# 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 puó 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 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_successi/N_totale

q = probabilità di insuccessi = 1-p

## Valore medio (proporzione):

X(successo) = 1
X(insuccesso) = 0

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

## dev.st:

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

# Stime campionarie di proporzioni

stima di p:

$$\bar{x} = \frac{N_{suc}}{N_T} = f_s$$

stima di $\sigma^2$

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

uso la stima di p per la stima della $\sigma$

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"

$H_o$: $f_m = p_o$

Frequenza osservata:

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

Errore stimato sulla frequenza:

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

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

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

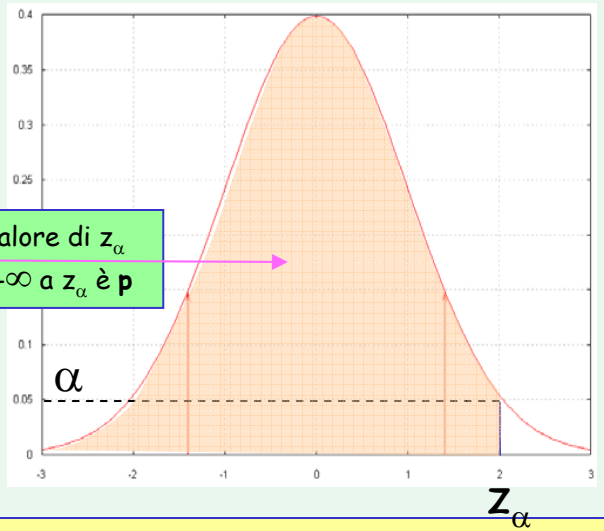
1) Scelgo un livello di confidenza $\alpha$=4%

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

3) lo confronto con il valore della $z_m$

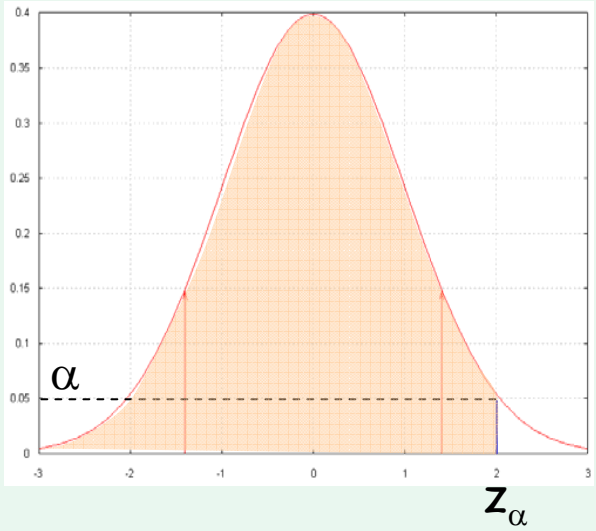
$$z_{100} = \frac{0.58 - 0.5}{\sqrt{0.58 \cdot 0.42/100}} = 1.62$$

INV.NORM.ST(**p**) = valore di $z_\alpha$ per cui l'integrale da -∞ a $z_\alpha$ è **p**




Con un livello di confidenza di 4% la maggiornaza del Si 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:

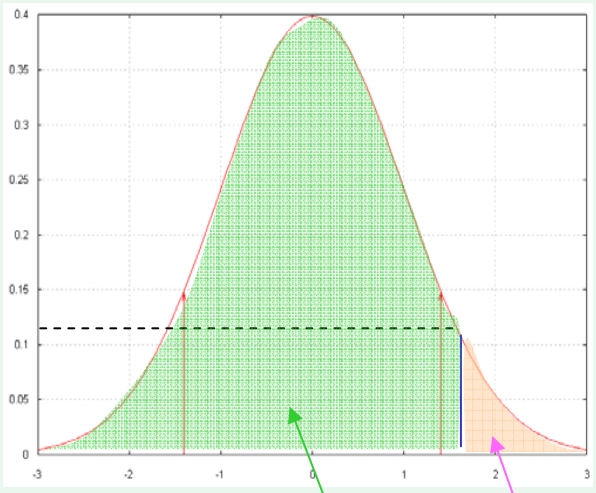
$$\frac{s_f}{|\Delta P|} < \frac{1}{z_\alpha}$$

| Nsucc | N | f_m | s | s_f |
|---|---|---|---|---|
| 58 | 100 | 0.58 | 0.494 | 0.049 |
| p_o | 0.5 | | | |
| z_m | 1.621 | | s_f/Dp | 0.617 |
| a | 0.050 | | | |
| z_a | 1.645 | | | |

| α | $z_\alpha$ 1 coda | $z_\alpha$ 2 code | $\frac{1}{z_\alpha}$ 1 coda | $\frac{1}{z_\alpha}$ 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: $P_A=P_B$ $H_1 : P_A>P_B$




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) = $\int_{-\infty}^{z_m} zdz$ =0.947

DISTRIB.NORM.ST(**z**) = probabilità di ottenere un valore minore di z.

$\int_{z_m}^{\infty} zdz$ = 1 - DISTRIB.NORM.ST(1.62) =0.052

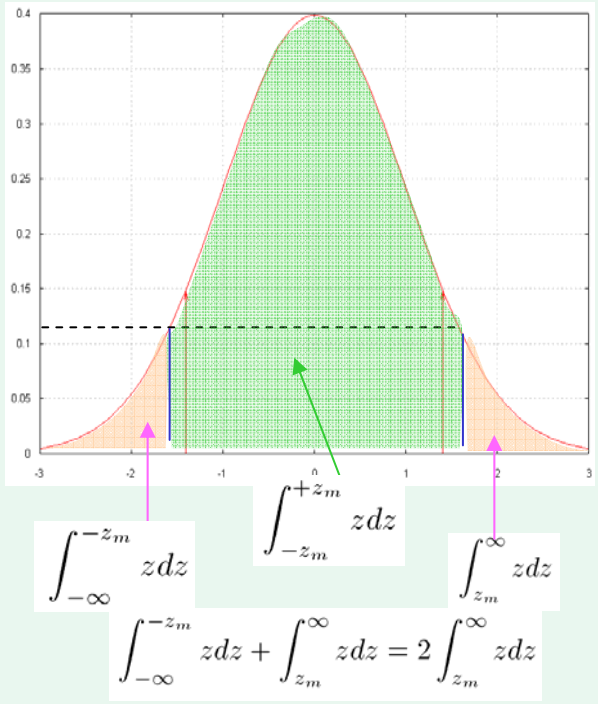
**1**-DISTRIB.NORM.ST(**z**) = probabilità di ottenere un valore maggiore di z.

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?

$H_0$: $P_A = P_B$ $H_1$ : $P_A \neq P_B$ (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) = $\int_{-\infty}^{z_m} zdz$ =0.947

$\int_{-\infty}^{-z_m} zdz + \int_{z_m}^{\infty} zdz = 2\int_{z_m}^{\infty} zdz$ = 2(1 - 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**: ($H_o$) i due esperimenti appartengono alla stessa popolazione (hanno lo stesso velore aspettato)

$$t_\nu = \frac{\text{differenza delle medie campionarie}}{\text{errore standard sulla differenza delle medie campionarie}}$$

$$t_\nu = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

$$z = \frac{\text{differenza delle frequenze campionarie}}{\text{errore standard sulla differenza delle frequenze campionarie}}$$

$$z = \frac{f_{m_1} - f_{m_2}}{s_{f_1 - f_2}}$$

$$s_{f_1 - f_2} = \sqrt{s_{f_1}^2 + s_{f_1}^2} = \sqrt{\frac{f_{m_1}(1 - f_{m_1})}{m_1} + \frac{f_{m_2}(1 - f_{m_2})}{m_2}}$$

$$H_o : f_{m_1} = f_{m_2}$$

$H_1 : f_{m_1} \neq f_{m_2}$ 1 coda

$H_1 : f_{m_1} > f_{m_2}$

$H_1 : f_{m_1} < f_{m_2}$

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:

