

Ulteriori Conoscenze di Informatica e Statistica

Carlo Meneghini

Dip. di fisica - via della Vasca Navale 84,
st. 83 (I piano) tel.: 06 55 17 72 17

meneghini@fis.uniroma3.it

Tassi e proporzioni

Classi nominali: non possono essere messe in relazione matematica quantitativa con una scala di riferimento.

Es.:

maschi/femmine,
bianco/nero,
mancini/destrorsi,...

45% maschi, 55% femmine



Processi Bernulliani: ammettono solo due possibilità

Ogni esperimento può avere solo due risultati V/F, 1/0, si/no,...
ovvero:

Ogni unità della popolazione appartiene solo a una delle due classi

gli esperimenti sono indipendenti,

ovvero

ogni unità del campione è determinata indipendentemente dalle altre

la probabilità p di un certo risultato è costante durante l'esperimento

ovvero

la proporzione delle classi è costante durante l'esperimento

Analisi di proporzioni

p = probabilità di successi = $N_{\text{successi}}/N_{\text{totale}}$

q = probabilità di insuccessi = $1-p$

Valore medio (proporzione):

$X(\text{successo}) = 1$
 $X(\text{insuccesso}) = 0$

$$\mu = \frac{1}{N_t} \sum_{i=1}^{N_t} X_i = \frac{N_{\text{succ}}}{N_t} = p$$

dev.st:

$$\sigma = \sqrt{p(1-p)}$$

Stime campionarie di proporzioni

stima di p :

$$\bar{x} = \frac{N_{\text{succ}}}{N_T} = f_s$$

stima di σ^2

$$s^2 = f_s(1-f_s)$$

uso la stima di p per la stima della σ

errore sulla stima di p :

$$s_f = \frac{\sigma}{\sqrt{N_T}} = \frac{\sqrt{f_s(1-f_s)}}{\sqrt{N_T}}$$

In un esperimento di 10 lanci di una moneta si ottengono 6 teste

$p_s = 0.6$

$s_p = 0.075$

il valore medio atteso (la probabilità di successo) è, con il 95% di probabilità, tra

$p-2s_p$ e $p+2s_p$

Test sulle frequenze di un "attributo"

$H_0: f_m = p_0$

Frequenza osservata:

$$f_m = \frac{N_{\text{succ}}}{m}$$

Errore stimato sulla frequenza:

$$\sigma_{f_m} = \frac{\sqrt{N_{\text{succ}}(m - N_{\text{succ}})}}{m} = \sqrt{\frac{f_m(1-f_m)}{m}}$$

La variabile aleatoria: $z_m = \frac{f_m - p_0}{\sigma_{f_m}}$ segue una distribuzione normale standard ($N(0,1)$)

Es.: su un campione si 100 intervistati ha risposto Sì il 58%. E' significativamente maggiore di 50%?

1) Scelgo un livello di confidenza $\alpha=4\%$

2) $\text{Inv.Norm.St}(0.98) = 1.75 = z_\alpha$

3) lo confronto con il valore della z_m

$$z_{100} = \frac{0.58 - 0.5}{\sqrt{0.58 \cdot 0.42/100}} = 1.62$$



Con un livello di confidenza di 4% la maggioranza del Sì ottenuta dal campione scelto non è significativa

$$z_m = \frac{f_m - p_o}{\sigma_{f_m}} = \frac{f_m - p_o}{s_f} = \frac{\Delta P}{s_f} > z_\alpha$$

La differenza osservata è significativa se:

$$\frac{s_f}{|\Delta P|} < \frac{1}{z_\alpha}$$

Nuovo	N	f _m	s	s _f
58	100	0.58	0.494	0.049
p _o	0.5			
z _m	1.621		s _f Cp	0.617
a	0.050			
z _a	1.645			

	α	1 coda	2 code	1 coda	2 code
0.050	1.645	1.960	0.61	0.51	
0.025	1.960	2.241	0.51	0.45	
0.010	2.326	2.576	0.43	0.39	
0.005	2.576	2.807	0.39	0.36	

Es.: su un campione si 100 intervistati ha risposto Si il 58%. Quale è il rischio di sbagliare affermando la vittoria del Si al referendum?

