Verifica dell'ipotesi di nullità della correlazione

Lezioni di Psicometria

Giovanni Battista Flebus

Il problema dell'inferenza statistica per il coefficiente di correlazione

- Abbiamo imparato a calcolare il coefficiente di correlazione di Bravais-Pearson, che ci permette di quantificare la relazione fra due variabili.
- Tuttavia, il valore del coefficiente che possiamo ottenere da un campione di numerosità N è veramente uguale a quello della popolazione? Abbiamo imparato che la variabilità stocastica produce degli effetti sulle medie. Che effetto ha la variabilità sul coefficiente di correlazione ?
- Progettiamo un esperimento …

Un esperimento con coppie senza relazione

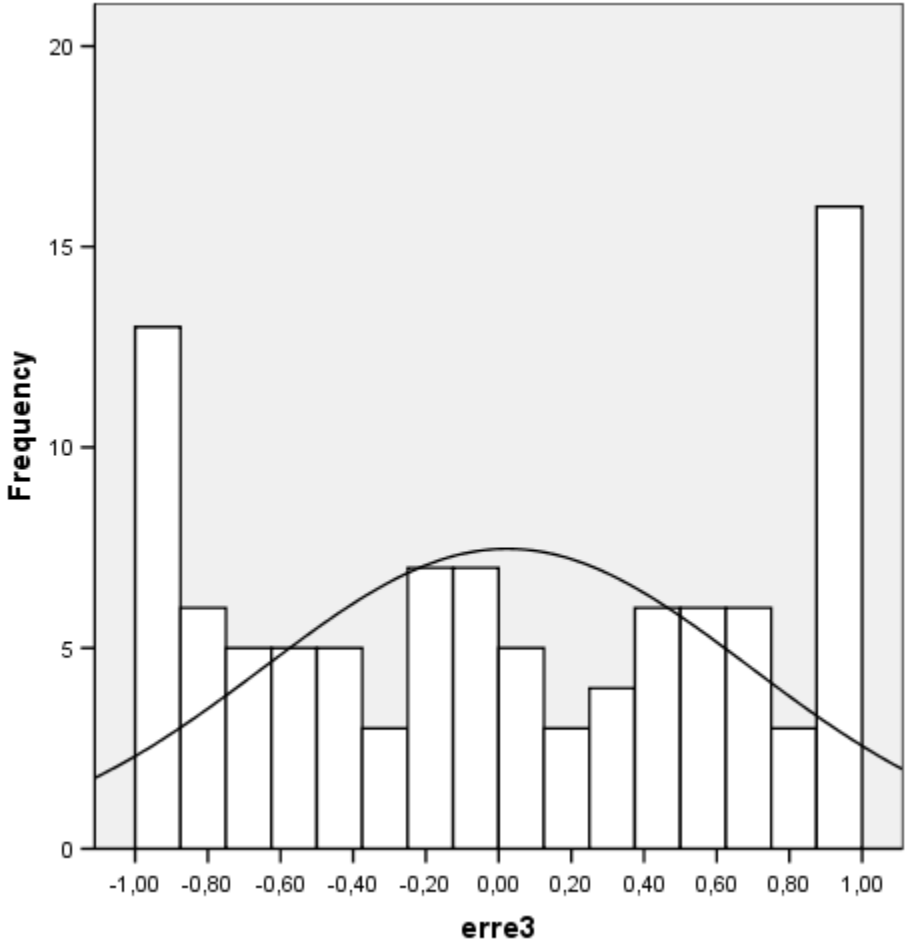
Supponiamo di chiedere a un certo numero di studenti due valori:

- Il numero di esami che hanno sostenuto
- Il numero civico in cui abitano

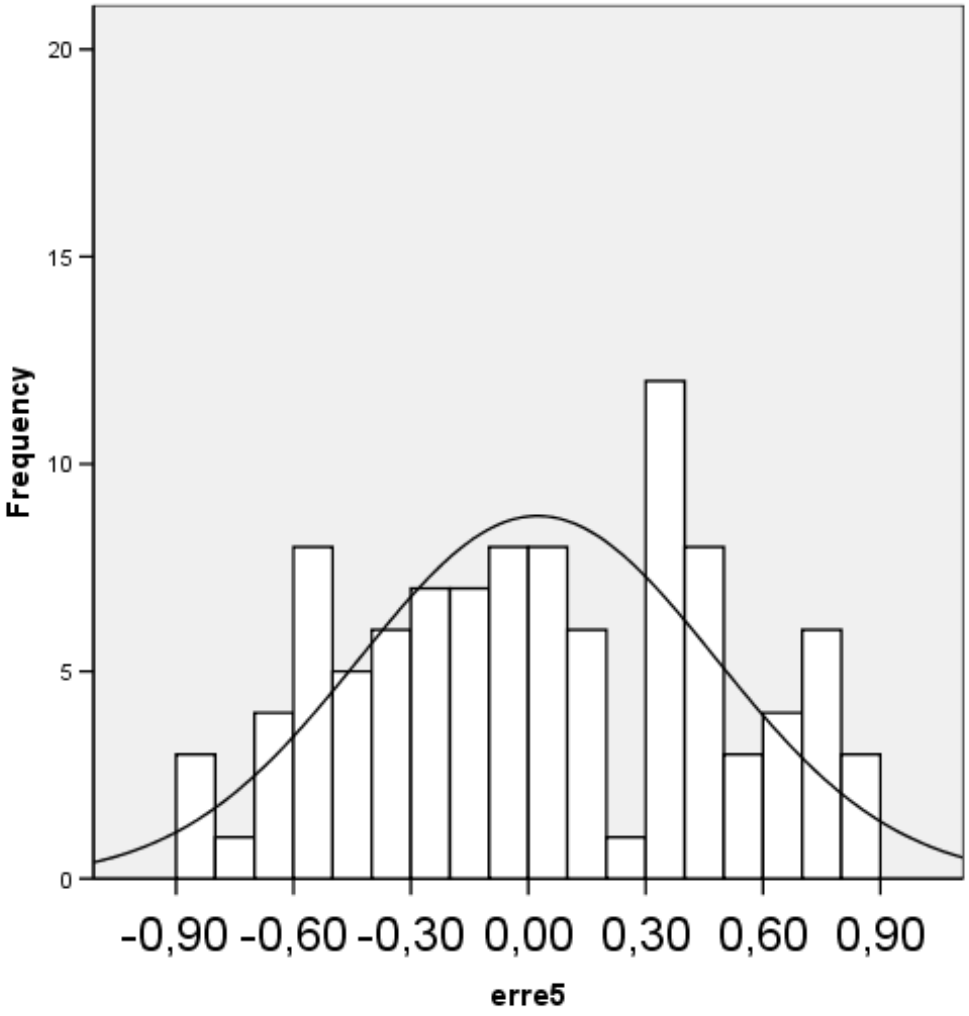
Non c'è nessuna relazione fra le due misurazioni. Se calcoliamo la correlazione in un gruppo di studenti, dovremmo trovare un coefficiente uguale a zero. Ma è sempre così? Verifichiamolo con un esperimento.

Chiederemo a un campione di tre studenti i loro dati e calcoleremo la correlazione. Non lo faremo una volta, ma cento volte. Poi presenteremo il grafico di questi 100 coefficienti. Ripeteremo l'esperimento con altri 100 campioni, ma questa volta con campioni di 5 studenti . Poi una terza volta con 10 studenti e una quarta volta con 40 studenti. Nell'ultimo esperimento quindi dovremo contattare 40 x 100= 4000 studenti. Naturalmente procederemo con una simulazione numerica. Vediamo i quattro grafici.