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

e la varianza calcolata sull'intero campione è:

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

$$z = \frac{f_{m_1} - f_{m_2}}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{m_1} + \frac{1}{m_2}\right)}}$$

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 m

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

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%

$m_1 = m_2 = 1000$

$f_1 = 0.42$

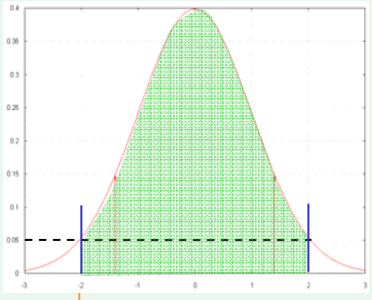
$f_2 = 0.46$

$\langle f \rangle = 0.44$

$s^2_f = 4.93 \cdot 10^{-4}$

$z = 1.8$

$z_{cor} = 1.76$



## Rischio: 7.9%

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

rischio: probabilità che, pur essendo valida l'ipotesi $f_1=f_2$, si osservi per caso un valore di z maggiore o uguale a quello trovato

Per credere alla differenza con un rischio minore del 5%

$\alpha = 5\%$

$z_{5\%}$ = inv.norm.st(1-$\alpha/2$) = 1.96

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

*Tabella sperimentale*

*frequenze assolute*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| 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 ($H_o$): 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

*Tabella sperimentale*

*frequenze relative*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| 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:

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

*frequenze relative*

*Tabella teorica*

*Ipotesi: i trattamenti hanno lo stesso effetto*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| 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 | $f_{E1}$ | $f_{E1}$ | 1 |

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

*frequenze assolute*

*Tabella teorica*

*Ipotesi: i trattamenti hanno lo stesso effetto*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| 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 | $E_1$ | $E_2$ | $N_T$ |

## *frequenze relative*

*Tabella sperimentale*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| Trattamento 1 | $n_{11}$ | $n_{12}$ | $N_1$ |
| Trattamento 2 | $n_{21}$ | $n_{22}$ | $N_1$ |
| totali B | $E_1$ | $E_2$ | $N_T$ |

*Tabella teorica*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| 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:**

## *frequenze assolute*

*Tabella sperimentale*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| Trattamento 1 | $f_{11}$ | $f_{12}$ | $f_{T1}$ |
| Trattamento 2 | $f_{21}$ | $f_{22}$ | $f_{T2}$ |
| totali B | $f_{E1}$ | $f_{E2}$ | 1 |

*Tabella teorica*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| 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 | $f_{E1}$ | $f_{E1}$ | 1 |

# Esempio

*Tabella sperimentale*

| Gruppi | effetto 1 | effetto 2 | totali A |
|---|---|---|---|
| Trattamento 1 | 18 | 7 | 25 |
| Trattamento 2 | 6 | 13 | 19 |
| totali B | 24 | 20 | 44 |

*Tabella teorica*

| 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 sognificative?

# La variabile $\chi^2$

frequenze attese

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

gradi di libertà

$\nu = (n_{col.}-1)(n_{righe}-1)$

frequenze osservate

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

exp

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

Th

| 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 |

$\nu$

$\alpha$

$\chi^2_1 = 7.1$

Funzione EXCEL: INV.CHI($\alpha$,$\nu$)

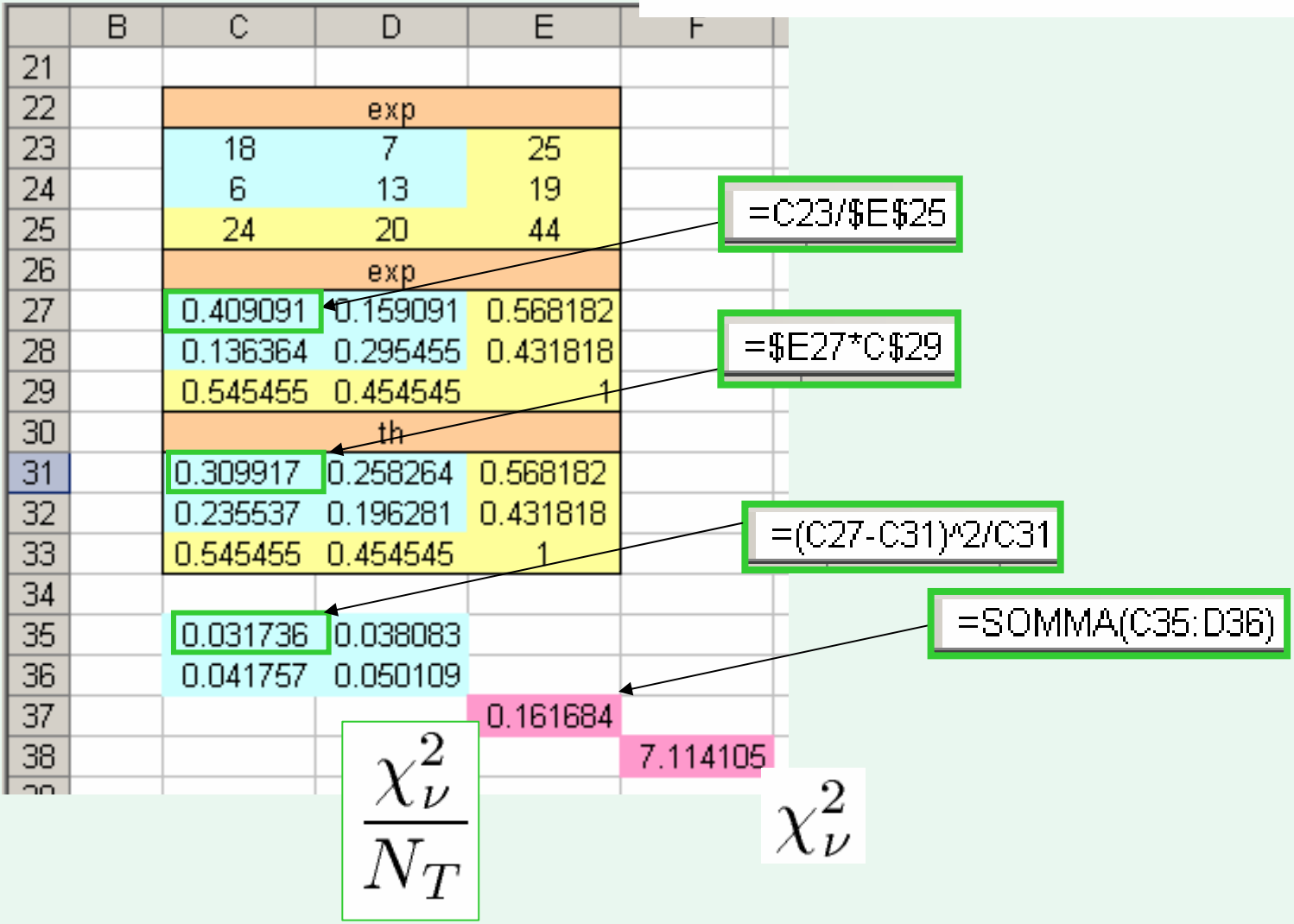
| | Level of Significance | | | | |
|---|---|---|---|---|---|
| **df** | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
| 1 | 3.84 | 5.02 | 6.63 | 7.88 | 10.83 |
| 2 | 5.99 | 7.38 | 9.21 | 10.60 | 13.82 |
| 3 | 7.81 | 9.35 | 11.34 | 12.84 | 16.27 |
| 4 | 9.49 | 11.14 | 13.28 | 14.86 | 18.47 |
| 5 | 11.07 | 12.83 | 15.09 | 16.75 | 20.51 |
| 6 | 12.59 | 14.45 | 16.81 | 18.55 | 22.46 |
| 7 | 14.07 | 16.01 | 18.48 | 20.28 | 24.32 |
| 8 | 15.51 | 17.53 | 20.09 | 21.95 | 26.12 |
| 9 | 16.92 | 19.02 | 21.67 | 23.59 | 27.88 |
| 10 | 18.31 | 20.48 | 23.21 | 25.19 | 29.59 |
| 11 | 19.68 | 21.92 | 24.73 | 26.76 | 31.26 |
| 12 | 21.03 | 23.34 | 26.22 | 28.30 | 32.91 |
| 13 | 22.36 | 24.74 | 27.69 | 29.82 | 34.53 |
| 14 | 23.68 | 26.12 | 29.14 | 31.32 | 36.12 |
| 15 | 25.00 | 27.49 | 30.58 | 32.80 | 37.70 |
| 16 | 26.30 | 28.85 | 32.00 | 34.27 | 39.25 |
| 17 | 27.59 | 30.19 | 33.41 | 35.72 | 40.79 |
| 18 | 28.87 | 31.53 | 34.81 | 37.16 | 42.31 |
| 19 | 30.14 | 32.85 | 36.19 | 38.58 | 43.82 |
| 20 | 31.41 | 34.17 | 37.57 | 40.00 | 45.31 |
| 21 | 32.67 | 35.48 | 38.93 | 41.40 | 46.80 |
| 22 | 33.92 | 36.78 | 40.29 | 42.80 | 48.27 |
| 23 | 35.17 | 38.08 | 41.64 | 44.18 | 49.73 |
| 24 | 36.42 | 39.36 | 42.98 | 45.56 | 51.18 |
| 25 | 37.65 | 40.65 | 44.31 | 46.93 | 52.62 |