Ho: P_A=P_B H₁: P_A>P_B

1) Calcolo il valore della z_m

$$z_{100} = \frac{0.58 - 0.5}{\sqrt{0.58 \cdot 0.42/100}} = 1.62$$

2) Calcolo la probabilità associata alla coda della distribuzione usando la funzione EXCEL:

DISTRIB.NORM.ST(1.62) = $\int_{-\infty}^{1.62} z dz$ = 0.947

$\int_{1.62}^{\infty} z dz = 1 - \text{DISTRIB.NORM.ST}(1.62) = 0.052$

Il rischio di sbagliare affermando la vittoria del Si al referendum è di 5.2%

Attenzione: test a una o due code

Es.: in un esperimento effettuato su un campione si 100 individui si osserva il 58% delle volte il carattere A e il 42% il carattere B. Quale è il rischio di sbagliare affermando che la popolazione non è equamente divisa?

H₀: P_A=P_B H₁: P_A≠P_B (maggiore o minore)

1) Calcolo il valore della z_m

$$z_{100} = \pm \frac{58 - 50}{\sqrt{58 \cdot 42/100}} = 1.62$$

DISTRIB.NORM.ST(1.62) = $\int_{-\infty}^{1.62} z dz$ = 0.947

$\int_{-\infty}^{-1.62} z dz + \int_{1.62}^{\infty} z dz = 2(1 - \text{DISTRIB.NORM.ST}(1.62)) = 0.104$

Il rischio di sbagliare affermando che la popolazione non è equamente divisa è di ~10%

Problema: si vogliono confrontare i risultati di due esperimenti in termini di frequenze relative

Ipotesi: (H₀) i due esperimenti appartengono alla stessa popolazione (hanno lo stesso valore aspettato)

$t_v = \frac{\text{differenza delle medie campionarie}}{\text{errore standard sulla differenza delle medie campionarie}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$

$z = \frac{\text{differenza delle frequenze campionarie}}{\text{errore standard sulla differenza delle frequenze campionarie}} = \frac{f_{m1} - f_{m2}}{s_{f1 - f2}}$

$s_{f1 - f2} = \sqrt{s_{f1}^2 + s_{f2}^2} = \sqrt{\frac{f_{m1}(1 - f_{m1})}{m_1} + \frac{f_{m2}(1 - f_{m2})}{m_2}}$

H₀: f_{m1} = f_{m2}

H₁: f_{m1} ≠ f_{m2} 1 coda

H₁: f_{m1} > f_{m2} 2 code

H₁: f_{m1} < f_{m2} 2 code

Nell'ipotesi (da verificare) che i due campioni provengono dalla stessa popolazione p_{m1}=p_{m2} la miglior stima della frequenza di successi ottenuta usando entrambi i campioni è:

$$\bar{f} = \frac{m_1 f_{m1} + m_2 f_{m2}}{m_1 + m_2}$$

e la varianza calcolata sull'intero campione è:

$$s_f^2 = \bar{f}(1 - \bar{f}) \left(\frac{1}{m_1} + \frac{1}{m_2} \right)$$

Correzione per la continuità (Yates):

Deriva dal fatto che la z può assumere solo valori discreti mentre la distribuzione normale è continua. La correzione diventa via via meno significativa man mano che aumenta il numero di prove n

$$z_{cor} = \frac{|f_{m1} - f_{m2}| - \frac{1}{2} \left(\frac{1}{m_1} + \frac{1}{m_2} \right)}{\sqrt{\bar{f}(1 - \bar{f}) \left(\frac{1}{m_1} + \frac{1}{m_2} \right)}}$$

Es.: in un sondaggio elettorale effettuato su un campione di 1000 persone, il 42% degli intervistati ha affermato di preferire la coalizione A.

In un secondo sondaggio effettuato su un egual numero di intervistati il 46% ha affermato di preferire la coalizione A.

Quale è il rischio che questa differenza sia dovuta al caso?

