNSVÄRDES SATSEN (CGS)

$\bar{X} \overset{\mathcal{E}}{\sim} N(\mu, \frac{\sigma^2}{n})$ detta ger då ett approximativt konf.intervall för μ då:

För σ känt

$$I_\mu = (\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

För σ okänd

$$I_\mu = (\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}})$$

KONFIDENSINTERVALL FÖR σ^2

kom ihåg: $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ → Vi behöver alltså hitta ett motsvarande uttryck för att få σ^2

DEFINITION χ^2-fördelning (chi-två fördelning)

Om $Z_i \sim N(0,1)$ så gäller $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$

χ^2-fördelad med n frihetsgrader.

Täthetsfunktionen för χ^2-fördelning

$$f(x) = \begin{cases} \frac{x^{n/2-1} \cdot e^{-x/2}}{2^{n/2} \cdot \Gamma(\frac{n}{2})} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Behöver inte kunna denna men bra att se.

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1) \quad \text{om} \quad X_i \sim N(\mu, \sigma^2)$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n)$$

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

Vi tappar en frihetsgrad ettersom vi skattar μ med $\bar{X}$

Låt $\chi^2_\alpha(n)$ vara α-kvantilen i $\chi^2(n)$-fördelningen.

$$P\left(\chi^2_{1-\alpha/2}(n-1) \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha$$

Vi vill lösa ut σ^2 för att få ut konf-intervallet för σ^2

$$P\left(\underbrace{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}} \leq \sigma^2 \leq \underbrace{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}}\right) = 1-\alpha$$

våra gränsvärden i konf, intervallet

SATS Om $X_i \sim N(\mu, \sigma^2)$ så är

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

ett konf, intervall för σ^2

av av

Man tar bara roten ur gränserna för att få σ

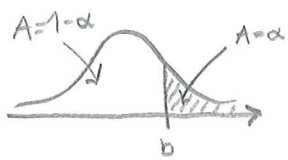
$$I_\sigma = \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}}\right)$$

konf, intervall för σ^2

Värden för χ^2-kvantiler finns i boken.

NOTERA! Dessa konfidensintervall för σ^2 måste vara normalfördelade.

ENSIDIGT KONFIDENSINTERVALL

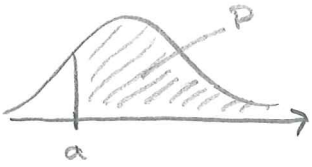


Tex. hur stor/liten kan fördelningen vara
Alltså finna övre/undre gräns

EX|

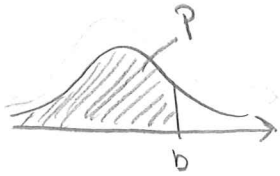
Ett undre begränsat konfidensintervall för θ ges av:

$(a(X_1 \dots X_n), \infty)$ Där $P(a < \theta_0) = 1-\alpha$



Ett övre begränsat konfidensintervall för θ ges av:

$(-\infty, b(X_1 \dots X_n))$ Där $P(\theta_0 < b) = 1-\alpha$



EX| Ett ensidigt konf.intervall för μ i $N(\mu, \sigma^2)$

Med känt σ^2 och övre begränsat

$I_\mu = (-\infty, \bar{X} + t_\alpha(n-1)\frac{S}{\sqrt{n}})$

Med okänt σ^2 och undre begränsat

$I_\mu = (\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$

Skillnaden mellan en- och tvåsidigt intervall är att man kvantilerna från $\alpha/2$ till α för ensidigt och man sätter undre gränsen till $-\infty$ om man söker ett övre begränsat ensidigt intervall, och tvärtom för undre begränsat.

MATEMATISK STATISTIK F9 5/5 2014

HYPOTESPRÖVNING

EXEMPEL: Hypotesprövningsfrågeställningar:

* Gör ett nytt läkemedel någon effekt?
* Dör rökare än icerökare?
* Har mättinstrument ett systematiskt fel?

Vi vill testa nollhypotesen $H_0: \theta = \theta_0$ mot mothypotesen $H_1: \theta \neq \theta_0$
Om testet tex bekräftar ett systemfel så förkastar vi H_0 till förmån för $H_1 \rightarrow \theta$ är signifikant skilld från θ_0

EX 1. En nyanställd mäter klorhalten 5 ggr

56,7 61,3 58,5 62,1 59,5

Sann klorhalt är 60 mg/l

Kan vi säga att den nyanställda systematiskt mäter fel?

$\bar{X} = 59,62$ $S^2 = 4,692$

Antag $X_i \sim N(\mu, \sigma^2)$

Vi vill testa: $H_0: \mu = 60$ $H_1: \mu \neq 60$

KONFIDENSINTERVALLMETODEN

Om vi bildar ett 95% konfidensintervall I_θ för θ kan vi direkt använda detta för att göra hypotestestet.

- Eftersom intervallet är beräknat så att det ska täcka det sanna värdet på θ i 95% av fallen så kan vi förkasta H_0 om intervallet inte innehåller θ_0
- Om intervallet däremot innehåller θ_0 kan vi inte förkasta H_0 dvs påstå $\theta \neq \theta_0$

Konfidensgraden vi använder när vi beräknar konfidens-intervallet är också konfidensgraden för hypotestestet vi gör.

- Det har blivit standard inom många områden att använda 95% konfidensgrad, men vi kan också använda högre konfidensgrad (tex 99%) eller en lägre konfidensgrad (tex 90%)

FORTS EX 1

Ett 95% konfidensintervall $\longrightarrow \alpha = 0{,}05$

$I_\mu = (\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}) = (56{,}93 \quad 62{,}31)$

$60 \in I_\mu \longrightarrow$ Vi kan ej förkasta H_0

DEFINITION Felrisk/signifikansnivå Typ 1. fel

Felrisken eller signifikansnivån definieras som:

$\alpha = P(H_0 \text{ förkastas} \mid H_0 \text{ sann})$

Översatt till konfidensintervall är signifikansnivån:

$$\alpha = P(\theta \notin I_\theta \text{ om } \theta = \theta_0) = P(\theta_0 \notin I_\theta)$$

Alltså fås signifikansnivån som 1 - konfidensgraden.