Se i trattamenti non avessero un effetto diverso, un valore di $\chi^2$ 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:

$$\frac{\chi^2_\nu}{N_T} = \sum_i \frac{(f_{exp} - f_{th})^2}{f_{th}}$$

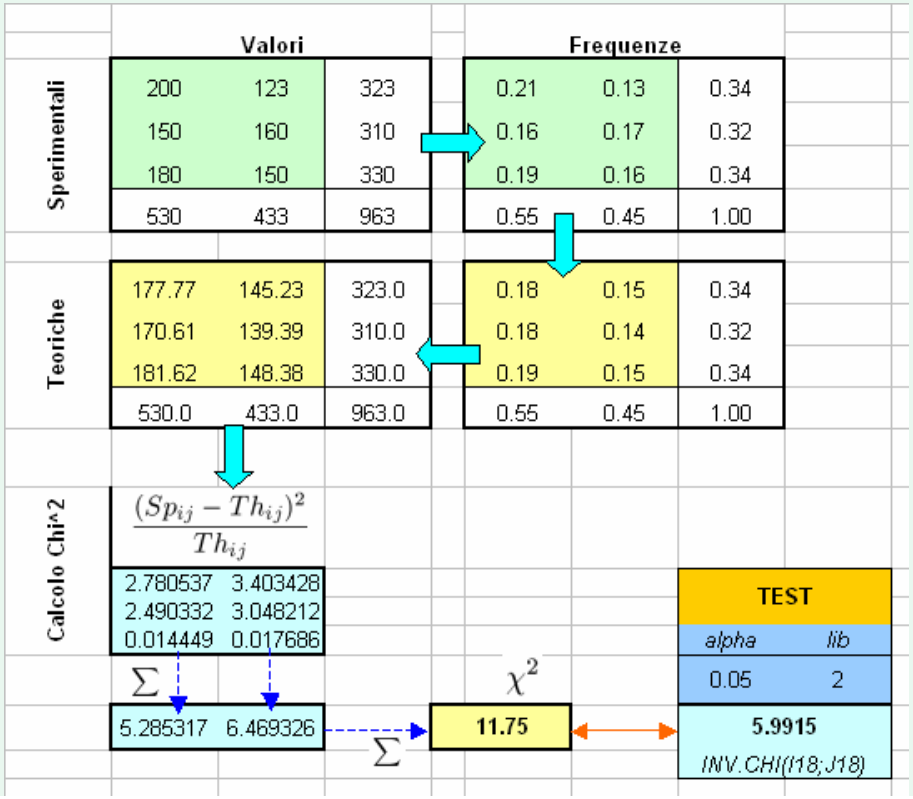


# Titanic

## 14-04-1912

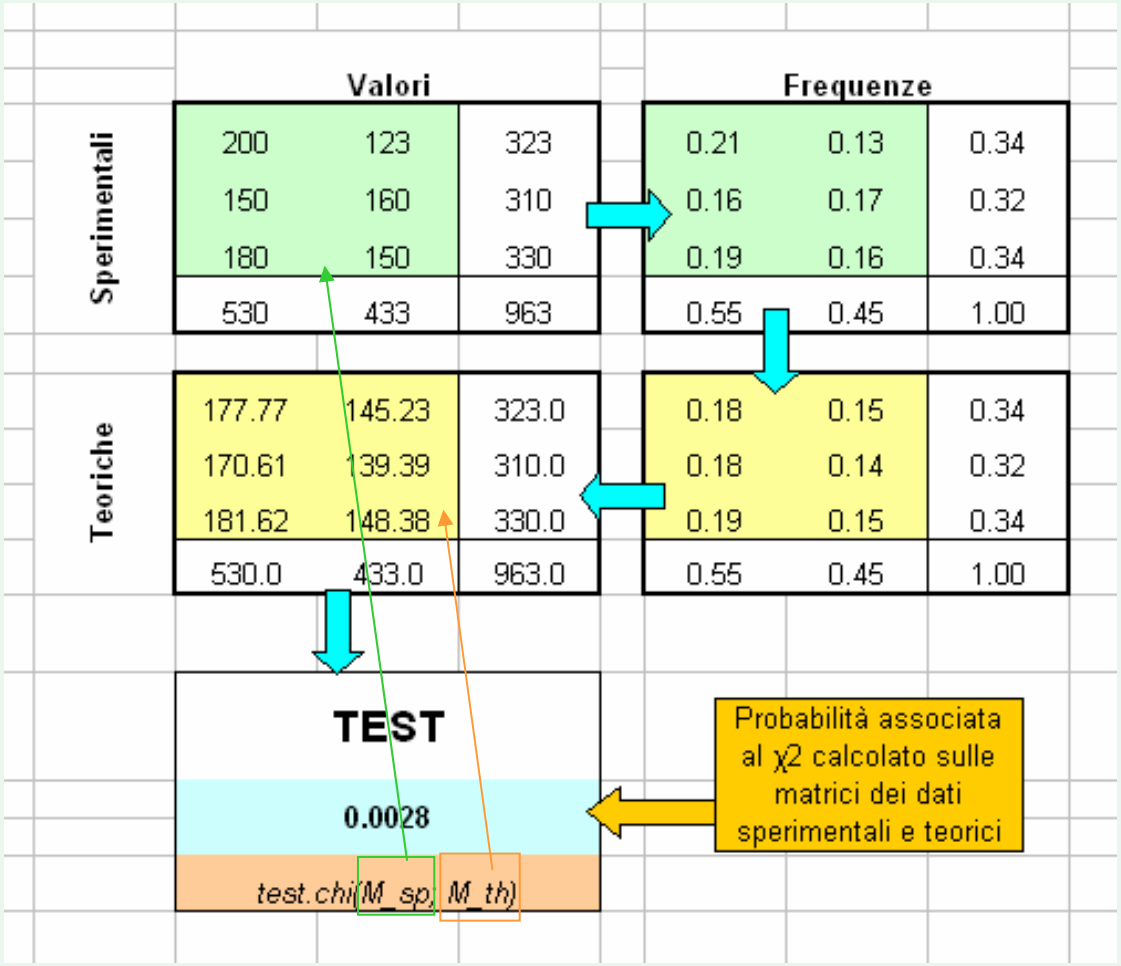
| Classe | sopravvissuti | deceduti | totali A |
|---|---|---|---|
| I | 200 | 123 | |
| II | 119 | 158 | |
| III | 181 | 528 | |
| | | | |

I decessi sono correlati con la classe?