Quale è il valore di Z per credere alla differenza con rischio di errore inferiore al 5%

m₁ = m₂ = 1000

f₁ = 0.42

f₂ = 0.46

<f> = 0.44

s_f² = 4.93 · 10⁻⁴

z = 1.8

z_{cor} = 1.76

Rischio: 7.9%

rischio = 2(1 - distrib.norm.st(z_{cor}))

z_{5%} = inv.norm.st(1 - α/2) = 1.96

Per credere alla differenza con un rischio minore del 5% α = 5%

rischio: probabilità che, pur essendo valida l'ipotesi f₁=f₂, si osservi per caso un valore di z maggiore o uguale a quello trovato

dal momento che z_{cor} < z_{5%} la probabilità di sbagliare affermando che i due risultati sono diversi è maggiore del 5%

Tabelle di contingenza				
<i>Tabella sperimentale</i>	Gruppi	effetto 1	effetto 2	totali A
<i>frequenze assolute</i>	Trattamento 1	n_{11}	n_{12}	$N_1 = n_{11} + n_{12}$
	Trattamento 2	n_{21}	n_{22}	$N_2 = n_{21} + n_{22}$
	totali B	E_1	E_2	$N_T = N_1 + N_2 = E_1 + E_2$

Ipotesi (H_0): valori trovati sono determinati da una distribuzione casuale.

Se l'ipotesi è vera i risultati in tabella (n_{ij}) sono scorrelati e quindi le distribuzioni dei valori trovati nelle righe e nelle colonne sono scorrelate

<i>Tabella sperimentale</i>	Gruppi	effetto 1	effetto 2	totali A
<i>frequenze relative</i>	Trattamento 1	f_{11}	f_{12}	f_{T1}
	Trattamento 2	f_{21}	f_{22}	f_{T2}
	totali B	f_{E1}	f_{E2}	1

Distribuzione congiunta $f_{ij} = n_{ij} / N_T$

Frazioni con diverso trattamento $f_{T1} = N_1 / N_T$
 $f_{T2} = N_2 / N_T$ $P_{Ti} = \frac{\sum_j n_{ij}}{N}$ Distribuzione dei trattamenti

Frazioni con diverso esito $f_{E1} = E_1 / N_T$
 $f_{E2} = E_2 / N_T$ $P_{Ej} = \frac{\sum_i n_{ij}}{N}$ Distribuzione degli effetti

Se effetto e trattamento sono scorrelati la probabilità di avere l'effetto j con il trattamento i è il prodotto:

$$p_{ij} = P_{Ti} P_{Ej} \sim f_{Ti} f_{Ej}$$

<i>Tabella teorica</i>	Gruppi	effetto 1	effetto 2	totali A
<i>Ipotesi: i trattamenti hanno lo stesso effetto</i>	Trattamento 1	$f_{E1} f_{T1}$	$f_{E2} f_{T1}$	f_{T1}
	Trattamento 2	$f_{E1} f_{T2}$	$f_{E2} f_{T2}$	f_{T2}
	totali B	f_{E1}	f_{E2}	1

I valori attesi nell'ipotesi di effetti indipendenti dal trattamento ($p_{Ei} p_{Tj}$) sono diversi dal valore sperimentali p_{ij}

<i>Tabella teorica</i>	Gruppi	effetto 1	effetto 2	totali A
<i>Ipotesi: i trattamenti hanno lo stesso effetto</i>	Trattamento 1	$f_{E1} f_{T1} N_T$	$f_{E2} f_{T1} N_T$	N_1
	Trattamento 2	$f_{E1} f_{T2} N_T$	$f_{E2} f_{T2} N_T$	N_2
	totali B	E_1	E_2	N_T

<i>Tabella sperimentale</i>	Gruppi	effetto 1	effetto 2	totali A
<i>frequenze relative</i>	Trattamento 1	n_{11}	n_{12}	N_1
	Trattamento 2	n_{21}	n_{22}	N_2
	totali B	E_1	E_2	N_T

oppure:

<i>Tabella sperimentale</i>	Gruppi	effetto 1	effetto 2	totali A
<i>frequenze assolute</i>	Trattamento 1	f_{11}	f_{12}	f_{T1}
	Trattamento 2	f_{21}	f_{22}	f_{T2}
	totali B	f_{E1}	f_{E2}	1

<i>Tabella teorica</i>	Gruppi	effetto 1	effetto 2	totali A
<i>frequenze assolute</i>	Trattamento 1	$f_{E1} f_{T1}$	$f_{E2} f_{T1}$	f_{T1}
	Trattamento 2	$f_{E1} f_{T2}$	$f_{E2} f_{T2}$	f_{T2}
	totali B	f_{E1}	f_{E2}	1

Esempio				
<i>Tabella sperimentale</i>	Gruppi	effetto 1	effetto 2	totali A
	Trattamento 1	18	7	25
	Trattamento 2	6	13	19
	totali B	24	20	44

<i>Tabella teorica</i>	Gruppi	effetto 1	effetto 2	totali A
	Trattamento 1	13.64	11.36	25
	Trattamento 2	10.36	8.64	19
	totali B	24	20	44

Le differenze osservate sono significative?