DEFINITION Typ 2 fel

Under hypotesprövning säger vi att vi gör ett typ 2 fel om vi inte förkastar H_0 trots att H_1 är sann. Sannolikheten för att göra ett typ 2 fel brukar betecknas med β

$$\beta = P(H_0 \text{ förkastas inte} \mid H_1 \text{ sann})$$

OLIKA UTFALL I HYPOTESTEST:

1. H_0 sann och hypotestestet förkastar inte H_0
2. H_0 sann och ——— " ——— förkastar H_0
   Detta är ett typ 1 fel och har enligt konstruktionen SLH α att inträffa
3. H_1 är sann och hypotestestet förkastar H_0
4. H_1 är sann och hypotestestet förkastar inte H_0.
   Detta är ett typ 2 fel och SLH för att detta inträffar brukar betecknas med β

> I boken kallar dom hypotestest med konfidensintervall för hypotesprövning och det med teststorhet för signifikansprövning

DEFINITION Teststorheter

En teststorhet $T = T(X_1 \dots X_n)$ är en funktion av observationerna, och alltså en slumpvariabel. $T_{obs} = T(x_1 \dots x_n)$ är ett observerat värde av teststorheten för givna observationer.

I allmänhet innehåller teststorheten en skattning θ^* av parametern θ. Tex för test av väntevärdet i en normalfördelning med:

Okänd varians:

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

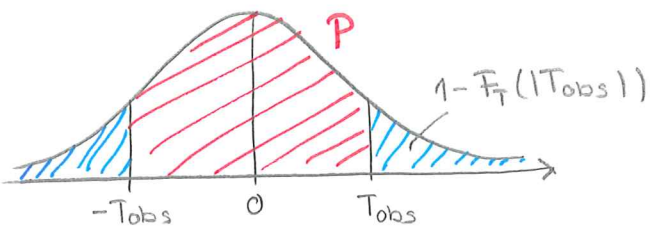
känd varians:

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

DEFINITION P-värde

P-värdet eller signifikanssannolikheten definieras som sannolikheten under nollhypotesen att vi får ett värde $|T|$ som är lika stort eller större än det observerade värdet $|T_{obs}|$

Symmetrisk fördelning



$$P = 2(1 - F_T(|T_{obs}|))$$

EX

- Om $0{,}01 < P < 0{,}05$ kallas det enstjärnigsignifikant (*)
- Om $0{,}001 < P < 0{,}01$ —— " —— två —— " —— (**)
- Om $P < 0{,}001$ —— " —— tre —— " —— (***)

Ju högre "stjärnfaktor" dessto större SLH att vi förkastar rätt

FORT EX 1

$$T = \frac{\bar{X} - 60}{s/\sqrt{n}} \sim t(n-1)$$

$$T_{obs} = \frac{59{,}62 - 60}{\sqrt{4{,}692}/\sqrt{5}} = -0{,}3933$$

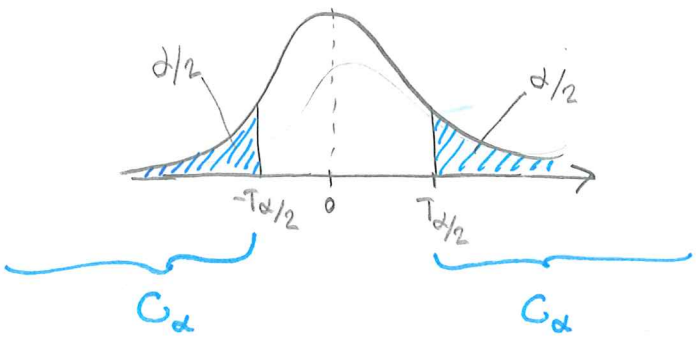
$P = 2(1 - F_{t(4)}(0{,}3933)) = 0{,}71 \longrightarrow$ minsta felrisken vi kan förkasta H_0 på är 71%

$P > 0{,}05 \longrightarrow$ förkastar ej H_0

DEFINITION Kritiskt område

Givet en signifikansnivå α definierar vi det kritiska området C_α som de värden på teststorheten T som leder till att man förkastar H_0 på nivå α.

Förkastas på nivå α $\quad p < \alpha \rightarrow 1 - F_T(|T_{obs}|) < \alpha/2$



C_α = kritiska området. Befinner sig T_{obs} innom detta området så förkastas H_0 vilket är ekvivalent med att säga att om vi hamnar utanför konfidensintervallet så förkastas H_0

KRITISKT OMRÅDE FÖR NORMALFÖRDELNING

Med test för väntevärdet hor normalfördelade data får vi

- Om variansen är känd: T under H_0 är $N(0,1)$-fördelad förkastas H_0 på nivån α om $|T| > z_{\alpha/2}$
- Om variansen är okänd: T under H_0 är $t(f)$-fördelad, förkasta H_0 på nivå α om $|T| > t_{\alpha/2}(f)$

$$C_\alpha = \{T : |T| > z_{\alpha/2}\} \rightarrow \text{Förkasta } H_0 \text{ om } |T| > z_{\alpha/2}$$

$$|T| = \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| \longrightarrow \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2} \longleftrightarrow (\bar{x} - \mu_0) > z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \longleftrightarrow$$

$$\longleftrightarrow \left\{\begin{array}{l} \mu_0 < \bar{x} - z_{\alpha/2}\,\sigma/\sqrt{n} \\ \mu_0 > \bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n} \end{array}\right\} \longleftrightarrow \mu \notin I_\mu = \left(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Här ifrån kommer kopplingen mellan hypotesprövningen och konfidensintervall

FORTS EX 1

$C_\alpha = \{T : |T| > t_{\alpha/2}(4)\}$ så $\alpha = 0{,}05 \rightarrow C_\alpha = \{T : |T| > 2{,}77\}$

ENSIDIGT TEST

Nollhypotesen $H_0: \theta = \theta_0$ testas mot antingen

Mothypotesen $H_1: \theta > \theta_0$ eller $H_1: \theta < \theta_0$

Använder vi konfidensintervall så:

- Om $H_1: \theta > \theta_0$ $I_\theta = \{a, \infty\}$
- Om $H_1: \theta < \theta_0$ $I_\theta = \{-\infty, b\}$

Förkastar H_0 om intervallet inte täcker θ_0

Använder vi teststorhet så:

- Om $H_1: \theta > \theta_0$ $P = P_{H_0}(T \geq T_{obs})$
- Om $H_1: \theta < \theta_0$ $P = P_{H_0}(T \leq T_{obs})$

Förkastar sedan H_0 om $p \leq \alpha$

KRITISKT OMRÅDE – ENSIDIGT

Om $H_1: \theta > \theta_0$

- Om σ är känd $\rightarrow$ förkasta H_0 på nivå α om $T > z_\alpha$
- Om σ är okänd $\rightarrow$ förkasta H_0 på nivå α om $T > t_\alpha(f)$

Om $H_1: \theta < \theta_0$

- Om σ är känd $\rightarrow$ ——— " ——— $T < -z_\alpha$
- Om σ är okänd $\rightarrow$ ——— " ——— $T < -t_\alpha(f)$