Funzione EXCEL: `INV.CHI(`$\alpha$,$\nu$`)` restituisce il valore associato ad un livello di confidenza $\alpha$ per una variabile $\chi^2$ con $\nu$ g.d.l.

la funzione: `DISTRIB.CHI(`$\chi^2$, $\nu$`)` restituisce la probabilità associata al valore di $\chi^2$ con $\nu$ g.d.l.




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**

**Osservatorio politico**

**Pubblicato il 5/4/2007.**

**Autore:**
Ekma S.r.l.

**Committente/ Acquirente:**
Clandestinoweb

**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:**
Interviste telefoniche - Metodologia C.A.T.I.

**Numero delle persone interpellate e universo di riferimento:**
1.000 casi - Universo di riferimento: 48.483.370 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**:

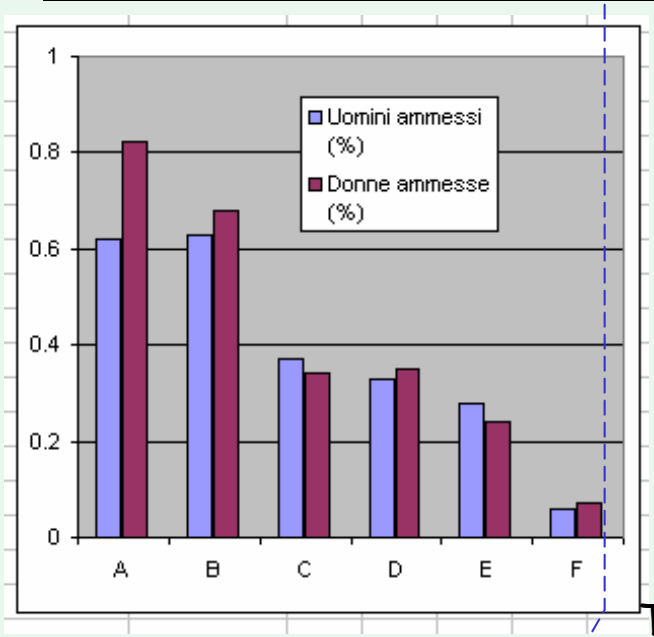
Le donne sono discriminate nei test di ammissione!

**V/F ?**

# Risultati dei Test di ammissione dettagli per dipartimento

| | Dip. A | | Dip. B | | Dip. C | | Dip. D | | Dip. E | | Dip. F | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | F | M | F | M | F | M | F | M | F | M | F |
| Amm. | 512 | 89 | 353 | 17 | 120 | 202 | 138 | 131 | 53 | 94 | 22 | 24 |
| Resp. | 313 | 19 | 207 | 8 | 205 | 391 | 279 | 244 | 138 | 299 | 351 | 317 |

| | Dip. A | | Dip. B | | Dip. C | | Dip. D | | Dip. E | | Dip. F | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | F | M | F | M | F | M | F | M | F | M | F |
| % Amm. | .62 | .82 | .63 | .68 | .37 | .34 | .33 | .35 | .27 | .24 | .06 | .07 |




Le differenza tra ammessi M o F sono significative?

No!

Test sulle frequenze

| | | | | | | |
|---|---|---|---|---|---|---|
| rischio (%) | 7.1e-3 | 77 | 40 | 60 | 49 | 69 |

Test $\chi^2$ tabelle di contingenza

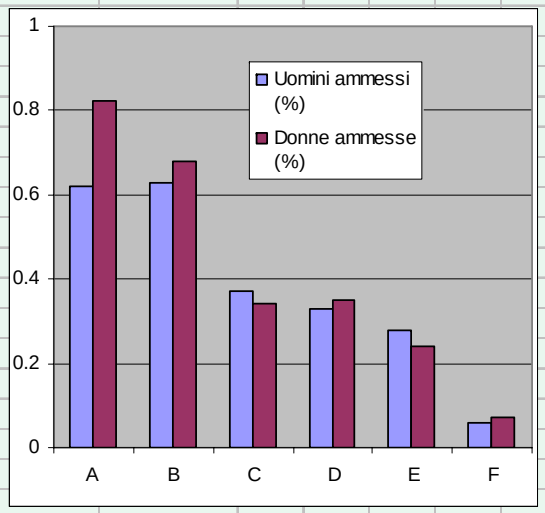
| | | | | | | |
|---|---|---|---|---|---|---|
| rischio (%) | 3.2e-3 | 61 | 39 | 58 | 31 | 53 |

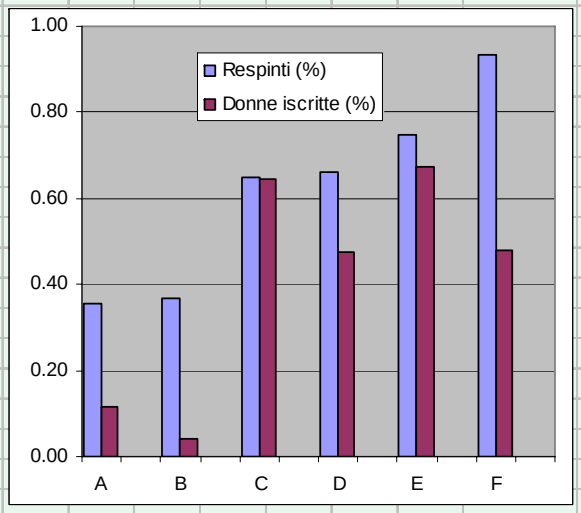
# Test: le iscrizioni ai vari dipartimenti sono casuali? La frazione di respinti per dipartimento è casuale?

## Ipotesi:

Le donne tendono a far domanda nei dipartimenti dove è maggiore la probabilità di essere respinti.

| | A | | B | | C | | D | | E | | F | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | F | M | F | M | F | M | F | M | F | M | F |
| Ammessi | 512 | 89 | 353 | 17 | 120 | 202 | 138 | 131 | 53 | 94 | 22 | 24 |
| Respinti | 313 | 19 | 207 | 8 | 205 | 391 | 279 | 244 | 138 | 299 | 351 | 317 |
| Totale (M/F) | 825 | 108 | 560 | 25 | 325 | 593 | 417 | 375 | 191 | 393 | 373 | 341 |
| Totale | 933 | | 585 | | 918 | | 792 | | 584 | | 714 | |
| % ammessi | 0.62 | 0.82 | 0.63 | 0.68 | 0.37 | 0.34 | 0.33 | 0.35 | 0.28 | 0.24 | 0.06 | 0.07 |
| % respinti | 0.36 | | 0.37 | | 0.65 | | 0.66 | | 0.75 | | 0.94 | |
| % donne | 0.12 | | 0.04 | | 0.65 | | 0.47 | | 0.67 | | 0.48 | |




Il grafico mostra qualitativamente una correlazione positiva tra frazione di respinti e numero di donne iscritte.

L'analisi qualitativa (visiva) può essere (e spesso è) fonte di errori soggettivi.

# Quantificare la correlazione

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

$F = m \cdot a$ ;  $x = v \cdot t$ ;  $S = b \cdot a$ ;  $V = I \cdot R$ ; .....

