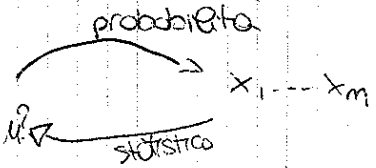
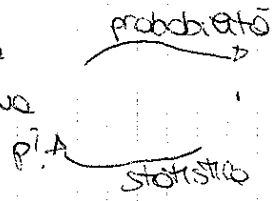
STATISTICA - STIMA

distribuzione di fondo (o popolazione) per esempio $N(20, 4^2)$ nell'esempio catastrofico



oppure data una Bernoulli $(\frac{1}{2})$ cioè moneta equa



Una statistica inferenziale va dal particolare (campione) al generale (popolazione o distribuzione), dal noto all'ignoto

Se non conosco un parametro, cioè una caratteristica della distribuzione, lo posso STIMARE
Per esempio, $\bar{X}$ stima (è stimatore di) μ.

Similmente, se non conosciamo σ^2, lo stimiamo tramite la varianza campionaria

$$S = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

E se $X_1, \dots X_m$ iid Bernoulli (p) e non conosciamo p, lo stimiamo con
$\hat{p}$ = frazione di successi sugli n test-prova

09-11-09

$X_1, \dots X_m$ iid con densità $f(x)$ si dice campione casuale

Una sua funzione è una STATISTICA.

Nella realtà spesso i parametri della $f(x)$ sono incognita. Possiamo usare delle statistiche per STIMARLI

Esempio: (METROLOGIA)

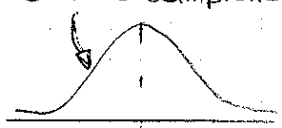
μ = misurando incognito (per esempio la distanza da una stella)

$X_1, \dots X_m$ n misurazioni indipendenti di μ, ciascuna con densità normale(μ, σ^2)

σ^2 è la varianza e dipende in maniera inversa dalla precisione $(= \frac{1}{\sigma^2})$ dello strumento di misura

In genere $\bar{X} = \frac{\sum X_i}{n}$, la media campionaria, è una stima di μ.

La densità campionaria di $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ ↑ c'è un fattore $\sqrt{n}$ che riduce la deviazione standard.



$\bar{X}$ è centrato su $\mu \rightarrow E(\bar{X}) = \mu \rightarrow$ si dice che $\bar{X}$ è uno stimatore non DISTORTO (NON POLARIZZ

la variabilità di $\bar{X}$ intorno a μ è misurata dalla deviazione standard della sua distribuzione campionaria $SD(\bar{X}) = \sqrt{var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$ chiamato anche ERRORE STANDARD di $\bar{X}$.

Esempio 2: supponiamo ora di non conoscere σ^2, lo possiamo stimare con la varianza campionaria

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

non si divide per n, ma per $n-1$ così il valore atteso di S^2 è centrato $E(S^2) = \sigma^2$ NON DISTORT

infatti $E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right] = \frac{1}{n-1}E\left(\sum(x_i^2-2x_i\bar{x}+\bar{x}^2)\right) =$

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

$E(x_i^2) - (E(x_i))^2 = \mathrm{Var}(x_i) = \sigma^2$

$\Downarrow$

$E(x_i)^2 = \sigma^2 + \mu^2$

$= \frac{1}{n-1}\left(\sum(\sigma^2+\mu^2) - 2nE(\bar{x}^2) + nE(\bar{x}^2)\right) =$

Analogamente $E(\bar{x}^2) = \mathrm{Var}(\bar{x}) + (E(\bar{x}))^2 = \frac{\sigma^2}{n} + \mu^2$

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

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

S^2 è stimatore corretto di σ^2

Se $x_1 \dots x_n$ i.i.d. Bernoulli(p), e p è incognito, lo stimiamo usando la proporzione campionaria di successi $\hat{p} = \frac{n^\circ \text{ di successi}}{n} = \frac{\sum x_i}{n}$

(caso particolare di $\bar{x}$ quando le x_i sono binarie)