DEFINITION Styrkefunktion

Vi vill testa $H_0: \theta = \theta_0$ och har bestämt en teststorhet T och en signifikansnivå α. Testets styrkefunktion $\pi(\theta)$ defineras som:

$$\pi(\theta) = P(T \in C_\alpha \mid \theta) = P(H_0 \text{ förkastas} \mid \text{parametervärdet är } \theta)$$

SATS

Antag att vi har n observationer av en normalfördelning $N(\mu, \sigma^2)$ med känd varians och att vi vill testa $H_0: \mu = \mu_0$

Testets styrka ges då av

Om $H_1: \mu > \mu_0$

$$\pi(\mu) = 1 - \Phi\left(z_\alpha - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right)$$

Om $H_1: \mu < \mu_0$

$$\pi(\mu) = \Phi\left(z_\alpha - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right)$$

Om $H_1: \mu \neq \mu_0$

$$\pi(\mu) = 1 + \Phi\left(-z_{\alpha/2} - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right) - \Phi\left(z_{\alpha/2} - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}\right)$$

SATS

Antag att vi har normalfördelade observationer $N(\mu, \sigma^2)$ och önskar testa $H_0: \mu = \mu_0$. Om vi vill att testet ska ge utslag för avvikelsen $\mu - \mu_0$ med sln $1-\beta$ ska antal observationer väljas som:

* känd varians:

Ensidigt $\quad n = \frac{\sigma^2}{(\mu-\mu_0)^2}(z_\alpha + z_\beta)^2$

Tvåsidigt $\quad n = \frac{\sigma^2}{(\mu-\mu_0)^2}(z_{\alpha/2} + z_\beta)^2$

* Om variansen skattas med s^2

Ensidigt $\quad n \approx \frac{\sigma^2}{(\mu-\mu_0)^2}(z_\alpha + z_\beta)^2 + \frac{z_\alpha^2}{2}$

Tvåsidigt $\quad n \approx \frac{\sigma^2}{(\mu-\mu_0)^2}(z_{\alpha/2} + z_\beta)^2 + \frac{z_{\alpha/2}^2}{2}$

MATEMATISK STATISTIK F10 8/5-2014

"JÄMFÖRELSER"

NOLLHYPOTES: $H_0: \mu = \theta$ eller $H_0: \mu \geq \theta$ osv.

MOTHYPOTES: $H_1: \mu \neq \theta$ $H_1: \mu < \theta$

Sannolikheten att inte göra ett **Typ 2 fel** kallas testets styrka och ges av $1-\beta$ där $\beta = P(H_0 \text{ förkastas ej} \mid H_1 \text{ sann})$

P-värdet ger minsta signifikansnivå som vi kan förkasta H_0 på.

OBEROENDE STICKPROV

Antag att vi har:

$X_{11} \dots X_{1n_1} \sim N(\mu_1, \sigma_1^2)$

Vi vill testa om $\mu_1 = \mu_2$!

$X_{21} \dots X_{2n_2} \sim N(\mu_2, \sigma_2^2)$

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} \qquad S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2$$

$$\bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i} \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2$$

Definiera $\theta = \mu_1 - \mu_2$

Skattning av $\theta \longrightarrow \theta^* = \bar{X}_1 - \bar{X}_2$

Vi vill testa: $H_0: \theta = 0$ $(\mu_1 = \mu_2)$ $H_1: \theta \neq 0$ $(\mu_1 \neq \mu_2)$

I regel har vi tre olika fall:

1) σ_1 och σ_2 kända

2) $\sigma_1 = \sigma_2 = \sigma$ okända

3) $\sigma_1 \neq \sigma_2$ samt okända

HUR GÖRS TESTET I DE OLIKA FALLEN?

1) σ_1 och σ_2 okända

$$V(\theta^*) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$ (ty oberoende)

$$I_\theta = (\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{V(\theta^*)})$$

Eftersom de var normalfördelade

För hypotestest

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{V(\theta^*)}} \sim N(0,1)$$ om H_0 är sann

Ex 1

Antag att vi har 2 stickprov

$\bar{X}_1 = 2{,}45 \quad n_1 = 10$

$\bar{X}_2 = 3{,}23 \quad n_2 = 20$

Vi känner till att $\sigma_1 = \sigma_2 = 1/2$

$$V(\theta^*) = \frac{(1/2)^2}{10} + \frac{(1/2)^2}{20} = 0{,}1936^2$$

$$I_\theta = (2{,}45 - 3{,}23 \pm 1{,}96 \times 0{,}1936) = (-1{,}9, -0{,}4)$$

$$T_{obs} = \frac{2{,}45 - 3{,}23}{0{,}1936} \approx -4$$

$$P = 2(1 - \Phi(4)) = 5 \times 10^{-5}$$

H_0 förkastas på väldigt hög signifikansnivå

2) $\sigma_1 = \sigma_2 = \sigma$ okänd

Om σ_1 och σ_2 är okända, skatta $V(\theta^*)$ med:

$$S^2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

Eftersom $\sigma_1 = \sigma_2$ är S_1 och S_2 2 olika skattningar av samma varians.

SATS

En väntevärdesriktig skattning av σ^2 från k normalfördelningar $N(\mu, \sigma^2)$ fås som.

$$S_p^2 = \frac{(n_1-1)\cdot S_1^2 + (n_2-1)\cdot S_2^2 + \ldots + (n_k-1)S_k^2}{(n_1-1)+(n_2-1)+\ldots+(n_k-1)}$$

Står för polad

totala antalet observationer $N = \sum_{i=1}^{k} n_i$

Dessutom gäller att $\frac{(N-k)S_p^2}{\sigma^2} \sim \chi^2(N-k)$

BEVIS

Visa att $E(S_p^2) = \sigma^2$ följer av att $E(S_i^2) = \sigma^2$

Vi har att $\frac{(n_i-1)S_i^2}{\sigma^2} \sim \chi^2(n_i-1)$

$\chi^2(N-k)$ följer av additionssatsen för χ^2-fördelningen

Satsen säger att $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$

Vi skattar $V(\theta^*)$ med: $S_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{1/n_1 + 1/n_2}}$ är t-fördelad ty skattad $t(n_1+n_2-2)$ under H_0

EX 2.) samma stickprov som i Ex 1 men vi vet även:

$s_1^2 = 0{,}57^2$ $s_2^2 = 0{,}48^2$ Vi vet att $\sigma_1 = \sigma_2 = \sigma$

Beräkna: $s_p^2 = \dfrac{9 \times 0{,}57^2 + 19 \times 0{,}48^2}{9+19} = 0{,}5107^2$

$V(\theta^*) = 0{,}5107^2 \left(\frac{1}{10} + \frac{1}{20}\right) = 0{,}1978^2$

$I_\theta = \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2\,(28)} \sqrt{V(\theta^*)} = (-1{,}18\ ,\ -0{,}38)$

$T_{obs} = \dfrac{2{,}45 - 3{,}23}{0{,}1978} = -3{,}94$

$P = 2\left(1 - F_{t\,(28)}(|T_{obs}|)\right) = 0{,}0004$ Väntevärdena är skillda

Vi kan förkasta H_0 på en hög signifikans nivå.