La variabile χ^2

$\chi^2 = \sum_i \frac{(n_{exp} - n_{th})^2}{n_{th}}$

gradi di libertà $v = (n_{col}-1)(n_{righe}-1)$

frequenze osservate

frequenze attese

$$\chi^2_1 = \frac{(18 - 13.64)^2}{13.64} + \frac{(7 - 11.36)^2}{11.36} + \frac{(6 - 10.36)^2}{10.36} + \frac{(13 - 8.64)^2}{8.64} = 7.1$$

Gruppi	effetto 1	effetto 2	totali A
Tratt. 1	18	7	25
Tratt. 2	6	13	19
totali B	24	20	44

Gruppi	effetto 1	effetto 2	totali A
Tratt. 1	13.64	11.36	25
Tratt. 2	10.36	8.64	19
totali B	24	20	44

$\chi^2_1 = 7.1$

Funzione EXCEL: INV.CHI(α, v)

Se i trattamenti non avessero un effetto diverso, un valore di χ^2 con 1 grado di libertà maggiore di 6.63 ha una probabilità di essere osservato minore del 1%. Possiamo quindi affermare che i due trattamenti hanno un diverso effetto sui due gruppi con una probabilità di errore minore dell'1%

http://faculty.vassar.edu/lowry/PDF/t_tables.pdf

Nota: $\chi^2_{NT} = \sum_i \frac{(f_{exp} - f_{th})^2}{f_{th}}$

exp

th

χ^2_{NT}

χ^2_{NT}

Titanic
14-04-1912

Classe	sopravvissuti	decaduti	totali A
I	200	123	
II	119	158	
III	181	528	

I decessi sono correlati con la classe?

Funzione EXCEL: INV.CHI(α, v) restituisce il valore associato ad un livello di confidenza α per una variabile χ^2 con v g.d.l.

la funzione: DISTRIB.CHI(χ^2, v) restituisce la probabilità associata al valore di χ^2 con v g.d.l.

Valori

Frequenze

TEST

Probabilità associata al χ^2 calcolato sulle matrici dei dati sperimentali e teorici

la funzione EXCEL: TEST.CHI(M_{exp}, M_{th}) restituisce la probabilità associata al valore di ottenuto da una matrice di dati sperimentali M_{exp} e da una matrice di dati teorici M_{th} ottenuta nell'ipotesi di distribuzione casuale dei risultati.

Sondaggio Politico-Elettorale

Autore: Elma S.r.l.

Comittente/Acquirente: Clandestino

Criteri seguiti per la formazione del campione: Le interviste sono state realizzate su un campione di 1.000 casi stratificato per sesso ed età

Metodo di raccolta delle informazioni: Intervista telefonica - Metodologia C.A.T.I.

Numero delle persone interpellate e universo di riferimento: 1.000 casi - Universo di riferimento: 48.493.370 adulti maggiorenni residenti in Italia (Fonte: Istat - Popolazione al 01/01/2005)