$\hat{p}$ è centrato su p, e la sua varianza è $\frac{p(1-p)}{n}$

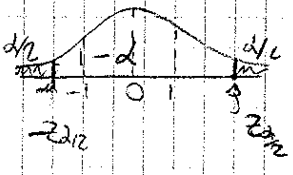
l'errore standard di $\hat{p}$ è $\sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{1}{4n}}$

INTERVALLO DI CONFIDENZA

Vogliamo arrivare ad una STIMA INTERVALLARE (INTERVALLO DI CONFIDENZA) di un parametro

Partiamo dalla distribuzione campionaria di $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ quando $x_1 \dots x_n$ i.i.d. $N(\mu, \sigma^2)$.

Possiamo scrivere $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$



$P\left(-z_{\alpha/2} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$ ragionato dalla distribuzione campionaria di $\bar{x}$

$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ (x esempio soft)

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

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

L'intervallo casuale (aleatorio)

$\left(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ contiene ("copre") μ con probabilità $1-\alpha$

lo chiamiamo INTERVALLO DI CONFIDENZA di livello $1-\alpha$

Brevemente possiamo scrivere $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ intervallo

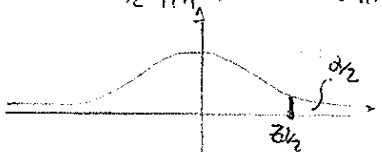
L'ampiezza dell'intervallo (che è centrato su $\bar{x}$, stimatore di μ) è $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

se n cresce, l'ampiezza decresce. Se σ cresce, anche l'ampiezza cresce

se è quello $1-\alpha$ cresce, cresce anche l'ampiezza

RIASSUNTO

Dato un campione casuale $X_1 \dots X_n$ normale $N(\mu, \sigma^2)$ con σ^2 nota, un i.c. (= intervallo di confidenza) per μ è $\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ abbreviato in $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ dove $z_{\alpha/2}$ è



12-11-09

Per esempio (8.5): il segnale μ è stimato 13.3 ± 1.86 con confidenza del 95%

Usiamo il termine confidenza per indicare il livello di affidabilità della procedura

Non possiamo parlare di "probabilità" perché μ non è aleatorio

Di tutti i possibili campioni che potrei osservare, il 95% (cioè $100(1-\alpha)\%$) contiene μ mentre il 5% ($100\alpha\%$) non lo becca.

L'ampiezza dell'intervallo è $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Tale ampiezza può servire per determinare n, la numerosità campionaria necessaria (n° di osservazioni).

Supponiamo di desiderare un'ampiezza D fissata:

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le D$$

$$2z_{\alpha/2}\frac{\sigma}{D} \le \sqrt{n} \quad \rightsquigarrow \boxed{\left(2z_{\alpha/2}\frac{\sigma}{D}\right)^2 \le n}$$

FORMULA PER PROGRAMMARE L'ESPERIMENTO

Esempio 8.9 pag 324

A volte conviene usare INTERVALLI UNILATERI, per esempio $(b, +\infty)$ per proprietà (e altre caratteristiche del tipo "più ce n'è, meglio è") oppure $(-\infty, a)$ per la situazione "meno ce n'è meglio è"

Per esempio, dalla disequazione probabilistica $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_\alpha\right) = 1-\alpha$ discende l'intervallo di confidenza unilaterale $\left(\underbrace{\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}}_{\text{si dice anche che } \bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}} \text{ è un estremo di confidenza inferiore di livello } 1-\alpha}, \infty\right)$ di livello $1-\alpha$

Simmetricamente $\left(-\infty, \bar{x} + z_\alpha\frac{\sigma}{\sqrt{n}}\right)$ i.c. unilaterale sinistro, ovvero $\bar{x} + z_\alpha\frac{\sigma}{\sqrt{n}}$ è un estremo di confidenza superiore di livello $1-\alpha$

Cosa succede se σ non è noto? Dobbiamo stimarlo, per esempio usando S, la radice della varianza campionaria

Purtroppo $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, ma $\frac{\bar{x}-\mu}{S/\sqrt{n}} \not\sim N(0,1)$ [vero approssimativamente se n grande, ma non per n piccolo.