3.) $\sigma_1 \neq \sigma_2$ och okända

$V(\theta^*) = \dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}$ Vi kan ej pola stickprovsvariansen ty $\sigma_1 \neq \sigma_2$

SATS

Om σ_1 och σ_2 okända så gäller att:

$T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{V(\theta^*)}}$ är approximativt t-fördelad med f frihetsgrader $t_{(f)}$

$$f = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$ = skattning av antalet frihetsgrader

* Om $S_1^2/n_1 \gg S_2^2/n_2$ blir antalet frihetsgrader $n_1 - 1$

* Om $S_1 \approx S_2$ och $n_1 \approx n_2 \approx n$ blir $f \approx 2(n-1)$

EX 3.

samma värden som i EX 2 fast $\sigma_1 \neq \sigma_2$

Skatta: $V(\theta^*) \approx \frac{0{,}57^2}{10} + \frac{0{,}48^2}{20} = 0{,}2098^2$

$$f = \frac{\left(\frac{0{,}57^2}{10} + \frac{0{,}48^2}{20}\right)^2}{\frac{\left(\frac{0{,}57^2}{10}\right)^2}{9} + \frac{\left(\frac{0{,}48^2}{20}\right)^2}{19}} = 15{,}5$$

Jämför med den polade stickprovsvariansskattningen $S_p^2 = 0{,}1978^2$, vilket är nära

Alltså hälften så många frihetsgrader som i EX 2.

$$I_\theta = (\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}(15) \cdot 0{,}2098) = (-1{,}23, -0{,}33)$$

$$T_{obs} = -3{,}718$$

$$P = 2(1 - F_{t(15)}(3{,}718)) = 0{,}002$$

kan förkasta H_0 på hög signifikansnivå

~~JÄMFÖRELSE AV VI~~

JÄMFÖRELSE AV VARIANSEN

Vi vill testa $H_0: \sigma_1^2 = \sigma_2^2$ $\quad H_1: \sigma_2^1 \neq \sigma_2^2$

Inför teststorheten: $T = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ (*)

Vi vet att $(n_1-1)S_1^2/\sigma^2$ är $\chi^2(n_1-1)$-fördelad

Alltså måste T-kvoten (*) vara en kvot mellan 2 χ^2-fördelningar.

DEFINITION

Om vi har $V_1 \sim \chi^2(f_1)$, $V_2 \sim \chi^2(f_2)$ så är $\frac{V_1/f_1}{V_2/f_2} \sim F(f_1, f_2)$ (F-fördelad)

täthetsfunktion finns på slidsen.

F-fördelningstabell finns i boken!

EX 4

Samma data som i EX3.

$$I_{\sigma_1^2/\sigma_2^2} = \left[\frac{0{,}57^2/0{,}48^2}{F_{\alpha/2}(9,19)}, \frac{0{,}57^2/0{,}48^2}{F_{1-\alpha/2}(9,19)}\right] = \underbrace{[0{,}48 ; 5{,}19]}$$

Alltså kvoten kan vara "lite varsom helst"

$I_{\sigma_1^2/\sigma_2^2}$ är mycket känsligt! Måste vara normalfördelat

$1 \in I_{\sigma_1^2/\sigma_2^2} \longrightarrow$ kan ej förkasta $H_0 \longrightarrow$

$\longrightarrow$ Rimmligt att använda polad stickprovsvarians

Om ej Normalfördelad $\longrightarrow$ vi måste veta i vilket fall (1,2,3) vi befinner oss i.

STICKPROV I PAR

Antag:

Hörihop

Hörihop

$X_{11} \ldots X_{1n} \sim N(\ ,\ )$

$X_{21} \ldots X_{2n} \sim N(\ ,\ )$

samman antal

Alltså, vi har förvandat 2 stickprov till 1

Bilda $D_i = X_{1i} - X_{2i} \sim N(\Delta, \sigma^2)$ ty samma variation

Vi vill testa $H_0: \Delta = 0$ $H_1: \Delta \neq 0$

tex. mäter vikt före/efter bantningskur hos n antal personer.

MATEMATISK STATISTIK F11 12/5-2014

"JÄMFÖRELSER OCH INFERENS"

PUNKTSKATTNING

$\theta^* \pm$ "Någon kvantil" $\cdot \sqrt{V(\theta^*)}$

POOLAD SKATTNING AV GEMENSAM VARIANS:

Om vi antar att variansen är lika kan vi vinna en del precision på att använda data från båda stickproven för att skatta den gemensama variansen. En skattning när vi använder data från flera stickprov för att skatta variansen kallas "pooled estimate"

SATS

En väntevärdesriktig skattning av ett gemensamt σ^2 från k olika normalfördelningar $N(\mu_j, \sigma^2)$ fås genom den sammanvägda skattningen.

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \ldots + (n_k-1)S_k^2}{(n_1-1)+(n_2-1)+\ldots+(n_k-1)}$$

Dessutom gäller att $(N-k)S_p^2/\sigma^2$ är $\chi^2(N-k)$-fördelad där $N = \sum_{i=1}^{k} n_i$

STICKPROV I PAR

Vanliga situationer då mätningar uppkommer i par:

* Studera hur mkt personer ökar i vikt då de slutar röka
* Systematiska skillnader mellan två mätmetoder och använder varje metod på var och ett av ett antal prover och jämför de två metoderna för varje prov

För varje prov kan vi bilda differensen D_i som antas vara $N(\mu, \sigma^2)$

$$D_i = X_i - Y_i \sim N(\Delta, \sigma^2)$$

Vi vill testa $\Delta = 0$, vilket är vanligt hos normalfördelning

FÖR- OCH NACKDELAR M. STICKPROV I PAR

Ofta är det mer effektivt att använda stickprov i par än oberoende stickprov, särskilt om variationen mellan mätningarna är stor.

Vi kan dela upp variationen i D_i som $\sigma^2 = \sigma_0^2 + \sigma_\Delta^2$ där σ_0^2 beskriver variationen mellan objekten.

Vid stickprov i par har vi:

$$V(\bar{D}) = 2\sigma_\Delta^2 / n$$

Vid oberoende stickprov har vi:

$$V(\bar{X} - \bar{Y}) = 2(\sigma_0^2 + \sigma_\Delta^2)/n$$

Fördelar

Vi vinner på metoden om $\sigma_0^2 > 0$

Nackdelar

Vi förlorar frihetsgrader
$\longrightarrow$ större osäkerhet

INFERENS FÖR DISKRETA DATA Kap 9

BINOMIALFÖRDELNINGEN

Antag att vi observerar $X \sim Bin(n, p)$

antal försök

Parametern vi söker

$p^* = \frac{x}{n}$ är en naturlig skattning av p.

