STATISTICA INDUSTRIALE ora 1

Note Title 03/03/2015

Probabilità → prerequisito

Ing. gest 9/6 cfu → statistica multivariata (avanzata)

5/6/9 cfu → statistica

Contenuti : statistica di base + prob } statistica (lea)

regressione

(ANOVA) analisi della varianza

test di adattamento

(DOE) design of experiment

} metodi e modelli a supporto delle decisioni (lea)

+ argomenti a seconda delle esigenze (text mining, big data)

... cluster analysis

... discriminant analysis

Bibliografia : Ross : "Prob & Stat per Ing e Scienze" Apogeo

Johnson Wichern : ...

→ Sleeper : "Design for Six Sigma Statistics" per "practitioner"

<u>Corso molto pratico</u> : poca teoria e poche dimostrazioni

poca astrazione e poca generalità

(Statistica per matematici : Mood : " ... "

Dispense Pratelli M. (Pisa))

Laboratorio : Excel / Minitab

Esame : scritto di laboratorio + orale (seminario?)

Storia della statistica

- Statistica descrittiva : informazioni complete sulla popolazione
  1600 → (Istat & co)
- Statistica inferenziale : dedurre info parziali da un campione
  ~1930 lab scientifici

  40 persone → 20 trattamento
  ↳ 20 controllo (placebo)

	g.	n.g.	
t.	16	4	20
c.	7	13	20

	g.	n.g.	
t.	14	6	20
c.	9	11	20

~1940/50 in azienda → NO

Deming → 1960 in Giappone (Toyota)
"qualità" (aziendale) = statistica

~1980 Motorola "six sigma"
green belt, black belt, master black belt

~oggi dipende dalle persone

STATISTICA DI BASE
(stimatori)
→ Intervalli di confidenza
→ Test statistici

STIMATORI
- Dati normali

$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ indipendenti

legge normale

media

varianza

"campione" : $X_1, \dots, X_n$
"popolazione" : $\mathcal{N}(\mu, \sigma^2)$

μ, σ^2 incognite (una o entrambe)

$x_1, x_2, \dots, x_n$ realizzazione numerica del campione nota

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

$\bar{X} \approx \mu$ è uno stimatore di μ

- Def una statistica è una funzione qualsiasi del campione (ad esempio $\bar{X}$)

- Def uno stimatore corretto di un parametro θ è una statistica Θ tale che

$$E(\Theta) = \theta$$

altrimenti si chiama distorto e $E(\Theta) - \theta$ si chiama bias

★ $\bar{X}$ è uno stimatore corretto di μ

$$E(\bar{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum E(X_i) = \mu$$

- Def uno stimatore consistente di un parametro θ è una statistica $\Theta_n = f_n(X_1, \dots, X_n)$ tale che

$$\Theta_n \xrightarrow[n\to\infty]{} \theta \quad \text{in probabilità o q.c.}$$

★ $\bar{X}$ è uno stimatore consistente di μ

la LGN (forte) dice che $X_1, X_2, \dots$ i.i.d. $X_i \in L^1$

allora $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} E(X_1)$ q.c.

Per una verifica diretta della convergenza in probabilità

(HW)

i. $E(\Theta_n) \xrightarrow[n\to\infty]{} \theta$

ii. $Var(\Theta_n) \xrightarrow[n\to\infty]{} 0$

} + disug di Chebyshev ⇒ conv in prob

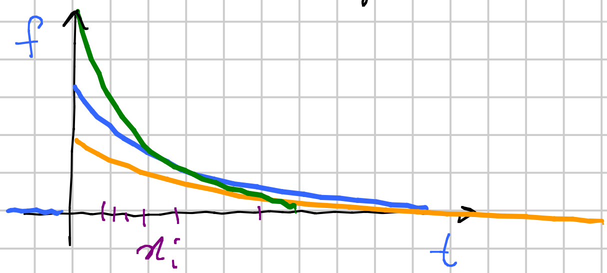
indipendenti

★ $Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} Var\left(\sum x_i\right) = \frac{1}{n^2}\sum Var(x_i) = \frac{\sigma^2}{n}$

- Stimatori di massima likelihood (= verosimiglianza)

esempio $x_1, x_2, \dots, x_n \sim expo(\lambda)$ λ incognito

$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0$



$X := (x_1, x_2, \dots, x_n)$

$f_X : \mathbb{R}^n \to \mathbb{R}_+$ densità congiunta

$f_X(t_1, t_2, \dots, t_n) = \prod_{i=1}^{n}\left(\lambda e^{-\lambda t_i}\right) = \lambda^n e^{-\lambda \sum t_i}$

indipendenza

$L(\lambda) := f_X(\underbrace{x_1, x_2, \dots, x_n}_{\text{parametri}}; \underset{\text{variabile}}{\lambda}) = \lambda^n e^{-\lambda \sum_1^n x_i}$ funzione di likelihood

cerco $\hat{\lambda}$ che massimizza L

$\ell(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_1^n x_i$ log-likelihood

derivo $\ell'(\lambda) = \frac{n}{\lambda} - \sum_1^n x_i$ e pongo $\ell'(\hat{\lambda}) = 0$

$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$

(HW?)

lo stimatore ML è $\Lambda = \frac{1}{\bar{X}}$ non credo sia corretto

★ Ogni stimatore ML è per forza consistente

HW: $X_1, X_2, \dots, X_n \sim \text{unif}[0, a]$ a incognito

i. trovare ML est di a

ii. verificare che non è corretto

iii. proporre una modifica affinché diventi corretto (restando consistente)

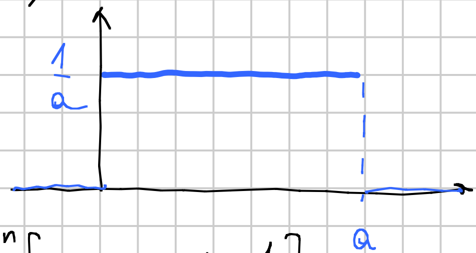
HW: $X_1, X_2, \dots, X_n \sim unif[0,a]$ a incognito

i. trovare ML est di a

ii. verificare che non è corretto

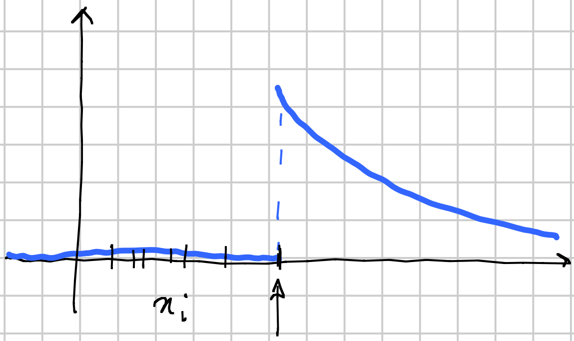
iii. proporre una modifica affinché diventi corretto (restando consistente)

$$f_{X_i}(t) = \mathbb{1}_{[0,a]}(t) \cdot \frac{1}{a}$$



$x_1, x_2, \dots, x_n$

$$L(a) = \prod_{i=1}^{n} f_{X_i}(x_i; a) = \prod_{i=1}^{n} \left[\mathbb{1}_{[0,a]}(x_i) \cdot \frac{1}{a}\right] = a^{-n}\, \mathbb{1}_{[0,a]}(\max_i x_i)$$



$L(\hat{a})$ è massima per $\hat{a} = \max x_i$

MLE $\hat{a} := \max_i X_i$ è consistente

$$\hat{a}_n \xrightarrow[n \to \infty]{P} a$$

Non è corretto: $E(\hat{a}) \neq a$

$$E(\max_i X_i) = ?$$

$$F_{X_1}(t) = \begin{cases} \frac{t}{a} & t \in [0,a] \\ 1 & t > a \\ 0 & t < 0 \end{cases}$$

$$F_{\hat{a}}(t) := P(\max_i X_i \le t) = P(X_i \le t \quad \forall i = 1,2,\dots,n)$$

$$= P\left(\bigcap_i \{X_i \le t\}\right) \underset{\text{indipendenza}}{=} \prod_i P(X_i \le t) = [F_{X_1}(t)]^n = \begin{cases} \frac{t^n}{a^n}, & t \in [0,a] \\ 1 & t > a \\ 0 & t < 0 \end{cases}$$

$$F_{\hat a} \longrightarrow f_{\hat a} \qquad f_{\hat a}(t) = \frac{n}{a^n} t^{n-1}$$

$$E(\hat a) = \int_0^a t \cdot \frac{n}{a^n} t^{n-1} dt = \frac{a^{n+1}}{n+1} \cdot \frac{n}{a^n} = a \frac{n}{n+1}$$

$$A := \frac{n+1}{n} \max_i X_i$$ corretto e consistente

$$A_n - \hat a_n = \frac{1}{n} \hat a_n \searrow 0 \qquad \hat a_n \xrightarrow{P} a$$

$$A_n \xrightarrow[n \to \infty]{P} a$$

$\hat a$ sottistima sistematicamente a

A è corretto

HW: 1942 carri armati tedeschi catturati

77 carri tra marzo e ottobre 1942

$|1042, 1717, 348, \ldots, 4434, \ldots, 2345|$ numeri di serie

inizio marzo — max 71° dato — fine ottobre

$\rightarrow$ stimare la capacità produttiva mensile

INTERVALLI DI CONFIDENZA

- Caso campione normale σ nota, $\boxed{\text{stimare}}$ μ

$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ indip

$\bar X$ è stimatore c&c di μ

$\mu \approx \bar x$ stima puntuale

$\mu = \bar x \pm r$ stima intervallo / forbice $\leftarrow$

Mi aspetto che $\mu \in [\bar x - r ; \bar x + r]$ con livello di confidenza $1-\alpha$ (95%, 90%, ...)

Il significato è che "$1-\alpha = P(\mu \in [\bar x - r ; \bar x + r])$"

Non è una vera probabilità $\quad \bar{x} = 10{,}1 \qquad r = 0{,}3$

$\underbrace{\mu \in [9{,}8 \,;\, 10{,}4]}$ al 95% di confidenza

tutto deterministico : o è vera o no

= fiducia nella affermazione

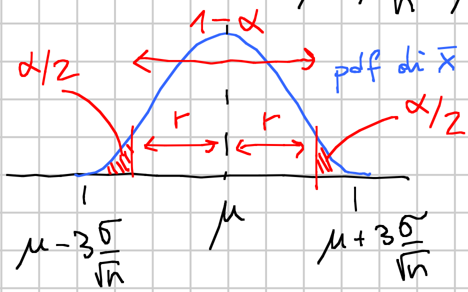
- Serve la legge di $\bar{X}$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad X_i \sim \mathcal{N}(\mu, \sigma^2) \text{ indip.}$$

☆ la classe $\mathcal{N}$ è riproducibile ovvero se sommo v.v.a.a. normali indipendenti (non occorre isonome) ottengo una v.a. $\mathcal{N}$

(more on that)

$$\rightarrow \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$



99,73%

Esiste $r > 0$ t.c.

$$1-\alpha = P\left(\bar{X} \in [\mu - r \,;\, \mu + r]\right)$$
$$= P\left(|\bar{X} - \mu| \leq r\right)$$
$$= P\left(\mu \in [\bar{X} - r \,;\, \bar{X} + r]\right)$$

$$1-\alpha = P\left(\mu \in [\bar{X} - r \,;\, \bar{X} + r]\right) \leftarrow \text{intervallo random}$$

$\mu \in [\bar{x} - r \,;\, \bar{x} + r]$ con livello di conf. $1-\alpha$

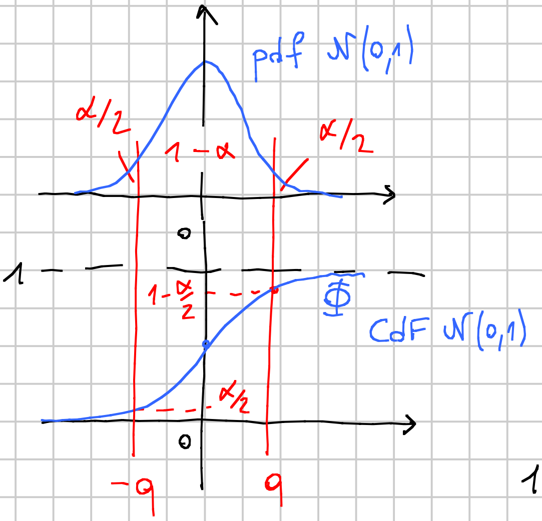
- Come trovo r t.c. $P(|\bar{X} - \mu| \leq r) = 1-\alpha$?

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$
$$\bar{X} - \mu \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

$$\boxed{\begin{array}{l} E(a + bX) = a + bE(X) \\ Var(a + bX) = b^2 Var(X) \end{array}}$$

$\mathcal{N}(0,1)$

$$1-\alpha = P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq \underbrace{\frac{r}{\sigma/\sqrt{n}}}_{q}\right)$$



$$q = F^{-1}_{\mathcal{N}(0,1)}\left(1-\frac{\alpha}{2}\right)$$
$$= \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \qquad \Phi = F_{\mathcal{N}(0,1)}$$

$$r = \frac{\sigma}{\sqrt{n}}\, q$$

$1-\alpha = 95\%$ $\quad q = 1{,}96 \approx 2$

$= 90\%$ $\quad q = 1{,}645$

$= 99{,}73\%$ $\quad q = 3$

* Valori di quantili sensati sono intorno a 2, o tra 1 e 3

$\mu \in \bar{x} \pm 1{,}96 \frac{\sigma}{\sqrt{n}}$ al 95% di conf.

$\boxed{\text{ora 4}}$

$$\Phi(t) := \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-s^2/2}\, ds \qquad \text{non ha scrittura analitica}$$

Funzioni Excel: Φ = NORMSDIST $\quad$ DISTRIB.NORM.ST

Φ^{-1} = NORMSINV $\quad$ ....

NORMDIST

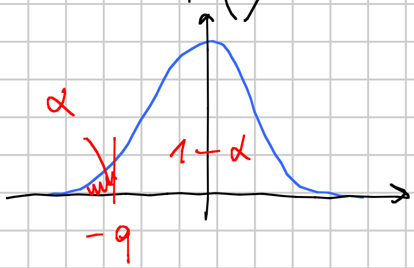
HW: prendere confidenza

HW: Ross Cap 7 pblm 8 e 9

- int. confid. unilaterale $\quad X_i \sim \mathcal{N}(\mu, \sigma^2)$ $\quad \sigma$ nota

voglio $\quad \mu \le UCL$ $\quad$ con lvl di conf $1-\alpha$

$$1-\alpha = P(\underbrace{\mu \le \bar{X}+r}_{\mu \le u}) = P(\bar{X}-\mu \ge -r) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \ge \overbrace{-r\frac{\sqrt{n}}{\sigma}}^{-q}\right)$$



$q = \Phi^{-1}(1-\alpha)$ $\quad$ HW: $\Phi(-x) = 1-\Phi(x)$

$-q = \Phi^{-1}(\alpha)$ $\quad$ $\Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$

- Generalizzazione : funzione ancillare

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

in generale è una funzione che dipende da :

i. i dati

ii. il parametro da stimare

iii. altre grandezze note

e la cui distribuzione sia nota

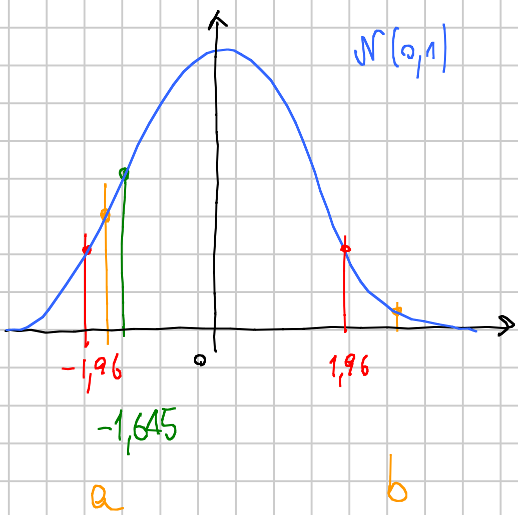
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \underbrace{\mathcal{N}(0,1)}_{} \quad \underbrace{1-\alpha}_{}$$

quantile / i

$$1-\alpha = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \in I_q\right) = \ldots = P\left(\mu \in I_{\bar{x},q}\right)$$

ricavo μ $\qquad$ $\mu \in I_{\bar{x},q}$ con lvl di conf...

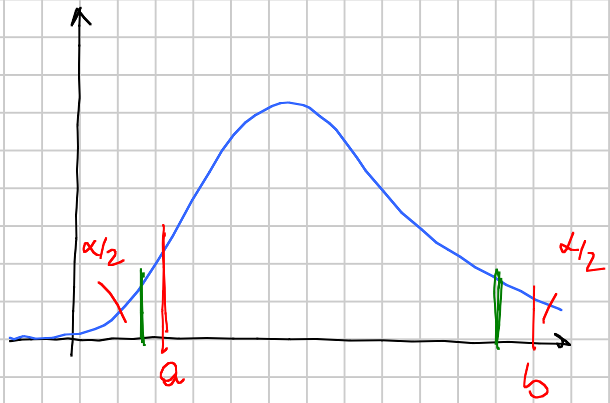
- Generalizzazione : altri intervalli



$1-\alpha$ fissato

q

∞



$$1-\alpha = P\left(a \le \mathcal{N}(0,1) \le b\right)$$

☆ L'intervallo di ampiezza minima e prob $1-\alpha$ è quello in cui il valore di f nei due estremi è uguale

- Caso Bernoulli / binomiale

faccio un sondaggio
n persone, chiedo
se seguono il calcio

$X \sim bin(n,p)$

$\varphi_X(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k=0,\dots,n$

$n=1$ Bernoulli

$P(X=1) = p \quad P(X=0) = 1-p$

$n \geq 1$ binomiale

$E(X) = np$

$Var(X) = np(1-p)$

Classe $bin(\cdot, p)$ è riproducibile

Bernoulliana:

$X_1, X_2, \dots, X_n$

$\downarrow \quad \downarrow$

$1 \quad 0 \quad 0 \dots 1 \dots \dots 0$

1: segue il calcio

0: no

binomiale: su n persone X sono quelle che seguono il calcio

→ Legame: $X = \sum_{i=1}^{n} X_i$ $\quad X_i \sim bin(1,p)$ indipendenti

$X \sim bin(n,p)$

dove p: probabilità che una persona a caso segua il calcio
p incognita, da stimare

☆ Esempio industriale: arriva fornitura 10000 pezzi con (popolazione)
frazione di difettosi $p \in [0,1]$ incognita

Scelgo un campione di n (say 20) li testo e
conto i difettosi → X

☆ Stimo p puntualmente e con intervallo di conf.

→ stimatore di p: $\hat{p} := \frac{X}{n}$ (HW: MLE?)

$\hat{p} := \bar{X}$ a livello delle X_i

→ legge di $\hat{p}$? $\quad X \sim bin(n,p) \quad \hat{p} = \frac{1}{n} X$

$X \dot\sim \mathcal{N}(np\,;\, np(1-p)) \quad \hat{p} \dot\sim \mathcal{N}\left(p\,;\, \frac{p(1-p)}{n}\right) \quad \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot\sim \mathcal{N}(0,1)$

$$\boxed{\text{TLC}: X_1, X_2, \ldots \in L^2 \text{ iid} \quad \frac{\sum_1^n X_i - nE(X_1)}{\sqrt{n\,Var(X_1)}} \xrightarrow[n\to\infty]{L} \mathcal{N}(0,1)}$$

- Approssimazione normale

$$\frac{\sum_1^n X_i - nE(X_1)}{\sqrt{n\,Var(X_1)}} \dot{\sim} \mathcal{N}(0,1)$$

approssimativamente distribuito come

$$\sum_1^n X_i \dot{\sim} \mathcal{N}(nE(X_1)\,;\, n\,Var(X_1))$$

→ Se $X_1, \ldots, X_n$ sono iid di media μ e var σ^2, se n è abbastanza grande,

$$\sum_1^n X_i \dot{\sim} \mathcal{N}(n\mu\,;\, n\sigma^2)$$

→ quanto grande n dipende fortemente dalla legge di X_i

$X_i \sim \mathcal{N}$ $\quad n = 1$

$X_i \sim bin(1, 10^{-10})$ $\quad n \gtrsim 10^{10}$

→ l'approssimazione $\sum X_i \dot{\sim} \mathcal{N}$ è migliore al centro della campana che sulle code

- Campione bernoulliano X : numero totale di "positivi" su n prove

p la prob. incognita di una prova di essere "positiva"

$$X \sim bin(n, p) \qquad E(X) = np \qquad Var(X) = np(1-p)$$

$$X \dot{\sim} \mathcal{N}(np\,;\, np(1-p))$$ approssimazione per il TLC

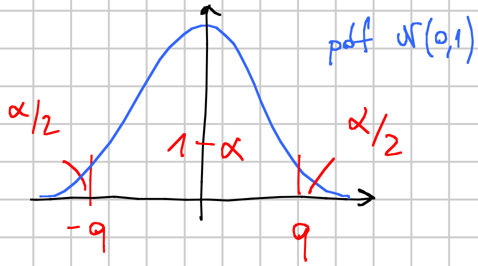
→ intervallo di conf. per p

$$p \approx \hat{p} := \frac{X}{n}$$ stimatore puntuale

$$\hat{p} \dot{\sim} \mathcal{N}\left(p\,;\, \frac{p(1-p)}{n}\right)$$ distrib stimatore (approx)

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} \mathcal{N}(0,1)$$ funz. ancillare (approx)

$$\boxed{\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \dot{\sim} \mathcal{N}(0,1)}$$ funz. ancillare (più approx)



$1-\alpha$ livello di conf assegnato
tipicamente 95%

$$\boxed{q = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)}$$

$$1-\alpha = P(-q \le \mathcal{N}(0,1) \le q) \approx P\left(-q \le \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le q\right) = P\left(p \in \hat{p} \pm q\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$\boxed{\text{con lvl di conf } 1-\alpha \qquad p \in \hat{p} \pm q\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

★ La II approssimazione non è necessaria, ma semplifica le formule

★ La I approx $bin(n,p) \dot{\sim} \mathcal{N}(np\,;\, np(1-p))$ $\quad np(1-p) \gtrsim 5$

- Stimatori per la varianza

(check)

$X \in L^2 \qquad Var(X) = E[(X - E(X))^2] \overset{\downarrow}{=} E(X^2) - E(X)^2$

$X_1, X_2, \dots X_n$ campione iid. $\quad \mu = E(X_i) \quad \sigma^2 = Var(X_i)$

$$\boxed{\sigma^2 \approx S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left[\sum_i X_i^2 - n\overline{X}^2\right]}$$ var. campionaria

(check)

17, 14, 14, 18, 15, ...

$S \approx 1{,}4 \qquad \sum X_i^2 \approx 3000 \qquad n\overline{X}^2 \approx 3000$

10,017, 10,014, 10,014, ... $\qquad S \approx 0{,}0014$

★ La II è mal condizionata

★ $\overline{X}(X_1, \dots, X_n) = \frac{1}{n}\sum_{1}^{n} X_i \qquad \overline{X}(a+bX_1, a+bX_2, \dots) = a + b\overline{X}$

$S^2(X_1, \dots, X_n) = \dots \qquad S^2(a+bX_1, a+bX_2, \dots) = b^2 S^2$

★ S^2 è uno stimatore corretto di σ^2

$$E(S^2) = \frac{1}{n-1}\sum_i \underbrace{E(X_i^2)}_{\sigma^2+\mu^2} - \frac{n}{n-1}\underbrace{E(\overline{X}^2)}_{Var(\overline{X}) + E(\overline{X})^2 = \frac{\sigma^2}{n} + \mu^2} = \frac{n}{n-1}\left\{\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\right\} = \sigma^2$$

→ Ho usato che $E(Y^2) = Var(Y) + E(Y)^2$

★ Nel caso (piuttosto raro) che si conosca μ

$$S_*^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

→ (stimatore corretto)

★ La media aritmetica $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ è il punto y che minimizza $\sum_{i=1}^{n}(x_i - y)^2$

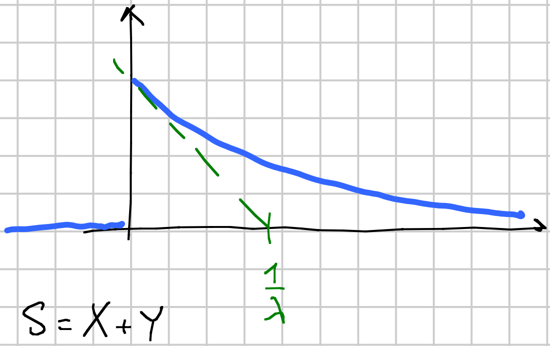
HW: Cosa si deve scegliere per minimizzare $\sum_{i=1}^{n} |x_i - y|$?

HW: S è uno stimatore distorto di σ (sottostima o sovrastima?)