Data in cui è stato realizzato il sondaggio: Tra il 2/4/2007 ed il 3/4/2007

<http://www.sondaggipoliticoelettorali.it/>

<http://www.sondaggipolitici.it/>

Risultati dei Test di ammissione

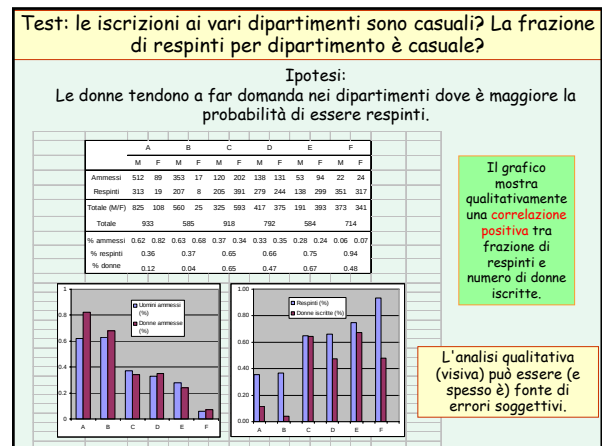
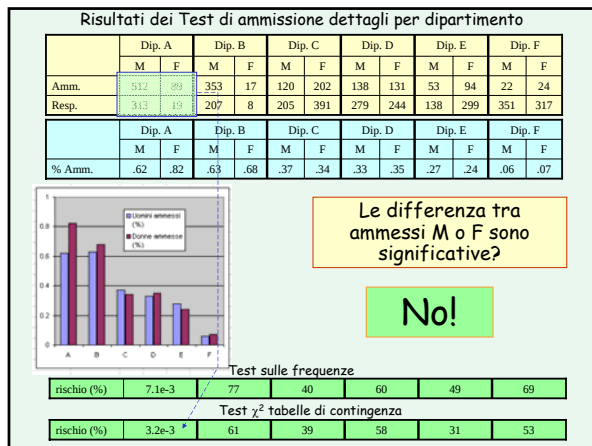
	M	F
Ammessi	1198	557
Respinti	1493	1278

	M	F
Ammessi	.44	.30
Respinti	.56	.70

Conclusione:

Le donne sono discriminate nei test di ammissione!

V/F ?



Quantificare la correlazione

$y = f(x)$ i valori della variabile y sono funzione dei valori assunti dalla variabile indipendente x (deterministico):

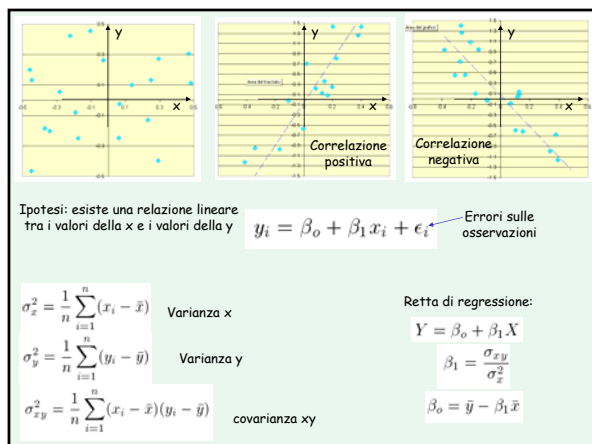
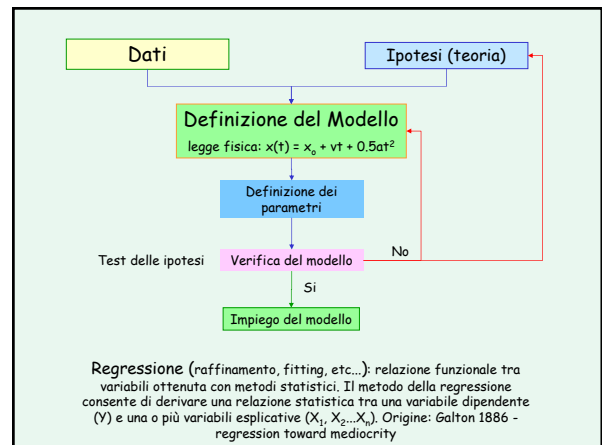
$F = ma$; $x = vt$; $S = ba$; $V = IR$;

Modelli statistici e inferenza statistica

Modelli: relazione funzionale tra ciò che si vuole spiegare (effetto) e le cause.