$E(X) = n \cdot p \Longrightarrow E(p^*) = \frac{np}{n} = p$ alltså en väntevärdesriktig skattning

$V(X) = n \cdot p(1-p) \Longrightarrow V(p^*) = \frac{p^*(1-p^*)}{n}$

Väldigt svårt att ta fram ett exakt konfidensintervall för Bin. Vanligt att normalapproximera, om n är stort

Tumregel: Om $np(1-p) > 10$ så är normalapproximation ok!

P^* är approximativt $N\left(p, \frac{p(1-p)}{n}\right)$ - fördelad

$$T = \frac{p^* - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{\sim}{\in} N(0,1)$$ (approximativt standard normal-fördelad)

$$I_p = \left[p^* \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]$$ är ett konfidensintervall med approximativ konfidensgrad $1-\alpha$

ENSIDIGA INTERVALL:

$$I_p = \left[0, p^* + z_\alpha\sqrt{\frac{p(1-p)}{n}}\right]$$

$$I_p = \left[p^* - z_\alpha\sqrt{\frac{p(1-p)}{n}}, 1\right] \qquad (0 \leq p \leq 1)$$

Vi vill testa: $H_0: p = p_0$ mot $H_1: p \neq p_0$ eller

$H_1: p < p_0$ eller $H_1: p > p_0$

Under H_0 gäller att $X \sim Bin(n, p_0)$

Så p-värdet beräknas som för:

$H_1: p > p_0 \longrightarrow p = P_{H_0}(X \leq x)$

$H_1: p < p_0 \longrightarrow p = P_{H_0}(X \geq x)$

$H_1: p \neq p_0 \longrightarrow$ om $x \geq np_0 \longrightarrow p = 2 \cdot P_{H_0}(X \geq x)$

$x \leq np_0 \longrightarrow p = 2 \cdot P_{H_0}(X \leq x)$

EX) $n = 1000$ $x = 2$

Antag defekta enheter max 0,1%

Vi vill testa: $H_0: p = 0{,}001$ $H_1: p > 0{,}001$

$$p = P(X \geq 2 \mid X \sim Bin(1000; 0{,}001)) = 1 - P(X \leq 1 \mid X \sim Bin(1000; 0{,}001)) =$$

$$= 1 - P(X=1) - P(X=2) = 1 - \binom{1000}{0} 0{,}001^0 \cdot (1-p)^{1000} - \binom{1000}{1} 0{,}001^1 \cdot (1-p)^{999} = 0{,}264$$

P-värdet är för stort. H_0 kan ej förkastas.

JÄMFÖRELSER AV BINOMIALFÖRDELNINGEN

Antag $X_1 \sim Bin(n_1, p_1)$ $\quad X_2 \sim Bin(n_2, p_2)$

Kan vi påstå att $p_1 = p_2$? Vi skattar p_1 & p_2 med p_1^* & p_2^*

$p_i^* = \frac{x_i}{n_i}$ Normalapproximation ok om $n_i p_i (1-p_i) > 10$

Vi skattar $p_1 - p_2$ med $p_1^* - p_2^*$

$E(p_1^* - p_2^*) = p_1 - p_2$ $\qquad V(p_1^* - p_2^*) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

$$T = \frac{p_1^* - p_2^* - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{\text{ej}}{\sim} N(0,1)$$

$$I_{p_1 - p_2} = \left[p_1^* - p_2^* \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right]$$

Om vi vill testa $H_0: p_1 = p_2$, så kan vi använda poolad variansskattning, eftersom $V_{H_0}(p_1^* - p_2^*) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

Ett gemensamt p skattas som:

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

Då blir teststorheten: $T = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\text{ej}}{\sim} N(0,1)$ under H_0

EX 1 Antag $n_1 = 100$ $\quad x_1 = 15$

$n_2 = 200$ $\quad x_2 = 18$

Vi har $p_1^* = \frac{15}{100} = 0{,}15$ $\qquad p_2^* = \frac{18}{200} = 0{,}09$

Vi vill testa: $H_0: p_1 = p_2$

$$p^* = \frac{15+18}{100+200} = 0{,}11 \longrightarrow T = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)\left(1/n_1 + 1/n_2\right)}} = 1{,}566$$

Vi har $|T_{obs}| < z_{\alpha/2} = 1{,}96$ Vi kan ej förkasta H_0

↑ kritiska värdet för $N(0,1)$

ICKEPARAMETRISKA METODER

Kom ihåg att medianen är ett tal M så att
$P(X<M) = P(X>M) = 1/2$

Antag att vi har ett stickprov av storleken n.
Vi vill testa: $H_0: M = M_0$ mot $H_1: M \neq M_0$ eller $H_1: M > M_0$
$H_1: M < M_0$

Låt $Q_+ = \#\{x_i > M_0\}$ $Q_- = \#\{x_i < M_0\}$

TECKENTEST FÖR MEDIANEN

Om H_0 är sann $\longrightarrow$ $Q_+ \sim Bin(n, 1/2)$ $Q_- \sim Bin(n, 1/2)$

H_0 för kastas om:

För $H_1: M < M_0 \longrightarrow Q_+$ är för liten

För $H_1: M > M_0 \longrightarrow Q_-$ är för liten

För $H_1: M \neq M_0 \longrightarrow \min(Q_+, Q_-)$ är för liten.

EX Antag $n=15$
Vi vill testa: $H_0: M=55$ mot $H_1: M<55$
Räkna antalet större än 55 $\longrightarrow$ säg att vi får 3.

$p = P(Q_+ \leq 3 \mid Q_+ \sim Bin(15, 1/2)) = P(X=0) + P(X=1) + P(X=2) + P(X=3) =$
$= 0{,}0176$

Väldigt litet p-värde $\longrightarrow$ H_0 förkastas.

Det enda vi antar är symmetrisk fördelning $\longrightarrow$
testet tappar i styrka.

WILCOXONS RANKTEST (Ett starkare test)

1) Beräkna alla $|X - M_0|$

2) Ordna dessa i ökande storleksordning

3) Beräkna R_i som ranken för $|X_i - M_0|$ gånger $\text{sign}(X_i - M_0)$

Definiera $W_+ = \sum_{i=R_i>0} R_i$ $\quad |W_-| = \sum_{i=R_i<0} |R_i|$

Teststorheten: är $W = \min(W_+, |W_-|)$ finns i tabell.