DISTRIBUZIONE GAMMA (CHI-QUADRO, ERLANG, ESPONENZIALE)

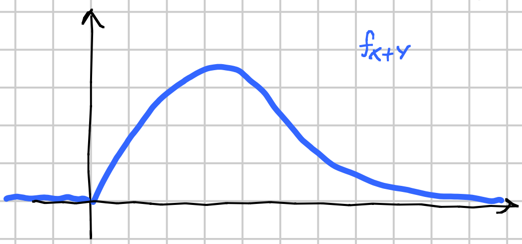
$X \sim \text{expo}(\lambda)$ $\qquad \lambda$ rate / intensità $\qquad \lambda = \dfrac{1}{E(X)}$

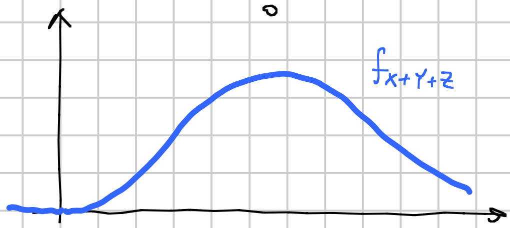
$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & \text{altrove} \end{cases}$$



$X, Y \sim \text{expo}(\lambda)$ indip. $\qquad S = X + Y$

$$f_S(t) = f_X * f_Y(t) = \int_{-\infty}^{\infty} f_X(s) f_Y(t-s)\, ds = \int_0^t \lambda^2 e^{-\lambda t} ds = \lambda^2 t e^{-\lambda t} \quad t \geq 0$$





$$f(t) = \frac{\lambda^3}{2} t^2 e^{-\lambda t}$$

$X_1, X_2, \dots, X_n \sim \text{expo}(\lambda)$ iid. $\qquad S = X_1 + \dots + X_n$

$$f_S(t) = c(n) \lambda^n t^{n-1} e^{-\lambda t}, \quad t \geq 0 \qquad \text{Erlang}(n, \lambda)$$

$$\int_0^\infty t^k e^{-t} dt = k! \quad \Rightarrow \quad c(n) = \frac{1}{(n-1)!}$$

- Legge gamma $\qquad X \sim \text{gamma}(\alpha; \lambda)$
  $\qquad X \sim \text{gamma}(\alpha; \beta)$ $\qquad \beta = \dfrac{1}{\lambda}$

$$\boxed{f_X(t) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t}, \quad t \geq 0}$$

$\alpha > 0 \qquad \beta > 0$

★ $E(X) = \alpha\beta \qquad Var(X) = \alpha\beta^2 \qquad aX \sim \text{gamma}(\alpha; a\beta)$

★ $\text{gamma}(\cdot, \beta)$ è riproducibile

- Legge chi-quadro $\qquad k = 1, 2, \dots$ numero di gradi di libertà

$$\boxed{X \sim \chi^2(k) \sim \text{gamma}\left(\frac{k}{2}; \underset{\beta}{2}\right)} \qquad \boxed{E(X) = k \qquad Var(X) = 2k}$$

$$\chi^2(2) \sim \text{expo}\left(\frac{1}{2}\right)$$

★ $X_1, X_2, \dots, X_n \sim \mathcal{N}(0,1)$ indipendenti

$$\boxed{\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)}$$

- Distribuzione della varianza campionaria per dati normali

$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2) \qquad \sigma$ incognita

$\rightarrow \mu$ nota $\qquad S_*^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$
$$X_i - \mu \sim \mathcal{N}(0, \sigma^2)$$
$$\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

★ $$\frac{S_*^2}{\sigma^2} n \sim \chi^2(n)$$

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

★ $$\boxed{\frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)}$$ funzione ancillare per σ

$$\rightarrow \frac{S^2}{\sigma^2}(n-1) \sim \text{gamma}\left(\frac{n-1}{2}; 2\right)$$

$$S^2 \sim \text{gamma}\left(\overset{\alpha}{\frac{n-1}{2}}; \overset{\beta}{\frac{2}{n-1}\sigma^2}\right)$$

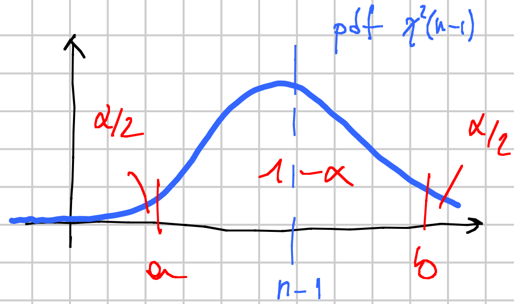
$E(S^2) = \alpha\beta = \sigma^2 \qquad Var(S^2) = \alpha\beta^2 = \frac{2}{n-1}\sigma^4 \qquad$ stim. c&c

• Intervallo di conf. per la varianza di una popolaz. normale

$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ $\qquad$ $1-\alpha$ livello di confidenza

$$\sigma^2 \approx S^2$$

$$\frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$$



$a < n-1$

$b > n-1$

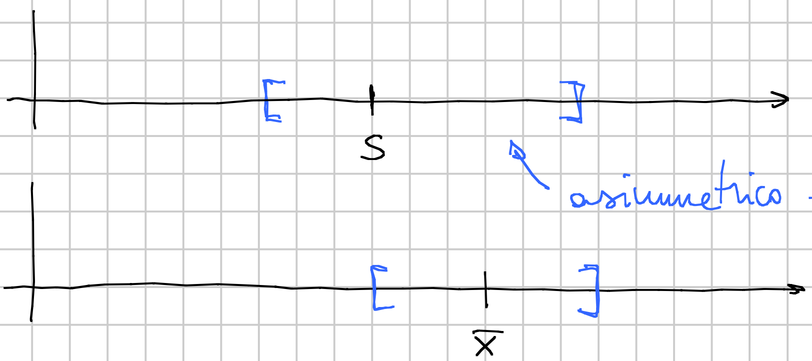
$$a = F^{-1}_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right) = \text{INV.CHI}\left(1-\frac{\alpha}{2}\,;\, n-1\right)$$

$$b = F^{-1}_{\chi^2(n-1)}\left(1-\frac{\alpha}{2}\right) = \text{INV.CHI}\left(\frac{\alpha}{2}\,;\, n-1\right)$$

$$1-\alpha = P\left(a \le \chi^2(n-1) \le b\right) = P\left(a \le \frac{S^2}{\sigma^2}(n-1) \le b\right) = P\left(\frac{n-1}{b}S^2 \le \sigma^2 \le \frac{n-1}{a}S^2\right)$$

$\sigma^2 \in \left[\frac{n-1}{b}S^2 \,;\, \frac{n-1}{a}S^2\right]$ con lvl di conf $1-\alpha$

$\sigma \in \left[\underbrace{\sqrt{\frac{n-1}{b}}\,S}_{<S} \,;\, \underbrace{\sqrt{\frac{n-1}{a}}\,S}_{>S}\right]$ con lvl di conf $1-\alpha$



asimmetrico e più largo a destra

HW: $Z \sim \mathcal{N}(0,1)$ $\qquad$ $W = Z^2$ $\qquad$ $f_W = ?$

$$F_W(t) = P(W \le t) = P(Z^2 \le t) = \dots\dots = \Phi\ ???$$

1) $f_W = F_W'$ $\qquad$ ci riconosco la gamma

2) W_1, W_2 iid $\qquad$ $f_{W_1+W_2} = ?$ $\qquad$ $\left(\text{expo}\left(\frac{1}{2}\right)\right)$

3) $Z_1, Z_2 \sim \mathcal{N}(0,1)$ $\qquad$ $f_{Z_1^2+Z_2^2} = ?$ $\qquad$ $\left(\text{expo}\left(\frac{1}{2}\right)\right)$

HW: Intervallo di confidenza per la media di un campione esponenziale

$X_1, X_2, \dots, X_n \sim expo\left(\frac{1}{\beta}\right)$

$\beta \approx \bar{X}$

eccetera ....

STATISTICA INDUSTRIALE ora 7

Note Title 11/03/2015

TEST STATISTICI

- Caso normale σ nota μ incognita

  $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ indip.

→ il test è sul parametro incognito (μ)

→ si fanno due ipotesi in opposizione e ci si domanda "quale sia vera"?

$H_0 : \mu = 10$ $H_1 : \mu \neq 10$

μ_0
valore target
AQL (acceptable quality level)

10,04   9,98   10,01   10,03   9,95   ...

Q: μ non può essere esattamente uguale a 10   NONSENSE

Non cerco quale è vera: so che è H_1

Cerco in effetti di capire:

se H_0 è plausibile, visti i dati / è compatibile con i dati } H_0 può essere accettata (test negativo)

se c'è evidenza statistica che H_1 sia vera } H_0 va rifiutata (test positivo)

★ Come si effettua il test?

dati → $\boxed{\text{black box}}$ → decisione

		test	
		0	1
caso	0	✓	fp
	1	fn	✓

fp: errori di I specie   fn: errori di II specie

α: prob. di un fp

β: prob. di un fn

$\boxed{bb} \longrightarrow \alpha \leq \overline{\alpha}$ ← tipicamente 5%

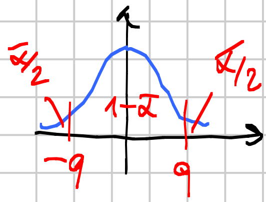
livello di significatività (assegnato)

① Primo modo: calcolo la statistica del test

$$Z := \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

la funz. ancillare era $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

$$q = \Phi^{-1}\left(1 - \frac{\overline{\alpha}}{2}\right)$$



$$\boxed{bb} \longrightarrow \begin{cases} \text{se } |Z| \leq q & \text{accetto } H_0 \\ \text{se } |Z| > q & \text{rifiuto } H_0 \end{cases}$$

vera H_0

$$\alpha = P(fp) = \text{"}P(\text{dire } H_1 \mid \text{vera } H_0)\text{"} = P(|Z| > q) = P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| > q\right) = \overline{\alpha}$$

$\mathcal{N}(0,1)$

→ Notazione: RA_T la regione di accettazione (di H_0) per una statistica T

① : $Z := \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ $\qquad RA_Z = [-q\,;\,q]$

② secondo modo: uso lo stimatore $\overline{X}$

$$RA_{\overline{X}} = \mu_0 \pm q\frac{\sigma}{\sqrt{n}}$$

→ Attenzione: assomiglia all'intervallo di conf

$$\mu \in \overline{x} \pm q\frac{\sigma}{\sqrt{n}}$$

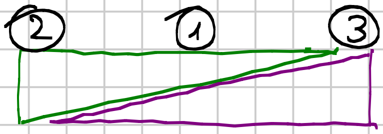
③ terzo modo: calcolo il p dei dati (p-value)

$$\alpha^* := 2 - 2\Phi(|z|) \quad \in [0,1]$$

$$RA_{\alpha^*} = [\bar{\alpha}\,;\,1]$$

HW: verificare che sia equivalente

complessità RA



complessità statistica

★ Con il p dei dati non occorre decidere prima $\bar{\alpha}$

→ se α^* è molto piccolo $\alpha^* = 10^{-4}$ → dico H_1

→ se α^* è grande $\alpha^* = 0{,}3$ → dico H_0

se α^* è vicino al 5% ovviamente occorre scegliere $\bar{\alpha}$.

- Curva operativa caratteristica
  permette di valutare le performance del test ($\boxed{bb}$)
  prima di utilizzarlo

$h : \mathbb{R} \to [0,1]$ $\qquad h : y \mapsto P(\text{accettare } H_0)$ quando $\mu = y$

spazio dove vive il parametro (μ)

$\mathcal{N}\left(y\,;\,\frac{\sigma^2}{n}\right)$

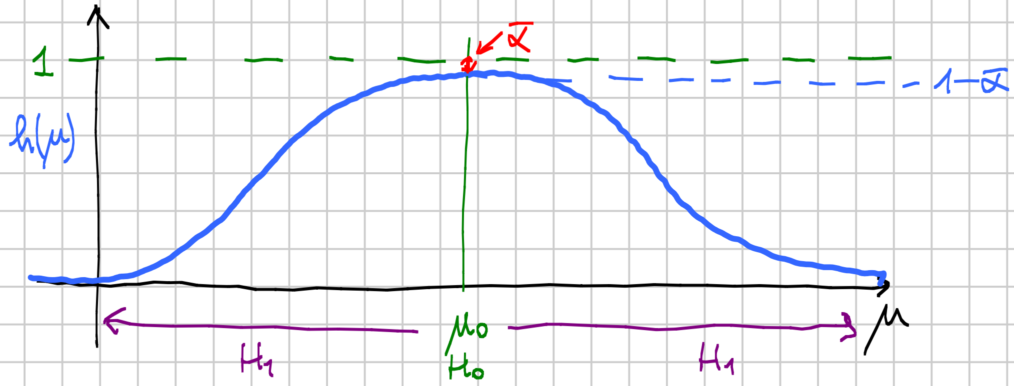
$$h(y) = P_{\mu=y}(\text{accetto } H_0) = P_{\mu=y}\left(\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| \leq q\right)$$

$$= P_{\mu=y}\Big(\underbrace{-q + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}}_{a} \leq \underbrace{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}_{\mathcal{N}(0,1)} \leq \underbrace{q + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}}_{b}\Big) = \int_a^b f_{\mathcal{N}(0,1)}\,dx$$

$$= \Phi(b) - \Phi(a)$$

$$b = b(y) = q + \frac{\mu_0 - y}{\sigma/\sqrt{n}}$$

$$a = a(y) = -q + \frac{\mu_0 - y}{\sigma/\sqrt{n}}$$



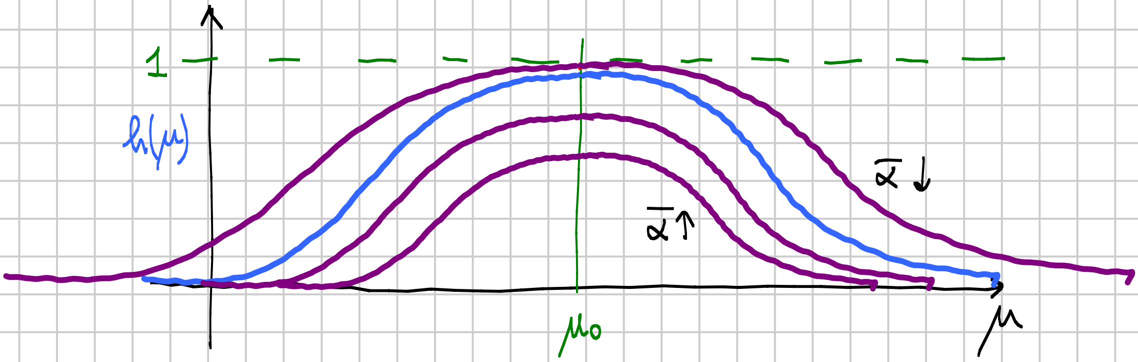
$\mu=\mu_0$ dico H_0 con prob $1-\bar{\alpha}$

$\mu\approx\mu_0$ dico H_0 con prob poco minore

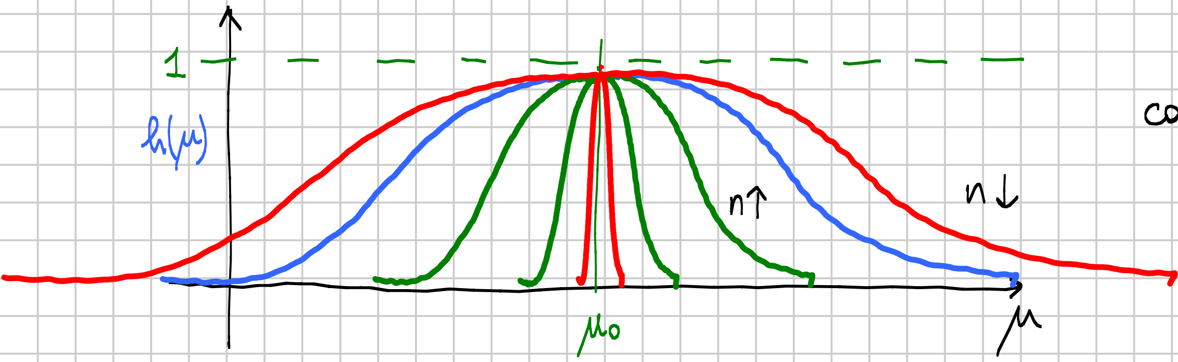
$\mu\neq\mu_0$ $h(\mu)=\beta(\mu)$ prob di errore di II specie (fn)

μ molto lontano da μ_0 dico H_1 con elevata probabilità

★ La potenza del test è $1-\beta(\mu)$ la "capacità" di dire H_1



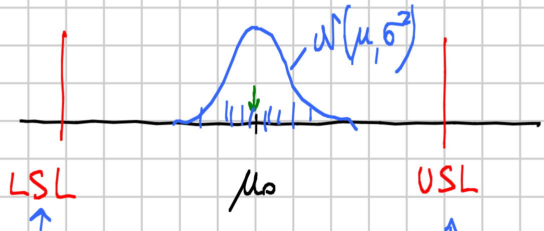
cambio $\bar{\alpha}$



cambio n

Se n aumenta il test è più potente, ma se è troppo potente ricado nel NONSENSE dell'ora 7

Process Capability



limiti di specifica

$$C_p = \frac{USL - LSL}{6\sigma}$$

$C_p = 1$ se i SL stanno a $\pm 3\sigma$
$\approx 2{,}7$ ‰ di difettosi

$C_p = \frac{4}{3}, \frac{3}{2}, \frac{5}{3}$ buoni

$C_p = 2$ SIX-SIGMA

$$C_{pk} = \min\left(\frac{USL-\mu}{3\sigma} ; \frac{\mu - LSL}{3\sigma}\right)$$

$C_{pk} \leq C_p$

Spesso $C_{pk}(\mu) \simeq C_{pk}(\hat{\mu})$

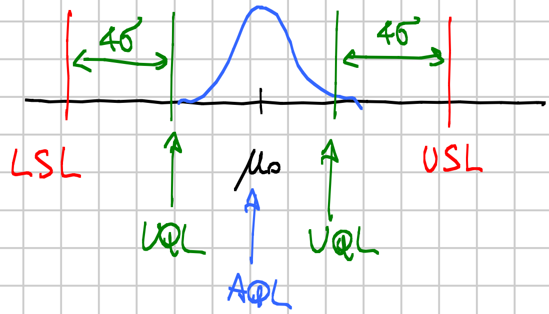
$\bar{x}$ stima di μ

★ Devo accorgermi che $\mu \neq \mu_0$ e dire H_1 tutte le volte che la gaussiana è così spostata che sto producendo troppi difettosi

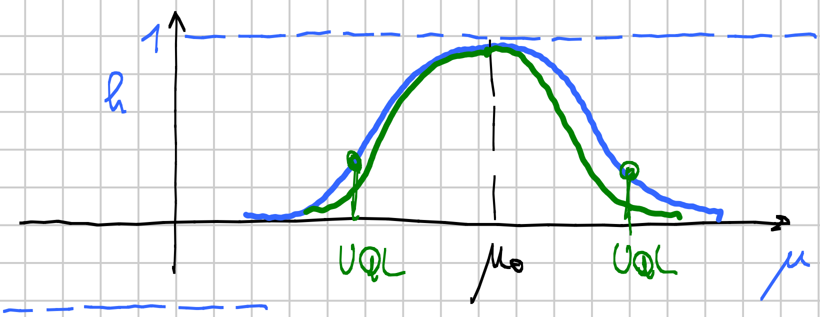
$C_{pk} < 1{,}33$ inaccettabile (mettiamo)

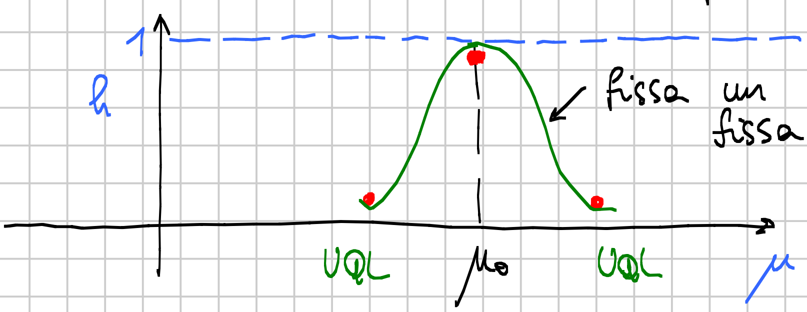
$\frac{USL-\mu}{3\sigma} < 1{,}33 \iff \mu > USL - 4\sigma$ UQL

$\frac{\mu - LSL}{3\sigma} < 1{,}33 \iff \mu < LSL + 4\sigma$ UQL



unacceptable quality level





fissa un n minimo
fissa α
fissa $\beta(UQL)$

★ Attenzione che nella realtà industriale non siano confusi i concetti : SL / UQL / $RA_{\bar{X}}$ tutti intervalli $\mu_0 \pm$ qualcosa

- Test per σ in caso di campione normale

$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

$\rightarrow \mu$ nota $\quad S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \qquad \boxed{\frac{S_*^2}{\sigma^2} n \sim \chi^2(n)}$

$\rightarrow \mu$ incognita $\quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \boxed{\frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)}$

★ Test bilaterale $\qquad H_0 : \sigma = \sigma_0$ (dato) $\qquad H_1 : \sigma \neq \sigma_0$

$\bar{\alpha}$ livello di significatività $\qquad$ ipotesi nulla $\qquad$ ipotesi alternativa

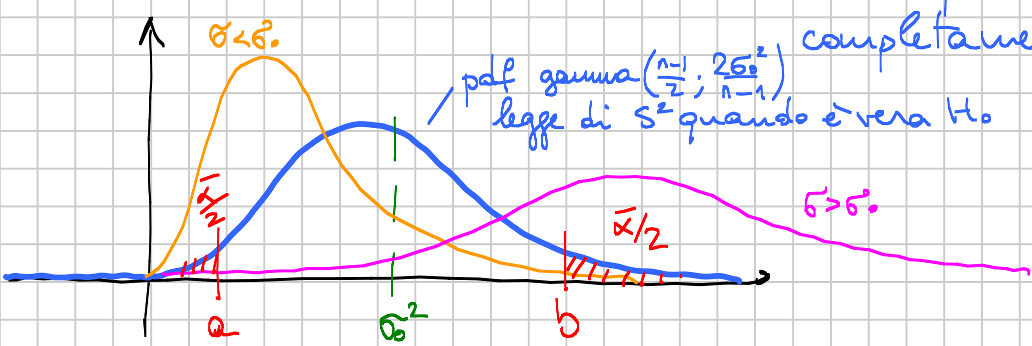
per costruire il test, si suppone vera H_0 e si studia la distribuzione di uno stimatore del parametro incognito determinandone i valori "verosimili"

$S^2 \approx \sigma^2 \qquad$ suppongo vera $H_0 : \sigma = \sigma_0$

$\frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1) \sim \text{gamma}\left(\frac{n-1}{2}; 2\right)$

$S^2 \sim \text{gamma}\left(\frac{n-1}{2}; \frac{2\sigma^2}{n-1}\right) \qquad S^2 \sim \text{gamma}\left(\frac{n-1}{2}; \frac{2\sigma_0^2}{n-1}\right)$

completamente specificata



$a = F^{-1}_{\text{gamma}}(\bar{\alpha}/2) \qquad RA_{S^2} = [a; b]$

$b = F^{-1}_{\text{gamma}}(1 - \bar{\alpha}/2)$

Se è vera H_0, $S^2 \in RA_{S^2}$ con prob $1 - \bar{\alpha}$

② Se $S^2 \in RA_{S^2}$ accetto H_0, altrimenti lo rifiuto perché è un forte indizio che sia vera H_1

① $\frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$ $\qquad W := \frac{S^2}{\sigma_0^2}(n-1) \overset{H_0}{\sim} \chi^2(n-1)$

$a' = F^{-1}_{\chi^2(n-1)}(\bar{\alpha}/2)$

$b' = F^{-1}_{\chi^2(n-1)}(1-\bar{\alpha}/2)$ $\qquad RA_W = [a' ; b']$

verifico equivalenza con ②

$$a' \leq \frac{S^2}{\sigma_0^2}(n-1) \leq b' \iff \frac{a'}{n-1}\sigma_0^2 \leq S^2 \leq \frac{b'}{n-1}\sigma_0^2$$

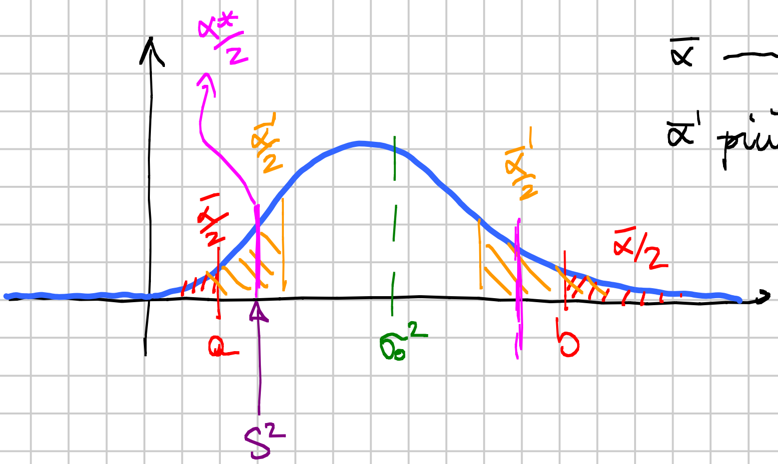
$$RA_{S^2} = \left[\frac{a'}{n-1}\sigma_0^2 ; \frac{b'}{n-1}\sigma_0^2\right]$$

devo verificare che $a = \frac{a'}{n-1}\sigma_0^2$ e analogo per b

$$a = F^{-1}_{\text{gamma}\left(\frac{n-1}{2};\frac{2\sigma_0^2}{n-1}\right)}\left(\frac{\bar{\alpha}}{2}\right) = \frac{\sigma_0^2}{n-1} \underbrace{F^{-1}_{\text{gamma}\left(\frac{n-1}{2};2\right)}}_{\chi^2(n-1)} = \frac{\sigma_0^2}{n-1} a'$$

- Il p-dei-dati

③

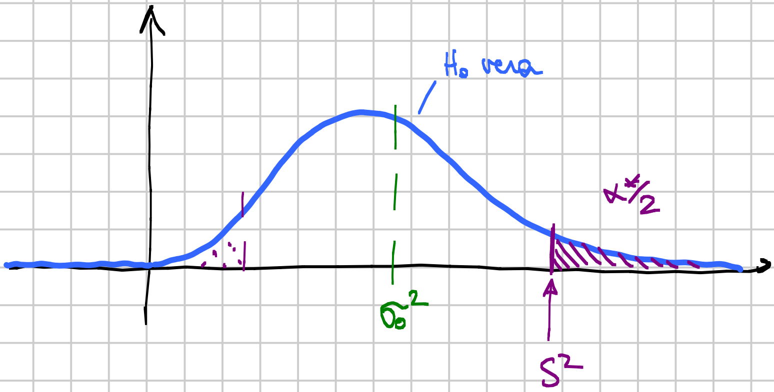


$\bar{\alpha} \rightsquigarrow$ dico H_0

$\bar{\alpha}'$ più grande $\rightsquigarrow$ dico H_1

Fissato il campione, dico H_0 se il lvl di sign è molto piccolo e dico H_1 se è molto grande.