σ è una costante, S è una v.a. legata al chi-quadro

Gosset (primi '900) scoprì in base alla distribuzione di S (vedi 7.6), che $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ ha una distribuzione non normale standard, bensì "t con $n-1$ gradi di libertà"

$\bar{x} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$ è un intervallo di confidenza per μ quando σ è incognito e n piccolo. $t_{n-1,\alpha/2}$ è un punto percentile analogo a $z_{\alpha/2}$ tratto dalla tavola di t

Analogamente, $\bar{X} - t_{m-1,\alpha}\frac{S}{\sqrt{m}}$ è un estremo INFERIORE di confidenza di quello $1-\alpha$

$\bar{X} + t_{m-1,\alpha}\frac{S}{\sqrt{m}}$ è un estremo SUPERIORE di confidenza di quello $1-\alpha$

- $X_1 \dots X_m$ iid Bernoulli(p). Pienso, vaglio stimare p.

Una stima puntuale di p è $\hat{p} = \frac{\sum x_i}{m}$ = proporzione campionaria di successi.

Adesso costruiamo un i.c. per p.

Abbiamo visto che, per il Teorema del limite centrale, se m grande,

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{m}\right).$$

Quindi, ripetendo la procedura paradigmatica, otteniamo

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{m}}$$

Problema: non conosciamo p, è proprio quello che stiamo stimando

Una prima soluzione è usare $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}$

che risente di almeno due approssimazioni: → T.L.C. (Teorema del limite centrale)
↳ usare doppiamente $\hat{p}$ al posto di p

Un'alternativa CONSERVATIVA è usare il massimo possibile di $\frac{p(1-p)}{m}$, cioè $\frac{1}{4m}$, ottenendo $\hat{p} \pm \frac{z_{\alpha/2}}{2\sqrt{m}}$ che dà intervalli molto ampi.

Un estremo inferiore di confidenza approssimato di quello $100(1-\alpha)\%$ è

$$\hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}.$$

Un estremo superiore di confidenza approssimato di quello $100(1-\alpha)\%$ è

$$\hat{p} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}$$

VERIFICA DI IPOTESI

19-11-09

Un'ipotesi è un'affermazione riguardo a un meccanismo probabilistico che non conosciamo completamente.

In particolare, un'ipotesi statistica è un'affermazione riguardo a un parametro incognito che regola la distribuzione di probabilità di un campione osservabile

Esempio: Una certa moneta è equa??

- ipotesi(nulla): la moneta è equa H_0
- ipotesi(alternativa): la moneta non è equa H_1 ← ipotesi da favorire testo

Noi possiamo osservare le campione $X_1 \dots X_n$ iid Bernoulli (p) con p incogn

$H_0: p = \frac{1}{2}$

$H_1: p \neq \frac{1}{2}$

Un test di ipotesi è una procedura che in base ai dati rifiuta o non rifiuta H_0

L'idea è di rifiutare H_0 se i dati forniscono sufficiente evidenza contro H_0

Procediamo con l'osservazione del campione: TCTTT

se $p = \frac{1}{2}$ (H_0 è vero) $P(\text{almeno 4 teste su } n=5 \text{ prove}) = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(1-\frac{1}{2}\right)^{5-4} + \binom{5}{5}\left(\frac{1}{2}\right)^5\left(1-\frac{1}{2}\right)^{5-5}$

$= P(X \geq 4 \text{ con } p=\frac{1}{2}) = 1 - P(X < 3 \text{ con } p=\frac{1}{2}) = 0.187$

1 - pbinom(3,5,0.5)

C'è il 19% di probabilità di osservare un numero di teste pari o più se H_0 è vero

Questa probabilità è chiamata p-value (valore-p)