EX 8.1.2

X_i:	115,1	117,8	116,5	121,0
$\lvert X_i - M_0 \rvert$:	4,9	2,2	3,5	1
Rank:	4	2	3	1
R_i :	-4	-2	-3	1

$W_+ = 1$ $\quad |W_-| = |-2-3-4| = 9$

$W = \min(1, 9) = 1$ slå upp kritiska värdet i tabell.

Det finns ett ekvivalent test för jämförelser → Wilcoxons Rank-sumtest

MATEMATISK STATISTIK F12 15/6 2014

EXEMPEL: STICKPROV I PAR

I fallet med parade observationer $(X_1, Y_1) \ldots (X_m, Y_m)$ bildar vi först differenserna $X_i - Y_i$ och använder sedan ett ranktest på differanserna för att testa om differansernas median är noll.

Från EX. 10.6.2 i boken →

Vi vill testa hur mycket minne 2 statistiska paket använder vid analys av ett dataset. Vi gör mätningar:

Vi vill testa: $H_0: M_X = M_Y$ $H_1: M_X \neq M_Y$

Vi ordnar differanserna och beräknar rankerna:

$\lvert D_i \rvert$:	0	10	12	25	25	40	50	100
Rank:	1	2	3	4,5	4,5	6	7	8
R_i:	-1	2	3	-4,5	4,5	6	7	8

Vi har nu

$$W_+ = 2 + 3 + 4{,}5 + 6 + 7 + 8 = 30{,}5$$

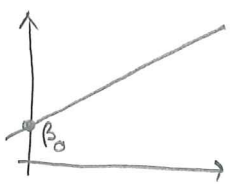
$$|W_-| = |-1 - 4{,}5| = 5{,}5$$

Eftersom vi har ett 2-sidigt test så använder vi:

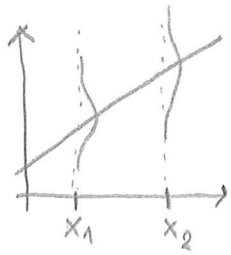
$W = \min(|W_-|, W_+) = 5{,}5$

från tabell ser vi då att den kritiska punkten blir: 6 för ett 2-sidigt test med $\alpha = 0{,}1$.

$W = 5{,}5 < 6 \rightarrow H_0$ förkastas.



$X = \beta_0 + \beta_1 X$



β_0 = interceptet

β_1 = lutningen

MINSTA KVADRAT METODEN (MK)

Antag den liten mer generella modellen att väntevärdet för Y ges av en funktion f.

$$E(Y_i) = f(\theta_1 \ldots\ldots \theta_k, x_i)$$

Parametrarna $\theta_1 \ldots \theta_k$ skattas enligt minsta kvadrat metoden genom att minimera kvadratfelet för den data vi har.

$$S(\theta_1 \ldots \theta_k) = \sum_{i=1}^{k} (y_i - f(\theta_1 \ldots \theta_k, x_i))^2 \qquad \text{m.a.p } \theta_1 \ldots \theta_k$$

Lösningen brukar fås som lösningen till:

$$\frac{\partial S}{\partial \theta_i} = 0 \qquad \text{för } i = 1 \ldots k$$

Ekvationssystemet löses av $\theta_1^* \ldots \theta_k^*$ som kallas M-k-skattningar av $\theta_1 \ldots \theta_k$

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\left.\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned}\right\} \quad \left.\begin{aligned} &\sum y_i - n\beta_0 - \beta_1 \sum x_i = 0 \\ &\sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 = 0 \end{aligned}\right\}$$

$$\left.\begin{aligned} &\beta_0 n + \beta_1 \sum x_i = \sum y_i \\ &\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \end{aligned}\right\} \quad \text{Multiplicera m. } -\frac{1}{n} \sum x_i \longrightarrow$$

$$\beta_1 \underbrace{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}_{S_{xx}} = \underbrace{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}_{S_{xy}}$$

$$\boxed{\beta_1 = \frac{S_{xy}}{S_{xx}}}$$

$S_{xx} = \sum x_i^2 - n\bar{x}^2 = \sum (x_i - \bar{x})^2$ "variationen i x-led"

$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = \sum (x_i - \bar{x})(y_i - \bar{y})$ "samvariation"

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i \xrightarrow{\times \frac{1}{n}} \beta_0 + \beta_1 \bar{x} = \bar{y} \rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Mk-skattningar

$\beta_1^* = \frac{S_{xy}}{S_{xx}}$ $\quad \beta_0^* = \bar{y} - \beta_1^* \bar{x}$

i boken skriver dom B_0 för β_0^* och B_1 för β_1^*

MK-SKATTNING AV σ^2

Skattas av $s^2 = \frac{Q_0}{n-2}$ ← ettersom vi har n-2 frihetsgrader

$Q_0 = \sum (y_i - \beta_0^* - \beta_1^* x_i)^2$

Om vi räknar för hand kan vi:

$Q_0 = S_{yy} - \beta_1^* S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

SATS

Vi har att: $E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$ $\quad V(\bar{Y}) = \frac{\sigma^2}{n}$

$E(\beta_1^*) = \beta_1$ $\quad V(\beta_1^*) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum (x_i - \bar{x})}$

BEVIS

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

$$E(\bar{Y}) = E\left(\frac{1}{n}\sum Y_i\right) = E\left(\frac{1}{n}\sum(\beta_0 + \beta_1 x_i + \varepsilon_i)\right) =$$

$$= E\Big(\underbrace{\beta_0 + \beta_1\left(\frac{1}{n}\sum x_i\right) + \frac{1}{n}\sum \varepsilon_i}_{\text{konstant, ej slumpmässig}}\Big) = \beta_0 + \beta_1 \bar{x} - \frac{1}{n}\sum \underbrace{E(\varepsilon_i)}_{N(0,\sigma^2)}$$

$$V(\bar{Y}) = V\left(\beta_0 + \beta_1 \bar{x} + \frac{1}{n}\sum \varepsilon_i\right) = V\left(\frac{1}{n}\sum \varepsilon_i\right) = \frac{1}{n^2}\sum \sigma^2 = \frac{\sigma^2}{n}$$

$$E(\beta_1^*) = E\left(\frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}\right) = \frac{\sum (x_i - \bar{x}) E(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} =$$

Eftersom det bara är y som är stokastisk

$$= \frac{\sum (x_i - \bar{x}) \beta (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \beta$$

Vi har att $\sum (x_i - \bar{x})(Y_i - \bar{Y}) = \sum (x_i - \bar{x}) Y_i - \sum (x_i - \bar{x}) \bar{Y} = 0$