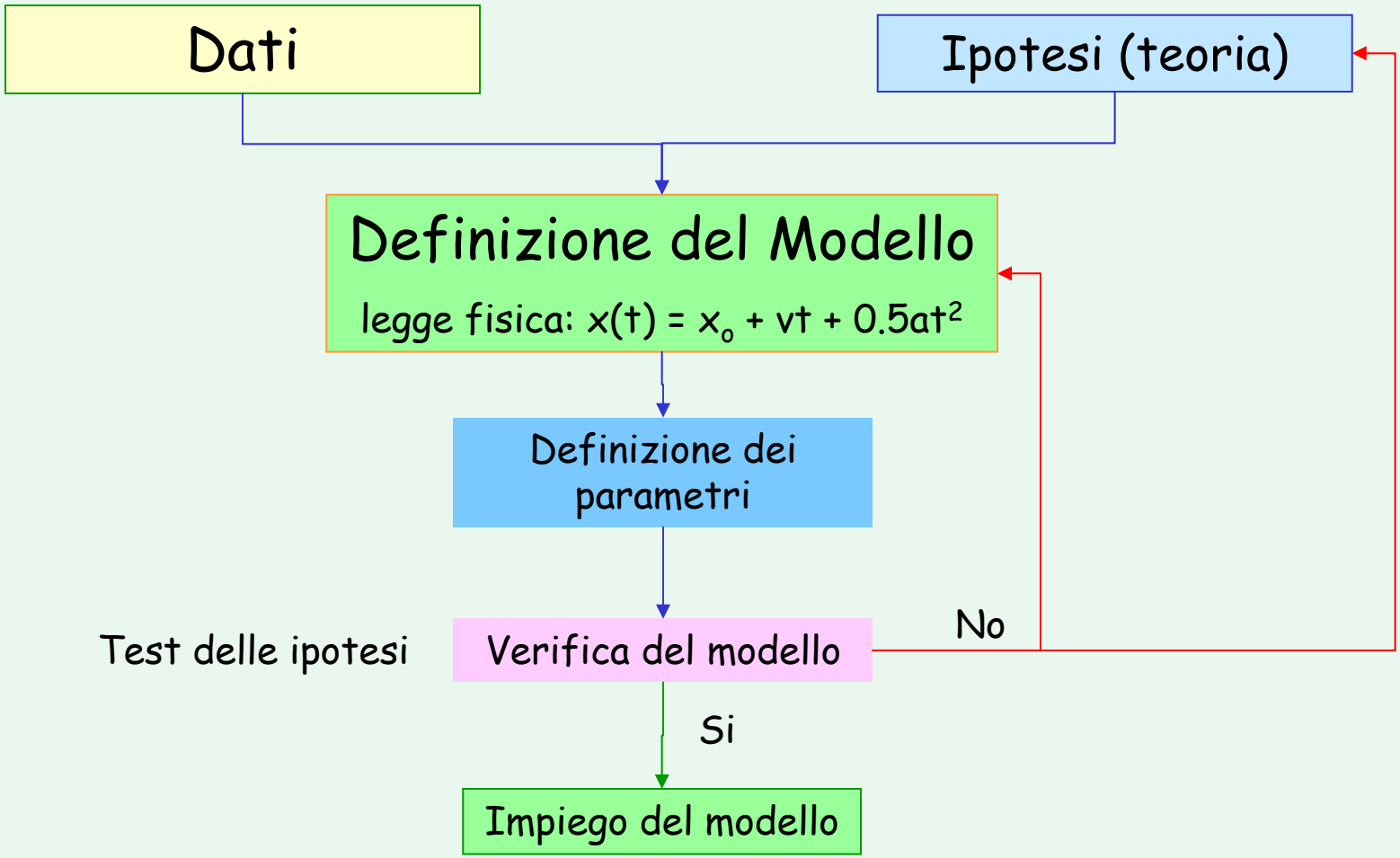
# Modelli statistici e inferenza statistica

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

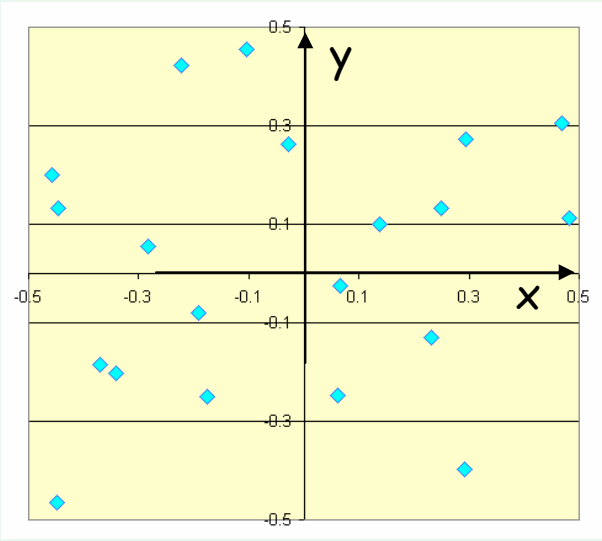
Un modello statistico è:

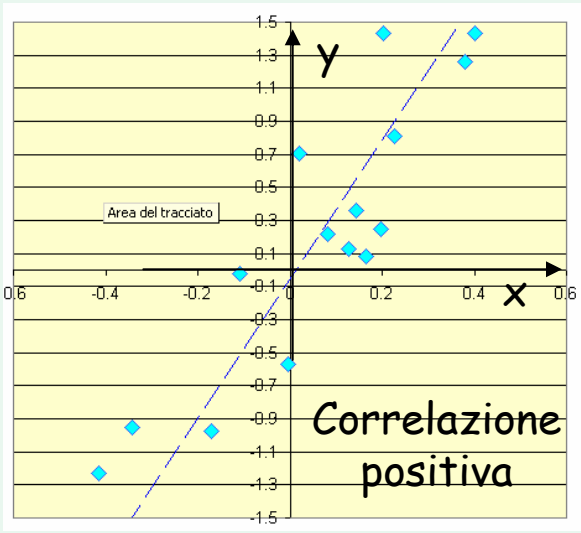
- una semplificazione della realtà (rasoio di Occam: scartare le ipotesi complesse se ne esistono di più semplici che portano allo stesso risultato)

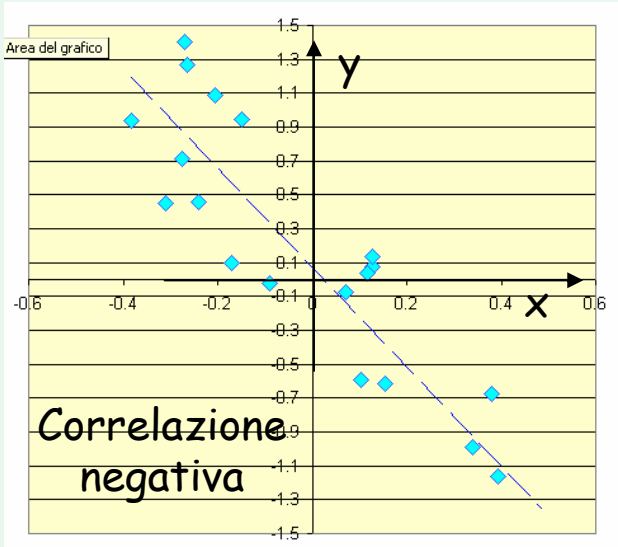
- un'analogia del fenomeno reale: il modello riproduce solo alcuni aspetti della realtà ma non è la realtà




**Regressione** (raffinamento, fitting, etc...): relazione funzionale tra variabili ottenuta con metodi statistici. Il metodo della regressione consente di derivare una relazione statistica tra una variabile dipendente (Y) e una o più variabili esplicative ($X_1$, $X_2$...$X_n$). Origine: Galton 1886 - regression toward mediocrity




Ipotesi: esiste una relazione lineare tra i valori della x e i valori della y

$$y_i = \beta_o + \beta_1 x_i + \epsilon_i$$

Errori sulle osservazioni

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

Varianza x

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})$$

Varianza y

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

covarianza xy

Retta di regressione:

$$Y = \beta_o + \beta_1 X$$

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\beta_o = \bar{y} - \beta_1 \bar{x}$$

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

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

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

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

r = coefficiente di correlazione di Pearson

$$r = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x - \bar{x})^2((y - \bar{y})^2)}}$$

Varianza degli errori

$$\sigma^2_{y_i} = \frac{1}{n}\sum \left(y_i - (\beta_o + \beta_1 x_i)\right)^2 = \frac{1}{n}\sum \left(\epsilon_i\right)^2$$