Un modello statistico è:

- una semplificazione della realtà (rasoio di Occam: scartare le ipotesi complesse se ne esistono di più semplici che portano allo stesso risultato)
- un'analogia del fenomeno reale: il modello riproduce solo alcuni aspetti della realtà ma non è la realtà



Usando le variabili standardizzate: $X_i = \frac{x_i - \bar{x}}{s_x}$ $Y_i = \frac{y_i - \bar{y}}{s_y}$

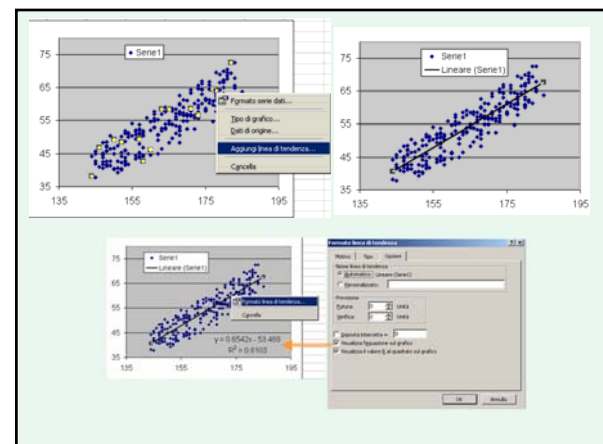
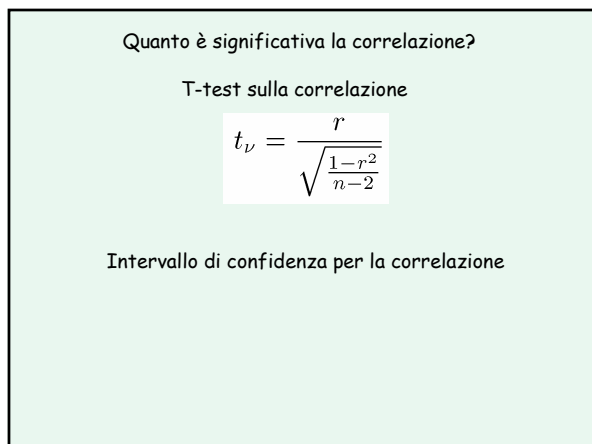
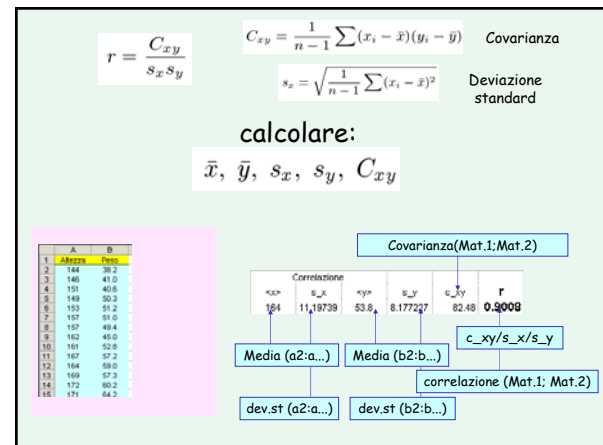
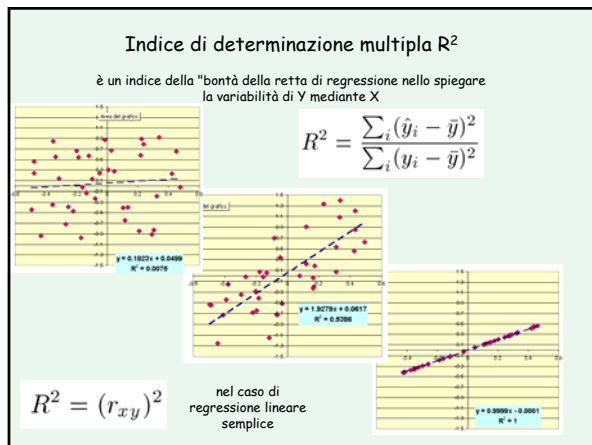
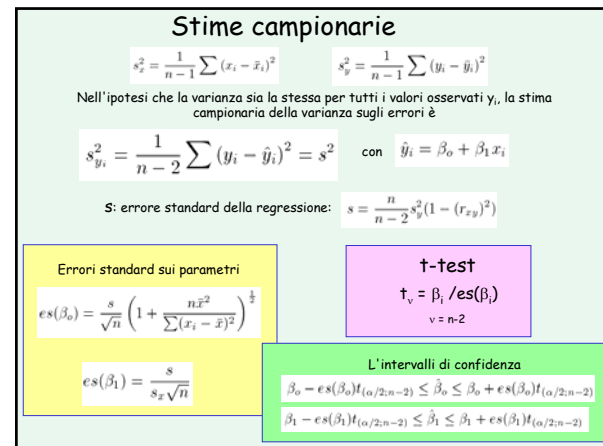
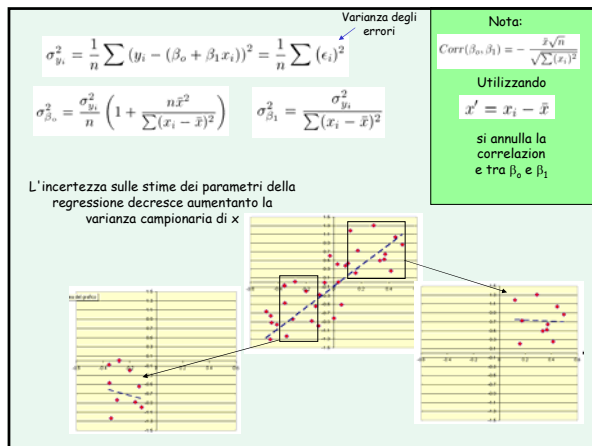
La retta di regressione diventa: $Y_i = \beta^* X_i$