$\rightarrow \sum (x_i - \bar{x}) Y_i = \sum (x_i - \bar{x}) \bar{Y}$

Detta gör det enklare att beräkna $V(\beta_1^*)$

$$V(\beta_1^*) = V\left(\frac{\sum (x_i - \bar{x}) Y_i}{\underbrace{\sum (x_i - \bar{x})^2}_{\text{konstant}}}\right) = \frac{1}{\left(\sum (x_i - \bar{x})^2\right)^2} \cdot V\left(\sum (x_i - \bar{x}) Y_i\right) =$$

$\boxed{\begin{array}{l} \forall\, Y_i \text{ oberoende ty} \\ \forall\, \varepsilon_i \text{ oberoende} \end{array}}$

$$= \frac{1}{\left(\sum (x_i - \bar{x})^2\right)^2} \sum (x_i - \bar{x})^2 \underbrace{V(x_i)}_{\sigma^2} =$$

$$= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma}{S_{xx}}$$

SATS

Om $\mu_Y^*(x_0) = \beta_0^* + \beta_1^* x_0$ är den skattade linjens värde för ett fixt x_0 så är:

$$E(\mu_Y^*(x_0)) = \beta_0 + \beta_1 x_0$$

Samt:

$$V(\mu_Y^*(x_0)) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

BEVIS

$$\mu_Y^*(x_0) = \beta_0^* + \beta_1^* x_0 = \bar{Y} + \beta_1^*(x_0 - \bar{x})$$

$$E(\mu_Y^*(x_0)) = E(\bar{Y}) + E(\beta_1^*)(x_0 - \bar{x}) = \beta_0 + \beta_1 \bar{x} + \beta_1(x_0 - \bar{x}) =$$

$$= \beta_0 + \beta_1 x_0$$

$$V(\mu_Y^*(x_0)) = \{C(\bar{Y}, \beta_1^*) = 0\} = V(\bar{Y}) + (x_0 - \bar{x})^2 \cdot V(\beta_1^*) =$$

$$= \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2 \cdot \sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

SATS

Om ε_i är normalfördelad gäller att $\bar{Y}$, β_0^*, β_1^* och $\beta_0^* + \beta_1^* x_0$ också är normalfördelade

BEVIS

$$\bar{Y} = \frac{1}{n} \sum Y_i \leftarrow N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\beta_1^* = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} = \underbrace{\frac{\sum (x_i - \bar{x})}{S_{xx}} \cdot \underbrace{(Y_i - \bar{Y})}_{\text{Normalfördelad}}} = \beta_1^* \in N(\beta_1, \sigma^2)$$

Normalfördelad ty konstant · normalfördelad = = normalfördelad

samt $\sum$ normalfördelad = normalfördelad

KONFIDENSINTERVALL OCH TEST

Låt θ vara någon av β_0, β_1 eller $\beta_0 + \beta_1 x_0$. Vi vet att dess skattningar är normalfördelade och vi har tagit fram variansen av skattningarna. Låt $d(\theta^*)$ vara standardavvikelsen för skattningen. Vi har då att:

$$T = \frac{\theta^* - \theta}{d(\theta^*)} \sim t(n-2)$$

Vilken på vanligt sätt används för att göra test och bilda konfidensintervall.

$$I_\theta = (\theta^* \pm t_{\alpha/2}(n-2)\, d(\theta^*))$$

För β_0

$$I_{\beta_0} = \left(\beta_0^* \pm t_{\alpha/2}(n-2)\, s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

För β_1

$$I_{\beta_1} = \left(\beta_1^* \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}\right)$$

För $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$

$$I_{\mu_Y(x_0)} = \left(\beta_0^* + \beta_1^* x_0 \pm t_{\alpha/2}(n-2)\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

MATEMATISK STATISTIK F13 19/5 2014

· Han skrev inget på tavlan!

MATEMATISK STATISTIK F 14 22/5 2014

"TENTAGENOMGÅNG"

16 JANUARI 2014 (Stefans.)

1. $P(A)=0{,}2 \quad P(B)=0{,}5 \quad P((A\cup B)^c)=0{,}4$

a) om $P(A\cap B)=P(A)P(B) \longrightarrow$ A och B oberoende

$P(A\cup B)=P(A)+P(B)-P(A\cap B)=1-P((A\cup B)^c)=0{,}6$

$\longrightarrow P(A\cap B)=0{,}5+0{,}2-0{,}6=0{,}1$

$P(A)\times P(B)=0{,}1 \longrightarrow$ <u>A och B är oberoende</u>

b) $P(A|B)$?

Två sätt: 1. Definition av betingad slh:

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{0{,}1}{0{,}5}=0{,}2$$

2. $P(A|B)=P(A)$ ty oberoende i a)

---

2. $X\sim Exp(\lambda) \qquad X_{min}=\min(X_1\dots X_{10})$

a) $P(X>x)=1-P(X\le x)=1-F(x)=1-(1-e^{-\lambda x})=e^{-\lambda x}$

b) $P(X_{min}>x)=P(\min(X_1\dots X_{10})>x)=P(X_1>x)\cap P(X_2>x)\cap\dots$
$\dots\cap P(X_{10}>x)=P(X_1>x)P(X_2>x)\dots P(X_{10}>x)=$
$=e^{-\lambda x}\cdot e^{-\lambda x}\cdot\dots\cdot e^{-\lambda x}=e^{-10\lambda x}$

c) $F_{X_{min}}(x)=P(X_{min}\le x)=1-P(X_{min}>x)=1-e^{-10\lambda x}$

d) $X_{min}\sim Exp(10\lambda)$

③ Är kass enligt david!

---

④

	n	$\bar{X}/\bar{Y}$	s
(x) Blyblandad	15	87	4
(Y) Vanlig	10	90	5

a) Man får anta $X \sim N(\mu_1, \sigma_1^2)$ $Y \sim N(\mu_2, \sigma_2^2)$

Testa $H_0: \sigma_1 = \sigma_2$ $H_1: \sigma_1 \neq \sigma_2$

$T_{obs} = \frac{S_Y^2}{S_X^2} = \frac{25}{16}$

Förkasta H_0 om $T_{obs} > F_{\alpha/2}(9,14) = 2{,}54$

$T_{obs} < 2{,}54$ vi kan ej förkasta H_0

$\longrightarrow$ Vi kan anta $\sigma_1 = \sigma_2$

I Facit har dom antagit detta utan att testa

$$I_{\mu_1-\mu_2} = \left(\bar{X} - \bar{Y} \pm \underbrace{t_{\alpha/2}(23)}_{2{,}069} \times S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\right)$$

Man kan använda poolad pga $\sigma_1 = \sigma_2$

$$S_p^2 = \frac{(n_x-1)S_x^2 + (n_y-1)S_y^2}{n_x+n_y-2} = 19{,}52$$

$\longrightarrow I_{\mu_1-\mu_2} = (-3 \pm 2{,}069 \times 1{,}803) = [-6{,}73 ; 0{,}73]$

b) statfan har inte specifiserat $H_0 \rightarrow$ antar $H_0: \mu_1 = \mu_2$

$0 \in I_{\mu_1-\mu_2} \rightarrow$ Vi kan ej förkasta H_0 på $\alpha = 0{,}05$.