Def: Il valore critico del lvl di sign. per cui cambio idea è il p dei dati e si denota con α^*



Altrimenti detto α^* è la probabilità che, sotto ipotesi H_0, si osservino dei dati "strani" come quelli del campione o più

Qui "strani" vuol dire che sembrano indicare H_1

$$\alpha^* = 2 \min\Big(\underbrace{F_{gamma}(s^2)}_{\text{coda SX}}\ ;\ \underbrace{1-F_{gamma}(s^2)}_{\text{coda DX}}\Big)$$

$\longrightarrow$ $\bar{\alpha} \leq \alpha^*$ accetto H_0

$\bar{\alpha} > \alpha^*$ rifiuto H_0

fissato $\bar{\alpha}$ questo vuol dire

$$RA_{\alpha^*} = [\bar{\alpha}\ ;\ 1]$$

★ Fatto importante: se è vera H_0, $\alpha^* \sim unif(0;1)$

$$\alpha = P(E_p) = P(\alpha^* \notin RA_{\alpha^*}) = P(\alpha^* < \bar{\alpha}) = \bar{\alpha}$$

se è vera H_1, α^* tende ad assumere valori piccoli con probabilità più elevata

<u>HW</u>: 1) Se F è la CdF di una v.a. X, allora $F(X) \sim unif(0;1)$

2) Se F è una CdF assegnata e $U \sim unif(0;1)$ allora $X := F^{-1}(U)$ è una v.a. con CdF F

Q: cosa si fa se F non è invertibile?

A: cercare Skorokhod sul Williams

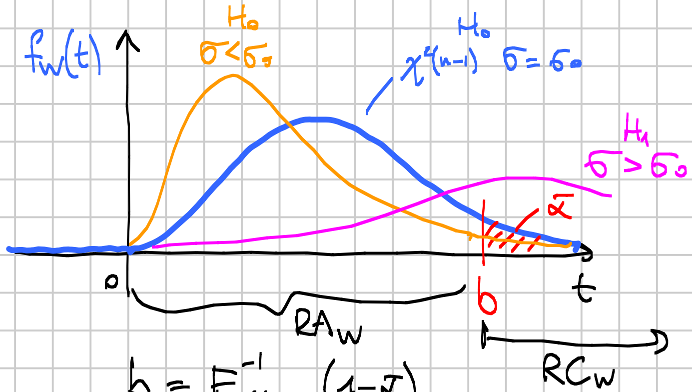
3) $\alpha^* \overset{H_0}{\sim} unif(0;1)$ nel caso dell'esempio

TEST UNILATERALI

$H_0: \sigma \leq \sigma_0 \qquad H_1: \sigma > \sigma_0$ (caso gaussiano, test su σ, μ incogn.)

1. funz. ancillare $\dfrac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$

$\rightarrow$ statistica $W := \dfrac{S^2}{\sigma_0^2}(n-1) \overset{\text{se } H_0 \text{ vera al bordo}}{\sim} \chi^2(n-1)$



diverse possibili densità di W a seconda di σ

$b = F^{-1}_{\chi^2(n-1)}(1-\bar{\alpha})$ $\qquad P(\chi^2(n-1) > b) = \bar{\alpha}$

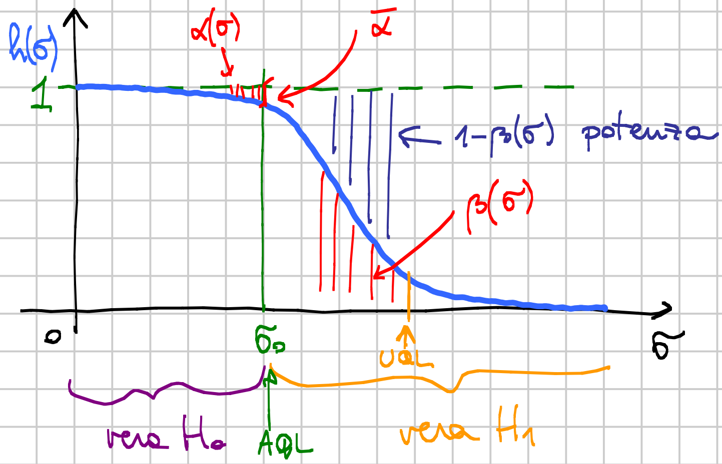
$$\alpha = P(f_p) = P(W > b) = P\Big(\underbrace{W\frac{\sigma_0^2}{\sigma^2}}_{\text{f.anc.}} > b\frac{\sigma_0^2}{\sigma^2}\Big) = P\Big(\chi^2(n-1) > b\frac{\sigma_0^2}{\sigma^2}\Big)$$

$H_0: \dfrac{\sigma_0^2}{\sigma^2} \geq 1 \Rightarrow \alpha \leq \bar{\alpha}$ uguaglianza al bordo

$H_1: \dfrac{\sigma_0^2}{\sigma^2} < 1 \Rightarrow \underbrace{P(\text{dire } H_1)}_{\text{potenza}} > \bar{\alpha}$

★ Curva O.C.

$$h(\sigma) = P(\text{dire } H_0) = P(W \leq b) = P\Big(\chi^2(n-1) \leq b\frac{\sigma_0^2}{\sigma^2}\Big)$$



Un test ben realizzato dice raramente H_1 se $\sigma \leq AQL$ e dice raramente H_0 se $\sigma \geq UQL$, mentre nella regione intermedia è indeciso

☆ Se scambio le ipotesi, AQL e UQL posso ottenere un test equivalente

$H_0: \sigma \leq AQL \quad H_1: \sigma > AQL$ lvl di sign. α assegnato

$\dots UQL \quad \beta(UQL) =: \gamma$

$H_0: \sigma \geq UQL \quad H_1: \sigma < UQL$ lvl di sign γ assegnato

$\longrightarrow$ Il primo test dice H_0 sse il secondo dice H_1

☆ Invece il test $H_0: \sigma \geq AQL \quad H_1: \sigma < AQL$ è completamente diverso.

HW: Curva OC per il test bilaterale su σ

■ TEST PER BERNOULLI/BINOMIALE (CONTROLLO QUALITÀ)

Ho un lotto di N pezzi e i difettosi dovrebbero essere una frazione inferiore a $3\% = AQL = p_0$. Per verificarlo, testo un campione di n pezzi e verifico che i difettosi non siano troppi.

X: # di difettosi nel campione

p: frazione difettosi nel lotto

assumo che $N \gg 1 \Rightarrow \boxed{X \sim bin(n, p)}$ (HW: perché? invece cos'è?)

$H_0: p \leq p_0 \qquad H_1: p > p_0$

■ TEST PER BERNOULLI / BINOMIALE (CONTROLLO QUALITÀ)

Ho un lotto di N pezzi e i difettosi dovrebbero essere una frazione inferiore a $3\% = AQL = p_0$. Per verificarlo, testo un campione di n pezzi e verifico che i difettosi non siano troppi.

X : # di difettosi nel campione

p : frazione difettosi nel lotto

assumo che $N \gg n \Rightarrow \boxed{X \sim bin(n, p)}$

$H_0 : p \le p_0$ $\quad H_1 : p > p_0$ $\quad \bar{\alpha}$ lvl di significatività

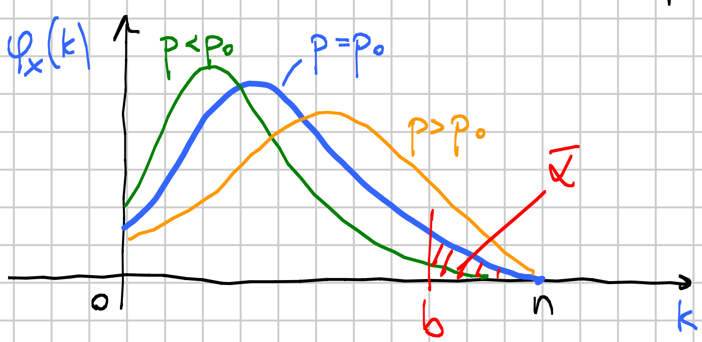
→ Approccio canonico tipo ① approssimato

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\cdot}{\sim} \mathcal{N}(0,1) \qquad Z := \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0 \text{ al bordo}}{\overset{\cdot}{\sim}} \mathcal{N}(0,1)$$

non occorre mettere $\hat{p}$

◉ Approccio moderno : X è la statistica del test

se H_0 è vera (al bordo) $p = p_0$, allora $X \sim bin(n, p_0)$



Scelgo il più piccolo b tale che $P(bin(n, p_0) \ge b) < \bar{\alpha}$

Se $X < b$ accetto H_0

Se $X \ge b$ rifiuto H_0

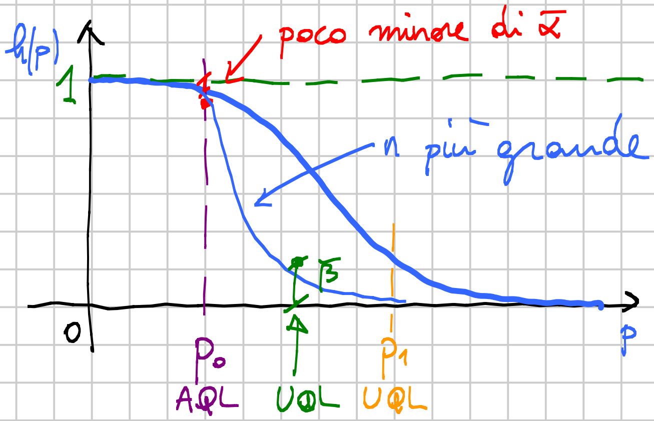
$fp \Rightarrow p \le p_0$

★ check : $P(fp) = P(X \ge b) = P(bin(X, p) \ge b) \le P(bin(X, p_0) \ge b) \le \bar{\alpha}$

$RA_X = \{0, 1, 2, \ldots, b-1\}$

- Curva OC

$$h = h(p) = P(\text{dire } H_0) = P(\text{bin}(n,p) \leq b-1) = F_{\text{bin}(n,p)}(b-1)$$



★ Diverse curve OC al variare di n e b hanno diverse performance che possono essere confrontate, ricavando a posteriori $\bar{\alpha}$ e $\bar{\beta}$ (e/o AQL e UQL che spesso non sono fissati in modo assoluto)

HW: formula del p-dei-dati per questo test

COLLEGAMENTO TRA INT. DI CONF. E TEST

Esempio int. conf. p Bernoulliana:

campione di n X: quelli che hanno la caratteristica

$X \sim \text{bin}(n, p)$ $\hat{p} = \frac{X}{n}$

$$p \in \hat{p} \pm q \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

quantile gaussiano

Quanti sono i mancini? $n = 5$ $X = 0$ non funziona

→ Si può dare almeno un upper bound? Sì:

90%? $p_0 = 90\%$

$H_0: p \geq p_0$ $H_1: p < p_0$ $\bar{\alpha} = 5\%$

$$\alpha^* = P(\text{bin}(n,p_0) = 0) = (1-p_0)^n = (0,1)^5 = 10^{-5} << \bar{\alpha} \quad H_1!$$

Modifico p_0 fino a che il test non dice H_0

$$(1-p_0)^5 = 5\% \qquad p_0 \approx 45\%$$

→ Questa procedura equivale a fare l'intervallo di confidenza $p \in [0 ; 45\%]$ con lvl di conf $1-\bar{\alpha} = 95\%$

★ Fatto generale: l'intervallo di confidenza di un parametro θ a lvl di conf $1-\alpha$ è l'insieme dei valori θ_0 tali che il test $H_0: \theta = \theta_0$ (o unilaterale) accetta H_0.

★ Fatto generale: se il test dice H_1 vuol dire che "è sicuro", indipendentemente da quanto piccolo sia n (tiene conto di quanto vale n, per dire H_1)

ora 12

● Funzioni ancillari

★ Campione $\mathcal{N}(\mu, \sigma^2)$

1) per μ, σ nota $\qquad \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

2) per μ, σ incognita $\qquad \boxed{\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)}$

3) per σ, μ nota $\qquad \dfrac{S_*^2}{\sigma^2} n \sim \chi^2(n)$

4) per σ, μ incognita $\qquad \boxed{\dfrac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)}$

Thm di Cochran

★ Campione Bernoulli p

5) approssimazione $\mathcal{N}$ $\qquad \dfrac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \mathrel{\dot\sim} \mathcal{N}(0,1)$

★ Campione esponenziale rate λ

6) $2n\lambda\bar{X} \sim \chi^2(2n)$

THM DI COCHRAN

- Sulla distribuzione normale multivariata

  - **Def** X v.a. a valori in $\mathbb{R}^n$ si dice avere legge normale se $\forall a \in \mathbb{R}^n$ $a \cdot X$ ha legge normale univariata

  - **Thm** se X ha legge normale multivariata $X : \Omega \to \mathbb{R}^n$ allora la sua densità congiunta è data da:

$$f : \mathbb{R}^n \to \mathbb{R}_+$$
$$f(x) = (2\pi)^{-n/2} \cdot \frac{1}{\sqrt{\det Q}} \exp\left\{\frac{1}{2}\langle x-\mu, Q^{-1}(x-\mu)\rangle\right\}$$

dove $\mu \in \mathbb{R}^n$ $\mu = E(X)$ $\mu = (\mu_1, \mu_2, \dots, \mu_n)$ $\mu_i = E(X_i)$

e dove $Q \in M_n$ simmetrica def positiva

$$\underbrace{Q = C(X)}_{\text{matrice di covarianza}} \qquad Q_{ij} := Cov(X_i ; X_j)$$

matrice di covarianza

★ Oppure X è concentrata su un sottospazio affine se $\det Q = 0$ e si può scrivere comunque la legge

  - **Cor** Se Q è diagonale, le componenti X_i sono indipendenti

    "per la legge $\mathcal{N}$ $Cov = 0 \Rightarrow$ indipendenza"

  - **Prop** Se $X \sim \mathcal{N}(\mu ; Q)$ e $N \in M_{k,n}$ allora

    1) NX è normale

    2) $NX \sim \mathcal{N}(N\mu ; NQN^T)$

★ Thm e Prop 1) discendono dal thm di Lévy sulla funz. caratteristica (trasf. di Fourier)

Dim 2) $Y = NX \qquad Y_i = \sum_{j=1}^{n} N_{ij} X_j$

$$E(Y_i) = \sum_{j=1}^{n} N_{ij} E(X_j) = \sum_j N_{ij}\mu_j = [N\mu]_i$$

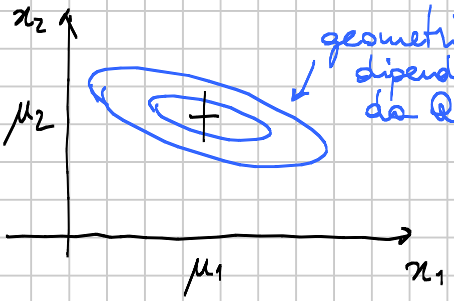
$$[C(Y)]_{i,j} := Cov(Y_i ; Y_j) = Cov\left(\sum_{h=1}^{n} N_{i,h} X_h ; \sum_{\ell=1}^{n} N_{j,\ell} X_\ell\right)$$

$$= \sum_{h,\ell} N_{i,h} N_{j\ell}\, Cov(X_h ; X_\ell) =: \sum_{h,\ell} N_{ih} N_{j,\ell} [C(X)]_{h,\ell} = \sum_{h,\ell} N_{ih} Q_{h,\ell} N_{j\ell} = [NQN^T]_{i,j}$$

HW: $v \in \mathbb{R}^n \qquad v + X \sim \mathcal{N}(v+\mu ; Q)$

☆ $n = 2 \qquad X \sim \mathcal{N}(\mu, Q) \qquad f : \mathbb{R}^2 \to \mathbb{R}_+ \qquad Q = \begin{pmatrix} \sigma_1^2 & c \\ c & \sigma_2^2 \end{pmatrix}$

curve di livello



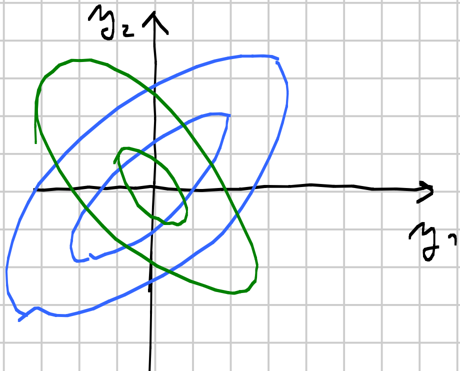
geometria dipende da Q

standardizzo :

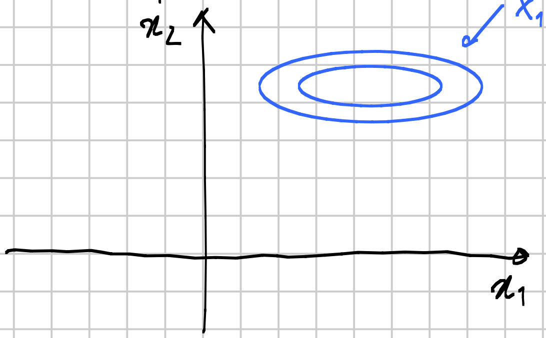
$$Y_i := \frac{X_i - \mu_i}{\sigma_i} \qquad i = 1,2$$

$E(Y_i) = 0 \qquad Var(Y_i) = 1$

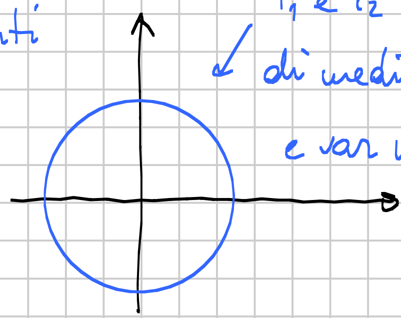
$$Y \sim \mathcal{N}\left(0 ; \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}\right)$$



X_1 e X_2 indipendenti



Y_1 e Y_2 indip. di media 0 e var uguale



☆ Dato $X \sim \mathcal{N}(\mu, Q)$ esiste una rotazione N di $\mathbb{R}^n$ tale che NX ha componenti indipendenti (thm spettrale)

☆ Se $X \sim \mathcal{N}(\mu, \sigma^2 I)$ allora qualunque rotazione non modifica l'indipendenza delle componenti

- Enunciato pratico del thm di Cochran

i. $X \sim \mathcal{N}(\mu ; \sigma^2 I)$ (μ e σ incognite)

ii. $\mu \in V$ sottospazio vettoriale di $\mathbb{R}^n$ $\qquad k = \dim V$

Allora 1) La proiezione ortogonale $\pi_V(X)$ di X su V è lo stimatore di massima verosimiglianza di μ e minimizza la distanza $|X - \pi_V(X)|$ è corretto e consistente

2) $SS := |X - \pi_V(X)|^2$ è una v.a. indipendente da $\pi_V(X)$ e inoltre $\frac{SS}{\sigma^2} \sim \chi^2(n-k)$

- Applicazione a campione normale

$X_1, X_2, \dots, X_n \quad X_i \sim \mathcal{N}(\mu, \sigma^2)$ indip, $\qquad \mu, \sigma$ incognite

$$X = (X_1, \dots, X_n) \sim \mathcal{N}(\underbrace{(\mu, \mu, \dots, \mu)}_{v \in \mathbb{R}^n} ; \sigma^2 I)$$

$v \in V := \mathrm{Span}((1,1,\dots,1)) \qquad k = \dim(V) = 1 \qquad \pi_V(X) = (Y, Y, \dots, Y)$

1) $$SS := |X - \pi_V(X)|^2 = \sum_{i=1}^{n} (X_i - [\pi_V(X)]_i)^2 = \sum_{i=1}^{n} (X_i - Y)^2$$

$$\frac{dSS}{dY} = \sum_i -2(X_i - Y) \underset{\uparrow}{=} 0 \qquad \sum X_i = nY \Rightarrow Y = \overline{X}$$

pongo (convessità)

$$\pi_V(X) = (\overline{X}, \overline{X}, \dots, \overline{X})$$

2) $$SS := \sum_{i=1}^{n} (X_i - \overline{X})^2$$ sum of squares

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

$$\frac{S^2}{\sigma^2}(n-1) = \frac{SS}{\sigma^2} \sim \chi^2(n-1) \text{ indip da } \overline{X}$$

Grazie all'indipendenza: $$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} =: \frac{Z}{\sqrt{\frac{W}{n-1}}}$$

dove $Z := \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

sono indipendenti

$$W := \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$$

Perciò la legge di $\dfrac{Z}{\sqrt{\frac{W}{n-1}}}$ è calcolabile e dipende solo da n

- <u>Def</u> Se $Z \sim \mathcal{N}(0,1)$ e $W \sim \chi^2(k)$ indipendenti

$\dfrac{Z}{\sqrt{\frac{W}{k}}} \sim t(k)$ ha legge t di Student con k gradi di libertà

STATISTICA INDUSTRIALE ora 13

Note Title 24/03/2015

★ Prossimamente facciamo qualche martedì 16:30 – 18:30

- Enunciato pratico del thm di Cochran

campione omoschedastico

i. $X \sim \mathcal{N}(\mu ; \sigma^2 I)$ (μ e σ incognite)

ii. $\mu \in V$ sottospazio vettoriale di $\mathbb{R}^n$ $k = \dim V$

Allora 1) La proiezione ortogonale $\pi_V(x)$ di x su V è lo stimatore di massima verosimiglianza di μ e minimizza la distanza $|x - \pi_V(x)|$ è corretto ~~e consistente~~

2) $SS := |x - \pi_V(x)|^2$ è una v.a. indipendente da $\pi_V(x)$ e inoltre $\frac{SS}{\sigma^2} \sim \chi^2(n-k)$

<u>Dim</u> $v_1, v_2, \dots, v_k$ base ortonormale di V

① $\pi_V(x) := \sum_{i=1}^{k} \langle x ; v_i \rangle v_i$ proiez. ortogonale su V

$\langle x - \pi_V(x) ; v \rangle = 0 \quad \forall v \in V$

$\forall v \in V \quad |x - v|^2 = |(x - \pi_V(x)) - \underbrace{(v - \pi_V(x))}_{\in V}|^2 = |x - \pi_V(x)|^2 + |v - \pi_V(x)|^2$

$$\geq |x - \pi_V(x)|^2$$

MLE $\ell(\overset{\mu}{\nu}) = \log f(x ; \mu = \nu) = c - \frac{1}{2} \langle x - \nu ; \overset{\sigma^2 I}{Q}(x - \nu) \rangle$

$= c - c_0 |x - \nu|^2 \qquad c_0 > 0$

$\pi_V(x)$ è MLE di μ

$E(\pi_V(x)) = \pi_V(E(x)) = \pi_V(\mu) = \mu$ stimatore corretto

② $R = X - \pi_V(X)$ $\qquad R = (R_1, R_2, \dots, R_n)$ cosiddetti residui

$v_{k+1}, \dots, v_n$ vettori che completano $v_1, \dots, v_k$ a base ortonormale di $\mathbb{R}^n$

$$N = \begin{pmatrix} - v_1 - \\ - v_2 - \\ \dots \\ - v_n - \end{pmatrix} \qquad [NX]_i = \langle v_i ; X \rangle \quad \forall i$$

$$[N\pi_V(X)]_i = \left[N \sum_{j=1}^{k} \langle v_j ; X \rangle v_j\right]_i = \begin{cases} \langle v_i ; X \rangle & i = 1, 2, \dots, k \\ 0 & i = k+1, \dots, n \end{cases}$$

$$[NR]_i = [NX - N\pi_V(X)]_i = \begin{cases} 0 & i = 1, 2, \dots, k \\ \langle v_i ; X \rangle & i = k+1, \dots, n \end{cases}$$

$$X \sim \mathcal{N}(\mu, \sigma^2 I) \Rightarrow NX \sim \mathcal{N}(N\mu ; \underset{N\sigma^2 I N^T}{\sigma^2 I}) \Rightarrow \text{componenti indipendenti}$$

$\Rightarrow N\pi_V(X)$ e NR indipendenti $\Rightarrow \pi_V(X)$ e R indip.

$$\frac{SS}{\sigma^2} = \frac{|R|^2}{\sigma^2} = \frac{|NR|^2}{\sigma^2} = \sum_{i=k+1}^{n} \frac{\langle v_i ; X \rangle^2}{\sigma^2} \sim \chi^2(n-k)$$

$$\langle v_i ; X \rangle \sim \mathcal{N}(\langle v_i ; \mu \rangle ; \sigma^2) \quad \text{indipendenti per } i = k+1, \dots, n$$

$\perp V \qquad \in V$

$$\frac{\langle v_i ; X \rangle}{\sigma} \sim \mathcal{N}(0, 1) \quad \text{iid per } i = k+1, \dots, n$$

HW $X_1, X_2, \dots, X_n$ $\quad X_i \sim \mathcal{N}(\mu_1, \sigma^2)$ $\quad \mu_1, \mu_2, \sigma^2$ incognite

$Y_1, Y_2, \dots, Y_m$ $\quad Y_i \sim \mathcal{N}(\mu_2, \sigma^2)$

i. $\overline{X} - \overline{Y}$ stimatore MLE, corretto e consistente di $\mu_1 - \mu_2$

S_X^2, S_Y^2 le varianze campionarie

$$S_p^2 := \frac{n-1}{m+n-2} S_X^2 + \frac{m-1}{m+n-2} S_Y^2$$ stimatore pooled di σ^2

ii. S_p^2 stimatore corretto (e consistente) di σ^2

iii. $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(n+m-2)$$ funz. ancillare per $\mu_1 - \mu_2$

iv. S_p^2 ha varianza minima tra tutte le combinazioni convesse di S_X^2 e S_Y^2

Cosa viene adesso:

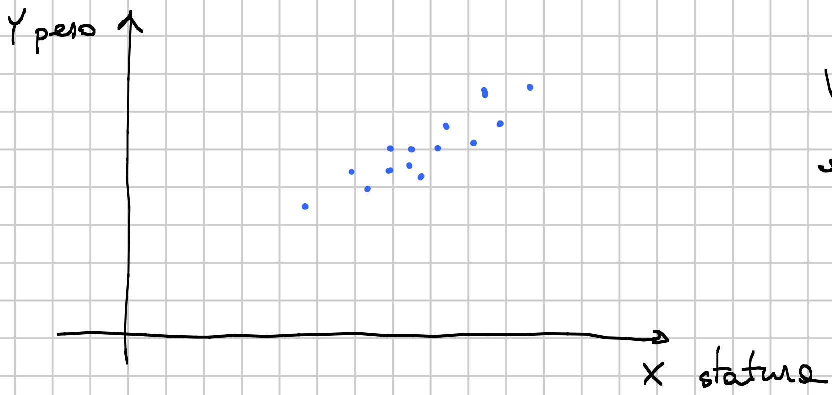
1. Regressione
2. Analisi della varianza