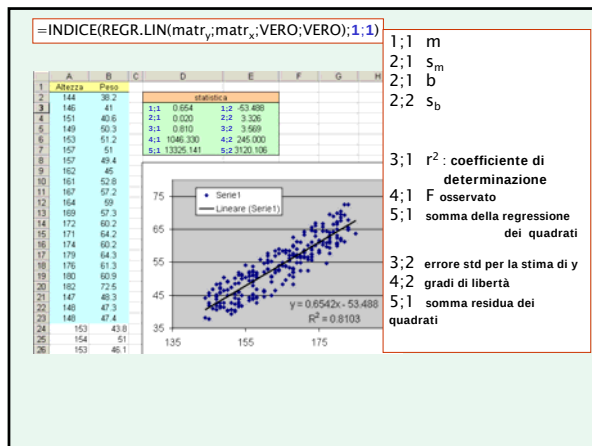
$y_i - \bar{y} = \beta^* \frac{s_y}{s_x} (x_i - \bar{x})$

$\beta_1 = \beta^* \frac{s_y}{s_x}$ $\beta^* = \beta_1 \frac{s_x}{s_y} = \frac{s_{xy}}{s_x s_y} = r_{xy}$

$r =$ coefficiente di correlazione di Pearson

$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$





help
regr.lin

Statistica	Descrizione
$\sigma_1, \sigma_2, \dots, \sigma_n$	I valori di errore standard per i coefficienti $m_1, m_2, \dots, m_n$
s_b	Il valore di errore standard per la costante b ($s_b = \text{MIN}(\text{D}$ quando cost = PALECO)
r^2	Il coefficiente di determinazione. Confronta i valori y previsti con quelli effettivi e può avere un valore compreso tra 0 e 1. Se è uguale a 1, significa che esiste una correlazione perfetta nel campione, vale a dire, non c'è alcuna differenza tra il valore previsto e il valore effettivo di y. Se invece il coefficiente di determinazione è uguale a 0, l'equazione di regressione non sarà di alcun aiuto nella stima di un valore y. Per ulteriori informazioni sul metodo di calcolo di r^2 , consultare "Osservazioni" più avanti in questo argomento.
σ_y	L'errore standard per la stima di y
F	La statistica F o il valore osservato di F. Utilizzare la statistica F per determinare se la relazione osservata tra le variabili dipendenti e indipendenti è casuale.
gdi	I gradi di libertà. Utilizzare i gradi di libertà per trovare i valori critici di F in una tabella statistica. Confrontare i valori trovati nella tabella con la statistica F restituita dalla funzione REGR.LIN per stabilire un livello di confidenza per il modello.
sqreg	La somma della regressione dei quadrati
sqres	La somma residua dei quadrati

La seguente illustrazione mostra l'ordine in cui vengono restituite le statistiche aggiuntive di regressione.

	A	B	C	D	E	F
1	m_1	$m_2, \dots$	m_n	b		
2	σ_{m_1}	$\sigma_{m_2, \dots}$	σ_{m_n}	s_b		
3	r^2	σ_y				
4	F	gdi				
5	sqreg	sqres				