Vi kan inte heller förkasta på starkare nivå (lägre nivå) $\longrightarrow$ Vi kan förkasta på $\alpha = 0{,}01$

5. Uppgiften visar en bra poäng.

	n	$\bar{\delta}$	S_δ
A	16	0,3	0,5
B	9	0,2	0,5

a) Vi vill testa $H_0: \delta = 0$ $H_1: \delta > 0$

"T-test" om $N(\delta, \sigma^2)$

$$T = \frac{\bar{\delta}}{S_\delta / \sqrt{n}} \quad \text{under } H_0 \quad T \sim t(n-1)$$

För A

$$T_{obs} = \frac{0,3}{0,5/\sqrt{16}} = 2,4$$

$$t_{0,05}(16-1) = 1,75$$

$T_{obs} > 1,75 \rightarrow$ förkasta H_0

För B

$$T_{obs} = \frac{0,2}{0,5/\sqrt{9}} = 1,2$$

$$t_{0,05}(9-1) = 1,86$$

$T_{obs} < 1,86 \rightarrow H_0$ kan ej förkastas

b) * pga olika n så har testen olika styrka

* Man borde göra ett test för skillnaden i väntevärde

$\rightarrow H_0: \delta_A = \delta_B$ $H_1: \delta_A > \delta_B$

Samma antaganden som i a). Använder poolad variansskattning

$$S_p^2 = 0,5^2$$

$$T = \frac{\bar{\delta}_A - \bar{\delta}_B}{S_p \sqrt{\frac{1}{16} + \frac{1}{9}}} = 0,24 \qquad t_{0,05}(23) = 1,71$$

$T < 1,71$ vi kan ej förkasta H_0

6. $H_0: \mu = 0$ $\quad H_1: \mu > 0$

$$T = \frac{\bar{X}}{\sigma/\sqrt{n_1}} = z_{0,05} = 1,645 \longrightarrow \frac{\bar{X}}{\sigma/\sqrt{n}} = 1,645 \longrightarrow \frac{\bar{X}}{\sigma} = \frac{1,645}{\sqrt{20}}$$

Om vi skulle gjort ett 2-sidigt test:

$$\frac{\bar{X}}{\sigma/\sqrt{n}} = z_{0,05/2} = 1,96 \longrightarrow \sqrt{n} = \frac{1,96}{\bar{X}/\sigma} = \boxed{\frac{1,96}{1,645/\sqrt{20}}}$$

$$\longrightarrow n = \left(\frac{1,96}{1,645/\sqrt{20}}\right)^2 = 28,4 \approx 29 \longrightarrow n \geq 29$$

---

7. $\bar{X} = 2,325$ $\quad \bar{Y} = 129,7$ $\quad I_{\beta_1} = [6,29 ; 24,99]$

$n = 16.$

a) $\mu^*_Y(x=0) = ?$ $\quad$ Hur ser I_{β_1} ut?

$$\beta_1^* = \frac{24,99 - 6,29}{2} = 15,64$$

$$\beta_0^* = \bar{Y} - \beta_1^* \cdot \bar{X} = 129,7 - 15,64 \times 2,325 = 93,337$$

$$\mu^*_Y(x=0) = \beta_0^* + \beta_1^* \cdot x = \beta_0^* = 93,337$$

b) $H_0: \beta_1 = 0$ $\quad H_1: \beta_1 \neq 0$

$$T = \frac{\beta_1^* - \beta_1}{S/\sqrt{S_{xx}}} = \frac{15,64}{S/\sqrt{S_{xx}}}$$

$$I_{\beta_1} = \left(\beta_1^* \pm t_{\alpha/2}(n-2) \cdot \frac{S}{\sqrt{S_{xx}}}\right)$$

Längden på intervallet $= 24,99 - 6,29 = 2 \cdot 2,145 \cdot \frac{S}{\sqrt{S_{xx}}} \longrightarrow$

$$\frac{S}{\sqrt{S_{xx}}} = 4,36$$

$$\longrightarrow T = \frac{15,64}{4,36} = 3,59$$

c), d) saknas info för att lösa detta

---

25 Maj 2012 (Aila)

Typ såhär kommer davids tentor se ut

2.

$$F(x) = \begin{cases} 0 & x<0 \\ 1/4 \cdot x^3 & 0 \le x < 1 \\ 3/4 \cdot x - 1/2 & 1 < x \le 2 \\ 1 & 2 < x \end{cases}$$

a) $P(0{,}5 \le X \le \le 1{,}5) = ?$

Finns 2 sätt att lösa detta!

Krånglig sätt

$$= \int_{0,5}^{1,5} f(x)dx = \int_{0,5}^{1,5} \frac{\partial F(x)}{\partial x} dx$$

Lättare sätt

$F(x) = P(X \le x) \longrightarrow$

$P(0{,}5 \le X \le 1{,}5) = P(X \le 1{,}5) - P(X < 0{,}5) = F(1{,}5) - F(0{,}5) =$

$= \frac{3}{4} \times 1{,}5 - \frac{1}{2} - \frac{0{,}5^3}{4} = 19/32$

b) $P(X \ge 1{,}5) = 1 - F(1{,}5) = 3/8$

$Y =$ "Antal $X_i > 1{,}5$" $\sim Bin(10, 3/8)$

$P(Y=2) = P_Y(2) = \binom{10}{2} (3/8)^2 (1 - 3/8)^8 = 0{,}147$

4. a) Minimera kvadratfel: $S = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$

m.a.p β_0 och R_i

b) Egentligen inga alls

c) $e_i = Y_i - \beta_0^* - \beta_1^* x_i$ Modellvalidering

---

5. $\overline{X} = 54{,}11$ $S^2 = 207{,}19$

a) $X \sim N(\mu, \sigma^2)$ $H_0: \mu = 50$ $H_1: \mu < 50$

Z-test $\longrightarrow$ om σ är känd

t-test $\longrightarrow$ om σ är okänd och skattas av s

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{54{,}11 - 50}{\sqrt{207{,}19}/\sqrt{20}} = 1{,}277$$

$t_\alpha(19) = 1{,}729$

H_0 förkastas pga $T < t_\alpha$

c) Jmf tecken- och t-test. Teckentestet borde vara bättre eftersom det ej ser ut som en jämn fördelning