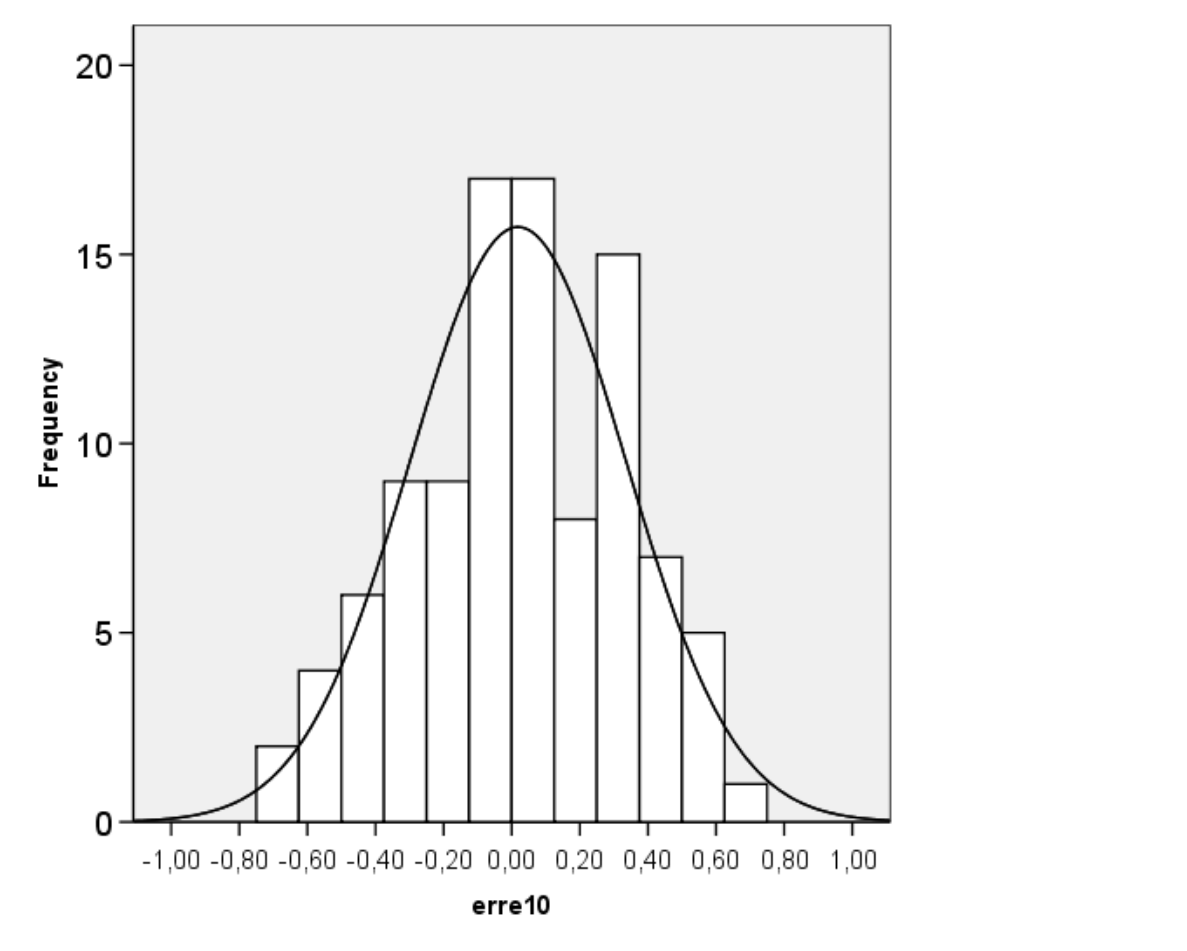
100 campioni di 3 coppie



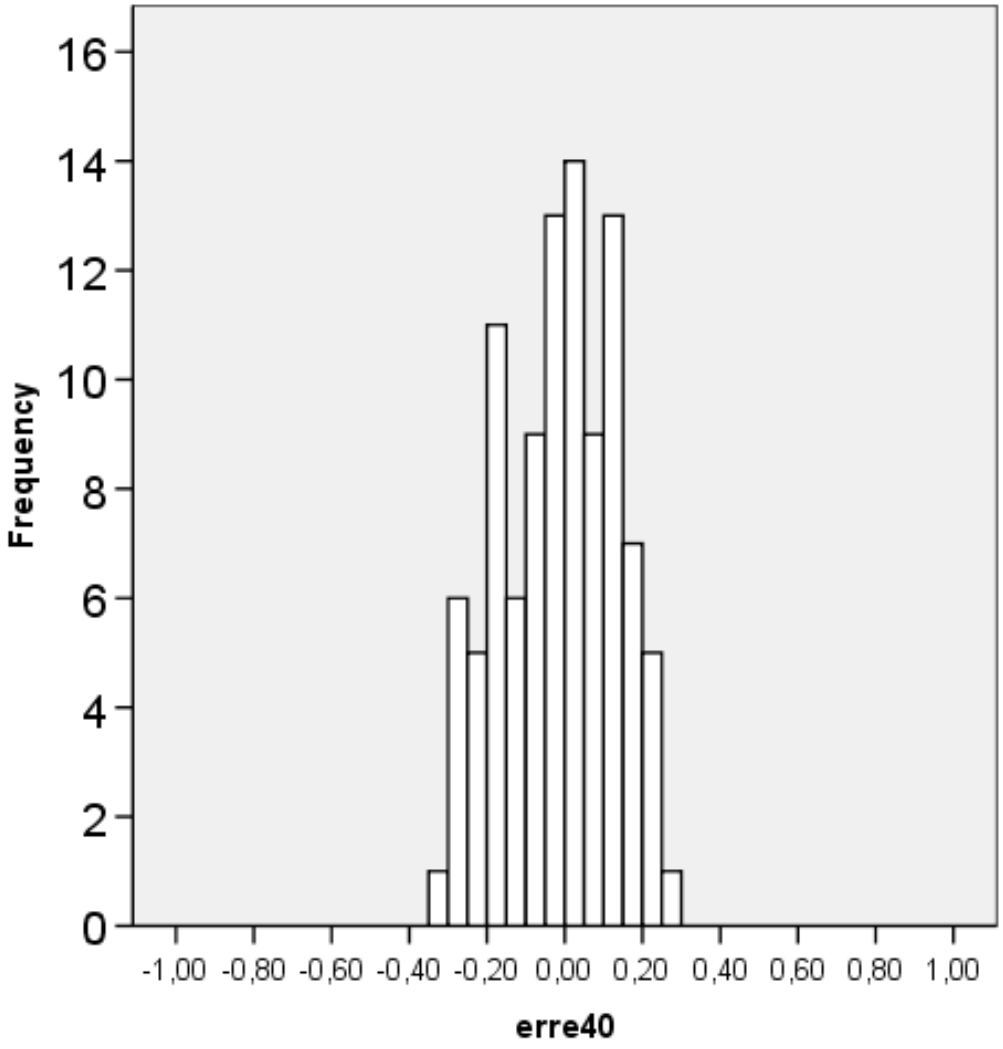
100 campioni di 5 coppie



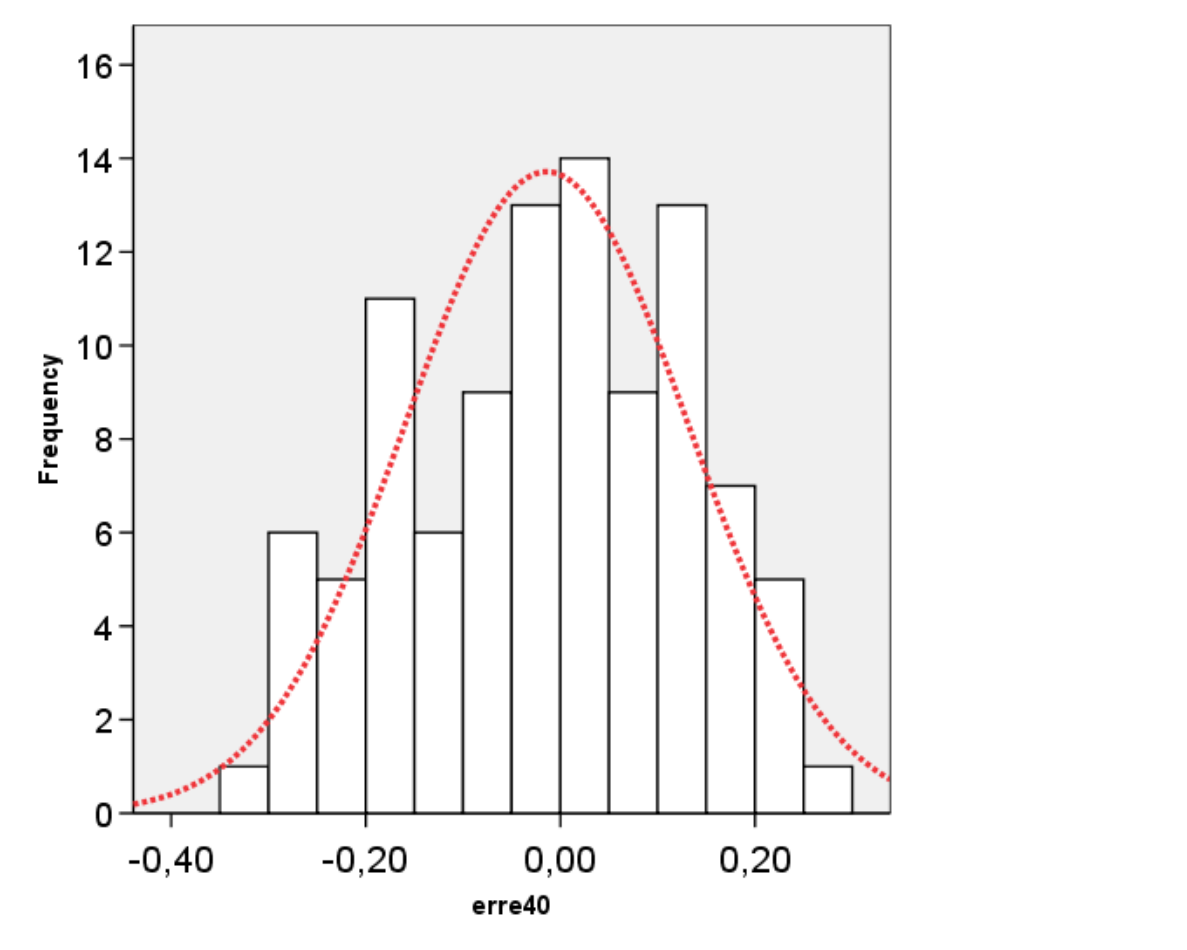
100 campioni di 10 coppie



100 campioni di 40 coppie



Ancora 40 coppie, ingrandita



Conclusioni dell'esperimento

- Anche se la correlazione reale è nulla, i coefficienti riscontrati sono **diversi** o anche **molto diversi** da 0
- Anche in totale assenza di correlazione, i coefficienti calcolati su piccoli campioni presentano una grandissima variabilità

STATISTICHE DESCRITTIVE DEI QUATTRO CAMPIONI (N=100)

	Minimum	Maximum	Mean	Std. Deviation
3 coppie	-1,00	1,00	,0235	,66727
5 coppie	-,90	,83	,0238	,45603
10 coppie	-,74	,69	,0183	,31711
40 coppie	-,31	,27	-,0141	,14544

Verifica dell'ipotesi di correlazione non nulla

- Come si stabilisce allora che un coefficiente di correlazione rappresenta veramente una relazione e non è invece il prodotto della variabilità stocastica?

- Come si stabilisce un intervallo di fiducia per un coefficiente di correlazione per escludere di aver trovato una relazione che invece appare solo come frutto di casualità?