???

ora 14

▣ REGRESSIONE

Campione bivariato $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$



Vogliamo studiare la relazione fra due variabili

REGRESSIONE LINEARE SEMPLICE

$$Y = \beta_0 + \beta_1 x + e$$

errore additivo $\quad e \sim \mathcal{N}(0, \sigma^2)$

relazione lineare

la variabile "causa" (aka "indipendente" o "di ingresso") è immaginata deterministica (a volte è impostata dallo sperimentatore)

Notazione $\quad \boxed{Y = \beta_0 + \beta_1 x + e}$

ipotesi di omoschedasticità

va intesa: $\quad (x_1, Y_1), \dots, (x_n, Y_n) \qquad Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i ; \sigma^2)$

indipendenti

oppure: $\quad Y_i = \beta_0 + \beta_1 x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma^2)$ iid. $\quad$ (equivalente)

→ ci sono 3 incognite β_0, β_1 e σ^2

→ possono essere stimati $\quad B_0$, B_1 e S_e^2 saranno gli stimatori

HW: trovare B_0 e B_1 MLE di β_0 e β_1 e usare Cochran:

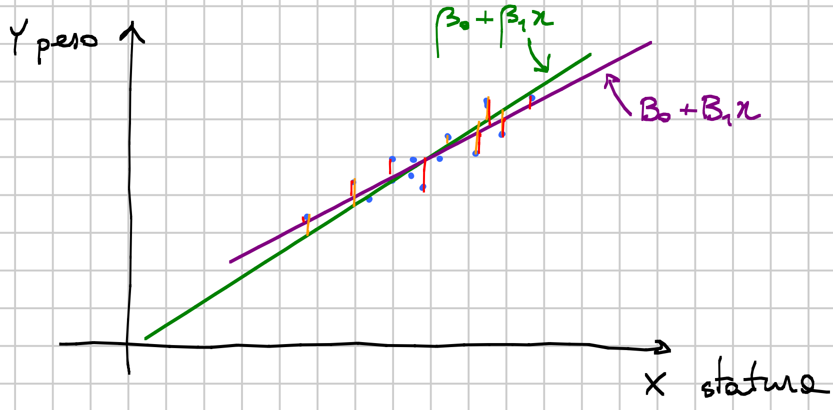
$$S_e^2 := \frac{\sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i)^2}{n-2} \qquad \frac{S_e^2}{\sigma^2}(n-2) \sim \chi^2(n-2)$$

indip. da B_0 e B_1

Si trova: $\quad \boxed{B_1 = \frac{\overline{xY} - \bar{x}\,\bar{Y}}{\overline{x^2} - \bar{x}^2}} \qquad B_0 = \bar{Y} - B_1 \bar{x} \qquad S_e^2 = \dots$

B_0 e B_1 sono funzioni lineari delle Y_i e quindi hanno legge normale

★ B_0 e B_1 non sono indipendenti: hanno covarianza non nulla



$\boxed{e_i} = Y_i - (\beta_0+\beta_1 x_i) \sim \mathcal{N}(0,\sigma^2)$

$\boxed{R_i} = Y_i - (B_0+B_1 x_i) \sim \mathcal{N}(0,?)$

complicate e non sono indipendenti

La retta B_0+B_1x si chiama retta di regressione oppure best fit (rispetto al campione)

$$SS=\sum_{i=1}^{n} R_i^2 \qquad S_e^2 := \frac{SS}{n-2} \qquad \sigma^2 \approx \frac{\sum e_i^2}{n}$$

se si dividesse per n si sottostimerebbe σ^2

- Quanto sono precisi gli stimatori? (intervalli di confidenza e test statistici, quindi: funzioni ancillari)

$$B_1 \sim \mathcal{N}(\beta_1 ; \sigma^2 k_1) \qquad B_0 \sim \mathcal{N}(\beta_0 ; \sigma^2 k_0)$$

dipende solo dalle x_i

faremo tutto per bene nel caso più generale

$$\mathrm{Cov}(B_0 ; B_1) = \sigma^2 k_{01}$$

dati; par. incognito

$$\frac{B_1-\beta_1}{\sigma\sqrt{k_1}} \sim \mathcal{N}(0,1)$$

incognita; dati; distrib. nota

non è f. anc.

$$\boxed{\frac{B_1-\beta_1}{S_e\sqrt{k_1}} \sim t(n-2)}$$

funz. ancillare

$$Z := \frac{B_1-\beta_1}{\sigma\sqrt{k_1}} \sim \mathcal{N}(0,1) \qquad W := \frac{S_e^2}{\sigma^2}(n-2) \sim \chi^2(n-2)$$

indipendenti per Cochran

$$\frac{B_1-\beta_1}{S_e\sqrt{k_1}} = \frac{Z}{\sqrt{W/(n-2)}} \sim t(n-2)$$

def di t di Student

$$\frac{B_0 - \beta_0}{\sigma\sqrt{K_0}} \sim \mathcal{N}(0,1) \qquad \boxed{\frac{B_0-\beta_0}{S_e\sqrt{K_0}} \sim t(n-2)}$$

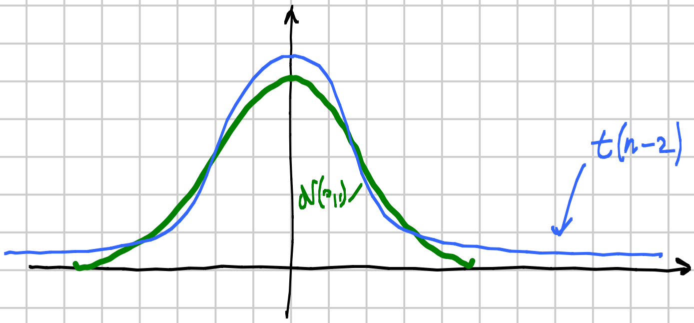
- Test per vedere se "c'è regressione" (= Y dipende da x)
  = le due variabili sono correlate / legate / ... ecc.

$H_0: \beta_1 = 0 \qquad H_1: \beta_1 \neq 0 \qquad \alpha$ liv. di significatività

retta orizzontale $\rightarrow Y_i \sim \mathcal{N}(\beta_0, \sigma^2)$ non dipendono da x

② $T := \dfrac{B_1}{S_e\sqrt{K_1}} \overset{H_0}{\sim} t(n-2) \qquad RA_T = [-q\ ;\ q]$

$$q = F^{-1}_{t(n-2)}\left(1 - \frac{\alpha}{2}\right)$$



i. è simile a $\mathcal{N}(0,1)$
ii. è simmetrica rispetto 0
iii. ha le code pesanti
(va a zero in modo polinomiale)
iv. per $n \to \infty$ $t(n) \to \mathcal{N}(0,1)$

☆ I quantili della t di Student si trovano come quelli della $\mathcal{N}(0,1)$ usando la CdF opportuna.
Vengono sempre più larghi e la differenza è maggiore quando n è piccolo

- Intervallo di confidenza per la risposta media

x generico, $Y \sim \mathcal{N}(\underbrace{\beta_0 + \beta_1 x}; \sigma^2)$ $1-\alpha$ lvl di confidenza

livello di ingresso — risposta associata a x — risposta media associata a x

→ $\beta_0 + \beta_1 x \in$ intervallo di confidenza

→ stimatore puntuale $\beta_0 + \beta_1 x \approx B_0 + B_1 x$ (corretto e consistente perché lo sono B_0 e B_1)

→ distribuzione stimatore $B_0 + B_1 x \sim \mathcal{N}(\beta_0 + \beta_1 x ; \sigma^2 p(x))$

$$Var(B_0 + B_1 x) = Cov(B_0 + B_1 x ; B_0 + B_1 x) = Cov(B_0;B_0) + 2x\,Cov(B_0;B_1) + x^2 Cov(B_1;B_1)$$

bilineare

$$= \sigma^2 \left[k_0 + 2x k_{01} + x^2 k_1\right]$$

polinomio di II grado
i coefficienti dipendono dalle x_i

→ funz. anc.

$$\frac{B_0 + B_1 x - (\beta_0 + \beta_1 x)}{\sigma \sqrt{p(x)}} \sim \mathcal{N}(0;1)$$

⇓ Cochran

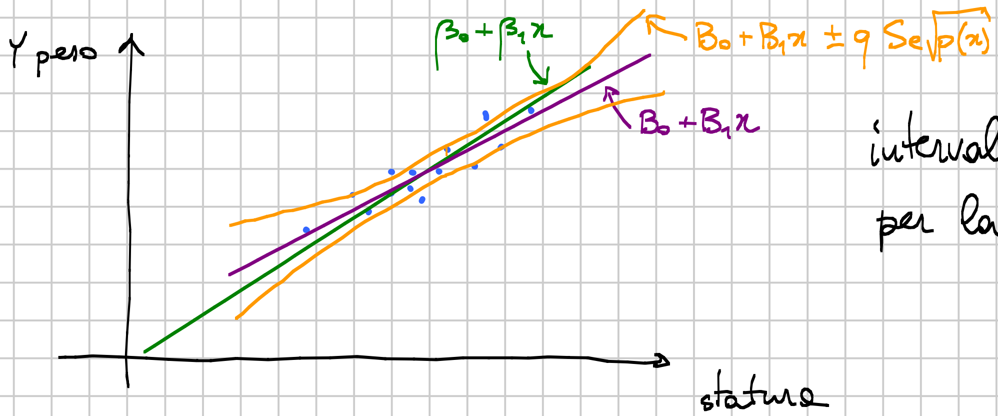
$$\frac{B_0 + B_1 x - (\beta_0 + \beta_1 x)}{S_e \sqrt{p(x)}} \sim t(n-2)$$

→ quantili $\pm q$ $\quad q = F^{-1}_{t(n-2)}\left(1 - \frac{\alpha}{2}\right)$

→ intervallo di conf

$$\beta_0 + \beta_1 x \in B_0 + B_1 x \pm q\, S_e \sqrt{p(x)}$$

$$p(x) = \frac{1}{n} + \frac{(x - \bar{x})^2}{n(\overline{x^2} - \bar{x}^2)}$$ (o qualcosa di simile)



intervallo di confidenza per la retta verde

- Intervallo di predizione per risposte future

x generico $\qquad Y \sim \mathcal{N}(\beta_0+\beta_1 x \,;\, \sigma^2)$

$\to$ $Y \in$ intervallo di predizione

$\to$ Y "viene previsto" da $\beta_0+\beta_1 x$ tenendo conto dell'incertezza σ

$\to$ $\beta_0+\beta_1 x \approx B_0+B_1 x \sim \mathcal{N}(\beta_0+\beta_1 x \,;\, \sigma^2 p(x))$

$\to$ $Y - (B_0+B_1 x) \sim \mathcal{N}(0 \,;\, \sigma^2(1+p(x)))$

dipendono da $Y_1, \dots, Y_n$

risposta di un esperimento futuro

] indipendenti

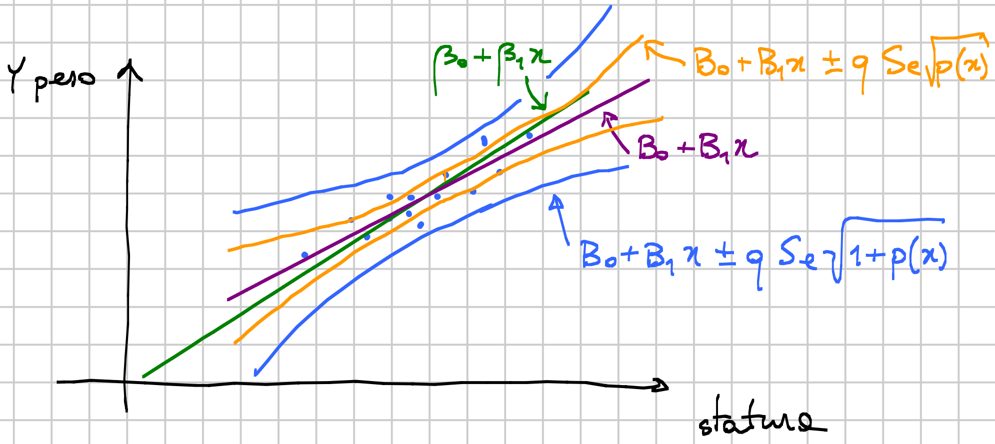
$$\boxed{Var(X-Y) = Var(X) + Var(Y) - 2\,Cov(X;Y)}$$

$\to$ f. ancillare (check)

$\to$ intervallo di predizione:

$$Y \in B_0+B_1 x \pm q\, S_e \sqrt{1+p(x)}$$

unica differenza con l'intervallo di confidenza per la risposta media



intervallo che contiene una frazione $1-\alpha$ dei punti

★ Siccome l'ipotesi di linearità di $Y = \beta_0+\beta_1 x + e$ è spesso una approssimazione valida localmente, è molto pericoloso usare questi intervalli al di fuori del range dei valori di x osservati nel campione.

HW: $X_1, X_2, \dots, X_n$ campione $\mathcal{N}(\mu, \sigma^2)$ (dati osservati)

incognite

$X_{n+1}, \dots, X_{n+m}$ esperimenti futuri

i. Intervallo di predizione per $\sum_{i=1}^{m} X_{n+i}$

ii. Intervallo di predizione per $\frac{1}{m}\sum_{i=1}^{m} X_{n+i}$

ora 16

mar 31 marzo 8:30–10:30
mer 1 aprile 11:30–13:30
LABORATORIO

sede scientifica di ingegneria

Regressione Lineare Multipla

$$\boxed{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + e} \qquad e \sim \mathcal{N}(0, \sigma^2)$$

↳ la variabile di ingresso diventa un vettore

$(x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,p}, Y_1)$ primo esperimento

.....

$(x_{n,1}, x_{n,2}, \dots \quad , x_{n,p}, Y_n)$ ultimo esperimento

$p < n$ necessario

$p \ll n$ preferibile

↳ si aggiunge di solito una x_0 "virtuale" (dummy variable) che vale sempre 1

$$x_{i,0} \equiv 1 \qquad i = 1, 2, \dots, n$$

→ scrittura matriciale : $Y = X\beta + E$

$$\boxed{Y \sim \mathcal{N}(X\beta; \sigma^2 I)}$$

dove : $Y = (Y_1, Y_2, \dots, Y_n)$ vettore colonna delle risposte

$X = (x_{i,j})_{i=1,\dots,n;\, j=0,\dots,p}$ matrice $n \times (p+1)$ dei valori di ingresso

$\beta = (\beta_0, \beta_1, \dots, \beta_p)$ vettore colonna dei coefficienti

$E \sim \mathcal{N}(0, \sigma^2 I)$ vettore gaussiano di $\mathbb{R}^n$

→ incognite : $\beta \in \mathbb{R}^{p+1}$, $\sigma > 0$

→ thm di Cochran i. $Y \sim \mathcal{N}(\underbrace{X\beta}_{\mu}; \sigma^2 I)$

ii. $\mu \in \mathrm{Im}(X) = X(\mathbb{R}^{p+1}) =: V$ ssv. di $\mathbb{R}^n$ di dimensione $p+1$

$\pi_V(Y) =: \hat{Y}$ trovo $\hat{Y}$ minimizzando $|R|^2$ dove $R = Y - \hat{Y}$

Sia $v \in V$ generico $v = Xb$ $\quad b \in \mathbb{R}^{p+1}$ qualsiasi

$$SS = SS(b) = |R|^2 = |Y - Xb|^2 = \langle Y - Xb; Y - Xb \rangle = (Y^T - b^T X^T)(Y - Xb)$$

$$\nabla_b SS = -2\,(Y^T - b^T X^T)\, X\, I \overset{!}{=} 0$$

$B \in \mathbb{R}^{p+1}$ quello che minimizza

$$Y^T X = B^T X^T X \quad \Leftrightarrow \quad X^T Y = X^T X B$$

$$\Uparrow$$

$$\boxed{B = (X^T X)^{-1} X^T Y}$$ stimatore per β

★ X di rango $p+1 \Rightarrow X^T X$ invertibile

$$\boxed{\hat{Y} = XB}$$ valori previsti per Y

→ $Y = xB$ equazione di regressione (analogo della retta viola)

$x \in \mathbb{R}^{p+1}$ vettore riga è un generico valore di ingresso e Y la risposta associata

→ $$SS = |R|^2 = \sum_{i=1}^{n} R_i^2 = |Y - \hat{Y}|^2 = \sum_{i=1}^{n} \Big(Y_i - \underbrace{\sum_{j=0}^{p} B_j x_{ij}}_{\hat{Y}_i \text{ previsto}} \Big)^2$$

$$\boxed{\frac{SS}{\sigma^2} \sim \chi^2(n-p-1)}$$

indipendente da $\hat{Y}$ e quindi da B

→ Stimatore per σ : errore standard S_e

$$\boxed{S_e := \sqrt{\frac{SS}{n-p-1}}}$$

$$\frac{S_e^2}{\sigma^2}(n-p-1) \sim \chi^2(n-p-1)$$

funz. anc.

HW: $E(S_e^2) = \sigma^2$; S_e^2 è consistente

→ Distribuzione di B

$$B = \underbrace{(X^T X)^{-1} X^T}_{=:N \text{ matrice } p+1 \times n} Y$$

$Y \sim \mathcal{N}(\mu ; Q)$ $\qquad$ $NY \sim \mathcal{N}(N\mu ; NQN^T)$

$$\boxed{B \sim \mathcal{N}\left(\beta ; \sigma^2 (X^T X)^{-1}\right)}$$

verifico: $N\mu = (X^T X)^{-1} X^T X \beta = \beta$

$$NQN^T = (X^T X)^{-1} X^T \sigma^2 I \, X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \cancel{X^T X} (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

- Esempio di inferenza

Test per vedere se la variabile x_j è significativa
(= se Y dipende da x_j)

$H_0: \beta_j = 0 \qquad H_1: \beta_j \neq 0 \qquad \alpha$ lvl di significatività

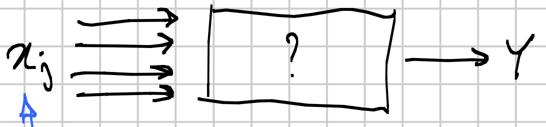
$\rightarrow B_j \approx \beta_j \qquad B_j \sim \mathcal{N}\left(\beta_j ; \sigma^2 \left[(X^T X)^{-1}\right]_{jj}\right)$

$\rightarrow$ funz. ancillare $\dfrac{B_j - \beta_j}{S_e \sqrt{\left[(X^T X)^{-1}\right]_{jj}}} \sim t(n-p-1)$

$\rightarrow$ statistica $T := \dfrac{B_j}{S_e \sqrt{\left[(X^T X)^{-1}\right]_{jj}}} \qquad RA_T = [-q ; q] \quad q = F^{-1}_{t(n-p-1)}\left(1 - \frac{\alpha}{2}\right)$

Selezione delle variabili

$$Y = \sum_{j=0}^{p} \beta_j x_j + e \qquad e \sim \mathcal{N}(0, \sigma^2)$$



candidati

x_j non impattante $\Rightarrow \beta_j = 0 \not\Rightarrow B_j = 0$

$B_j \neq 0$ sempre non nullo

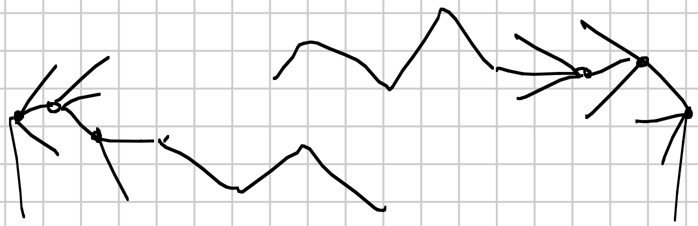
$$Y = B_0 + B_1 x_1 + \ldots + B_j x_j + \ldots + B_p x_p$$

★ Le variabili non impattanti andrebbero tutte tolte dal modello per evitare "rumore" se si usa per prevedere Y su dati futuri
→ inoltre tipicamente la precisione sugli altri stimatori migliora, togliendo le variabili ininfluenti

Lo strumento base per la selezione è il test $H_0: \beta_j = 0$ $H_1: \beta_j \neq 0$ che va fatto per tutte le variabili, tranne il termine noto.

★ Se tolgo una variabile il risultato dei test per le altre può cambiare
- → si toglie sempre una sola variabile alla volta
- → l'ordine in cui si tolgono è influente: non per forza se x_j ha $\alpha^* = 80\%$ e x_k ha $\alpha^* = 70\%$ è giusto togliere prima x_j
- → il metodo di selezione stepwise backward però funziona proprio così: toglie iterativamente la variabile con α^* max

→ il metodo di selezione stepwise forward inserisce una var alla volta, scegliendo quella con impatto maggiore (confronta α^* globale della regressione su modelli diversi)



non sempre si incontrano

→ la forward si fa male con Excel: nel caso si scelga, occorre uno strumento più sofisticato

★ Analisi della varianza di regressione

$Y = X\beta + e$ notazione matriciale

$Y = XB + R$

↓ tolgo alcune variabili : $\tilde{B} = (\tilde{B}_0, \tilde{B}_1, \dots, \tilde{B}_k, 0, \dots, 0)$ ($\tilde{B}_p$)

$Y = X\tilde{B} + \tilde{R}$ $\quad |R| \leq |\tilde{R}|$ perché B è ottimo

$|R|^2 = SS = SS_R = \sum_{i=1}^{n} R_i^2 \qquad SS_{\tilde{R}} = \sum_{i=1}^{n} \tilde{R}_i^2 \qquad SS_D = SS_{\tilde{R}} - SS_R \geq 0$

$$\boxed{\text{Ipotesi} : \beta = (\beta_0, \beta_1, \dots, \beta_k, 0, \dots, 0)}$$

$S_e^2 = \dfrac{SS_R}{n-p-1} \approx \sigma^2$ stimatore corretto

$\tilde{S}_e^2 = \dfrac{SS_{\tilde{R}}}{n-k-1} \approx \sigma^2$ stimatore corretto

$$SS_D = (n-k-1)\tilde{S}_e^2 - (n-p-1)S_e^2 \approx (p-k)\sigma^2$$

$S_D^2 := \dfrac{SS_D}{p-k} \approx \sigma^2$ stimatore corretto

$H_0: \beta_{k+1} = \beta_{k+2} = \dots = \beta_p = 0$ $\qquad$ H_1: non tutti nulli

$$F := \frac{S_D^2}{S_e^2}$$ statistica del test

→ in ipotesi H_0 fa circa 1

→ in ipotesi H_1 viene più grande

☆ È sempre possibile usare questo test per confrontare due modelli di regressione con set di variabili contenute uno nell'altro.

ora 18

☆ Gli strumenti di calcolo che fanno la stepwise backward e forward usano tipicamente test F.

Ad esempio: forward, passo $k+1$

ho già inserito x_i $\quad i \in I_k \subseteq \{1, 2, \dots, n\}$ $\qquad \# I_k = k$

$\forall j \in \{1, \dots, n\} \setminus I_k$ confronto la regressione con variabili I_k con quella con var $I_k \cup \{j\}$ e ottengo un valore di F_j per il test (e un valore di α_j^*)

Scelgo j che massimizza F_j (o minimizza α_j^*) equivalente

$I_{k+1} := I_k \cup \{j\}$

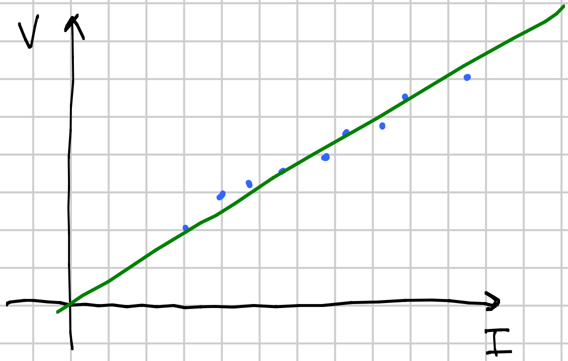
Mi fermo quando $\max_j F_j \leq$ soglia o $\min_j \alpha_j^* \geq$ soglia

☆ Stepwise generale: parte come forward ma ad ogni step verifica anche se ci sono variabili da togliere (backward step)

☆ Regola gerarchica: se nel modello includo un termine, devo includere anche tutti quelli "che lo dividono"

→ in particolare tengo il termine noto sempre

→ eccezione: se ho motivi teorici astratti per supporre che alcuni coefficienti siano nulli



legge di Ohm : $V = I \cdot R$

→ Esempio : $p = 3$ $\qquad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

r.g.

dummy : x_1^2 , $x_1 x_2$, $x_1^2 x_2$, x_2^2 , $x_1^2 x_2^2$

non significativi

● Metodi globali di selezione delle variabili

Si definisce uno "score" globale di regressione e poi si cerca il modello tra tutti quelli possibili che lo rende massimo

→ Tutti i modelli sono tanti : 2^p come minimo
($2^{2p+\binom{p}{2}}$ se includo un po' di dummy)

★ Esempio di "score" che non funziona : minimizzare $|R|$

→ con tutte le variabili il valore di $|R|$ è sempre minimo

★ Uno "score" che funziona è $\boxed{S_e}$ (minimizzato)

★ Un altro citato spesso si chiama AIC

- Coefficiente di determinazione (corretto o no)

(R^2) $\qquad R_d^2 := 1 - \dfrac{SS_R}{SS_Y}$ $\qquad\qquad SS_Y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$

$= SS_{\tilde{R}}$ con $k=0$

$0 \le R_d^2 \le 1$ è grande se l'ordine di grandezza dei residui è molto minore degli scarti di Y_i da $\overline{Y}$ (stima senza variabili x_j di regressione)

★ massimizzare $R_d^2 \iff$ minimizzare $|R|$

non si può usare per la selezione delle variabili

★ $\qquad R_a^2 := 1 - \dfrac{S_e^2}{S_Y^2}$ $\qquad$ varianza campionaria $S_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$

$= \tilde{S}_e^2$ con $k=0$

HW: $R_a^2 \le R_d^2$

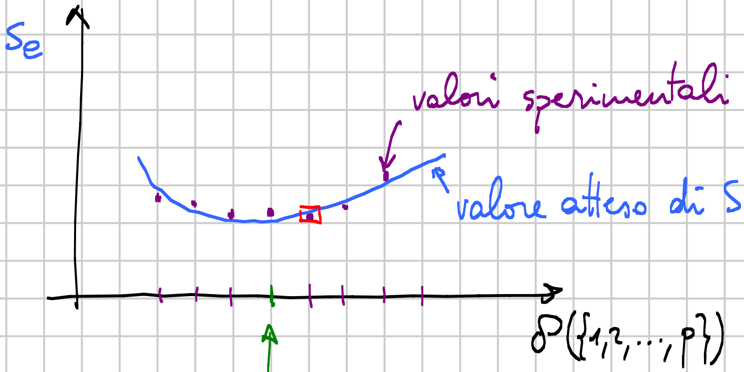
$\rightarrow$ misura la qualità della regressione

★ massimizzare $R_a^2 \iff$ minimizzare S_e

si può usare per la selezione delle variabili

!!!! Sono cruciali !!!!