Quindi per ora non rifiutiamo H_0

In generale,

p-value = probabilità di osservare dati così estremi o più estremi di quelli disponibili se H_0 fosse vero.
valore-p

pag. 386 es 3.8

$H_0: p = \frac{1}{2}$ (meglio $H_0: p \leq \frac{1}{2}$) $H_1: p > \frac{1}{2}$

p = proporzione di preoccupati

Spesso:
H_0 = "status quo"
H_1 = quello che il ricercatore vuole dimostrare

$n = 920$ osserviamo $x = 478$ sì → 1 - pbinom(477, 920, .5)

p-value $= P(X \geq 478 / p = \frac{1}{2}) = 0.1242$ è abbastanza alto ⇒ non rifiutiamo H_0

Il p-value è una misura probabilistica di compatibilità dei dati con l'ipotesi nulla H_0

Nell'esempio 3.8, p-value = 0.12, ancora abbastanza alto

Idea del test di ipotesi: identifico una regione di rifiuto R che mi porta a rifiutare H_0 se i dati sono in R, a non rifiutarla altrimenti.

R sarà una regione di BASSA COMPATIBILITÀ DEI DATI con H_0 (basso p-value)

	H_0 vero	H_0 non vero → H_1
rifiuto H_0	errore di prima specie	OK
non rifiuto H_0	OK	errore di seconda specie

mie opzioni basate sui dati

In genere si fissa la probabilità di errore di 1ª specie α a un livello convenzionale ($\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$)

Nell'esempio 8.8 esteso, se scelgo come soglia 486, cioè

$R = \{m : m \geq 486\}$

allora l'errore di I° specie è $P(\text{rifiuto } H_0 \mid H_0 \text{ vero}) = 0.046 \approx 0.05$

Rifiuterò se $x \geq 486$, ovvero se p-value ≤ 0.05

Vediamo come applicare questi concetti al caso del test di ipotesi sul valore atteso di una distribuzione normale

Il sistema di ipotesi che consideriamo è: $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$

dove μ_0 è un certo valore fissato di interesse

Esempio 8.1: $\mu_0 = 10$ $H_0 : \mu = 10$, $H_1 : \mu \neq 10$

La regione di rifiuto corrisponderà a $\bar{x}$ molto lontano da μ_0, cioè:

(1) $|\bar{x} - \mu_0|$ grande

(2) cioè $|z| = \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right|$ è grande

Quanto grande? Tanto grande che $P\left(\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \mid H_0 \right) = \alpha$ (probabilità errore di 1ª specie)

Ragionare con (2) invece che con (1) ci permette di individuare $z_{\alpha/2}$ come la soglia che cercavamo

Ponendo α = probabilità dell'errore di 1ª specie (o livello di significatività) uguale a un valore basso

23-11-09

Se $x_1 \dots x_n$ iid $N(\mu, \sigma^2)$ σ^2 noto

oppure

se n è grande e $x_1 \dots x_n$ iid con valore atteso incognito μ,

possiamo costruire un test dell'ipotesi

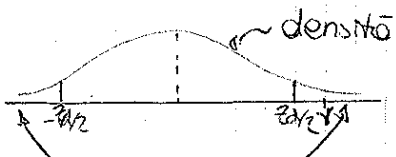
$H_0 : \mu = \mu_0$ contro l'alternativa $H_1 : \mu \neq \mu_0$ con μ_0 valore di "status quo"

Calcoliamo la statistica del test (ST):

$$ST = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

e rifiutiamo quando $|ST| \geq z_{\alpha/2}$ → REGOLA DI DECISIONE (la probabilità di errore di 1ª specie è α)

Il p-value è invece



regioni di rifiuto di R: insieme di valori con poco supporto per H_0

Supponiamo ora di fare il test e osservare il valore γ per la ST
↳ nel caso in figura $\gamma > z_{\alpha/2}$ → quindi rifiuto H_0

il (p-value) è valore-p è $P(z > \gamma)$

→ p-value = il più piccolo valore di α che porterebbe al rifiuto di H_0

Test di ipotesi unilaterale: $H_0: \mu = \mu_0$ ovvero $H_0: \mu \leq \mu_0$
$H_1: \mu > \mu_0$

Esempio 9.5

$x_1 \dots x_m$ sono le variazioni del livello di colesterolo in 40 pazienti.