Verifica dell'ipotesi di correlazione non nulla

Qual è la realtà?

$$H_0 : \rho_{xy} = 0$$

$$H_1 : \rho_{xy} \neq 0$$

Diversi modi per fare la verifica.
Questo usa il *t* di Student:

Distribuzione campionaria di r

- Se r =0, la distribuzione campionaria di r può essere descritta da un t di Student con N-2 gradi di libertà e si calcola con la formula

$$t = \frac{r_{xy} \cdot \sqrt{N-2}}{\sqrt{1-r^2_{xy}}}$$

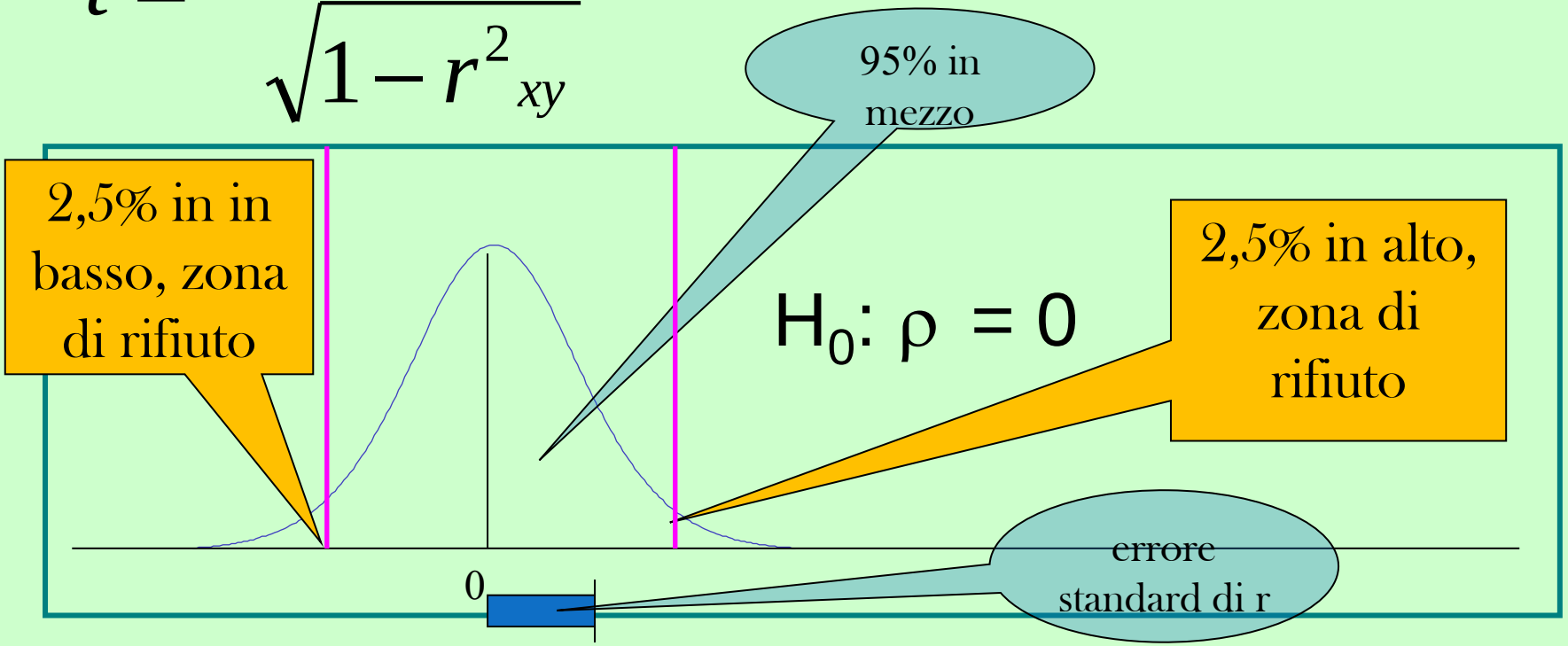


Tabella dei valori critici per p=0,05

$$P(t_{n-1} \geq t_{\alpha/2}) = \alpha/2$$

$$\text{ovvero } P(|t_{n-1}| \leq t_{\alpha/2}) = \alpha$$

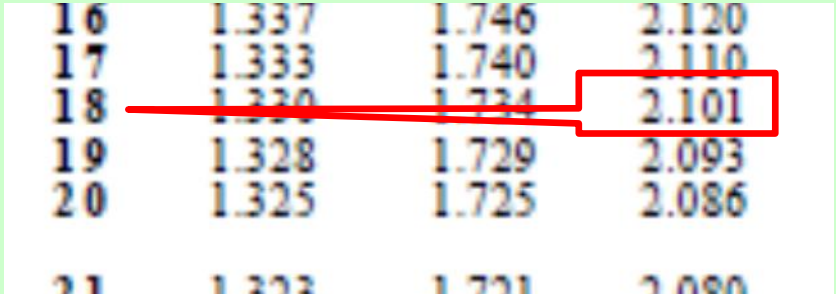
n	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571
6	1.440	1.943	2.447
7	1.415	1.895	2.365
8	1.397	1.860	2.306
9	1.383	1.833	2.262
10	1.372	1.812	2.228
11	1.363	1.796	2.201
12	1.356	1.782	2.179
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120
17	1.333	1.740	2.110
18	1.330	1.734	2.101
19	1.328	1.729	2.093
20	1.325	1.725	2.086
21	1.323	1.721	2.080
22	1.321	1.717	2.074
23	1.319	1.714	2.069
24	1.318	1.711	2.064
25	1.316	1.708	2.060
26	1.315	1.706	20.56
27	1.314	1.703	2.052
28	1.313	1.701	2.048
29	1.311	1.699	2.045
30	1.310	1.697	2.042
40	1.303	1.684	2.021
60	1.296	1.671	2.000
120	1.289	1.658	1.980
∞	1.282	1.645	1.960

Primo esempio

- Il coefficiente di correlazione fra abilità di calcolo e interessi matematici in un gruppo di 20 studenti è risultato pari a 0,30.
- Si può affermare che esiste una correlazione nella popolazione?
- Applicando la formula

$$t = \frac{r_{xy} \cdot \sqrt{N-2}}{\sqrt{1-r^2_{xy}}} \qquad \frac{0{,}30 \cdot \sqrt{18}}{\sqrt{1-0{,}09}} = 1{,}33$$

- Il valore 1,33 è un valore molto comune, al di sotto del valore critico (2, 10 con 18 gl. Anche se esiste una relazione moderata, il valore ottenuto è anche probabile senza che vi sia correlazione. Si decide perciò che non si può escludere l'ipotesi nulla:



16	1.337	1.746	2.120
17	1.333	1.740	2.110
18	1.330	1.734	2.101
19	1.328	1.729	2.093
20	1.325	1.725	2.086
21	1.323	1.721	2.080

Secondo esempio

- Se lo stesso coefficiente fosse stato calcolato su 60 studenti, si sarebbe arrivati alla stessa conclusione?
- Applicando la formula

$$t = \frac{r_{xy} \cdot \sqrt{N-2}}{\sqrt{1-r^2_{xy}}}$$

$$\frac{0{,}30 \cdot \sqrt{58}}{\sqrt{1-0{,}09}} = 2{,}39$$

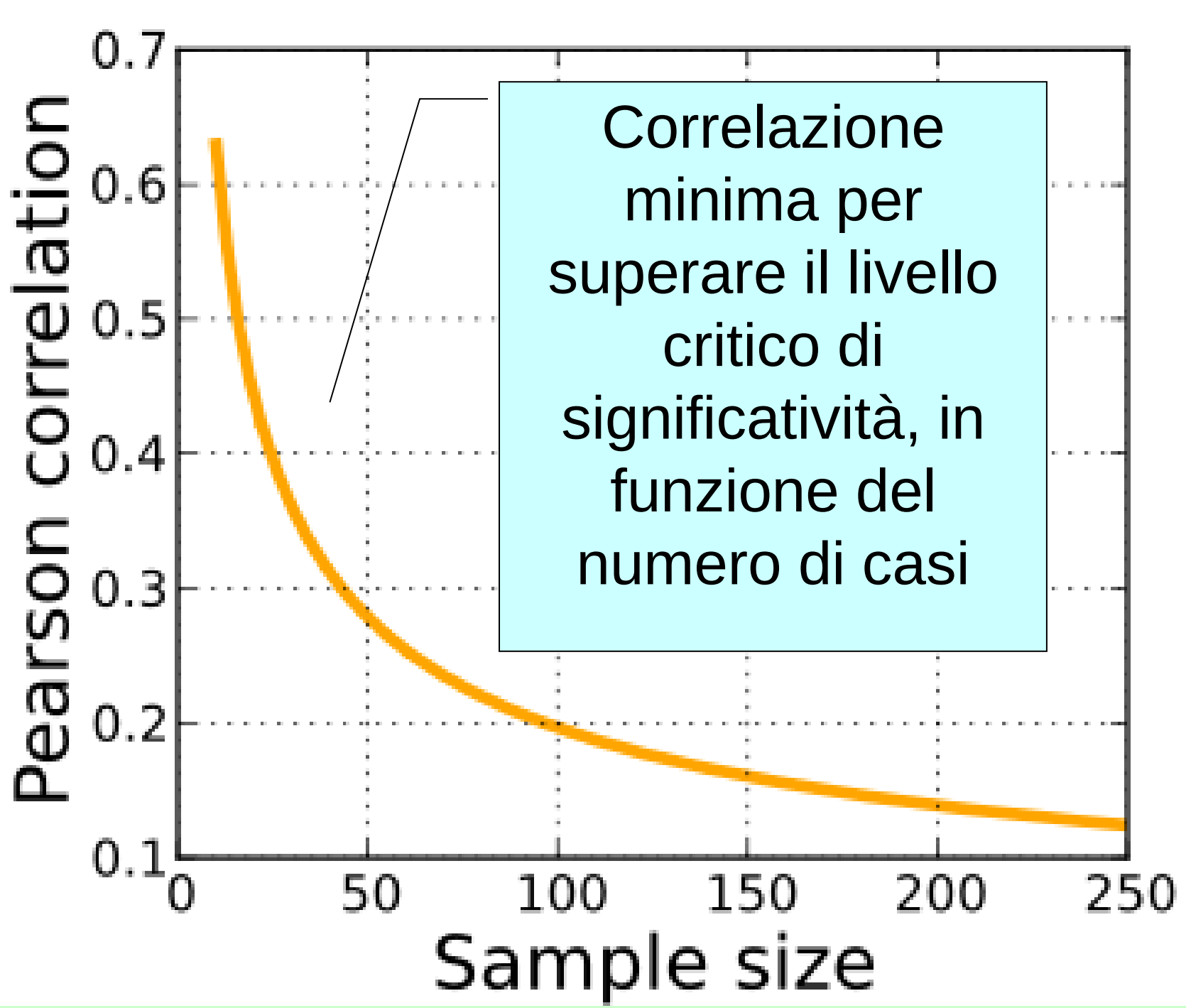
Il valore 2,39 è un valore molto raro, al di sopra del valore critico. Se non ci fosse correlazione nella popolazione, tale risultato avrebbe ben poche probabilità di comparire. Questa volta si decide di escludere la possibilità che la correlazione nella popolazione sia nulla.