- Selezione variabili con metodi globali



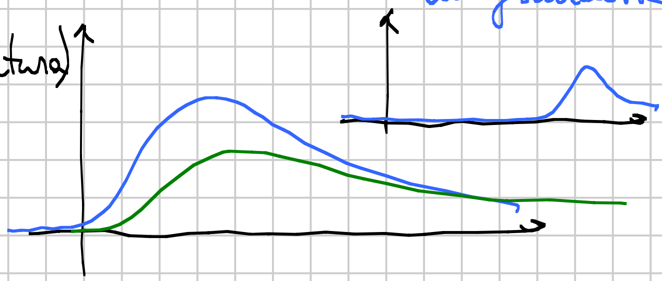
$$\frac{Se^2}{\sigma^2}(n-k-1) \sim \chi^2(n-k-1)$$

☆ Non è opportuno cercare il minimo di una funzione usando una sua stima casuale

☆ Molte volte tutti i metodi concordano.

■ GESTIONE DELLE VARIABILI

- categoriche nominali (regione, sesso)
- categoriche ordinali (titolo di studio)
  - niente
  - licenza media
  - diploma
  - laurea
- numeriche rapporto (reddito, popolazione città) diversi ordini di grandezza
- numeriche differenza (temperatura)



- Come si trattano le categoriche nominali

→ se le categorie sono solo 2, si codifica con 0/1 o −1/+1

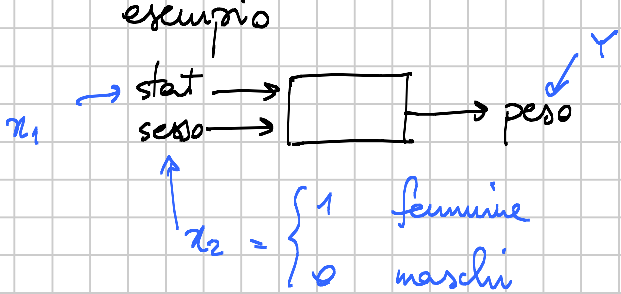
↳ diventa una variabile numerica

↳ i numeri scelti per la codifica non hanno nessuna importanza

★ Consiglio comunque 0/1 usando lo 0 per la categoria più comune (di default) oppure per il "no" se le categorie sono sì/no

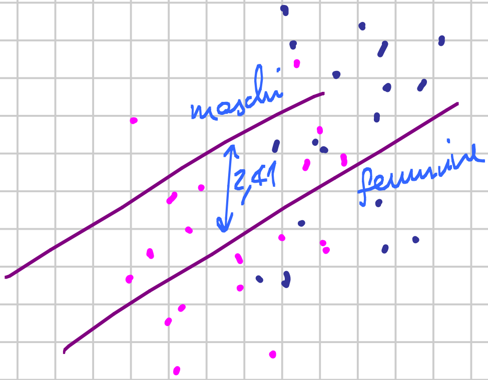
↳ Interpretazione risultati

esempio



$$Y = B_0 + B_1 x_1 + B_2 x_2$$

$$B_2 = -2{,}41$$



i maschi pesano in media e a parità di altezza 2,41 Kg in più.

→ se le categorie sono $k > 2$

★ Errore classico: codificare con i numeri da 1 a k

$$\text{regione} \rightarrow x_1 = \begin{cases} 1 & \text{Valle d'Aosta} \\ 2 & \\ \vdots & \\ 20 & \text{Sardegna} \end{cases}$$

$$Y = B_0 + B_1 x_1$$

l'ordine è del tutto arbitrario

★ Esplodo in dicotomiche

↳ scelgo una categoria di default
(a volte non è facile e occorre fare diverse prove)

↳ le altre $k-1$ diventano altrettante variabili 0/1

zona: nord/centro/sud

centro: default

id	zona	nord	sud
1	centro	0	0
2	sud	0	1
⋮	centro	0	0
⋮	nord	1	0
⋮	sud	0	1
⋮	⋮	⋮	⋮
n	centro	0	0

default $\rightarrow \overbrace{0, 0, 0, \ldots, 0}^{k-1}$

j-esima categ. $\rightarrow \overbrace{0, 0, \ldots, 1, 0, \ldots 0}^{k-1}$

posiz j

★ I coefficienti della regressione corrispondenti esprimono la differenza media attesa nella var di risposta tra la categoria in esame e quella di default

★ La scelta della codifica non è arbitraria: cambiare il default o cambiare altro (tipo imporre somma 0 su ogni riga: $-\varepsilon, -\varepsilon, \ldots, 1, \varepsilon, \varepsilon, \ldots, \varepsilon$) fa cambiare i risultati,

★ Selezione variabili → si collassano categorie assieme a quella di default semplicemente togliendo variabili dicotomiche non significative.
Eventualmente cambiare il default per collassare altri gruppi.

$x \longrightarrow \boxed{} \longrightarrow Y$

categorico nominale

si studia (meglio) anche con la ANOVA ad una via.

★ sono abbastanza equivalenti

- Categoriche ordinali

1) come le nominali, ignorando l'ordine

2) codificare rispettando l'ordine

la scala non importa, ma i rapporti tra le differenze sì: la regressione può venire diversa

tit di studio	C_1	C_2	C_3	C_4	C_5
niente	0	1	1	0	5
medie	1	2	3	1	8
diploma	2	3	5	3	13
laurea	3	4	7	4	17

equivalenti (C_1, C_2, C_3)

diversa (C_4)

- Variabili numeriche

→ attenzione che non sia una categorica nascosta (tipo di pompa: 0, 1, 2)

- Trasformazioni lineari di singole variabili

$$x_j \longmapsto a_j + b_j x_j = x_j' \qquad j = 1, 2, \dots, p$$

$$Y \longmapsto a_0 + b_0 Y = Y'$$

$$Y = B_0 + \sum_{j=1}^{p} B_j x_j + e$$

$$Y' = a_0 + b_0 Y = a_0 + b_0 B_0 + \sum_{j=1}^{p} b_0 B_j \frac{x_j' - a_j}{b_j} + b_0 e$$

$$= \underbrace{a_0 + b_0 B_0 - b_0 \sum \frac{B_j a_j}{b_j}}_{B_0'} + \sum_{j=1}^{p} \underbrace{\frac{b_0 B_j}{b_j}}_{B_j'} x_j' + \underbrace{b_0 e}_{e'} = B_0' + \sum_{j=1}^{p} B_j' x_j' + e'$$

$$e' \sim \mathcal{N}(0, b_0^2 \sigma^2)$$

$B_j = 0 \iff B_j' = 0$ la selez delle var non cambia

HW: R_a^2 non cambia

★ trasformazioni lineari non hanno essenzialmente alcun effetto (Equivalenze viste sopra 0/1 – −1/+1 e $c_1 \sim c_2 - c_3$)

★ Standardizzazione (o normalizzazione) delle variabili

$x_j : x_{1,j}, x_{2,j}, \dots, x_{n,j}$ $\qquad \bar{x}_j := x_{*,j} := \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$

devianza $\rightsquigarrow SS_j := \sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)^2$

varianza $\rightsquigarrow s_j^2 := \frac{1}{n-1} SS_j$

$x_{i,j} \longmapsto x'_{i,j} := \frac{x_{i,j} - \bar{x}_j}{s_j}$ $\qquad x'_{*,j} = 0 \qquad s_j' = 1$

→ ottengo coefficienti di regressione confrontabili

$B_j' = 2{,}4 \qquad B_k' = 6{,}1$ significa che x_k ha impatto maggiore su Y di x_j

↳ in realtà per questo basta $\frac{x_{i,j}}{s_j}$.

★ Forse da evitare sulle dicotomiche e sulle variabili rapporto.

Trasformazioni nonlineari

→ se ne applicate una, cambia tutto

→ i modelli possono essere confrontati solo qualitativamente

↳ grafici dei residui

↳ numero di variabili selezionate alla fine

→ se si trasformano (alcune delle) x_j ma non Y, i valori di R^2_d, R^2_a ed S_e sono confrontabili

★ trasformazioni tipiche: $\log(x)$, $\sqrt{x}$, $\frac{1}{x}$, x^a

power transformation

→ In natura oltre a $\mathcal{N}$ (contributi indipendenti additivi) si trovano spesso anche leggi lognormali (contributi indipendenti moltiplicativi): sono le classiche var. rapporto

→ conviene trasformarle

Def X ha legge lognormale di parametri μ, σ se $\log X \sim \mathcal{N}(\mu, \sigma^2)$

HW: $E(X) = e^{\mu + \sigma^2/2}$

→ Anche in tutti i casi in cui mi aspetto che gli errori e siano proporzionali a Y, ha senso fare il $\log(Y)$

$$Y' = \log(Y) \qquad Y' = \beta_0 + \beta_1 x_1 + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

omoschedastico

$$Y = \exp Y' = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdot e^{\varepsilon}$$

siccome è moltiplicata l'errore è proporzionale a Y

STATISTICA INDUSTRIALE ora 21

16.30 – 18.30 lab inf base 4 sede didattica ing.

HW dall'ore 7

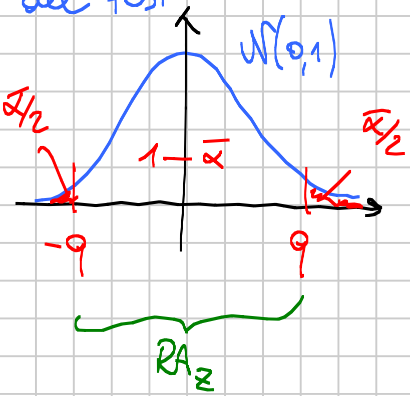
$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ σ nota μ_0 valore target $\bar{\alpha}$ liv signi.

$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$ bilaterale

→ test con funzione ancillare : tre livelli / modi per farli

① $Z := \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \overset{H_0}{\sim} \mathcal{N}(0,1)$ statistica del test

$RA_Z : [-q ; +q]$ $q = \Phi^{-1}\left(1 - \frac{\bar{\alpha}}{2}\right)$



③ $\alpha^* = 2 - 2\Phi(|z|)$

$RA_{\alpha^*} : [\bar{\alpha} ; 1]$

devo verificare che $z \in RA_Z \iff \alpha^* \in RA_{\alpha^*}$

$z \in RA_Z \iff |z| \leq q \iff \Phi(|z|) \leq \Phi(q) = 1 - \frac{\bar{\alpha}}{2}$ Φ monotona

$\iff \frac{\bar{\alpha}}{2} \leq 1 - \Phi(|z|)$

$\iff \bar{\alpha} \leq 2 - 2\Phi(|z|) =: \alpha^* \iff \alpha^* \in RA_{\alpha^*}$

* In generale si possono sempre fare i passaggi da $S \in RA_S$ generico per ricavare $\bar{\alpha}$: la statistica confrontata con $\bar{\alpha}$ è il p-dei-dati

HW dall'ora 11

$X \sim bin(n, p)$ $\quad$ p_0 valore AQL $\quad$ $\bar{\alpha}$ lvl di sign.

$H_0 : p \leq p_0 \qquad H_1 : p > p_0$ $\quad$ test unilaterale $\quad$ no funz. ancillare

① Scelgo il più piccolo b tale che $P(bin(n, p_0) \geq b) < \bar{\alpha}$

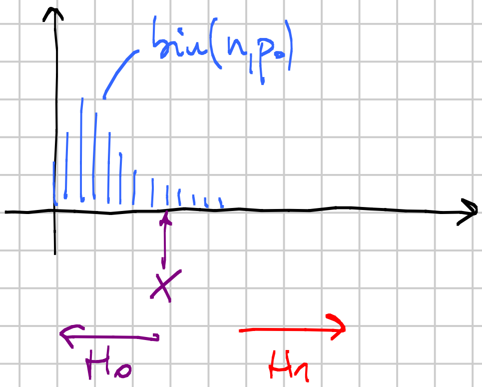
X statistica $\quad RA_X = \{0, 1, \dots, b-1\}$ $\quad$ def di b

② Calcolo il p-dei-dati

$$X \leq b-1 \iff F_{bin(n,p_0)}(X-1) \leq F_{bin(n,p_0)}(b-2) \leq 1-\bar{\alpha}$$

$$P(bin(n,p_0) \geq b-1) \geq \bar{\alpha} \iff F_{bin(n,p_0)}(b-2) \leq 1-\bar{\alpha}$$

$$\boxed{\bar{\alpha} \leq 1 - F_{bin(n,p_0)}(X-1) =: \alpha^*}$$



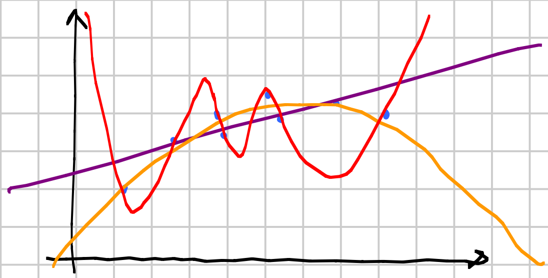
Se suppongo vera H_0, valori di X grandi sono "strani".
α^* è la prob (sotto H_0) di osservare valori strani come X o più strani

$$\alpha^* = P(bin(n,p_0) \geq X)$$

OVERFITTING NELLA REGRESSIONE

Esempio: campione bivariato, fit polinomiale

id	x	x^2	$x^3 \dots$	Y
1				
⋮				
n				



$n = 7$ se arrivo a x^6 trovo una curva che passa per tutti i punti

Overfitting: quando il modello interpola bene i punti del campione ma male quelli futuri

Diagnosi possibile: cross-validation

dìvido il campione in due sottocampioni A, B

fisso p (grado massimo)

regressione su A, calcolo i residui dai punti di B

regressione su B, calcolo i residui dai punti di A

sommo i quadrati di tutti i residui $\rightarrow SS_T$

$SS_T = SS_T(p)$ è uno score

cambio p e rifaccio

alla fine scelgo p che massimizza lo score

regressione su tutto

* SS_T è comunque una v.a. quindi va massimizzata con buon senso, tenendo conto che non sappiamo la sua variabilità e che comunque a parità di tutto un modello con meno variabili è preferibile

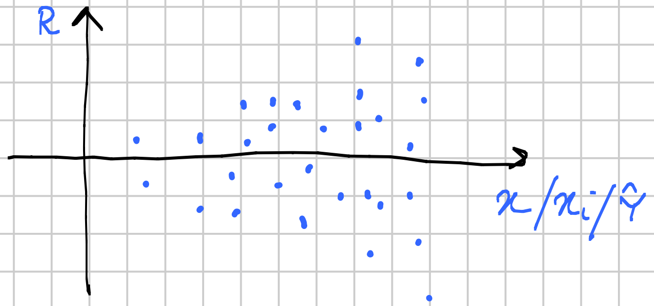
ora 22

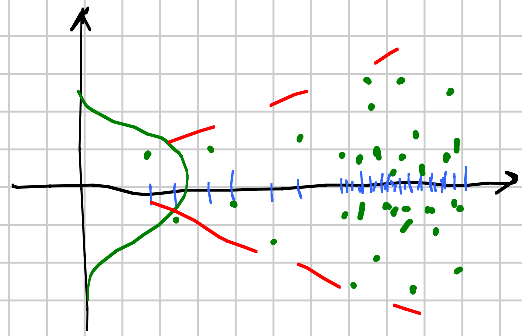
- Più in generale c'è overfitting nella regr. multipla tutte le volte che p è dell'ordine di grandezza di n
  - $\rightarrow$ se $p+1 = n$ si dice che il modello è saturo (residui nulli, impossibile fare inferenza: $n-p-1=0$)
  - $\rightarrow$ rule of thumb: $\frac{n}{p} \geq 5$ dovrebbe garantire che non ci sia overfitting
  - $\rightarrow$ in caso di overfitting l'inferenza è inaffidabile per questo la stepwise forward andrebbe preferita a quella backward.

Regressione pesata

generalizzazione per quando il modello omoschedastico non è ritenuto adeguato

- Diagnosi: analisi dei residui





- regr. lin. semplice → x in ascissa
- regr. lin. multiple → vanno fatti molti grafici

$x_1, x_2, \ldots, x_p, \hat{Y}$ ← previsti

- si cerca di "vedere" un andamento e proporre un modello:

$\sigma \propto \sqrt{x}$

$\sigma \propto \sqrt{x_i}$ per un i particolare

$\sigma \propto \hat{Y}$

questi casi hanno a volte una motivazione teorica

x distanza da percorrere in città

Y tempo di percorrenza in auto con traffico

n: # semafori $\propto x$ $\quad Y = T_1 + T_2 + \ldots + T_n \quad Var(Y) \propto n \propto x$

$\sigma \propto \sqrt{x}$

nei casi di "Y log-normale" o "misti" può essere ragionevole $\cancel{\sigma \propto Y}$ $\quad \sigma \propto \hat{Y}$

★ Attenzione alla diversa densità dei punti in ascisse che può ingannare

★ In genere i risultati non dipendono molto dal modello scelto.

- Modello eteroschedastico

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + e \qquad e \sim \mathcal{N}(0, \sigma^2(x_1, x_2, \dots, x_p))$$

σ ancora incognita, ma ipotizzo il tipo di dipendenza

$v = (x_1, \dots, x_p)$ modello noto

$\sigma \propto v$ ipotesi

→ definisco dei pesi : $i = 1, 2, \dots, n$

$$w_i := \frac{1}{v_i^2} = (v(x_{i,1}, \dots x_{i,p}))^{-2}$$

peso del punto i

→ somma quadrati residui pesati

$$SS_w := \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} B_j x_{ij}\right)^2 w_i$$

da minimizzare per trovare B_j

→ alternativa equivalente più pratica

occhio!

$$Y_i' = Y_i \sqrt{w_i} \qquad x_{i,j}' = x_{ij} \sqrt{w_i} \qquad i = 1, \dots, n \quad j = 0, 1, \dots, p$$

$$Y_i' = \beta_0 \sqrt{w_i} + \beta_1 x_{i1}' + \dots + \beta_p x_{ip}' + e_i' \qquad e_i' \sim \mathcal{N}(0, \sigma^2)$$

$$\Updownarrow \qquad\qquad \Updownarrow$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \underbrace{e_i' / \sqrt{w_i}}_{=: e_i} \qquad e_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{w_i}\right)$$

★ Studiare il modello "primato" come regressione omoschedastica è equivalente a studiare il modello originale con quella pesata

STATISTICA INDUSTRIALE ora 23

Note Title 22/04/2015

▣ ANALISI DELLA VARIANZA (cap 10 Ross)

x_i numeriche $\rightarrow$ $\boxed{?}$ $\longrightarrow$ $Y \sim \mathcal{N}(? ; \sigma^2)$ regressione

1 var categorica $\rightarrow$ $\boxed{?}$ $\longrightarrow$ $Y \sim \mathcal{N}(? ; \sigma^2)$ ANOVA a 1 via

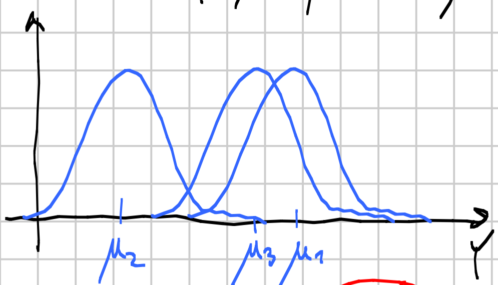
2 var categorica $\rightarrow$ $\boxed{?}$ $\longrightarrow$ $Y \sim \mathcal{N}(? ; \sigma^2)$ ANOVA a 2 vie

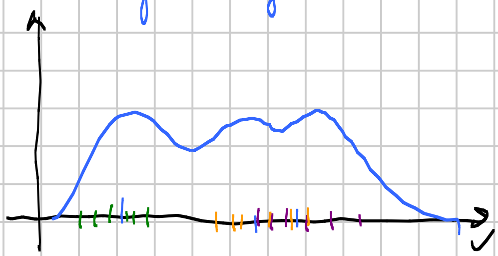
▣ ANOVA a 1 VIA

x categorica $\quad x \in \{1, 2, \dots, m\}$ codifica come etichette

m il numero di categorie $\quad Y \sim \mathcal{N}(\mu(x) ; \sigma^2)$

$x = 1, 2, \dots, m \quad \mu(x) = \mu_x \in \mathbb{R}$ ogni categoria ha la sua media





Campione $\cancel{(x_i ; Y_i)}$ $\rightarrow$

$$Y_{1,1}, \dots Y_{1,n_1} \sim \mathcal{N}(\mu_1 ; \sigma^2) \quad \text{iid}$$
$$Y_{2,1}, \dots Y_{2,n_2} \sim \mathcal{N}(\mu_2 ; \sigma^2) \quad \text{iid}$$
$$\dots$$
$$Y_{m,1}, \dots Y_{m,n_m} \sim \mathcal{N}(\mu_m ; \sigma^2) \quad \text{iid}$$

★ tipicamente Y sull'intero campione ha distribuzione multimodale

id	x_1	x_2	x_3 (Y)	...	x_K (categorica x)
1					
⋮					
n					

- Test fondamentale

H_0: Y non dipende da x H_1: Y dipende da x

$H_0: \mu_1 = \mu_2 = \dots = \mu_m$ H_1: non tutte uguali

- Diverse "varianze campionarie"

1) Varianza within (aka "entro i campioni")

Si stima σ^2 dentro ogni campione e poi si fa la media

→ Not. standard

$i = 1, 2, \dots, m$ $\quad Y_{i,*} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$ media campionaria

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{i,j} - Y_{i,*})^2$$ var. campionaria

$$\boxed{S_W^2 = \sum_{i=1}^{m} \pi_i S_i^2 \qquad \pi_i = \frac{n_i - 1}{N - m} \qquad N = \sum_{i=1}^{m} n_i}$$

HW: Per ogni scelta di π_i : $\sum_i \pi_i = 1$ la media pesata è uno stimatore corretto di σ^2. La scelta

$\pi_i \propto$ gdl della categoria i

è quella che minimizza la varianza

HW: Thm Cochran ⟹ distribuzione di S_W^2

$$\frac{S_W^2}{\sigma^2}(N-m) \sim \chi^2(N-m)$$

S_W^2 indipendente da $Y_{1,*}, Y_{2,*}, \dots Y_{m,*}$

2) Varianza between (aka "fra i campioni")

$\mu_i \approx Y_{i,*}$ $\qquad \frac{1}{m-1}\sum (Y_{i,*} - Y_{*,*})^2$ non proprio ma quasi

$$Y_{*,*} = \cancel{\frac{1}{m}\sum_{i=1}^{m} Y_{i,*}} \qquad Y_{*,*} = \frac{1}{N}\sum_{i,j} Y_{i,j} = \sum_{i=1}^{m} \frac{n_i}{N} Y_{i,*}$$

(check)

$$Y_{i,*} \sim \mathcal{N}\left(\mu_i ; \frac{\sigma^2}{n_i}\right) \qquad (Y_{i,*} - \mu_i)^2 \approx \operatorname{Var}(Y_{i,*}) = \frac{\sigma^2}{n_i}$$

$$\boxed{S_B^2 = \sum_{i=1}^{m} \frac{n_i}{m-1} (Y_{i,*} - Y_{*,*})^2}$$

☆ Se H_0 è vera

S_B^2 è uno stimatore corretto di σ^2 indipendente da S_W^2

☆ Se H_0 è falsa

S_B^2 è tendenzialmente più grande (e comunque indip. da S_W^2)

→ Verifichiamo la prima

$H_0 : \mu_i \equiv \mu \quad i = 1, 2, \dots, m \qquad Y_{i,j} \sim \mathcal{N}(\mu, \sigma^2)$

$$Y_{i,*} = \frac{1}{n_i}\sum_j Y_{i,j} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n_i}\right)$$

$$Y_{*,*} = \frac{1}{N}\sum_{i,j} Y_{i,j} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

$$X_i = \sqrt{n_i}\, Y_{i,*} \sim \mathcal{N}(\mu\sqrt{n_i} ; \sigma^2)$$

$$X = (X_1, \dots, X_m) \sim \mathcal{N}(\nu ; \sigma^2 I) \qquad \nu \in \operatorname{Span}(\sqrt{n_1}, \sqrt{n_2}, \dots, \sqrt{n_m})$$