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

$$\sigma^2_{\beta_1} = \frac{\sigma^2_{y_i}}{\sum (x_i - \bar{x})^2}$$

**Nota:**

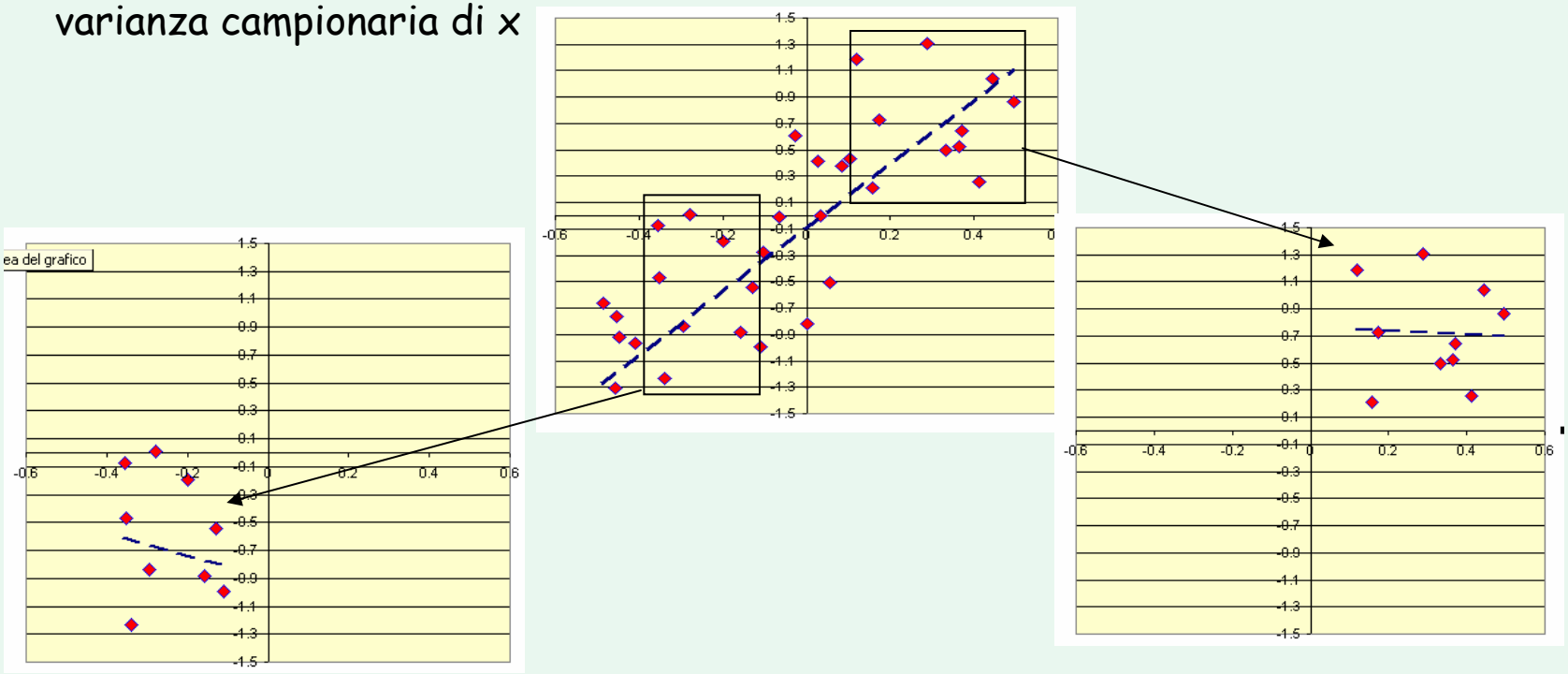
$$Corr(\beta_o, \beta_1) = -\frac{\bar{x}\sqrt{n}}{\sqrt{\sum (x_i)^2}}$$

Utilizzando

$$x' = x_i - \bar{x}$$

si annulla la correlazione e tra $\beta_o$ e $\beta_1$

L'incertezza sulle stime dei parametri della regressione decresce aumentanto la varianza campionaria di x



# Stime campionarie

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x}_i)^2 \qquad s_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y}_i)^2$$

Nell'ipotesi che la varianza sia la stessa per tutti i valori osservati $y_i$, la stima campionaria della varianza sugli errori è

$$s_{y_i}^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2 = s^2 \qquad \text{con} \quad \hat{y}_i = \beta_o + \beta_1 x_i$$

**s**: errore standard della regressione: $s = \frac{n}{n-2} s_y^2 (1 - (r_{xy})^2)$

Errori standard sui parametri

$$es(\beta_o) = \frac{s}{\sqrt{n}} \left(1 + \frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)^{\frac{1}{2}}$$

$$es(\beta_1) = \frac{s}{s_x \sqrt{n}}$$

**t-test**

$t_\nu = \beta_i / es(\beta_i)$

$\nu = n-2$

L'intervalli di confidenza

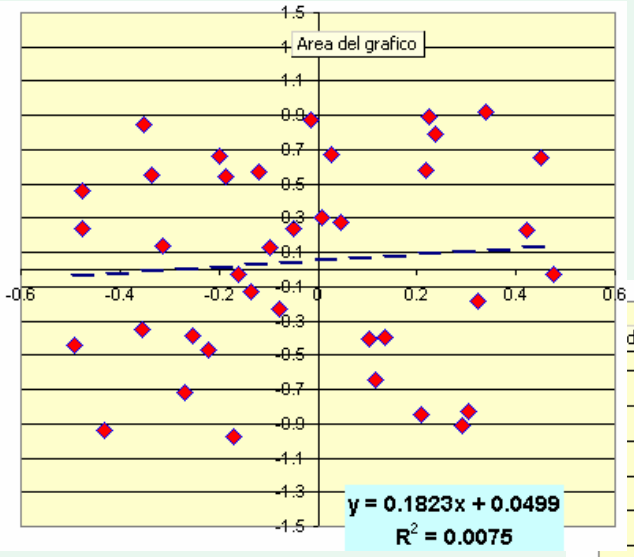
$$\beta_o - es(\beta_o) t_{(\alpha/2;n-2)} \leq \hat{\beta}_o \leq \beta_o + es(\beta_o) t_{(\alpha/2;n-2)}$$

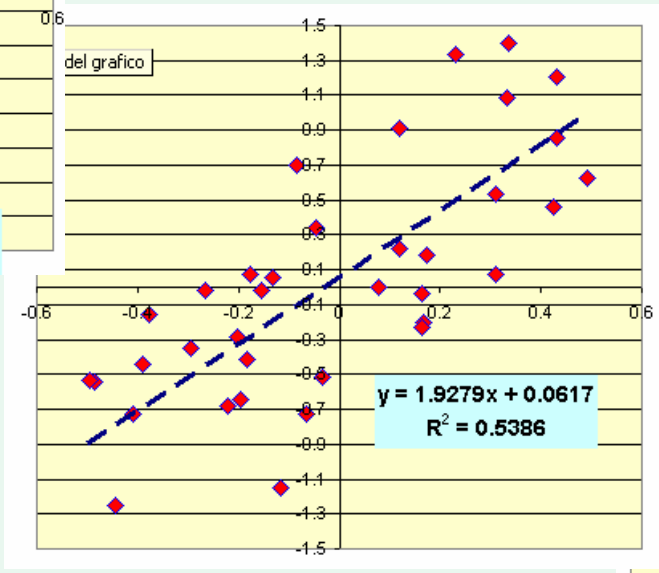
$$\beta_1 - es(\beta_1) t_{(\alpha/2;n-2)} \leq \hat{\beta}_1 \leq \beta_1 + es(\beta_1) t_{(\alpha/2;n-2)}$$