x_i = colesterolo dopo + colesterolo prima del trattamento

$H_0: \mu = 0$

6.8 = valore medio osservato, valore osservato di $\bar{x}$

12.1 = valore osservato di S

$$t = \frac{\sqrt{m}(\bar{x} - \mu_0)}{S} = 3.55 \qquad \text{p-value} = 0.0001$$

CAPITOLO 10 (10.2, 10.3)

$X_1 \dots X_m$ iid $N(\mu_X, \sigma_X^2)$
$Y_1 \dots Y_n$ iid $N(\mu_Y, \sigma_Y^2)$ > indipendenti

Vogliamo testare $H_0: \mu_X = \mu_Y$

La statistica test $\bar{X} - \bar{Y}$ se σ_X^2 e σ_Y^2 sono note

(1) $\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$

oppure (2) $\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}$ se m e n grandi, per il TLC

Se H_0 è vero, la ST si comporta come $z \sim N(0,1)$

Quindi → tabella pag 403 per (1)
→ tabella pag. 416 per (2)

Rifiuto H_0 se ST è "estremo", cioè su uno delle due code per il test bilaterale, o su uno delle due per il test unilaterale

Esempio 10.3

In questo esempio poiché: $H_0: \mu_X \leq \mu_Y$ rifiuto H_0 per valori $\bar{x}$ molto
$H_1: \mu_X > \mu_Y$

più grandi di $\bar{y}$, cioè per valori $\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}$ sulla coda destra

poiché osservo $\gamma = 1.34$ ~D p-value = 0.09.

Poiché p-value è alto rifiuterei H_0 solo per valori di α maggiori di 0.09, troppo alti
Quindi non rifiuto H_0 (la prova è non significativa)

Relazione tra i.c. e test: spesso un test su un parametro θ con ipotesi nulla della forma $H_0: \theta = \theta_0$ rifiuta H_0 se e solo se un appropriato i.c. per θ non contiene θ_0

Esempio: $H_0: \mu_X = \mu_Y$ ovvero $H_0: \mu_X - \mu_Y$
$H_1: \mu_X \neq \mu_Y$

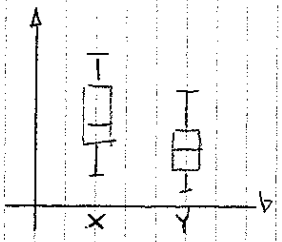
Test (a quello 2) rifiuta H_0 se un i.c. (di quello $1-\alpha$) per $\mu_X - \mu_Y$ non contiene 0

REGRESSIONE - CAPITOLO 12

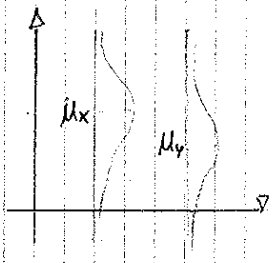
30 - 11 - 05

RIPASSO

Confronto tra 2 campioni $x_1 \dots x_m$, $y_1 \dots y_m$



osservazioni reali



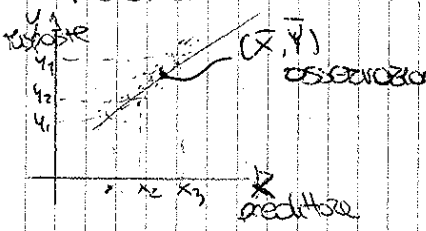
modello teorico

Abbiamo visto che:
1) i.c. per $\mu_X - \mu_Y$
2) test dell'ipotesi $H_0: \mu_X - \mu_Y$

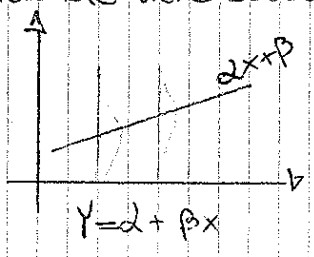
Oggi parleremo di dipendenza di una variabile risposta Y da una variabile predittore x.