$\nu_i := \mu\sqrt{n_i}$ $\qquad$ ssv di $\mathbb{R}^m$ di dim 1

$$\sum_{i=1}^{m} (X_i - \mu\sqrt{n_i})^2 \text{ lo minimizzo } \quad \hat{\mu} = \frac{\sum \sqrt{n_i}\, X_i}{\sum n_i} = \frac{\sum n_i Y_{i,*}}{\sum n_i} =: Y_{*,*}$$

$$0 = \sum_i \sqrt{n_i}\,(X_i - \hat{\mu}\sqrt{n_i})$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}\left(x_i - Y_{*,*}\sqrt{n_i}\right)^2 \sim \chi^2(m-1) \quad \text{indipendente da } Y_{*,*}$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{m} n_i\left(Y_{i,*} - Y_{*,*}\right)^2 =: \boxed{\frac{S_B^2}{\sigma^2}(m-1) \sim \chi^2(m-1)}$$

quindi S_B^2 è uno stimatore corretto di σ^2

ANOVA A UNA VIA

- Se H_0 è falsa S_B^2 è tendenzialmente maggiore di σ^2

$$S_B^2 = \frac{1}{m-1}\sum_{i=1}^{m} n_i (Y_{i*} - Y_{**})^2 \qquad SS_B = (m-1) S_B^2$$

devianza between

$$SS_B = \sum_{i=1}^{m} n_i (Y_{i*} - \mu_i + \mu_i - \mu + \mu - Y_{**})^2$$

dove $\mu := \mu_* = \frac{1}{N}\sum_{i,j}\mu_i = \frac{1}{N}\sum_{i=1}^{m} n_i\mu_i$

$$= \sum_i n_i\left[(Y_{i*}-\mu_i)^2 + (\mu_i-\mu)^2 + (\mu - Y_{**})^2\right]$$

$$+ 2\sum_i n_i (Y_{i*}-\mu_i)(\mu_i-\mu) + 2\sum_i n_i (Y_{i*}-\mu_i)(\mu - Y_{**}) + 0$$

$$E\,SS_B = \sum_i n_i \underbrace{Var(Y_{i*})}_{\sigma^2} + \sum_i n_i(\mu_i-\mu)^2 + N\,Var(Y_{**}) + 0 - 2N\,Var(Y_{**})$$

$$(\mu - Y_{**})\sum n_i (Y_{i*}-\mu_i) = (\mu - Y_{**})(N Y_{**} - N\mu) = -N(Y_{**}-\mu)^2$$

$$Y_{i*} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \qquad Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \qquad Y_{i*} \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right)$$

$$Y_{**} = \frac{1}{N}\sum_{i=1}^{m} n_i Y_{i*} \qquad Y_{**} \sim \mathcal{N} \qquad E(Y_{**}) = \frac{1}{N}\sum_{i=1}^{m} n_i\mu_i =: \mu$$

$$Var(Y_{**}) = \sum_{i=1}^{m} \frac{n_i^2}{N^2}\frac{\sigma^2}{n_i} = \frac{\sigma^2}{N^2}\sum_i n_i = \frac{\sigma^2}{N}$$

$$E\,SS_B = m\sigma^2 + \sum_{i=1}^{m} n_i(\mu_i-\mu)^2 - \sigma^2$$

se è vero H_1

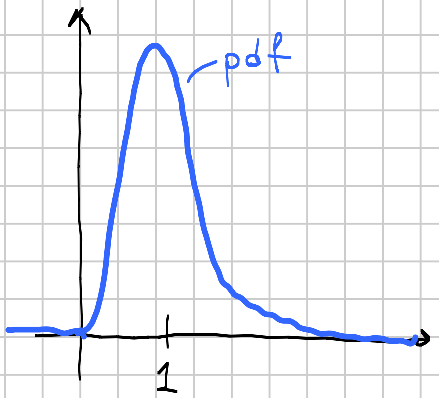
$$E\,S_B^2 = \frac{1}{m-1}E\,SS_B = \sigma^2 + \frac{1}{m-1}\sum_{i=1}^{m} n_i(\mu_i-\mu)^2 > \sigma^2$$

★ Prima di fare il test introduciamo la F di Fisher

Distribuzione F di Fisher

Def: se $V \sim \chi^2(m)$ $W \sim \chi^2(n)$ indipendenti

$$\frac{V/m}{W/n} \sim F(m;n)$$ legge F di Fisher con m gdl al numeratore e n gdl al denominatore



★ Confronto delle varianze di due campioni normali (Cap 8)

$X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu, \sigma^2)$
$Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\nu, \tau^2)$ indipendenti

$H_0 : \sigma = \tau$ $\qquad H_1 : \sigma \neq \tau$ $\qquad$ (altri esempi: $H_0\ \sigma \geq \tau$, $H_0\ \sigma \leq 10\tau$)

$\frac{\sigma}{\tau} \approx \frac{S_X}{S_Y}$ $\qquad \frac{S_X^2}{\sigma^2}(m-1) \sim \chi^2(m-1)$ $\qquad \frac{S_Y^2}{\tau^2}(n-1) \sim \chi^2(n-1)$ indip

$$\frac{\frac{S_X^2}{\sigma^2}}{\frac{S_Y^2}{\tau^2}} \sim F(m-1; n-1) \qquad \frac{S_X^2}{S_Y^2}\left(\frac{\sigma}{\tau}\right)^{-2} \sim F(m-1; n-1)$$ funz. ancillare

$H_0 : \frac{\sigma}{\tau} = 1$ $\qquad$ statistica del test $R := \frac{S_X^2}{S_Y^2} \overset{H_0}{\sim} F(m-1; n-1)$

$\bar{\alpha}$ livello di sign. assegnato

$a = F^{-1}_{F(m-1;n-1)}\left(\frac{\bar{\alpha}}{2}\right) = INV.F\left(1-\frac{\bar{\alpha}}{2}; m-1; n-1\right)$

$b = F^{-1}_{F(m-1;n-1)}\left(1-\frac{\bar{\alpha}}{2}\right) = INV.F\left(\frac{\bar{\alpha}}{2}; m-1; n-1\right)$ (check)

$a^* = F^{-1}\left(\frac{\alpha^*}{2}\right)$ $\qquad b^* = F^{-1}\left(1-\frac{\alpha^*}{2}\right)$ $\qquad R = a^*$ o $R = b^*$

$F(R) = \frac{\alpha^*}{2}$ o $F(R) = 1-\frac{\alpha^*}{2}$

$\alpha^* = 2F(R)$ o $\alpha^* = 2 - 2F(R)$

$\alpha^* = 2\min\left(F(R)\,;\,1-F(R)\right)$ p dei dati

HW: curva OC

★ ANOVA a una via

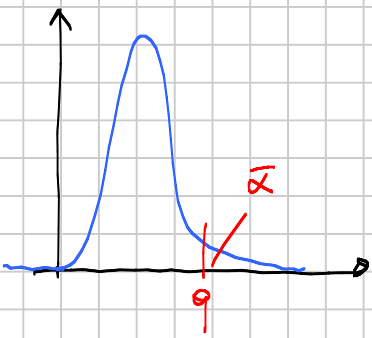
$R = \dfrac{S_B^2}{S_W^2}$ statistica del test

- se è vera H_0 $\dfrac{S_B^2}{\sigma^2}(m-1) \sim \chi^2(m-1)$ $\dfrac{S_W^2}{\sigma^2}(N-m) \sim \chi^2(N-m)$

$$\Rightarrow R = \frac{\frac{S_B^2}{\sigma^2}}{\frac{S_W^2}{\sigma^2}} \sim F(m-1; N-m)$$

- se è vera H_1 R è tendenzialmente più grande

→ allora si imposta un test unilaterale

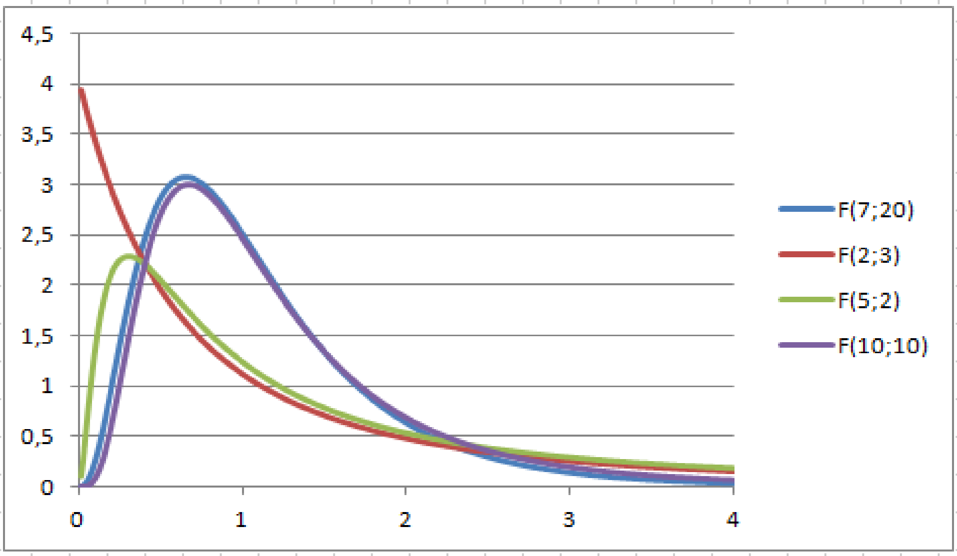


$q = F^{-1}_{F(m-1;N-m)}(1-\bar{\alpha}) = \mathrm{INV.F}(\bar{\alpha}\,; m-1\,; N-m)$

$RA_R = [0\,; q]$

$\alpha^* = 1 - F(R)$ (check)

ora 25



$$S_B^2 = \frac{1}{m-1}\sum_{i=1}^{m} n_i (Y_{i*} - Y_{**})^2 \qquad S_W^2 = \sum_{i=1}^{m} \frac{n_i - 1}{N-m} s_i^2$$

- Identità delle devianze (algebrica)

$$\boxed{SS_Y = SS_W + SS_B}$$

dove $S_Y^2 = \frac{1}{N-1}\sum_{i,j}(Y_{i,j} - Y_{**})^2$ var. camp. complessiva

$gdl_Y = N-1$

dove $SS_i = gdl_i \cdot S_i^2$

★ SS_Y e SS_W sono più veloci da ricavare di SS_B

Dim $SS_B + SS_W = \sum_{i=1}^{m} n_i (Y_{i*} - Y_{**})^2 + \sum_{i=1}^{m}(n_i - 1)\, s_i^2$

$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i*})^2$

$$= \sum_{i,j}(Y_{i*} - Y_{**})^2 + \sum_{i,j}(Y_{ij} - Y_{i*})^2$$

$$\overset{?}{=} \sum_{i,j}(Y_{i,j} - Y_{**})^2 = SS_Y$$

B.D.C. $\sum_{i,j}(Y_{i*} - Y_{**})(Y_{ij} - Y_{i*}) \overset{?}{=} 0$

$$\sum_{i,j}\left[Y_{i*}Y_{ij} - Y_{i*}^2 - Y_{**}Y_{ij} + Y_{**}Y_{i*}\right]$$

si semplificano sommando su j si semplificano sommando su j

- Altra inferenza relativa a questo modello (in ipotesi H_1)

→ inferenze su μ_i

$$\mu_i \approx Y_{i*} \sim \mathcal{N}\left(\mu_i ; \frac{\sigma^2}{n_i}\right) \qquad \frac{Y_{i*} - \mu_i}{\sigma/\sqrt{n_i}} \sim \mathcal{N}(0,1)$$

$$\boxed{\frac{Y_{i*} - \mu_i}{S_W/\sqrt{n_i}} \sim t(N-m)}$$

→ int conf

→ test statistici

→ inferenza su $\mu_i - \mu_j$

$$\mu_i - \mu_j \approx Y_{i*} - Y_{j*} \sim \mathcal{N}\left(\mu_i - \mu_j \, ; \, \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_j}\right)$$

$$\boxed{\frac{Y_{i*} - Y_{j*} - (\mu_i - \mu_j)}{S_W \cdot \sqrt{n_i^{-1} + n_j^{-1}}} \sim t(N-m)}$$

→ int conf
→ test statistici

★ Se il test fondamentale dice H_0 (e se ci credo: dipende dalla potenza / numerosità) allora si fa inferenza più elementare:

$$Y_{i,j} \sim \mathcal{N}(\mu \, ; \sigma^2)$$ un campione normale

$$\mu \approx Y_{**} = \overline{Y} \qquad \frac{Y_{**} - \mu}{S / \sqrt{N}} \sim t(N-1)$$

$$\sigma^2 \approx S^2 \qquad \frac{S^2}{\sigma^2}(N-1) \sim \chi^2(N-1)$$

• Altre considerazioni pratiche

1) Fare sempre se possibile l'analisi dei residui
   → va fatta a mano
   → controllare: outliers ; omoschedasticità ; normalità

$$R_{i,j} = Y_{i,j} - \underset{\uparrow}{Y_{i*}}$$ e graficarli in qualche modo

miglior predittore per Y_{ij}

2) In caso di eteroschedasticità: provare trasformazioni nonlineari
   a) dati lognormali → $\log(Y_{ij})$
   b) dati conteggi assimilabili alla legge di Poisson → $\sqrt{Y_{ij}}$

★ Se $X \sim \mathrm{Pois}(\nu)$ $P(X=k) = \frac{\nu^k}{k!} e^{-\nu}$, $k = 0, 1, \ldots$ $E(X) = \nu$, $\underline{Var(X) = \nu}$

$\boxed{\sqrt{X} \text{ ha media che dipende da } \nu \text{ e varianza } \approx \frac{1}{4}}$

$$Y = \sqrt{X} \qquad X \sim \mathrm{Pois}(\nu)$$

$$\mathrm{Var}(Y) = E(Y^2) - E(Y)^2 = E(\overset{\nu}{X}) - E(Y)^2$$

$$E(Y) = E(\sqrt{X}) = \sum_{k=0}^{\infty} \sqrt{k}\, \frac{\nu^k}{k!} e^{-\nu}$$

$$E(Y)^2 = \sum_{i,j=0}^{\infty} \sqrt{i}\sqrt{j}\, \frac{\nu^{i+j}}{i!\,j!} e^{-2\nu} = \sum_{k=0}^{\infty} \sum_{i+j=k} \sqrt{i(k-i)}\, \frac{(2\nu)^k}{k!} \binom{k}{i} e^{-2\nu} \cdot 2^{-k}$$

$$= \sum_{k=0}^{\infty} e^{-2\nu} \frac{(2\nu)^k}{k!} \underbrace{\sum_{i=0}^{k} \binom{k}{i} i^{1/2} (k-i)^{1/2} \cdot 2^{-k}}_{\frac{k}{2}+\frac{1}{4}}$$

$X \sim \text{Pois}(\nu) \qquad Y = \sqrt{X}$

$$(E(Y))^2 = \underbrace{\sum_{k=0}^{\infty} e^{-2\nu} \frac{(2\nu)^k}{k!}}_{\text{legge Poisson } 2\nu} \sum_{i=0}^{k} \underbrace{\binom{k}{i} \cdot 2^{-k}}_{\text{legge bin}(k,\frac{1}{2})} i^{1/2}(k-i)^{1/2}$$

$$\sum_{i=0}^{k} \binom{k}{i} \cdot 2^{-k} i^{1/2}(k-i)^{1/2} = E\left[\sqrt{Z(k-Z)}\right] \qquad \text{dove } Z \sim \text{bin}\left(k; \tfrac{1}{2}\right)$$

$$\approx \frac{k}{2} \qquad \overset{!}{=} \frac{k}{2} E\left[\sqrt{\frac{2}{k}Z\left(2-\frac{2}{k}Z\right)}\right] \qquad W := \frac{2}{k}Z - 1$$

$$\overset{!}{=} \frac{k}{2} E\left[\sqrt{(1+W)(1-W)}\right] = \frac{k}{2} E\left[\sqrt{1-W^2}\right]$$

$k \gg 1 \quad Z \,\dot{\sim}\, \mathcal{N}\left(\frac{k}{2}; \frac{k}{4}\right) \qquad W \,\dot{\sim}\, \mathcal{N}\left(0; \frac{1}{k}\right)$

$k \gg 1 \quad \sqrt{1-W^2} \approx 1 - \frac{1}{2}W^2 \quad \Rightarrow$

$$\sum_{i=0}^{k} \binom{k}{i} \cdot 2^{-k} i^{1/2}(k-i)^{1/2} \approx \frac{k}{2} E\left(1 - \frac{1}{2}W^2\right)$$

$$= \frac{k}{2} - \frac{k}{2} \cdot \frac{1}{2} \cdot \frac{1}{k} = \frac{k}{2} - \frac{1}{4}$$

HW: formalizzare con un limite

$$(E(Y))^2 = \sum_{k=0}^{\infty} e^{-2\nu} \frac{(2\nu)^k}{k!} \left(\frac{k}{2} - \frac{1}{4} + r(k)\right) = \frac{1}{2} \cdot 2\nu - \frac{1}{4} + \sum_{k=0}^{\infty} e^{-2\nu} \frac{(2\nu)^k}{k!} r(k)$$

tende a zero

$$\text{Var}(Y) = \nu - E(Y)^2 = \frac{1}{4} - \underbrace{\sum_{k=0}^{\infty} e^{-2\nu} \frac{(2\nu)^k}{k!} r(k)}_{\text{tende a 0 per } \nu \to \infty}$$

$V_\nu \sim \text{Pois}(2\nu) \,\dot{\sim}\, \mathcal{N}(2\nu; 2\nu)$

$$\text{termine corr} = E(r(V_\nu)) \xrightarrow[\nu \to \infty]{} 0$$

HW: formalizzare

ANOVA A DUE VIE

$x_1, x_2 \to \boxed{?} \to Y$

categoriche

continua

condiz. normale

x_1 m categorie $\{1, 2, \dots, m\}$

x_2 n categorie $\{1, 2, \dots, n\}$

$m \times n$ combinazioni

per ciascuna ho un certo numero $l_{i,j}$ di osservazioni

Struttura dei dati

id	x_1	x_2	Y
1			
⋮			
N			

$\longrightarrow$ $Y_{i,j,k}$ $\quad i = 1, 2, \dots, m$

$j = 1, 2, \dots, n$

$k = 1, 2, \dots, l_{i,j}$

✱ Condizione di ortogonalità / buona decomposizione delle devianze

$l_{i,j}$ non dipende da i,j $\quad \equiv l$

			x_1		
	1	2			m
1	7	4			
x_2 2		2		0	
			5		
n					

$l_{i,j}$ $\longrightarrow$ devo buttare via dati e imporre

$$l := \min_{i,j} l_{i,j}$$

✱ Se una delle caselle è vuota $l_{i,j} = 0$ l'unica possibilità è buttare via la colonna o la riga.

✱ Altri approcci generali (si trovano nei software) vanno considerati qualitativi.

$$\boxed{Y_{i,j,k} \sim \mathcal{N}(\mu_{i,j}\,;\,\sigma^2)}$$

indipendenti

$\longrightarrow$ due casi : $l = 1$ (senza repliche) $\quad l > 1$ (con repliche)

1WAY ANOVA $\longrightarrow$ 2WAY ANOVA W/O R $\longrightarrow$ 2WAY ANOVA W/R

☒ SENZA REPLICHE $(\ell=1)$

• Ipotesi fondamentale di linearità

$$\mu_{i,j} = \underbrace{\mu}_{\text{media globale}} + \underbrace{\alpha_i}_{\text{effetto riga}} + \underbrace{\beta_j}_{\text{effetto colonna}}$$

→ ipotesi tecnica (wlog) : $\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0$

HW: check μ, α_i, β_j qualsiasi $\longrightarrow \mu', \alpha_i', \beta_j' := \dots$ t.c. $\mu_{ij} = \mu'_{ij}$ + ip. tecn.

$$\boxed{Y_{i,j} = \mu + \alpha_i + \beta_j + e_{i,j}} \qquad e_{i,j} \sim \mathcal{N}(0, \sigma^2)$$

$\mu, \alpha_i, \beta_j, \sigma^2$ incognite

• Test principali

a) $H_0 : \alpha_i \equiv 0$ — Y non dipende da x_1 | H_1 : non tutti zero — Y dipende da x_1 | test per l'effetto riga

b) $H_0 : \beta_j \equiv 0$ — Y non dipende da x_2 | H_1 : non tutti zero — Y dipende da x_2 | test per l'effetto colonna

ora 27

• Stimatori

$$\mu + \alpha_i \approx Y_{i*} \sim \mathcal{N}\left(\mu + \alpha_i ; \frac{\sigma^2}{m}\right) \quad \times$$

$$\mu + \beta_j \approx Y_{*j} \sim \mathcal{N}\left(\mu + \beta_j ; \frac{\sigma^2}{n}\right) \quad \times$$

$$\mu \approx Y_{**} \sim \mathcal{N}\left(\mu ; \frac{\sigma^2}{mn}\right) \quad \checkmark$$

$$\alpha_i \approx \underbrace{Y_{i*} - Y_{**}}_{\text{NON sono indipendenti}} \sim \mathcal{N}\left(\alpha_i ; \frac{n-1}{mn}\sigma^2\right) \quad \checkmark$$

vedi prossima pagina

$$Y_{i*} - Y_{**} = Y_{i*} - \frac{1}{n}\sum_{a=1}^{n} Y_{a*} = Y_{i*}\left(1-\frac{1}{n}\right) - \frac{1}{n}\sum_{a\neq i} Y_{a*}$$

INDIPENDENTI

$$\left(\frac{n-1}{n}\right)^2 \frac{\sigma^2}{m} + \frac{1}{n^2}(n-1)\frac{\sigma^2}{m} = \boxed{\frac{n-1}{mn}\sigma^2}$$

$$\beta_j \approx Y_{*j} - Y_{**} \sim \mathcal{N}\left(\beta_j ; \frac{m-1}{mn}\sigma^2\right) \quad ✓$$

* Per σ^2, come nella regressione (e nell'ANOVA a 1 VIA) si fa una media quadratica dei residui

↳ previsti? $Y_{i,j} \approx \mu + \alpha_i + \beta_j$ miglior preditore

$\approx Y_{i*} + Y_{*j} - Y_{**}$ HW: trovare la varianza

$$R_{i,j} = Y_{i,j} - Y_{i*} - Y_{*j} + Y_{**}$$

$$R_{i,j} = Y_{i,j}\left(1 - \frac{1}{n} - \frac{1}{m} + \frac{1}{mn}\right) + \sum_{b\neq j} Y_{ib}\left(-\frac{1}{n}+\frac{1}{mn}\right) + \sum_{a\neq i} Y_{aj}\left(-\frac{1}{m}+\frac{1}{mn}\right)$$

$$+ \sum_{\substack{a\neq i \\ b\neq j}} Y_{a,b} \cdot \frac{1}{mn}$$

$$Var(R_{i,j}) = \frac{\sigma^2}{m^2n^2}\left\{(m-1)^2(n-1)^2 + (m-1)^2(n-1) + (m-1)(n-1)^2 + (m-1)(n-1)\right\}$$

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

$$SS_e := \sum_{i,j} R_{i,j}^2 \qquad E(SS_e) = mn\, Var(R_{i,j}) = (m-1)(n-1)\sigma^2$$

$$\boxed{S_e^2 := \frac{SS_e}{(m-1)(n-1)}}$$ stimatore corretto di σ^2

HW: Cochran $\Rightarrow$ $\frac{S_e^2}{\sigma^2}(m-1)(n-1) \sim \chi^2((m-1)(n-1))$

indipendente da Y_{i*}, Y_{*j}

- Falsi stimatori di σ^2

★ $S_R^2 := \frac{1}{m-1} \sum_{i=1}^{m} n (Y_{i*} - Y_{**})^2$

$H_0\ a) \Rightarrow S_R^2 \approx \sigma^2 \qquad \frac{S_R^2}{\sigma^2}(m-1) \sim \chi^2(m-1)$

$H_1\ a) \Rightarrow S_R^2 \gtrsim \sigma^2$

★ Analogo per le colonne

- Come si fanno i test

★ $\boxed{U_R := \frac{S_R^2}{S_e^2} \overset{H_0 a)}{\sim} F(m-1; (m-1)(n-1))}$

$H_1 a) \Rightarrow U_R$ tipicamente grande $\longrightarrow$ test unilaterale

$\alpha_R^* = \text{DISTRIB.F}\left(U_R\ ;\ m-1\ ;\ (m-1)(n-1)\right)$

★ Analogo per le colonne

- Identità delle devianze (conseguenza ipotesi di ortogonalità)

$SS_Y = SS_R + SS_C + SS_e$

sono tutte indipendenti

$\frac{SS_Y}{\sigma^2} \overset{H_0 a) b)}{\sim} \chi^2(mn-1) \qquad \frac{SS_R}{\sigma^2} \overset{H_0 a)}{\sim} \chi^2(m-1) \qquad \frac{SS_C}{\sigma^2} \overset{H_0 b)}{\sim} \chi^2(n-1) \qquad \frac{SS_e}{\sigma^2} \sim \chi^2((m-1)(n-1))$

$mn - 1 = (m-1) + (n-1) + (m-1)(n-1)$

- Altre inferenze :

★ $\alpha_i \approx Y_{i*} - Y_{**}$

$$\boxed{\frac{Y_{i*} - Y_{**} - \alpha_i}{S_e \sqrt{\frac{m-1}{mn}}} \sim t((m-1)(n-1))}$$

$\star$ μ, β_j, $\mu+\alpha_i$, $\mu+\beta_j$ .... analoghe

$H_0: \alpha_1 = \alpha_2 \qquad H_1: \alpha_1 \neq \alpha_2$

$$\alpha_1 - \alpha_2 \approx Y_{1*} - Y_{2*} \sim \mathcal{N}(\alpha_1 - \alpha_2 \, ; \, ???) \rightarrow \frac{Y_{1*} - Y_{2*} - (\alpha_1 - \alpha_2)}{S_e \sqrt{???}} \sim t((m-1)(n-1))$$

$\longrightarrow$ e così via ...