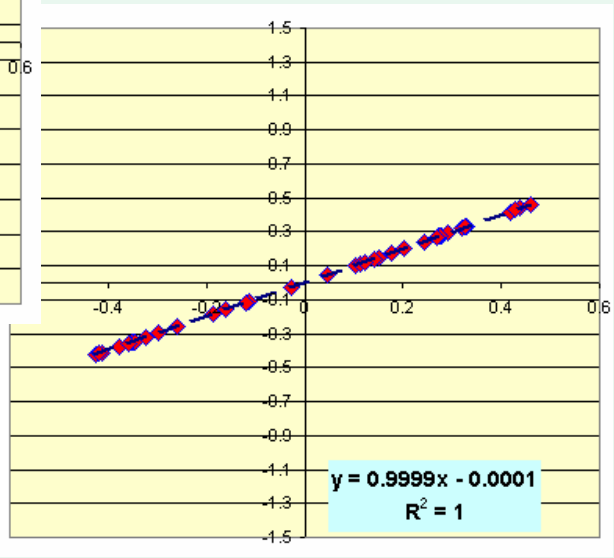
# Indice di determinazione multipla $R^2$

è un indice della "bontà della retta di regressione nello spiegare la variabilità di Y mediante X

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$




$$R^2 = (r_{xy})^2$$

nel caso di regressione lineare semplice

$$r = \frac{C_{xy}}{s_x s_y}$$

$$C_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

Covarianza

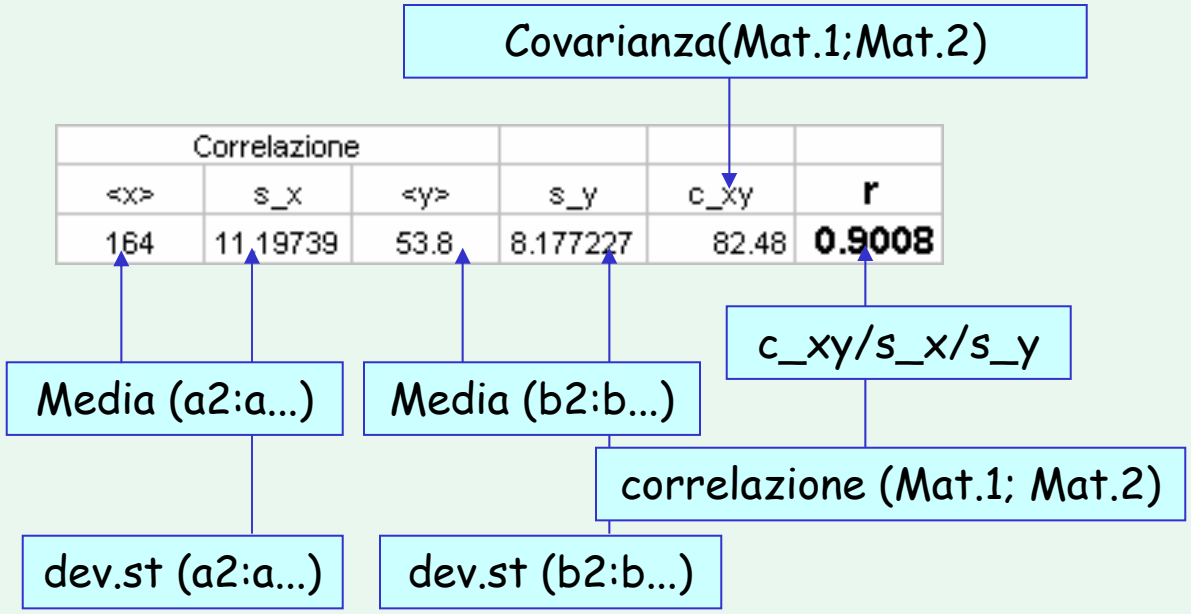
$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

Deviazione standard

# calcolare:

$$\bar{x},\ \bar{y},\ s_x,\ s_y,\ C_{xy}$$

| | A | B |
|---|---|---|
| 1 | Altezza | Peso |
| 2 | 144 | 38.2 |
| 3 | 146 | 41.0 |
| 4 | 151 | 40.6 |
| 5 | 149 | 50.3 |
| 6 | 153 | 51.2 |
| 7 | 157 | 51.0 |
| 8 | 157 | 49.4 |
| 9 | 162 | 45.0 |
| 10 | 161 | 52.8 |
| 11 | 167 | 57.2 |
| 12 | 164 | 59.0 |
| 13 | 169 | 57.3 |
| 14 | 172 | 60.2 |
| 15 | 171 | 64.2 |

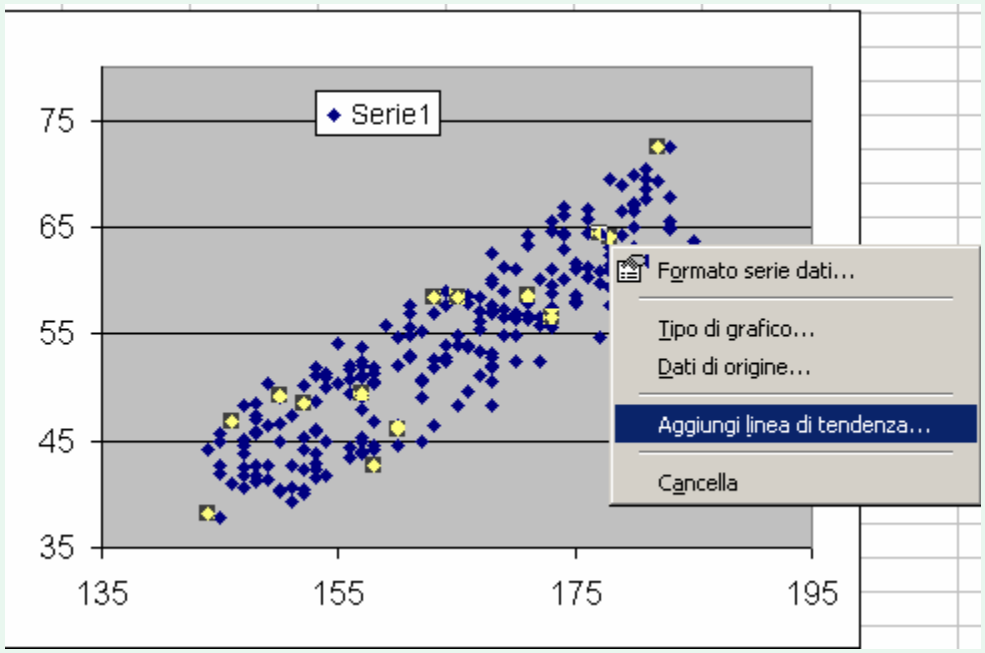


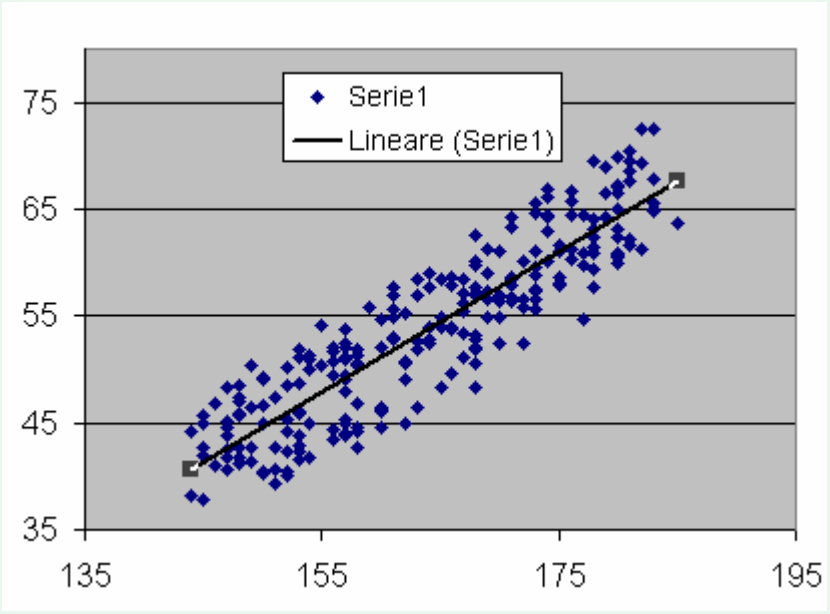
Quanto è significativa la correlazione?