30	1.310	1.697	2.042
40	1.303	1.684	2.021
60	[illegible]	[illegible]	2.000
120	1.289	1.658	1.980
∞	1.282	1.645	1.960

Tabella per rilevare r significativo (ipotesi bidirezionale, p=0,05)

R_{xy} *è significativamente* $\neq 0$ *se* $|R_{xy}| > r_{2.5\%}$

n	$r_{2.5\%}$
3	0.997
4	0.95
5	0.88
6	0.81
7	0.75
8	0.71
9	0.67
10	0.63
11	0.60
12	0.58
13	0.55
14	0.53
15	0.51
16	0.50
17	0.48
18	0.47
19	0.46
20	0.44
21	0.43
22	0.42
23	0.41
24	0.40
25	0.40
26	0.39
27	0.38
28	0.37
29	0.37
30	0.36
60	0.25
120	0.18

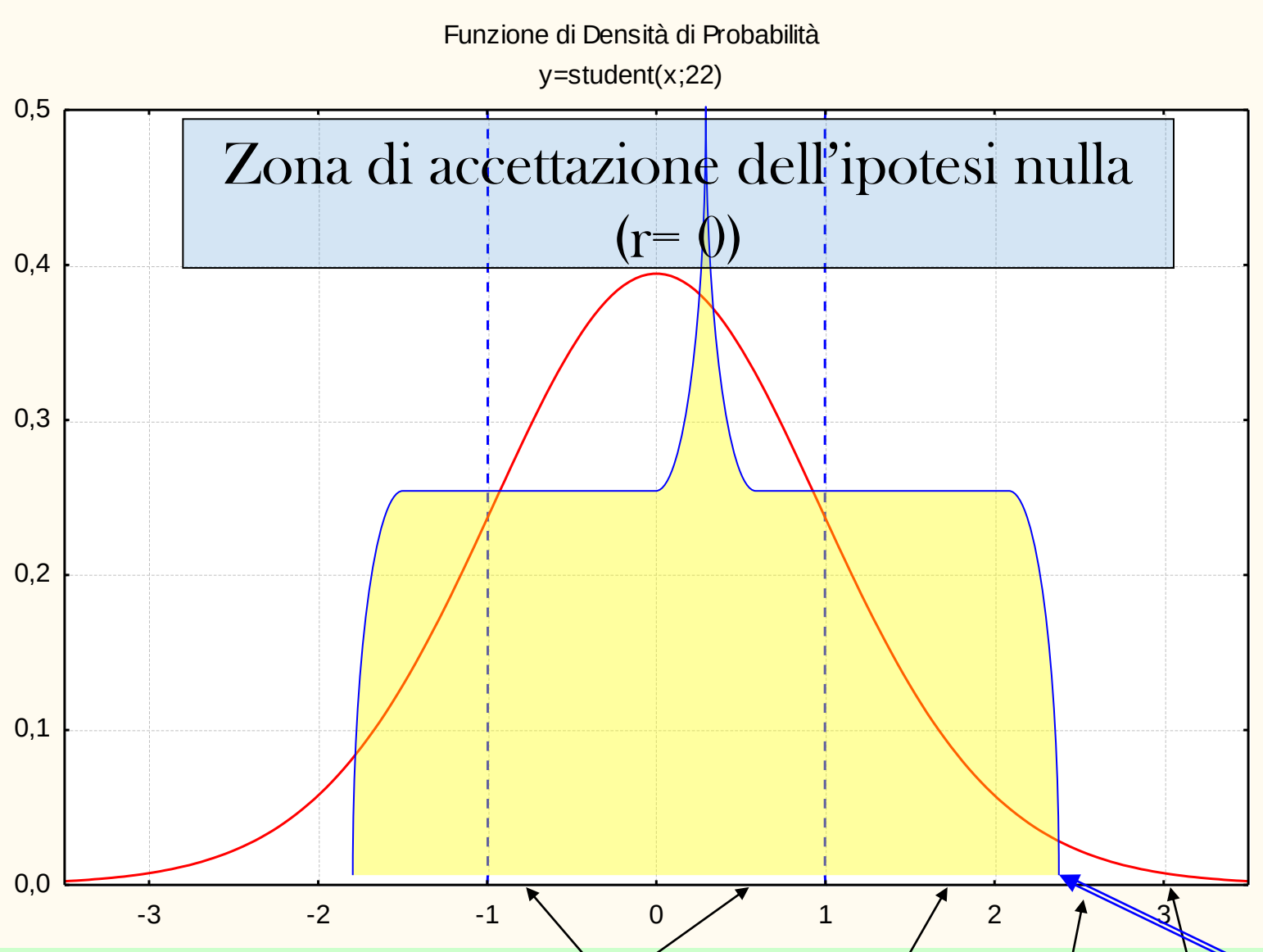

Correlazione minima per superare il livello critico di significatività, in funzione del numero di casi
Pearson correlation
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
Sample size

Otteniamo un coefficiente di 0,40 in un campione di 12 , di 24, di 48, di 96 coppie. In quali casi esiste una correlazione nella popolazione?

SIGNIFICATIVITÀ DI ALCUNI R SECONDO LA NUMEROSITÀ DEL CAMPIONE

	Valore di r						
N	-0,1	0,1	-0,2	0,3	0,4	0,5	t crit 0,05
12	-0,348	0,348	-0,707	1,089	1,512	2	2,228
24	-0,492	0,492	-1,000	1,541	2,138	2,828	2,074
48	-0,696	0,696	-1,414	2,179	3,024	4	2,021
96	-0,985	0,985	-2,000	3,081	4,276	5,657	1,990
192	-1,393	1,393	-2,828	4,358	6,047	8	1,960

Per stabilire la significatività di r utilizzo il test t. Nell'area in rosso sono evidenziati i valori t superiori al t critico.



GL	-0,1	0,1	-0,2	0,3	0,4	0,5	t crit 0,05
24	-0,492	0,492	-1,000	1,541	2,138	2,828	2,074

La probabilità che indica significatività del coefficiente di correlazione…

- È la probabilità di ottenere quel valore se non c'è correlazione nella popolazione
- Non è la probabilità di ottenere un'altra volta quel valore
- Non è la probabilità di ottenere esattamente quel valore

Se la probabilità è bassa allora …

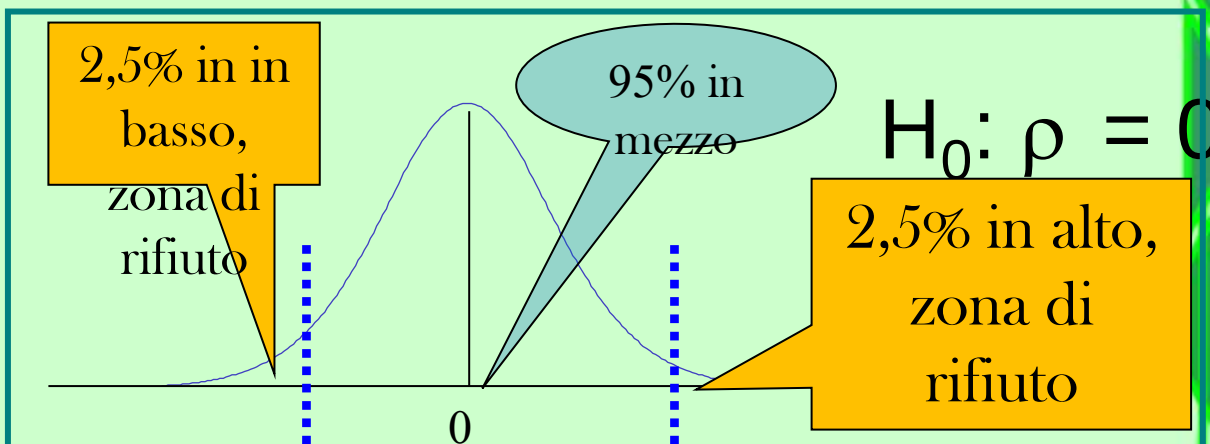
- Il valore ottenuto è molto infrequente (compare meno di una volta su 20)

Come si spiega questo valore raro? Con il ragionamento seguente:

- Non è vero che abbiamo ottenuto un valore infrequente con una correlazione nulla nella popolazione. Abbiamo ottenuto un valore (presumibilmente) comune perché la correlazione della popolazione è diversa da zero.