$\star$ Intervallo di predizione per osservazioni ulteriori in $(i;j)$

$\tilde{Y}_{i,j} \sim \mathcal{N}(\mu + \alpha_i + \beta_j \, ; \, \sigma^2)$ indip da $Y_{a,b}$ $a = 1, \dots, m$, $b = 1, \dots, n$

$\tilde{Y}_{ij} \approx \mu + \alpha_i + \beta_j \approx Y_{i*} + Y_{*j} - Y_{**}$

$$\longrightarrow \tilde{Y}_{ij} - \underbrace{Y_{i*} - Y_{*j} + Y_{**}}_{\frac{m+n-1}{mn}\sigma^2} \sim \mathcal{N}(0 \, ; \, ???)$$

$\sigma^2 +$ $\quad$ $\frac{mn+m+n-1}{mn}\sigma^2$

$$\tilde{Y}_{ij} \in Y_{i*} + Y_{*j} - Y_{**} \pm q \cdot S_e \cdot \underbrace{\sqrt{\frac{mn+m+n-1}{mn}}}_{>1}$$

quantile $t((m-1)(n-1))$

ANOVA A DUE VIE CON REPLICHE

(aka "CON INTERAZIONI", "NONLINEARE")

$Y_{i,j,k} \sim \mathcal{N}(\mu_{i,j} \,;\, \sigma^2) \quad i=1,2,\dots,m \,,\ j=1,2,\dots,n \,,\ k=1,2,\dots,\ell$

$\to \ell = 1 \qquad \mu_{i,j} = \mu + \alpha_i + \beta_j$ ipotesi di linearità

$\to \ell \geq 2 \qquad \mu_{i,j}$ qualsiasi

$\mu_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$

interazioni

$$0 = \sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{i,b} = \sum_j \gamma_{a,j} \qquad \forall a \forall b$$

HW: check che dati μ_{ij}, gli altri parametri sono univocamente determinati, grazie ai vincoli

$Y_{i,j,k} \sim \mathcal{N}(\mu + \alpha_i + \beta_j + \gamma_{ij} \,;\, \sigma^2)$

- Test di linearità

$H_0 : \gamma_{ij} = 0 \ \forall i \forall j \qquad H_1$: non tutti nulli

di nuovo si basa sul rapporto di certe "varianze" e sulla F di Fisher.

$$SS_E := \sum_{i,j,k} (Y_{i,j,k} - Y_{i,j,*})^2 \qquad S_E^2 := \frac{SS_E}{(\ell-1)mn}$$

$$\boxed{\frac{S_E^2(\ell-1)mn}{\sigma^2} \sim \chi^2((\ell-1)mn)}$$

$S_E^2 \approx \sigma^2$ sempre corretto

$SS_R + SS_C + SS_{IN} + SS_E = SS_Y$ identità delle devianze

$(m-1)+(n-1) + (m-1)(n-1) + (\ell-1)mn = \ell mn - 1$ gradi di libertà

→ Test linearità $\quad \dfrac{S_{IN}^2}{S_E^2} \overset{H_0}{\sim} F\big((m-1)(n-1)\,;\,(\ell-1)mn\big)$

→ Test effetto riga $\quad \dfrac{S_R^2}{S_E^2} \overset{H_0}{\sim} F\big(m-1\,;\,(\ell-1)mn\big)$

→ Test effetto colonna $\quad \dfrac{S_C^2}{S_E^2} \overset{H_0}{\sim} F\big(n-1\,;\,(\ell-1)mn\big)$

- Altra inferenza

$\mu_{ij} \approx Y_{ij*} \sim \mathcal{N}\left(\mu_{ij}, \frac{\sigma^2}{\ell}\right) \qquad \dfrac{Y_{ij*} - \mu_{ij}}{S_E / \sqrt{\ell}} \sim t\big((\ell-1)mn\big)$

(esempio)

▣ RELAZIONE TRA I TRE TIPI DI ANOVA

In genere si passa da 2WA con repliche a 2WA senza a 1W

$Y_{ijk} \quad i=1,\dots,m,\ j=1,\dots,n,\ k=1,\dots,\ell$

→ test di linearità

↳ se viene H_1 mi fermo

↳ se viene H_0 : $Y'_{ij} := Y_{ij*} \sim \mathcal{N}\left(\mu+\alpha_i+\beta_j\,;\,\frac{\sigma^2}{\ell}\right)$

e rifaccio come 2WA senza repliche

★ in questo modo posso verificare due volte se vi sia effetto riga o colonna e il secondo metodo in generale è più potente

↳ se effetto riga ed effetto colonna sono significativi mi fermo

↳ se almeno uno non lo è : $Y''_i := Y'_{i*} \sim \mathcal{N}\left(\mu+\alpha_i\,;\,\frac{\sigma^2}{\ell n}\right)$

oppure $Y''_j := Y'_{*j} \sim \mathcal{N}\left(\mu+\beta_j\,;\,\frac{\sigma^2}{\ell m}\right)$

e rifaccio come 1WA

★ se nella 2WA era significativa solo una variabile (say 'R') cancello l'altra e studio Y''_i

NON per vedere se vi è dipendenza da 'i' (lo so già)
ma per rendere più accurate le stime e più precisi/potenti
gli ALTRI tipi di inferenza

* se nella 2WA non era significativa nessuna delle due, ne cancello una (poi provo con l'altra) e provo 1WA (tipo stepwise backward nella regressione)

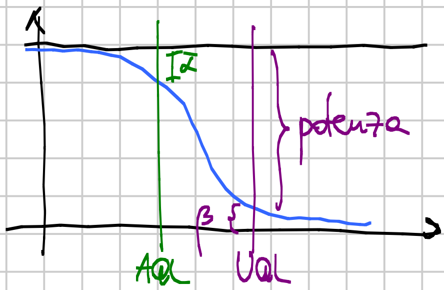
ora 29

▣ CORREZIONI PER TEST MULTIPLI

→ Se si fanno n test, ciascuno con livello di significatività $\bar{\alpha}$
la probabilità che almeno uno dia H_1, anche se sono vere tutte le ipotesi H_0, è più alta di $\bar{\alpha}$

↳ esempio jelly beans — acne (xkcd/882)

↳ esempio Alzheimer — 80 geni



* Devo calare $\bar{\alpha}$ ma questo diminuisce la potenza

↳ esempio ANOVA di tre tipi → 3 test effetto riga
↳ esempio: stepwise backward ~ $\frac{p^2}{2}$ test
(anche nella forward)
↳ esempio: $X_i \sim \mathcal{N}(\mu, \sigma^2)$ $Y_j \sim \mathcal{N}(\nu, \sigma^2)$
$H_0: \mu = \nu$ $H_1: \mu \neq \nu$ (cap 8)
test $\frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \overset{H_0}{\sim} t(m+n-2)$

→ $X_i \sim \mathcal{N}(\mu, \sigma^2)$ $Y_j \sim \mathcal{N}(\nu, \sigma^2)$ $Z_k \sim \mathcal{N}(\xi, \sigma^2)$

$\mu = \nu$ poi $\mu = \xi$ (poi $\nu = \xi$)

SBAGLIATO!!

↳ ANOVA 1 VIA è il test giusto

- Correzione di Bonferroni

Se devo fare n test e voglio un livello di significatività globale $\bar{\alpha}$ devo eseguire i singoli test con $\bar{\alpha}' = \frac{\bar{\alpha}}{n}$

$i = 1, 2, \dots, n$ i diversi test $H_0^{(i)}$ $H_1^{(i)}$ le ipotesi

uso il p dei dati per confrontarli in modo semplice

$\{U_i\}_{i=1,\dots,n}$ i diversi p dei dati

se $H_0^{(i)}$ è vera, $U_i \sim unif[0;1]$

se $H_0^{(i)}$ è falsa, U_i è "piccolo"

sia t il lvl di significatività sui singoli test e α la probabilità di errore di I specie globale (almeno un test dà H_1, anche se tutte le $H_0^{(i)}$ sono vere)

$$\alpha = P(\exists i : U_i < t) = P(\min_i U_i < t) = P\left(\bigcup_i \{U_i < t\}\right) \overset{\text{subadditività}}{\leq} \sum_i P(U_i < t) \overset{H_0^{(i)} \text{ tutte vere}}{=} nt$$

$$\alpha \leq nt$$

Bonferroni: $t := \frac{\bar{\alpha}}{n} \Rightarrow \alpha \leq \bar{\alpha}$ okay!

- Caso dei test indipendenti

$$1 - \alpha = 1 - P\left(\bigcup_i \{U_i < t\}\right) = P\left(\bigcap_i \{U_i \geq t\}\right) \overset{\text{indipendenza}}{=} \prod_i P(U_i \geq t) \overset{H_0^{(i)} \text{ tutte vere}}{=} (1-t)^n$$

Ipotesi di indipendenza: $t := 1 - (1-\bar{\alpha})^{1/n} \Rightarrow \alpha = \bar{\alpha}$

In genere non è ragionevole

★ Per $\bar{\alpha}$ piccolo $1 - (1-\bar{\alpha})^{1/n} \approx 1 - \left(1 - \frac{\bar{\alpha}}{n}\right) = \frac{\bar{\alpha}}{n}$

- Correzione di Holm

Si ordinano gli U_i dal minore al maggiore

$U_{(1)} \le U_{(2)} \le \dots \le U_{(n)}$ notazione standard statistiche di ordine

dico H_1 se $U_{(i)} < \frac{\bar{\alpha}}{n-i+1}$ per qualche i

$$\alpha = P\left(\bigcup_i \left\{U_{(i)} < \frac{\bar{\alpha}}{n-i+1}\right\}\right) = \underbrace{\dots \le \dots}_{} = \bar{\alpha}$$

HW: provate (domani lo faccio)

★ Uniformemente migliore di Bonferroni (più potente)

B.: $H_1^{(B)}$ se $\min_i U_i < \frac{\bar{\alpha}}{n} \iff U_{(1)} < \frac{\bar{\alpha}}{n}$ cfr ora 30

H: $H_1^{(H)}$ se $U_{(1)} < \frac{\bar{\alpha}}{n}$ o $U_{(2)} < \frac{\bar{\alpha}}{n-1}$ o .... o $U_{(n)} < \bar{\alpha}$

$H_1^{(B)} \Rightarrow H_1^{(H)}$

$\alpha^{(B)} \le \alpha^{(H)} \le \bar{\alpha}$

- Prossimo argomento: Design of Experiment

→ pianificazione esperimenti
Sleeper (Cap 10)
Montgomery (tanti libri sul DoE)

→ esperimenti: soprattutto regressione, ma anche ANOVA
→ due "livelli" per ogni variabile (ipotesi di linearità)
→ si pianificano quali combinazioni fare (e quante, e come, e quando)

★ è un elenco di "best practices" che escono dal mondo aziendale

- Correzione di Holm

Si ordinano gli U_i dal minore al maggiore

$U_{(1)} \leq U_{(2)} \leq \dots \leq U_{(n)}$ notazione standard statistiche di ordine

(1) $$U_{(i)} \underset{H_0^{(i)}}{\overset{H_1^{(i)}}{\leq}} \frac{\bar{\alpha}}{n-i+1}$$

☆ Sia h il minimo i per cui dico $H_0^{(i)}$

Il criterio di Holm prevede di dire

$$\underbrace{H_1^{((1))}, H_1^{((2))}, \dots, H_1^{((h-1))}, H_0^{((h))}}_{\text{secondo la disug. (1)}}, \underbrace{H_0^{((h+1))}, \dots, H_0^{((n))}}_{\text{forzati } H_0 \text{ indipendentemente da (1)}}$$

Dim (che $\alpha \leq \bar{\alpha}$)

Sia $I \subseteq \{1, 2, \dots, n\}$ ipotesi in cui è vera H_1

1) $\#I = 0$ $\quad I = \emptyset$ vera $H_0^{(i)}$ per ogni i

$$\alpha = P(\text{dire } H_1^{(i)} \text{ per qualche } i) = P\left(U_{(1)} < \frac{\bar{\alpha}}{n}\right) \leq \bar{\alpha}$$

Bonferroni

2) $\#I = 1$ $\quad I = \{j\}$ vera $H_0^{(i)}$ $i \neq j$ vera $H_1^{(j)}$

$$\alpha = P(\text{dire } H_1^{(i)} \text{ per qualche } i \neq j) \leq P\left(U_i \leq \frac{\bar{\alpha}}{n-1} \text{ per qualche } i \neq j\right)$$

se $(1) \neq j$ $\quad fp \Leftrightarrow$ dire $H_1^{((1))} \Rightarrow \min_{i \neq j} U_i \leq \frac{\bar{\alpha}}{n} < \frac{\bar{\alpha}}{n-1}$

se $(1) = j$ $\quad fp \Leftrightarrow$ dire $H_1^{((2))} \Leftrightarrow \min_{i \neq j} U_i \leq \frac{\bar{\alpha}}{n-1}$

$$= P\left(\min_{i \neq j} U_i \leq \frac{\bar{\alpha}}{n-1}\right) \leq \bar{\alpha}$$

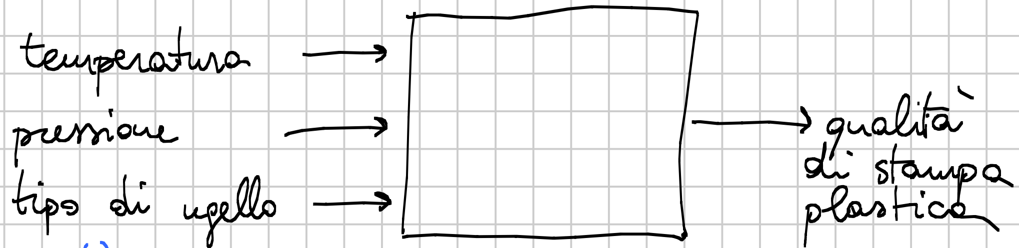
Bonferroni

3) Analogamente (check) #I qualsiasi

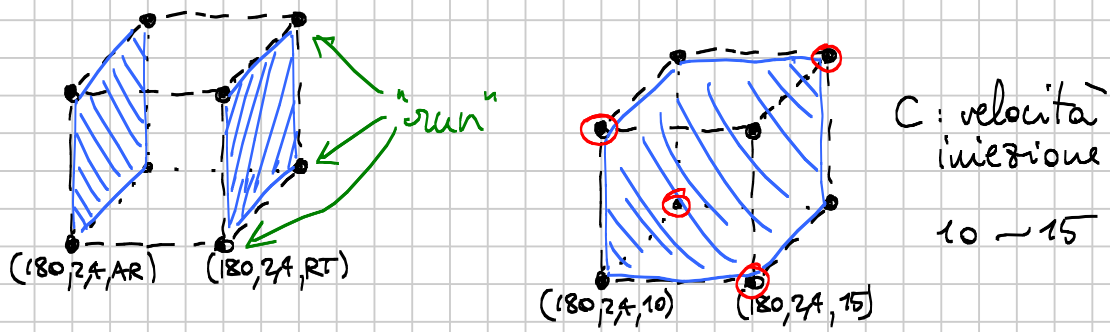
DESIGN OF EXPERIMENT

- pianificare esperimenti di regressione (decido X)
  - ↳ quanti esperimenti ($)
  - ↳ quante variabili
  - ↳ quali variabili
  - ↳ quali esperimenti (imposto i valori delle x_i)
- nei casi più comuni tutte le var di ingresso vengono forzate ad assumere solo due valori ($T=180$, $T=200$)
  (categoriche solo con 2 scelte : se ho più valvole da testare ne scelgo solo 2 di rappresentative)
- fissati i valori delle x_i si fanno alcune prove (repliche), mai una sola e mai fatte consecutivamente
- idea base : provare tutte le combinazioni dei due livelli delle variabili di ingresso

↳ Esempio :



(uncoded)							
temp :	180 ~ 200	fattore	A :	-1 (180)	$+1$ (200)	(coded)
pressione :	2,4 ~ 2,8		B :	-1	$+1$	
ugello :	AR , RT		C :	-1	$+1$	(no valori intermedi)

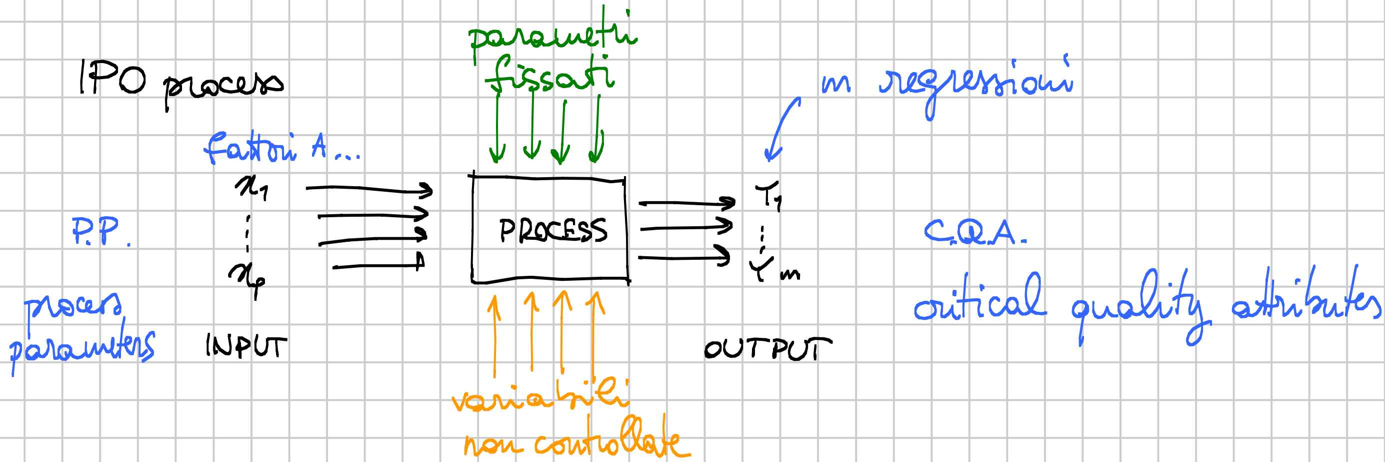


(due esempi di design space)

- Ogni configurazione che sperimento è una run.
  Per ogni run faccio un po' (lo stesso numero!) di repliche
  ↳ es: 3 fattori → $8 = 2^3$ run → (5 repliche) → 40 esper.
- Siccome 2^p cresce troppo si introducono i design frazionari che riducono il numero delle run (di $\frac{1}{2}$, $\frac{1}{4}$, ... ecc)

ora 31

Scelta delle variabili



- CQA : devono essere numeriche, informative, precise (con errore gaussiano) NON dicotomiche o discrete
- PP : QUANTI ? si distinguono
  - ☆ esperimenti di screening → tanti fattori, poche run
    ↳ pesantemente frazionari
    ↳ meno informativi
    → servono a selezionare le variabili più importanti
  - ☆ esperimenti di modeling (meno fattori, meno fraz, più inf)
    → servono per trovare un modello $Y = \sum_j \beta_j x_j$ più preciso

- PP: livelli

Criteri (si usano con l'esperto/ingegnere per decidere)

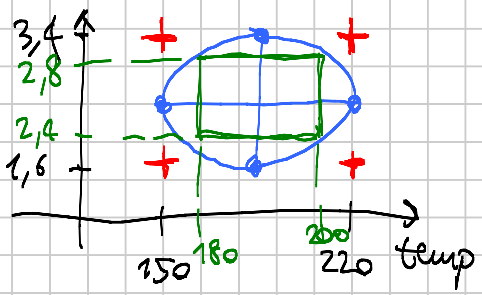
a) linearità (se spazio è troppo grande fallisce)

b) SNR (signal-to-noise ratio)

da −1 a +1 → effetto su Y se è $\ll \sigma$ non lo vede

(se spazio è troppo piccolo fallisce)

c) funzionamento ai bordi (sst grande f)



d) utilità del design space (sst piccolo f)

- Scelta del design

Vari tipi: $\boxed{\text{Taguchi}}$, Plackett-Burman, ...

buoni solo per esperimenti di screening

Taguchi: $\underbrace{L4, L8, L16, L32}_{\#\text{ di run}}$, ...

Design L8 di Taguchi

5-7 fattori	A	B	C	D	E	F	G
4 fattori	A	B	C				D
3 fattori	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	1	1	1	-1
2	1	-1	-1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1
4	1	1	-1	1	-1	-1	-1
5	-1	-1	1	1	-1	-1	1
6	1	-1	1	-1	1	-1	-1
7	-1	1	1	-1	-1	1	-1
8	1	1	1	1	1	1	1

si usano solo le colonne che servono

→ con la dummy $x_0 \equiv 1$ fattore I si ottiene la matrice X

• Proprietà: i vettori colonna sono a due a due ortogonali (sono base ortonormale di $\mathbb{R}^n$, n = run, riscalata)

$X \in M_{n,p}$ — n: run, p: 1+ fattori

$$X^T X = \begin{pmatrix} 8 & & 0 \\ & 8 & \\ & \ddots & \\ 0 & & 8 \end{pmatrix} = nI$$

$$\boxed{(X^T X)^{-1} = \frac{1}{n} I}$$

$B = (X^T X)^{-1} X^T Y$

$B \sim \mathcal{N}(\beta ; \underbrace{\sigma^2 (X^T X)^{-1}}_{\text{matrice di covarianza di } B})$

✱ Questa ortogonalità implica che le stime e i test sui coefficienti B_j sono indipendenti le une dalle altre → selez. variabili più facile

⇛ gli effetti dei FP sono ben determinati

(normalmente nella regressione multipla il coefficiente B_j dipende dal set delle variabili inserite nel modello)

Available Factorial Designs (with Resolution)

Runs	Factors 2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	Full	III												
8		Full	IV	III	III	III								
16			Full	V	IV	IV	IV	III	III	III	III	III	III	III
32				Full	VI	IV	IV	IV	IV	IV	IV	IV	IV	IV
64					Full	VII	V	IV	IV	IV	IV	IV	IV	IV
128						Full	VIII	VI	V	V	IV	IV	IV	IV

Available Resolution III Plackett-Burman Designs

Factors	Runs	Factors	Runs	Factors	Runs
2-7	12,20,24,28,...,48	20-23	24,28,32,36,...,48	36-39	40,44,48
8-11	12,20,24,28,...,48	24-27	28,32,36,40,44,48	40-43	44,48
12-15	20,24,28,36,...,48	28-31	32,36,40,44,48	44-47	48
16-19	20,24,28,32,...,48	32-35	36,40,44,48		

- Termini nonlineari e interazioni

L8 di Taguchi (ad esempio)

run	I	A	B	C	$\cancel{A^2}$	AB	AC	BC	ABC
1	1	-1	-1	-1	1	1	1	1	-1
2	1	1	-1	-1		-1	-1	1	1
3	1	-1	1	-1		-1	1	-1	1
4	1	1	1	-1		1	-1	-1	-1
5	1	-1	-1	1		1	-1	-1	1
6	1	1	-1	1		-1	1	-1	-1
7	1	-1	1	1		-1	-1	1	-1
8	1	1	1	1	1	1	1	1	1

tutte diverse e ortogonali

I termini quadratici vanno bene solo con DoE a ≥ 3 livelli

5-7 fattori	A	B	C	D	E	F	G
4 fattori	A	B	C				D
3 fattori	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	1	1	1	-1
2	1	-1	-1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1
4	1	1	-1	1	-1	-1	-1
5	-1	-1	1	1	-1	-1	1
6	1	-1	1	-1	1	-1	-1
7	-1	1	1	-1	-1	1	-1
8	1	1	1	1	1	1	1

☆ Se uso 3 fattori (full factorial) posso stimare i coefficienti di tutte le interazioni

☆ Se uso 4 fattori (o più) alcune interazioni hanno la stessa colonna di un fattore

	ABCD	BCD	ACD	ABD	ABC	AB	AC	BC
run	I	A	B	C	D	CD	BD	AD
1	1	−1	−1	−1	−1	1	1	1
2	1	1	−1	−1	1	−1	−1	1
3	1	−1	1	−1	1	−1	1	−1
4	1	1	1	−1	−1	1	−1	−1
5	1	−1	−1	1	1	1	−1	−1
6	1	1	−1	1	−1	−1	1	−1
7	1	−1	1	1	−1	−1	−1	1
8	1	1	1	1	1	1	1	1

* C'è una struttura di ALIAS

I + ABCD
A + BCD
B + ACD
C + ABD
D + ABC
AB + CD
AC + BD
AD + BC

5f D = AB E = AC

I + + +
A + BD + CE +
B + AD + +
C + AE + +
D + AB + +
E + AC + +
BC + DE + +
BE + CD + ABC +

→ # righe alias structure = run (talvolta < run)

→ in ogni riga lo stesso numero di termini:

1 se il design è full factorial (no alias)

2 se il design è $\frac{1}{2}$ − frazionario (ad es $L8 = \frac{1}{2} \cdot 16 = 2^4$)

4 se il design è $\frac{1}{4}$ − frazionario (ad se $L8 = \frac{1}{4} \cdot 32 = 2^5$)

....

```
Design Generators: D = AB; E = AC

Alias Structure

I + ABD + ACE + BCDE

A + BD + CE + ABCDE     3
B + AD + CDE + ABCE     3
C + AE + BDE + ABCD     3
D + AB + BCE + ACDE     3
E + AC + BCD + ABDE     3
BC + DE + ABE + ACD     4
BE + CD + ABC + ADE     4
```

III risoluzione

- Risoluzione di un design frazionario

Def: è il minimo al variare della riga dell'alias structure della somma dei gradi dei due termini di grado più basso in alias fra loro

$\boxed{III}$: ci sono fattori singoli in alias con interazioni fra due

$\boxed{IV}$: ci sono interazioni tra due fattori in alias fra loro (o 1+3)

$\boxed{\geq V}$: non ci sono fattori singoli o interazioni tra due in alias fra loro

- Significato pratico della alias structure

★ Ipotesi: interazioni a tre o più non capitano nel mondo industriale

→ Se due termini sono in alias, il coefficiente stimato corrispondente sarà la somma delle stime dei due coefficienti → non è possibile separarli

→ In pratica di solito si assume che uno dei due sia zero e che la stima sia quella dell'altro.

↳ un po' arbitrario ma non assurdo: in molte situazioni pratiche è proprio così

★ Esperimenti di modeling vanno fatti di IV o V o più in generale non ci può essere aliasing tra due termini potenzialmente non nulli

★ Esperimenti di screening si possono fare anche di III (con qualche falso positivo da mettere in conto in caso di interazioni a due rilevanti e aliasing)

Available Factorial Designs (with Resolution)

Runs	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Factors													
4	Full	III												
8		Full	IV	III	III	III								
16			Full	V	IV	IV	IV	III	III	III	III	III	III	III
32				Full	VI	IV	IV	IV	IV	IV	IV	IV	IV	IV
64					Full	VII	V	IV	IV	IV	IV	IV	IV	IV
128						Full	VIII	VI	V	V	IV	IV	IV	IV

Available Resolution III Plackett-Burman Designs

Factors	Runs	Factors	Runs	Factors	Runs
2-7	12,20,24,28,...,48	20-23	24,28,32,36,...,48	36-39	40,44,48
8-11	12,20,24,28,...,48	24-27	28,32,36,40,44,48	40-43	44,48
12-15	20,24,28,36,...,48	28-31	32,36,40,44,48	44-47	48
16-19	20,24,28,32,...,48	32-35	36,40,44,48		

★ Design di Plackett-Burman

$\cancel{B + AD}$ $\longrightarrow$ $\boxed{B + \frac{1}{2}AD + \frac{1}{2}CE}$

aliasing — confounding

↳ ancora meglio per lo screening

(approfondimento per l'esame?)