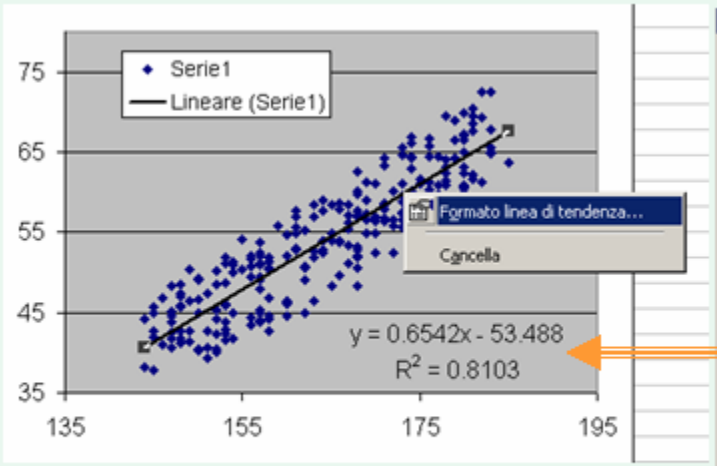
T-test sulla correlazione

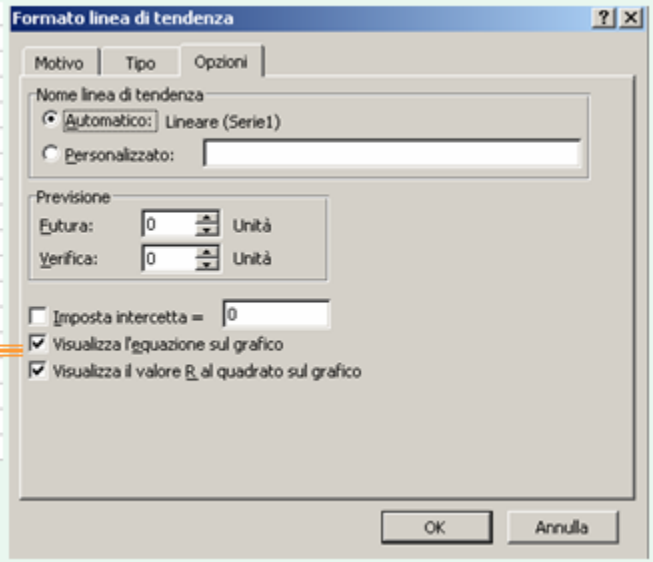
$$t_\nu = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Intervallo di confidenza per la correlazione


Serie1
75
65
55
45
35
135
155
175
195
Formato serie dati...
Tipo di grafico...
Dati di origine...
Aggiungi linea di tendenza...
Cancella

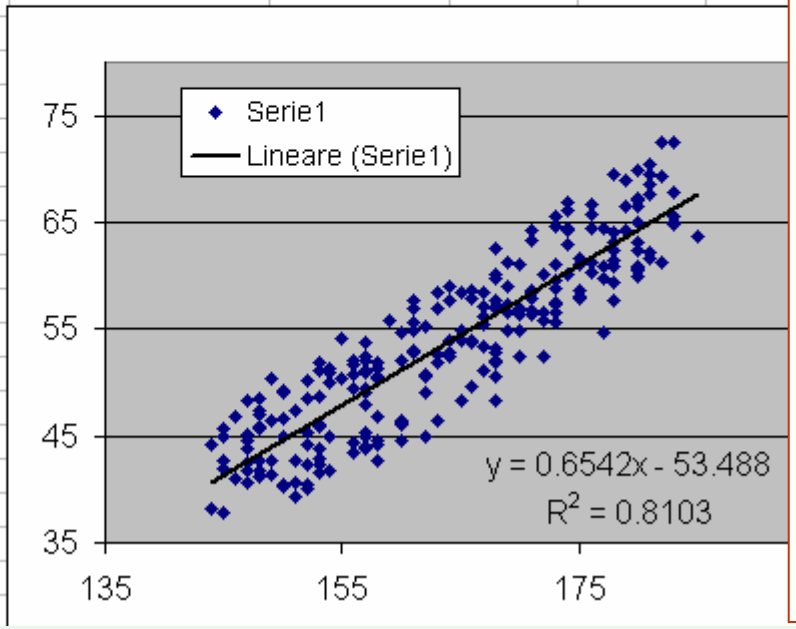

Serie1
Lineare (Serie1)
75
65
55
45
35
135
155
175
195


Serie1
Lineare (Serie1)
75
65
55
45
35
135
155
175
195
Formato linea di tendenza...
Cancella
y = 0.6542x - 53.488
R2 = 0.8103


Formato linea di tendenza
Motivo
Tipo
Opzioni
Nome linea di tendenza
Automatico:
Lineare (Serie1)
Personalizzato:
Previsione
Futura:
0
Unità
Verifica:
0
Unità
Imposta intercetta =
0
Visualizza l'equazione sul grafico
Visualizza il valore R al quadrato sul grafico
OK
Annulla

=INDICE(REGR.LIN(matr$_y$;matr$_x$;VERO;VERO);**1**;**1**)

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Altezza | Peso | | | | | | |
| 2 | 144 | 38.2 | | statistica | | | | |
| 3 | 146 | 41 | | 1;1 | 0.654 | 1;2 | -53.488 | |
| 4 | 151 | 40.6 | | 2;1 | 0.020 | 2;2 | 3.326 | |
| 5 | 149 | 50.3 | | 3;1 | 0.810 | 3;2 | 3.569 | |
| 6 | 153 | 51.2 | | 4;1 | 1046.330 | 4;2 | 245.000 | |
| 7 | 157 | 51 | | 5;1 | 13325.141 | 5;2 | 3120.106 | |
| 8 | 157 | 49.4 | | | | | | |
| 9 | 162 | 45 | | | | | | |
| 10 | 161 | 52.8 | | | | | | |
| 11 | 167 | 57.2 | | | | | | |
| 12 | 164 | 59 | | | | | | |
| 13 | 169 | 57.3 | | | | | | |
| 14 | 172 | 60.2 | | | | | | |
| 15 | 171 | 64.2 | | | | | | |
| 16 | 174 | 60.2 | | | | | | |
| 17 | 179 | 64.3 | | | | | | |
| 18 | 176 | 61.3 | | | | | | |
| 19 | 180 | 60.9 | | | | | | |
| 20 | 182 | 72.5 | | | | | | |
| 21 | 147 | 48.3 | | | | | | |
| 22 | 148 | 47.3 | | | | | | |
| 23 | 148 | 47.4 | | | | | | |
| 24 | 153 | 43.8 | | | | | | |
| 25 | 154 | 51 | | | | | | |
| 26 | 153 | 46.1 | | | | | | |




1;1 m

2;1 $s_m$

2;1 b

2;2 $s_b$

3;1 $r^2$ : **coefficiente di determinazione**

4;1 F **osservato**

5;1 **somma della regressione dei quadrati**

3;2 **errore std per la stima di y**

4;2 **gradi di libertà**

5;1 **somma residua dei quadrati**

help

**regr.lin**

| Statistica | Descrizione |
|---|---|
| s1;s2;...;sn | I valori di errore standard per i coefficienti m1;m2;...;mn |
| sb | Il valore di errore standard per la costante b (sb = #N/D quando cost è FALSO) |
| r2 | 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 sussiste 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 r2, consultare "Osservazioni" più avanti in questo argomento. |
| sy | 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. |
| gdl | 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 | $m_n$ | $m_{n-1}$ | ... | $m_2$ | $m_1$ | b |
| 2 | $se_n$ | $se_{n-1}$ | ... | $se_2$ | $se_1$ | $se_b$ |
| 3 | $r_2$ | $se_y$ | | | | |
| 4 | F | $d_f$ | | | | |
| 5 | $ss_{reg}$ | $ss_{resid}$ | | | | |