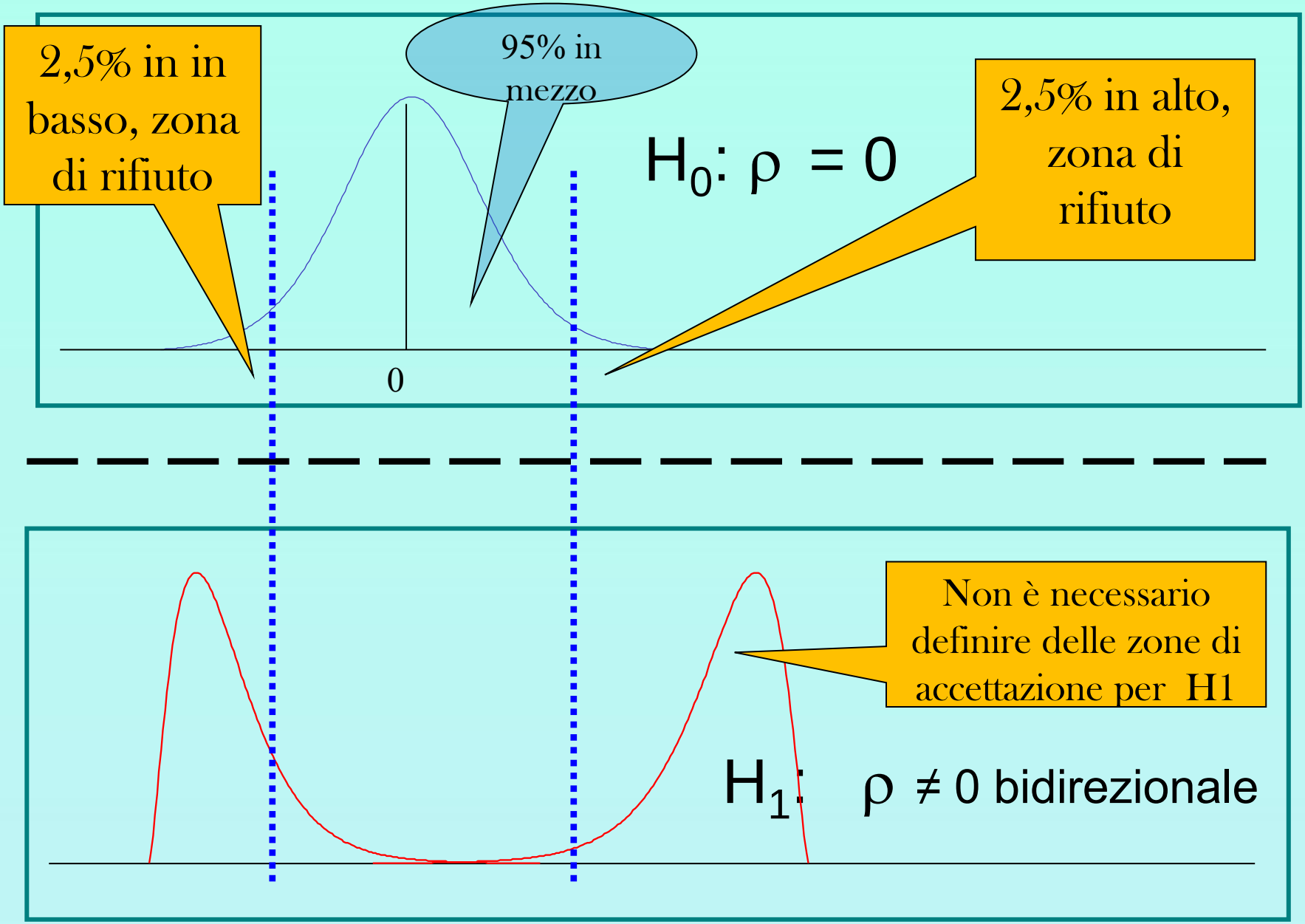
Inferenza statistica sul coefficiente di correlazione

- Si estrae un campione di numerosità N
- si calcola il coefficiente di correlazione
- Si definiscono le due ipotesi (nulla e alternativa)
- si individua la distribuzione teorica dell' r campionario
- si stabilisce il valore critico di r
- si opera il confronto
- si traggono le conclusioni.



Le due ipotesi

- Secondo l'ipotesi nulla, il coefficiente di correlazione è zero, i valori comuni sono quelli attorno allo zero. Valori rari sono molto lontani da zero.
- L'ipotesi alternativa prevede che r sia diverso da zero, quindi elevato, e che campioni di r elevato siano comuni

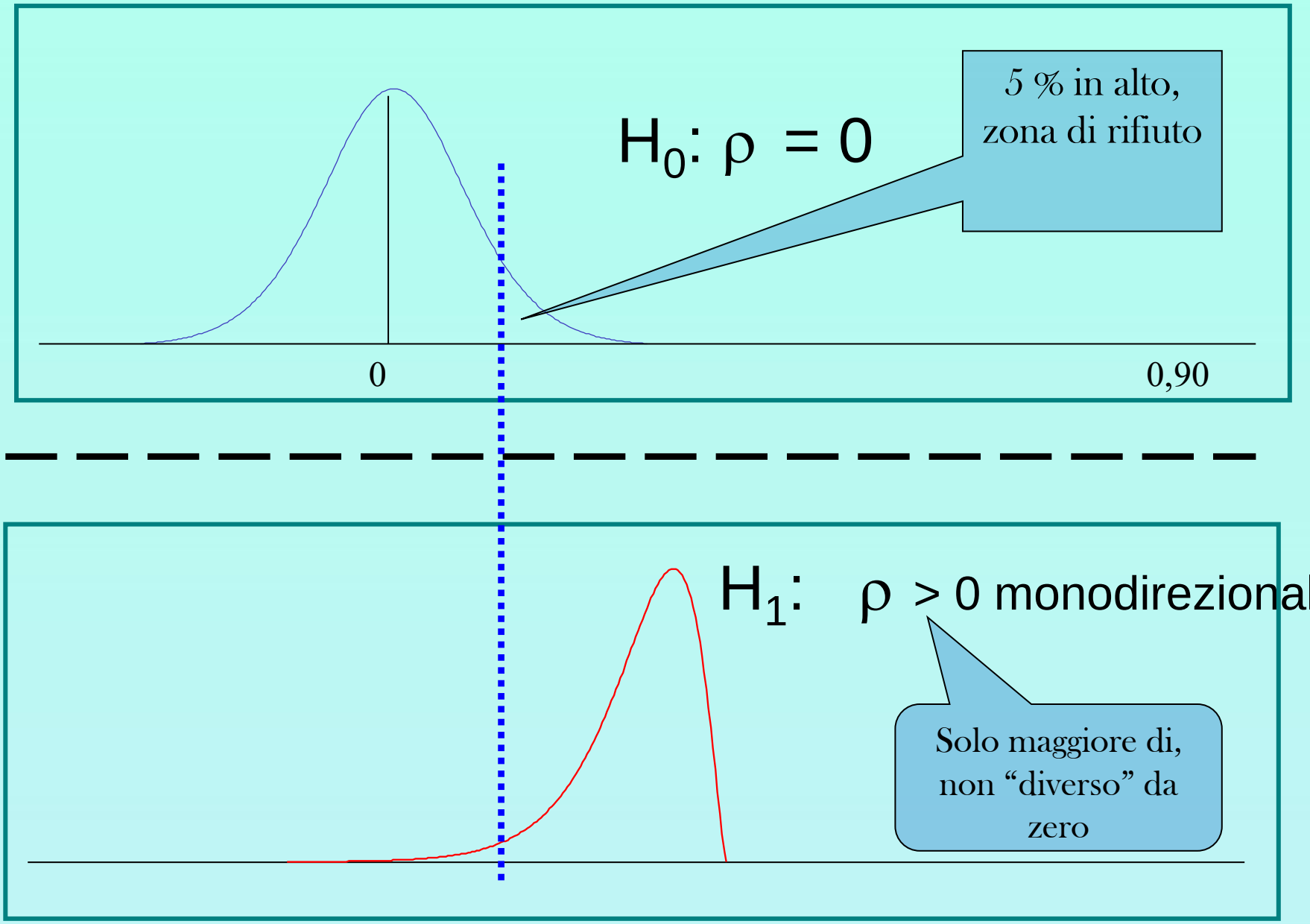

2,5% in in basso, zona di rifiuto
95% in mezzo
H_0: ρ = 0
2,5% in alto, zona di rifiuto
0
Non è necessario definire delle zone di accettazione per H1
H_1: ρ ≠ 0 bidirezionale

Ipotesi a una o due direzioni ?