- Efficienza di un design

I	+ ABCDEF
A	+ BCDEF
B	+ ACDEF
C	+ ABDEF
D	+ ABCEF
E	+ ABCDF
F	+ ABCDE
AB	+ CDEF
AC	+ BDEF
AD	+ BCEF
AE	+ BCDF
AF	+ BCDE
BC	+ ADEF
BD	+ ACEF
BE	+ ACDF
BF	+ ACDE
CD	+ ABEF
CE	+ ABDF
CF	+ ABDE
DE	+ ABCF
DF	+ ABCE
EF	+ ABCD
ABC	+ DEF
ABD	+ CEF
ABE	+ CDF
ABF	+ CDE
ACD	+ BEF
ACE	+ BDF
ACF	+ BDE
ADE	+ BCF
ADF	+ BCE
AEF	+ BCD

ok

15 coefficienti quasi tutti nulli

10 coefficienti sprecati

L32 6 fattori $\frac{22}{32} = \frac{11}{16} \approx 0{,}688$

I	+	ABCE	+	ADEF	+	BCDF
A	+	BCE	+	DEF	+	ABCDF
B	+	ACE	+	CDF	+	ABDEF
C	+	ABE	+	BDF	+	ACDEF
D	+	AEF	+	BCF	+	ABCDE
E	+	ABC	+	ADF	+	BCDEF
F	+	ADE	+	BCD	+	ABCEF
AB	+	CE	+	ACDF	+	BDEF
AC	+	BE	+	ABDF	+	CDEF
AD	+	EF	+	ABCF	+	BCDE
AE	+	BC	+	DF	+	ABCDEF
AF	+	DE	+	ABCD	+	BCEF
BD	+	CF	+	ABEF	+	ACDE
BF	+	CD	+	ABDE	+	ACEF
ABD	+	ACF	+	BEF	+	CDE
ABF	+	ACD	+	BDE	+	CEF

ok

2 o 3 inter. in alias

2 coeff sprecati

L16 6 fattori $\frac{10}{16} \approx 0{,}625$

★ Def: l'efficienza generale di un design è la media per righe dello "score" seguente

$$\begin{cases} 1 & \text{se c'è un solo EOI} \\ \frac{1}{2} & \text{se ci sono due EOI} \\ 0 & \text{se ci sono 0 o} \geq 3 \text{ EOI} \end{cases}$$

↳ EOI = effect of interest

sono i termini che vogliamo inserire nel modello

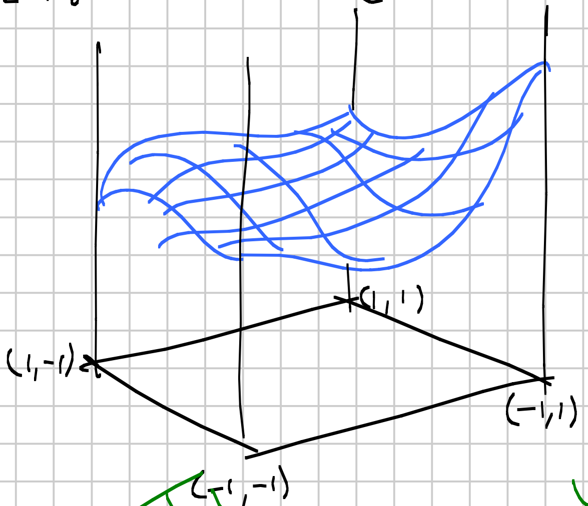
quindi in generale: tutti i fattori singoli e le interaz. a 2

★ L'efficienza specifica si calcola partendo da un set di EOI che sia un sottoinsieme di quelli sopra, dipendente dal particolare esperimento

Treatment Structure	Number of Factors 3	4	5	6	7	8	9
L4	0.375						
L8	0.875	0.813	0.500	0.125	0.125		
L16		0.688	$\boxed{1.000}$	0.625	0.500	0.563	0.313
L32			0.500	0.688	0.766	0.781	0.750
L64				0.344	0.453	0.578	0.641
L128					0.227	0.289	0.359

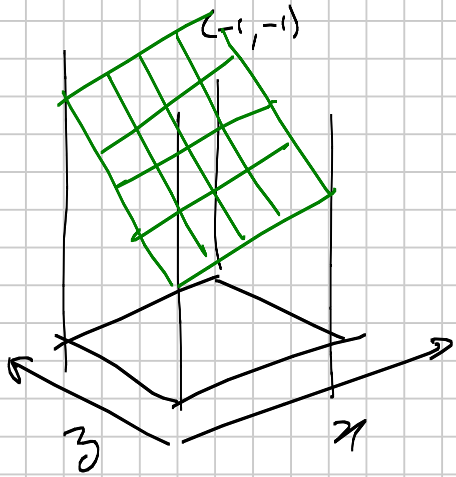
- Sulle interazioni

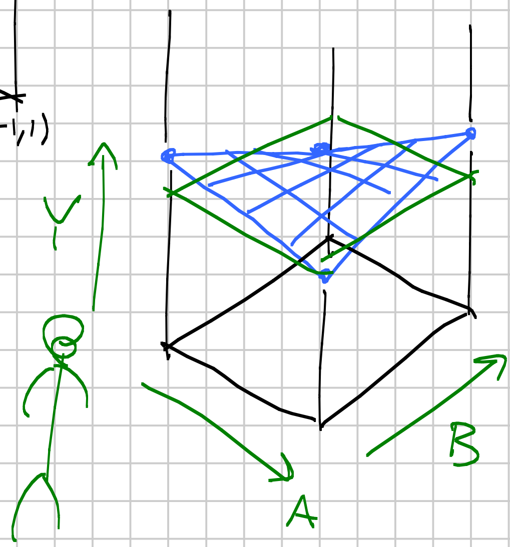
$f : [-1,1]^2 \to \mathbb{R}$ (due soli fattori)



$$f(x,y) = \alpha + \beta x + \gamma y$$

$$f(x,y) = \alpha + \beta x + \gamma y + \delta x y$$

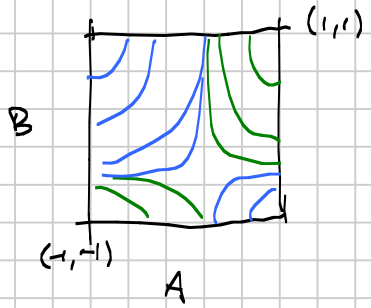




no interazioni:
→ effetto di A non dipende da B e viceversa

interazioni
→ effetto di A dipende dal valore di B e viceversa

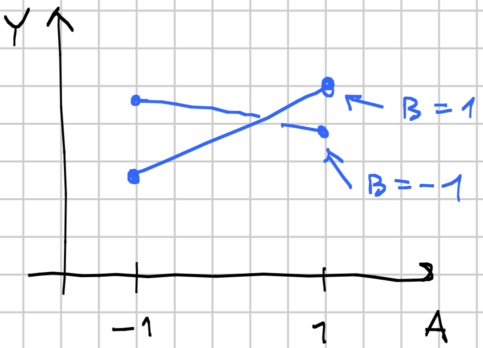
* Contour plot:



curve di livello di

$c_1 A + c_2 B + c_3 AB$

* Altra rappresentazione:

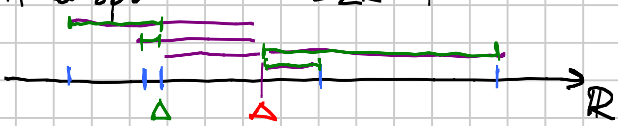


HW: Cosa si deve scegliere per minimizzare $\sum_{i=1}^{n} |x_i - y|$?

$x_1, x_2, \dots, x_n$ reali (dall'ora 5)

$y = \arg\min f(y)$ $\qquad f(y) = \sum_{i=1}^{n} |x_i - y|$

- n dispari $\qquad n = 2k-1$



$x_1 < x_2 < \dots < x_{2k-1}$ wlog

x_k mediana campionaria

$x_i < y < x_{i+1}$ $\qquad i \geq k$ wlog

$$f(y) - f(x_k) = \sum_{j=1}^{k} \left(|x_j - y| - |x_j - x_k| \right) + \sum_{k+1}^{i} \left(|x_j - y| - |x_j - x_k| \right) + \sum_{i+1}^{2k-1} \left(\qquad \right)$$

$$= \sum \left(y - \cancel{x_j} - x_k + \cancel{x_j} \right) + \sum \left(y - x_j - x_j + x_k \right) + \sum \left(\cancel{x_j} - y - \cancel{x_j} + x_k \right)$$

$$\geq (y - x_k)(\underbrace{k - 2k + 1 + i}_{i+1-k}) - (y - x_k)(i - k) \geq y - x_k$$

$$|x_j - y| - |x_j - x_k| \geq -|y - x_k|$$

- n pari (analogo) $\qquad n = 2k$

$\forall x \in [x_k ; x_{k+1}]$ $\qquad x = \arg\min_{y} f(y)$

☒ Torniamo al DoE

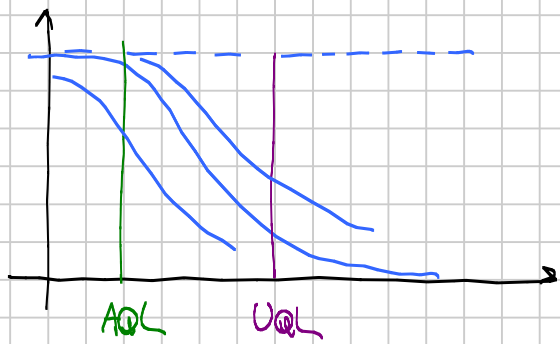
- Numero di repliche
  - più numerose sono, più precisa è l'analisi e più alti i costi
  - sarebbe bene che fossero almeno 3 per ogni run

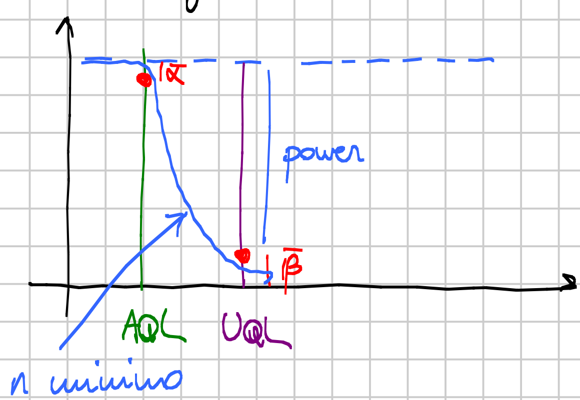
≥ 3 rilevo e identifico outliers

$= 2$ rilevo ma non identifico outliers

$= 1$ non rilevo gli outliers

– per determinare il numero minimo adeguato





$\boxed{n} \leftarrow$ { $\boxed{\text{effetto minimo}}$ di un fattore che vogliamo poter rilevare; $\boxed{\alpha}$ (5%) lvl sign.; $\boxed{1-\beta}$ potenza (80%); $\boxed{\sigma}$ dev. std di Y }

★ Complicato, richiede σ, effetto minimo (?)

↳ Scorciatoia : "rule of thumb"

$$\left\lceil \frac{run + 32}{run} \right\rceil = \text{repliche per ogni run}$$

L4 → 9 repliche

L8 → 5 repliche

L16 → 3 repliche

≥L32 → 2 repliche

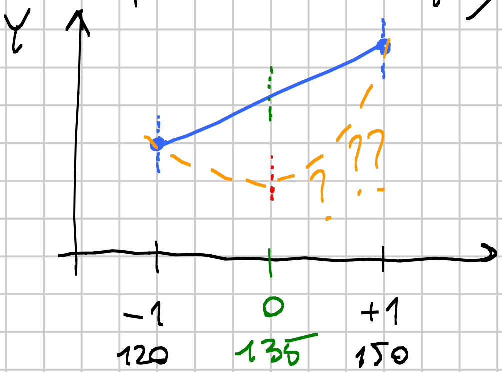
- Center point

I	A	B	C	D
1	−1	−1	−1	−1
...	...	...	...	...
1	1	−1	−1	1
1	0	0	0	0 ← center point

★ Punto aggiuntivo   L4 → run = 5 → 8 repliche × 5 run = 40

↳ va replicato anche lui

★ Se vi sono variabili categoriche non si fa (a meno di complicare il design)



★ Serve a verificare (a posteriori) che il modello lineare sia adeguato

● Randomizzare le repliche

→ es: L4 3 fattori 4 repliche

run	I	A	B	C	r1	r2	r3	r4	— order —				— stessa cosa —			
1	1	1	1	1	~	~	~	~	16	1	13	4	1	4	13	16
2	1	1	-1	-1	~	~	~	~	14	10	9	6	6	9	10	14
3	1	-1	-1	1	~	~	~	~	2	7	3	15	2	3	7	15
4	1	-1	1	-1	~	~	~	~	8	12	5	11	5	8	11	12

★ si spreca tempo e denaro per randomizzare

★ Serve a "schermare" l'effetto del tempo

tempo = ordine degli esperimenti = $x_?$ variabile nascosta

↳ spesso viene significativa

i. umani → curve d'apprendimento

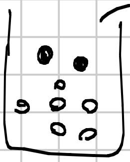
ii. usura, logorio, sporco

iii. variazioni macroscopiche (pezzo sostituito)

★ Se non randomizzo, qualunque effetto di questo tipo si può ripercuotere su uno o più fattori (nell'esempio, sicuramente A e forse C)

- Parentesi: HW alla fine dell'ora 10

→ se estraggo n pezzi da una popolazione di N (senza rimessa) in cui vi è una frazione p di difettosi, il numero di difettosi estratti X ha legge ipergeometrica che può essere approssimata con quella binomiale



N di cui a_N •

$n \to \begin{cases} X = \bullet \\ n - X = \circ \end{cases}$

$p \approx \frac{a_N}{N} = p_N \qquad p_N \xrightarrow[N\to\infty]{} p$

$$P(X=k) = \frac{\binom{a_N}{k}\binom{N-a_N}{n-k}}{\binom{N}{n}} = \frac{n!(N-n)!}{N!} \cdot \frac{a_N!}{k!(a_N-k)!} \cdot \frac{(N-a_N)!}{(n-k)!(N-a_N-n+k)!}$$

$$= \binom{n}{k} \underbrace{\frac{a_N(a_N-1)\cdot\ldots\cdot(a_N-k+1)}{N(N-1)\cdot\ldots\cdot(N-k+1)}}_{k \text{ fattori}} \cdot \underbrace{\frac{(N-a_N)\cdot\ldots\cdot(N-a_N-n+k+1)}{(N-k)\cdot\ldots\cdot(N-n+1)}}_{n-k \text{ fattori}}$$

$\frac{a_N-c}{N-c} \xrightarrow[N\to\infty]{} p \qquad \frac{N-a_N-c}{N-k-c} \xrightarrow[N\to\infty]{} 1-p$

$$\approx \binom{n}{k} p^k (1-p)^{n-k}$$

- DoE: si fanno gli esperimenti
  - → presenza
  - → documentazione (diario)
  - → esperimenti, misurazioni, dati
- Analisi dei dati
  - regressione con alcune differenze

$X^T X = cI$

B_i hanno covarianza nulla

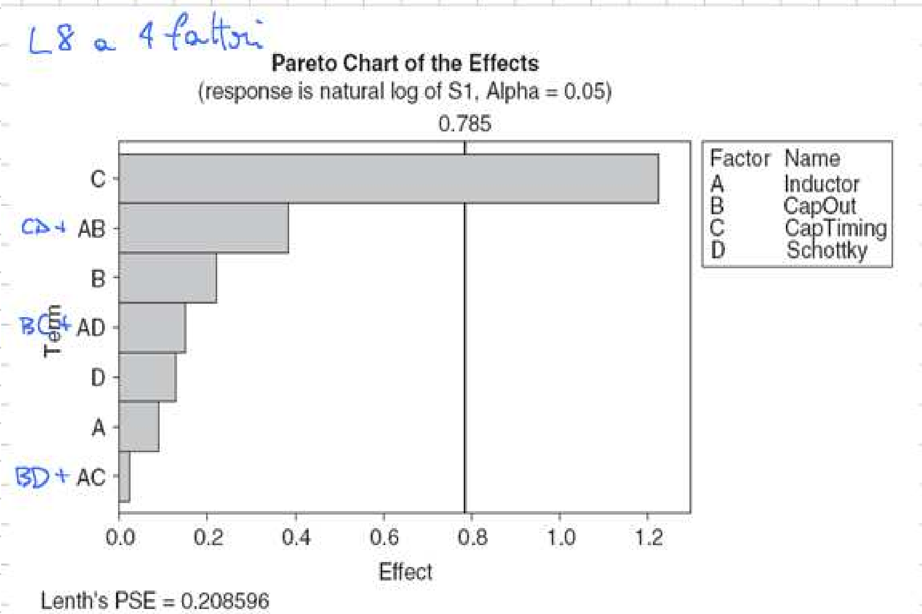
se cambio le variabili non variano i B_i calcolati

i B_i sono tutti confrontabili

$Y = 12{,}4 + 1{,}2A - 0{,}7B + 2{,}5C - 2{,}1AB$ coded: ± 1

effetti : C 5 ⟸ valore assoluto del doppio del coeff.
AB 4,2
A 2,4
B 1,4

★ Selezione delle variabili



→ Si fa qualitativo, tenendo conto di cosa misura Y e di qual è un effetto rilevante

→ Scelte le variabili (regola gerarchica + aliasing) si riduce il modello e si ricalcolano i residui e si controllano

TEST DEL CHI-QUADRO

φ legge vera di $X_1, \dots, X_n$ iid (legge discreta con pochi valori k)

φ_0 legge ipotetica $\boxed{P(X_i = j) = \varphi(j)}$

Test : $H_0 : \varphi = \varphi_0$ $H_1 : \varphi \neq \varphi_0$

$O_j = \#\{ i \leq n : X_i = j \}$ $j = 1, 2, \dots, k$ wlog

→ Che legge ha O_j ? $O_j \sim bin(n ; \varphi(j))$

$E(O_j) = n\varphi(j)$ vero $n\varphi_0(j) = A_j = E(O_j)$ sotto ipotesi H_0

★ Che legge ha $(O_1, O_2, \dots, O_k)$?

le componenti sono binomiali ma non indipendenti

$\sum_j O_j = n$ ad esempio

→ Si definisce la legge congiunta multinomiale

$$P\big((O_1, \dots, O_k) = (h_1, \dots, h_k)\big) = \binom{n}{h_1, h_2, \dots, h_k} \varphi(1)^{h_1} \varphi(2)^{h_2} \cdot \ldots \cdot \varphi(k)^{h_k}$$

$\sum_j h_j = n$

$$\binom{n}{h_1, h_2, \dots, h_k} = \frac{n!}{h_1! h_2! \cdot \ldots \cdot h_k!}$$

coefficiente multinomiale

(conta gli anagrammi)

"ABRACADABRA" ha $\binom{11}{5,2,2,1,1}$ anagrammi

★ Se prendo $O_1, \dots, O_k$ indipendenti con $O_j \sim Pois(\nu_j)$

la legge di $(O_1, \dots, O_k)$ condizionata all'evento $\{\sum_j O_j = n\}$

è quella multinomiale

- Multinomiale e Poisson

$S \sim Pois\left(\sum_j \nu_j\right)$

$S = O_1 + \dots + O_k$

$O = (O_1, \dots, O_k)$ $\quad O_j \sim Pois(\nu_j)$ indipendenti

$$P(O = (a_1, \dots, a_k)) \underset{\text{indip}}{=} \prod_{j=1}^{k} P(O_j = a_j) \underset{\text{Pois}}{=} \prod_{j=1}^{k} \frac{\nu_j^{a_j}}{a_j!} e^{-\nu_j}$$

$\nu := \sum_{j=1}^{k} \nu_j$

$$= \underbrace{\frac{(\sum a_j)!}{a_1! a_2! \dots a_k!}}_{\binom{\sum a_j}{a_1, \dots, a_k}} \frac{(\sum_j \nu_j)^{\sum_j a_j}}{(\sum_j a_j)!} \cdot \frac{\nu_1^{a_1} \cdot \dots \cdot \nu_k^{a_k}}{(\sum_j \nu_j)^{a_1} \cdot \dots \cdot (\sum_j \nu_j)^{a_k}} e^{-\sum_j \nu_j}$$

$n = \sum_{j=1}^{k} a_j$

$$= \underbrace{\binom{n}{a_1, \dots, a_k} \prod_{i=1}^{k} \left(\frac{\nu_i}{\nu}\right)^{a_i}}_{\text{legge multinomiale}} \underbrace{\frac{\nu^n}{n!} e^{-\nu}}_{\text{legge di Poisson}}$$

$$= P(S = n) \, P\left(\text{multin}\left(n; \frac{\nu_1}{\nu}, \frac{\nu_2}{\nu}, \dots, \frac{\nu_k}{\nu}\right) = (a_1, \dots, a_k)\right)$$

$$= P(S = n) \, P(O = (a_1, \dots, a_k) \mid S = n)$$

$$P(O = (a_1 \dots a_k)) = P(\{O = (a_1, \dots, a_k)\} \cap \{S = n\}) = P(O = (a_1 \dots a_k) \mid S = n) P(S = n)$$

$$= f(a_1 \dots a_k; n) \, g(n)$$

★ Se $\sum_n g(n) = 1$ e $\sum_a f(a; n)$ $\forall n$ allora $g(n) = P(S = n)$

e $f(a; n) = P(O = a \mid S = n$

(HW: check)

★ $O = (O_1, \dots, O_k)$ $\quad O_j \sim Pois(\nu_j)$ indipendenti $\quad S = \sum_{j=1}^{k} O_j \quad \nu = \sum_{j=1}^{k} \nu_j$

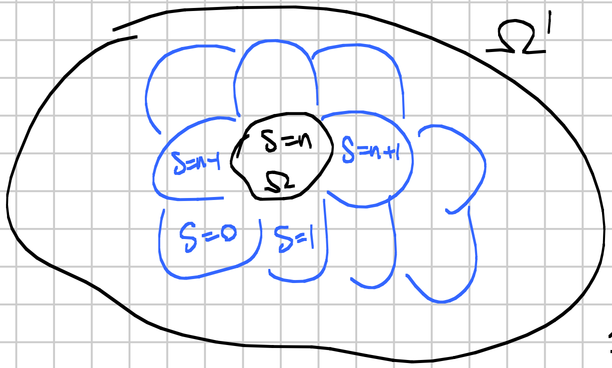
La legge di O condizionata a $\{S = n\}$ è multinomiale $\left(n; \frac{\nu_1}{\nu}, \dots \frac{\nu_k}{\nu}\right)$

• Chi-quadro.

$X_1, \dots, X_n$ con legge φ indip. $\qquad \varphi$ ha k valori possibili

$O_1, \dots, O_k \qquad O_j = \#\{i \leq n : X_i = k\} \qquad S = \sum_{j=1}^{k} O_j \qquad S = n$ q.c.

$(O_1, \dots, O_k) \sim \text{multin}(n; \varphi(1), \dots, \varphi(k))$



Su Ω' $O_1, \dots, O_k$ sono Poisson indipendenti di medie $\nu_1, \dots, \nu_k$ proporzionali a $\varphi(1) \dots \varphi(k)$

Per comodità: $\nu_j = \varphi(j) n$

in modo che $E(S) = n$, $S = n$ credibile

★ Statistica: $\displaystyle W = \sum_{j=1}^{k} \frac{(O_j - A_j)^2}{A_j}$ $\qquad$ dove $A_j = n \varphi_0(j)$

sotto ipotesi $H_0 : \varphi = \varphi_0$

$$W = \sum_{j=1}^{k} \frac{(O_j - \nu_j)^2}{\nu_j} = \sum_{j=1}^{k} Z_j^2 \qquad Z_j := \frac{O_j - \nu_j}{\sqrt{\nu_j}} \qquad E(Z_j) = 0$$

$$Var(Z_j) = 1$$

TLC: $\nu_j \gg 1 \Rightarrow Z_j \approx \mathcal{N}(0;1)$

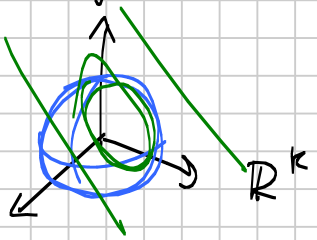
Nello spazio Ω', Z_j sono indipendenti: se $\nu_j \gg 1 \ \forall j$

$$\boxed{W \approx \chi^2(k)}$$

La legge di W nello spazio Ω è la stessa che in Ω' ma condizionata a $\{S = n\}$

$$S = n \Leftrightarrow \sum_{j=1}^{k} O_j = n \Leftrightarrow \sum_{j=1}^{k} (\nu_j + \sqrt{\nu_j} Z_j) = n \Leftrightarrow \sum_j \alpha_j Z_j = c$$

relazione lineare



$\boxed{W \approx \chi^2(k-1)}$ su Ω

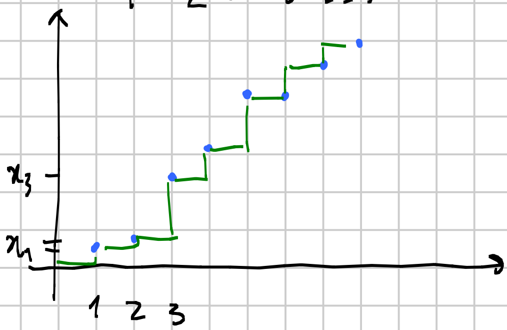
Vedi pdf 2011

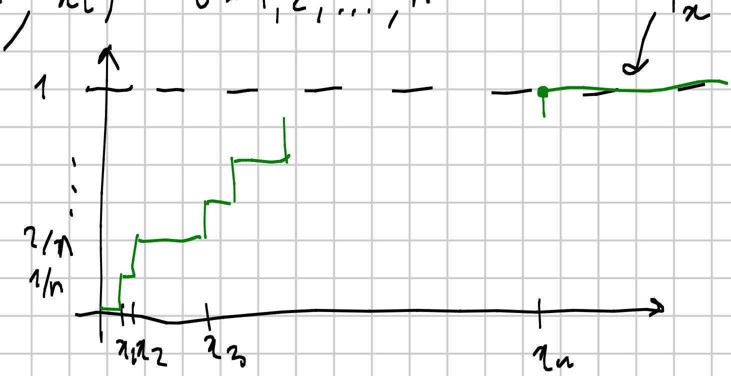
FINE ?

$x_1, x_2, \dots, x_n$

$x_1 \leq x_2 \leq x_3 \dots$

$(i ; x_i) \quad i = 1, 2, \dots, n$





$$F_x(t) = \frac{1}{n} \# \left\{ i \leq n : x_i \leq t \right\} \approx P(X_1 \leq t) = F_{X_1}(t)$$

Funzione di ripartizione empirica