Relazione più semplice: lineare $\rightarrow Y = a + bx$

In statistica una relazione funzionale viene osservata a meno di un errore (variabile)



$\Rightarrow$



Esempi:

x	y
altezza padre	altezza figlio
altezza	consumo

$E(Y) = \alpha + \beta x$ ovvero $Y = E(Y) + \text{errore} = \alpha + \beta x + e$

REGRESSIONE LINEARE SEMPLICE

dipendenza statistica $E(Y) = f(x)$ — $f(x)$ lineare — $f(x)$ funzione della sola x (non stocastica = non aleatoria)

Il predittore x è una variabile sotto il controllo dello sperimentatore

La risposta Y è una variabile aleatoria (stocastica) il cui valore atteso è funzione di x, $E(Y) = f(x)$

Esempio 12.4

DATI: osservazioni $y_1 \dots y_n$ corrispondenti a diversi livelli del predittore $x_1 \dots x_n$

OGGETTO DI INFERENZA: i parametri α e β e la variabilità dell'errore

METODO DI STIMA DI α e β: metodo dei minimi quadrati

↓ trova α e β in modo da minimizzare

$$\sum (y_i - (\alpha + \beta x_i))^2$$

(SCARTO DELLA RETTA)

la somma dei quadrati degli scarti dalla retta

$$\min_{\alpha,\beta} \sum (y_i - (\alpha + \beta x_i))^2$$

Si ottiene la seguente soluzione in forma chiusa:

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{cases}$$

la retta stimata che ne risulta $\hat{y} = \hat{\alpha} + \hat{\beta}x$ è chiamata RETTA DI REGRESSIONE STIMATA O RETTA DEI MINIMI QUADRATI

Il MODELLO di REGRESSIONE LINEARE SEMPLICE

$E(Y) = \alpha + \beta x$ cioè $Y = \alpha + \beta x + e$

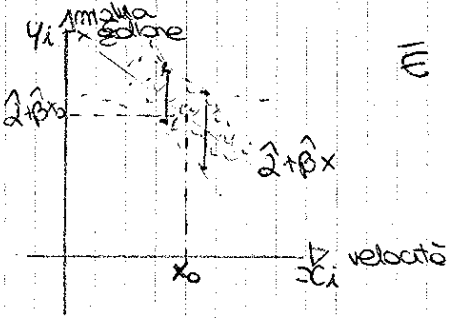
→ stima dei minimi quadrati $\hat{\alpha}$ e $\hat{\beta}$

Se, in aggiunta, supponiamo che gli Y sono normali, con la stessa varianza incognita σ^2

possiamo stimare σ^2 $\quad$ MSE = $\hat{\sigma}^2 = \dfrac{\sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n-2} = \dfrac{SS_E}{n-2}$

↳ mean square error

parametro incognito	stima
α	$\hat{\alpha}$
β	$\hat{\beta}$
σ^2	$\hat{\sigma}^2$
$\alpha + \beta x_0$	$\hat{\alpha} + \hat{\beta} x_0$



È possibile anche costruire degli i.c. per i parametri incogniti

$E(Y; x_0) = \alpha + \beta x_0$ = valore atteso di Y corrispondente al nuovo valore di x_0

? Tutti i.c. non sono nel testo ma sono ampiamente disponibili

INTERVALLI DI PREDIZIONE

02-12-08

Le differenze $y_i - \hat{\alpha} - \hat{\beta} x_i$ sono chiamate residui

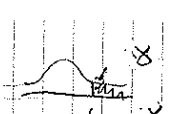
Una volta osservati i valori y_i di Y_i i residui diventano i numeri $y_i - \hat{\alpha} - \hat{\beta} x_i$

Test di ipotesi: $H_0: \beta = 0$ (x non ha influenza su Y)

Rifiutiamo H_0 se $\hat{\beta}$ è troppo distante da 0, ovvero

rifiutiamo H_0 se $\sqrt{\dfrac{(n-2)SS_x}{SS_E}}\hat{\beta} = ST$ (statistico test) è più grande, in valore assoluto,