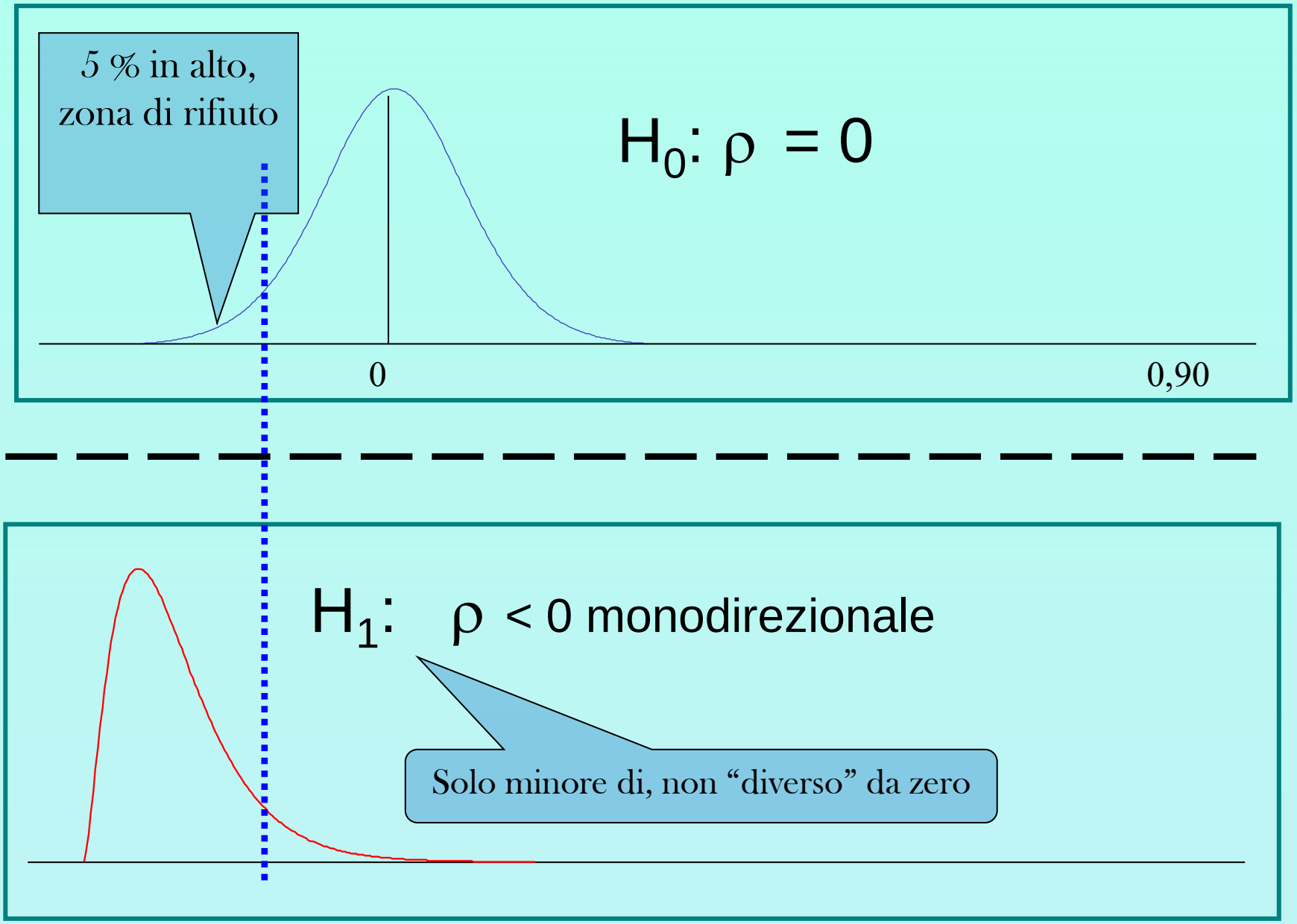
- Se abbiamo qualche motivo di ipotizzare che il coefficiente può essere solo positivo (ma non negativo), formuliamo l'ipotesi alternativa in questo modo:
- *H1: Il coefficiente di correlazione è elevato e positivo nella popolazione*
- In tal caso si parla di ipotesi monodirezionale (o unidirezionale) o ipotesi a una coda: il 5 % si trova solo su una coda. Questo significa che il valore critico è più vicino allo zero, e quindi si può accettare l'ipotesi alternativa con coefficienti più bassi. Ma dobbiamo avere buoni motivi per supporre che il coefficiente possa essere solo positivo e non negativo .

Ipotesi a una o due direzioni

- Supponendo che il valore di alfa (quel valore di probabilità che definisce il valore critico) sia pari a 0,05…
- Nell'ipotesi bidirezionale è diviso nella parte bassa (0,025 in basso) e nella parte alta della distribuzione (0,025 in alto)
- Nell'ipotesi unidirezionale invece si trova solo su una coda (o sinistra o destra).


H_0: ρ = 0
5 % in alto, zona di rifiuto
0
0,90
H_1: ρ > 0 monodirezionale
Solo maggiore di, non "diverso" da zero

- L’ipotesi unidirezionale si può formulare anche con segno algebrico opposto (se ci aspettiamo che la correlazione possa essere solo negativa)


5 % in alto,
zona di rifiuto
H_0: ρ = 0
0
0,90
H_1: ρ < 0 monodirezionale
Solo minore di, non "diverso" da zero

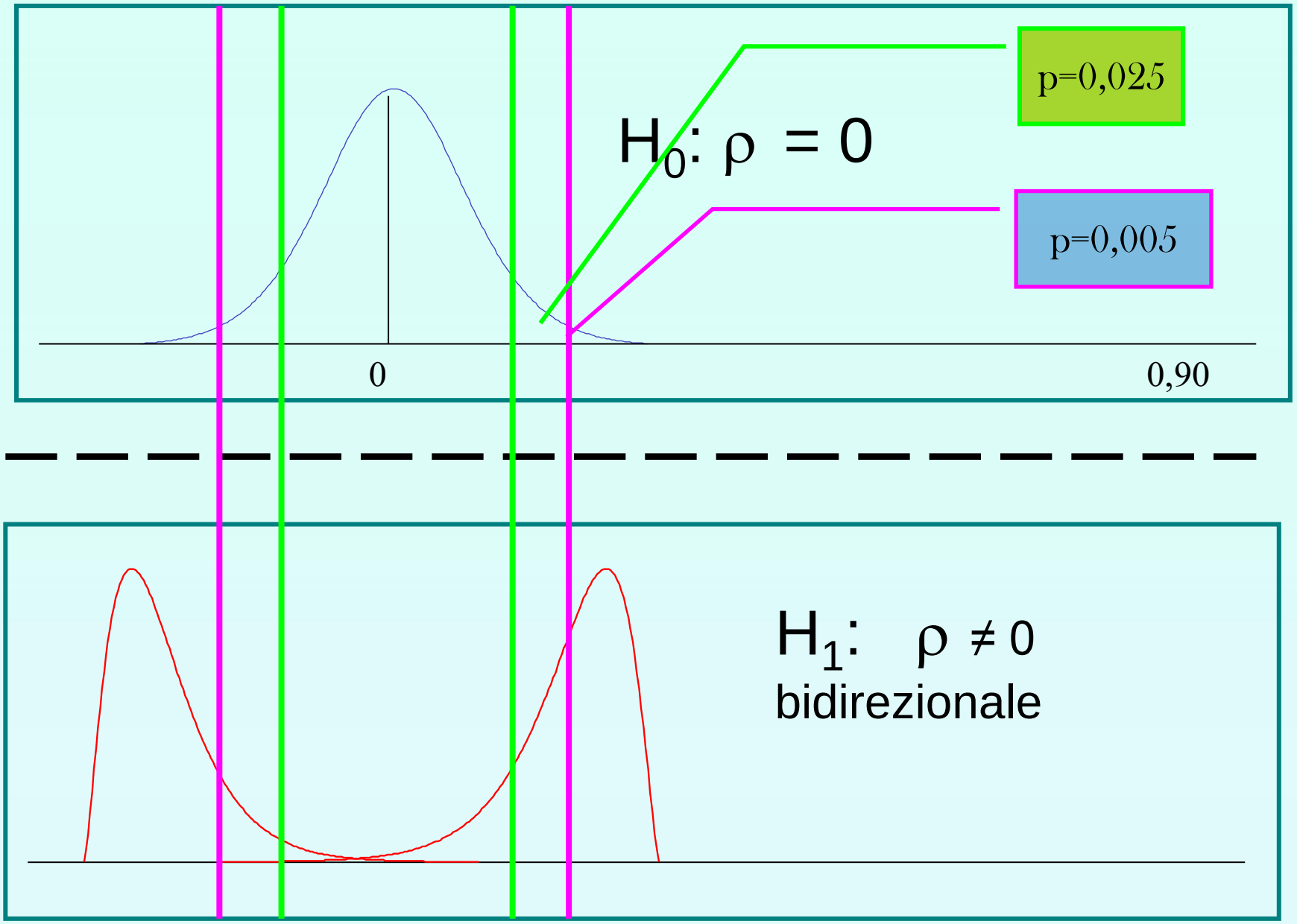
Potenza di un test statistico

Che altro significa il valore p = 0,05?

- E' pur sempre possibile che nella popolazione il coefficiente di correlazione sia realmente nullo e che realmente il campione estratto sia molto raro.
- Il valore 0,05 è la probabilità di rifiutare l'ipotesi nulla mentre essa è vera: si afferma che esiste correlazione, mentre in realtà non esiste.

Si può diminuire p da 0,05 a 0,001 per essere più sicuri?

- Si può farlo, ma in tal caso aumentiamo il rischio inverso, quello di rifiutare l'ipotesi alternativa quando essa è vera.

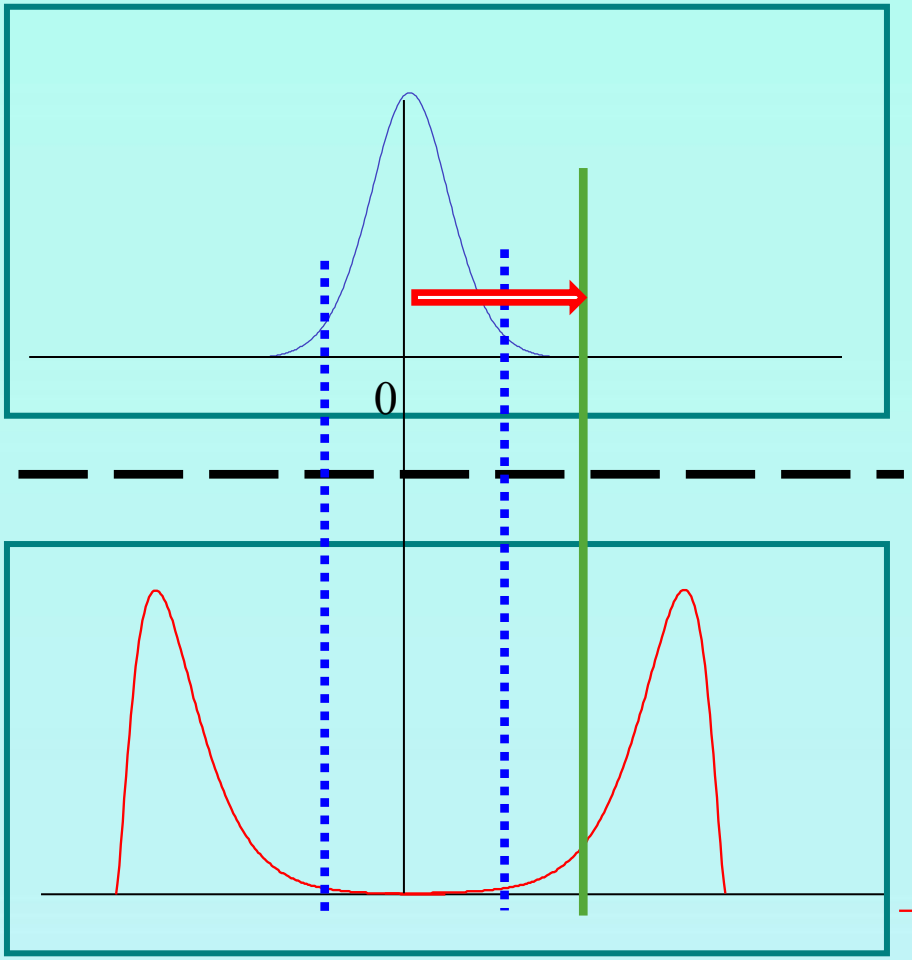

p=0,025
H_0: ρ = 0
p=0,005
0
0,90
H_1: ρ ≠ 0
bidirezionale

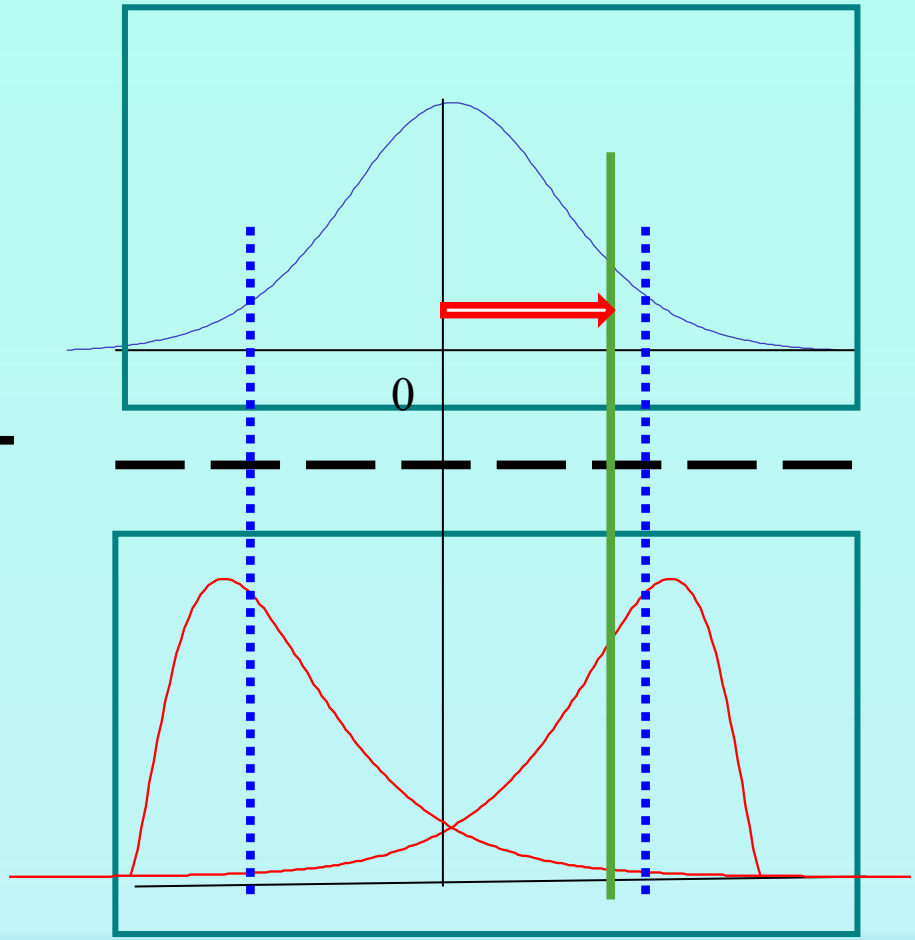
Si possono diminuire entrambi i rischi? Come?

- Si possono diminuire entrambi i rischi aumentando la numerosità di N, facendo in modo di diminuire la sovrapposizione dei valori delle due ipotetiche popolazioni.
- Questo è il modo migliore per aumentare la potenza di un test (ossia la capacità di un test statistico di individuare il valore reale che si vuole stimare.

N elevato, valori concentrati

N basso, valori dispersi





La correlazione con SPSS …

- Menu Analizza

poi Correlazione

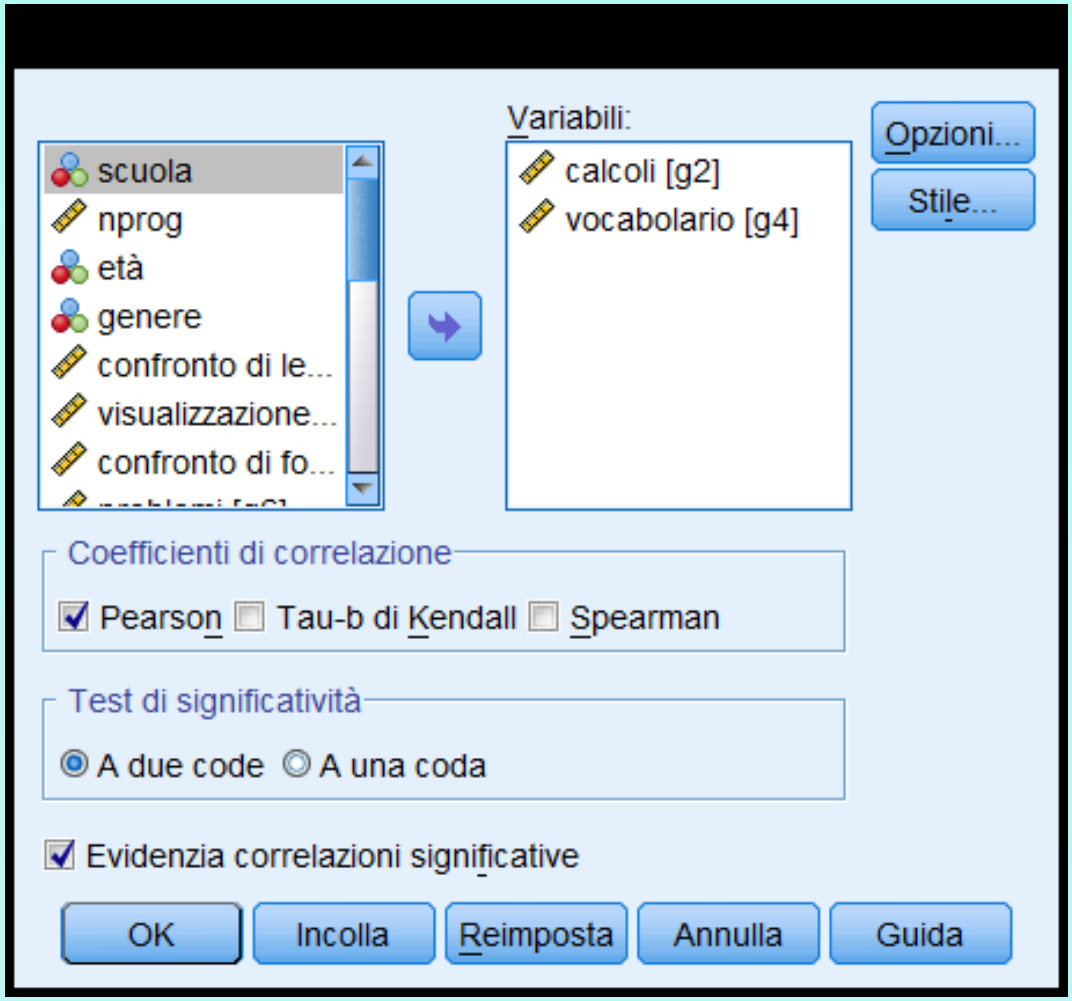
poi Bivariata

Poi Si scelgono le variabili dalla finestra di sinistra

Correlazione→ Bivariata

Opzione di default

Di solito la significatività è a due code



Opzioni ...