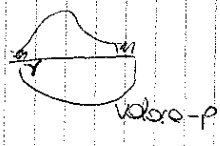
di t_{n-2}, γ



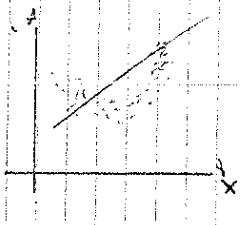
Equivalentemente, il valore-p del test è il seguente. Supponiamo di osservare

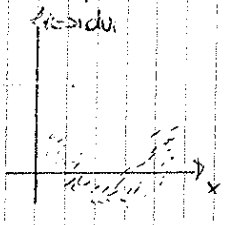
$$ST = \gamma$$

Allora valore-p $= P(|T_{n-2}| \geq |\gamma|)$



Rifiutiamo H_0 se valore-p $< \delta$





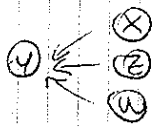
esempio di grafico dei residui che denuncia una deficienza dell'interpolazione lineare

Sospetto: Y dipende in maniera non lineare da x

Per esempio $y = e^{\beta x}$ oppure $\log y = x$ $\rightarrow$ trasformazioni dei dati

oppure $y = a + bx + cx^2$ due predittori = regressione multipla

Se Y dipende da x e da altre variabili z, w



$Y = a + bx + cz + dw + \text{errore}$

REGRESSIONE MULTIPLA

REGRESSIONE MULTIPLA:

$$(*) \quad Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Y: variabile risposta; β_0: intercetta; β_1: coefficienti di regressione; x_i: k predittori

Partendo da una relazione di regressione multipla del tipo $(*)$ possiamo chiederci

1) Ci sono alcuni predittori inutili? (e quindi $\beta = 0$?)
2) abbiamo dimenticato qualche predittore importante?
3) è la relazione lineare sufficientemente buona?

1) $\rightarrow$ test di ipotesi sui singoli β: $H_0: \beta_1 = 0$

Sulla base dei dati calcoliamo le stime dei minimi quadrati

$\hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_k$ di $\beta_0, \beta_1 \dots \beta_k$

minimizzando $\sum (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik})$

con x_{ij} = valore che il predittore j-esimo (j-mo coefficiente) assume nell'unità statistica i-esima (i-esimo rigo)

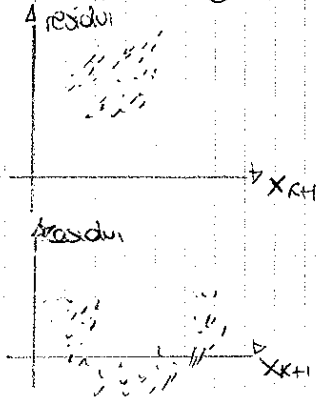
Il test dell'ipotesi $H_0: \beta_1 = 0$ viene fatto usando $\hat{\beta}_1$ e altre informazioni...

Il valore-p per questa ipotesi può essere letto nell'output del programma di regressione

2) è più difficile. Usiamo i residui. I residui sono:

$$(Y_i - B_0 - B_1 X_{i1} - \dots - B_K X_{iK} \qquad i = 1 \dots n$$

Potrei fare un grafico di questi residui su un eventuale valore predittore X_{K+1}



no pattern = X_{K+1} non spiega nulla in più

andamento costante: X_{K+1} "spiega" i residui

COEFFICIENTE DI DETERMINAZIONE

3) i residui servono anche per valutare la bontà di un fit di regressione tramite il coefficiente di determinazione

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = \text{frazione di variabilità di } Y \text{ spiegata dall'interpolazione lineare}$$

dove $S_{YY} = \sum (Y_i - \bar{Y})^2$ variabilità totale di Y

$SS_R = \sum (Y_i - \hat{Y}_i)^2$ variabilità di Y non "spiegata" dall'interpolazione $\hat{Y}$

Più R^2 si avvicina a 1, più buona è la regressione

R^2 si usa anche per la regressione multipla. Se c'è un solo predittore X,

R^2 = quadrato del coefficiente di correlazione lineare tra Y e X

$-1 \le R \le 1$ solo univariato

$0 \le R^2 \le 1$ anche multivariato