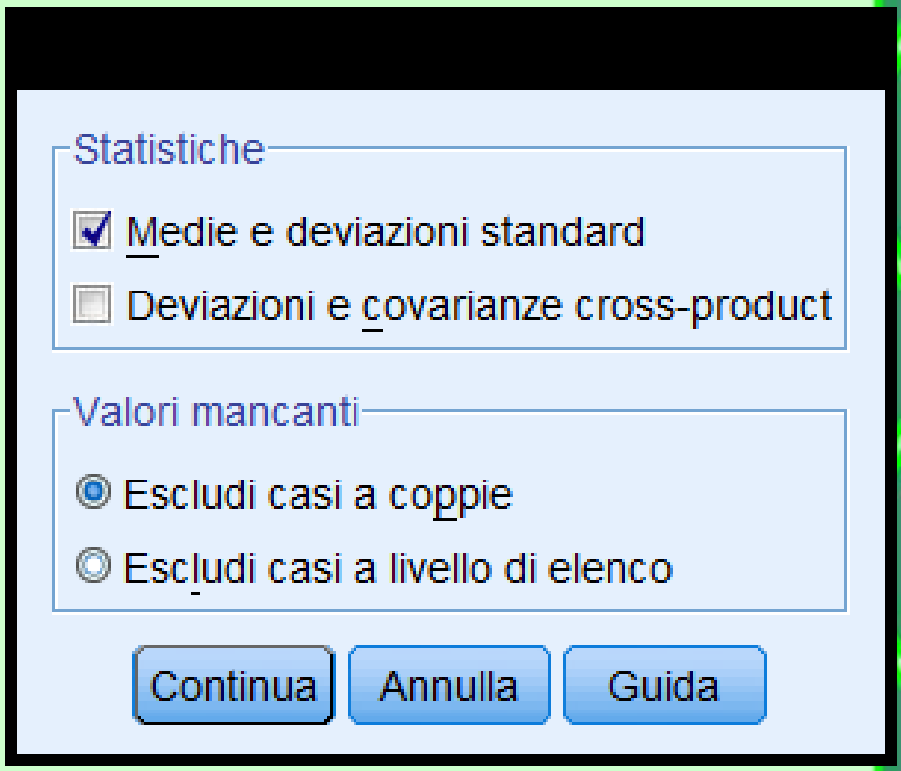
Nelle opzioni si può scegliere medie e ds se servono

Si può scegliere il modo di trattare i dati mancanti

a coppie: utilizza per ogni coppia di variabili tutti i dati disponibili (l'n può variare per le coppie se ci sono dati mancanti)

elenco: usa solo le righe dei dati che contengono dati disponibili per tutte le coppie richieste. Se una riga ha anche un solo dato mancante, provoca l'eliminazione di tutti i dati della riga e diminuisce l'N totale.

L'N totale è uguale per tutte le correlazioni richieste



Esame dell'output di SPSS

Correlazioni

		calcoli	vocabolario
calcoli	Correlazione di Pearson	1	,407**
	Sign. (a due code)		,000
	N	635	635
vocabolario	Correlazione di Pearson	,407**	1
	Sign. (a due code)	,000	
	N	635	635

**. La correlazione è significativa a livello 0,01 (a due code).

La diagonale riporta il valore 1 (correlazione della variabile con sé stessa e il numero di casi disponibili

Esame dell'output di SPSS

La matrice dei dati è simmetrica: sopra la diagonale sono riportati gli stessi valori in forma speculare che si trovano nella parte sotto la diagonale

Correlazioni

		calcoli	vocabolario
calcoli	Correlazione di Pearson	1	,407**
	Sign. (a due code)		,000
	N	635	635
vocabolario	Correlazione di Pearson	,407**	1
	Sign. (a due code)	,000	
	N	635	635

**. La correlazione è significativa a livello 0,01 a due code).

Correlazione
Significatività
Numero di casi usati

FINE

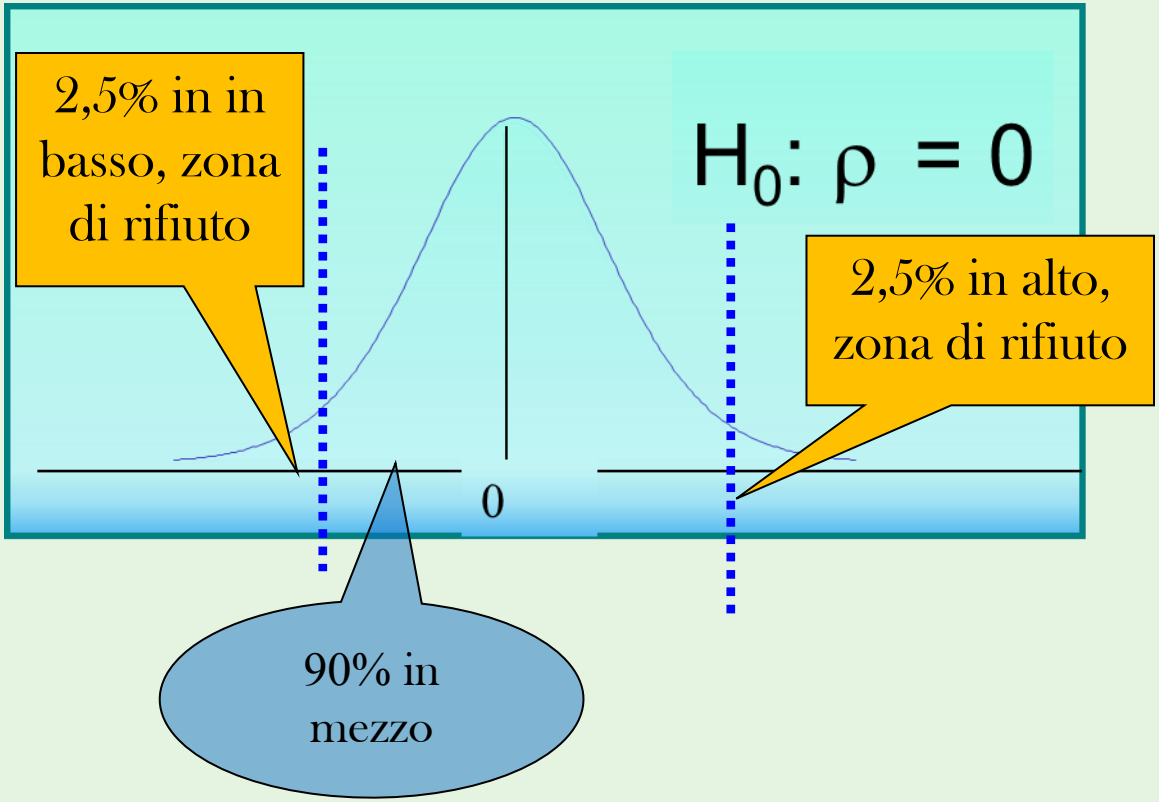
Esame dell'output di SPSS

Correlazioni

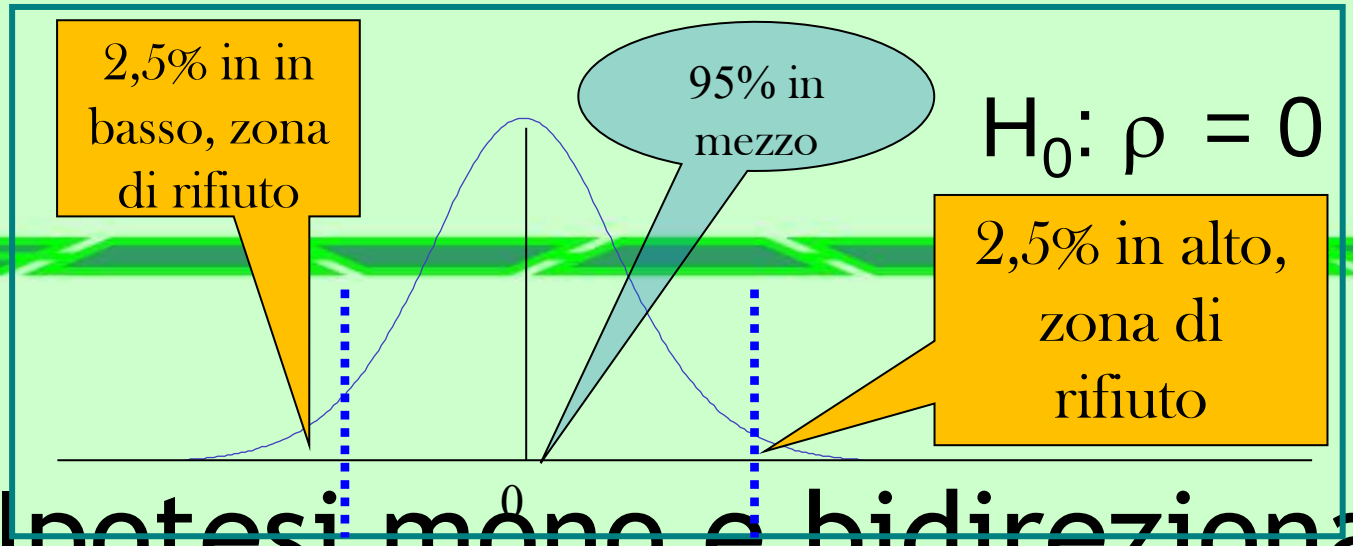
		calcoli	vocabolario
calcoli	Correlazione di Pearson	1	,407**
	Sign. (a due code)		,000
	N	635	635
vocabolario	Correlazione di Pearson	,407**	1
	Sign. (a due code)	,000	
	N	635	635

**. La correlazione è significativa a livello 0,01 (a due code).

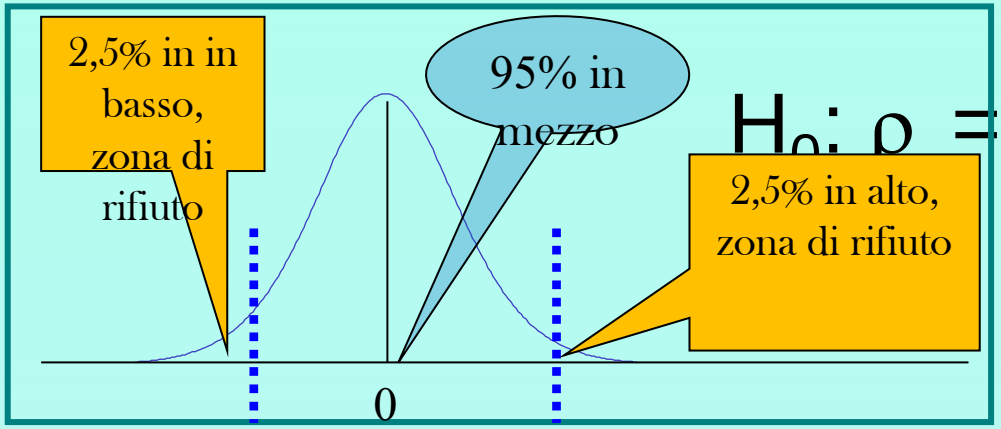
Quando si riportano i dati in una ricerca, di solito si usano i livello 0,01 e 0,05 di significatività

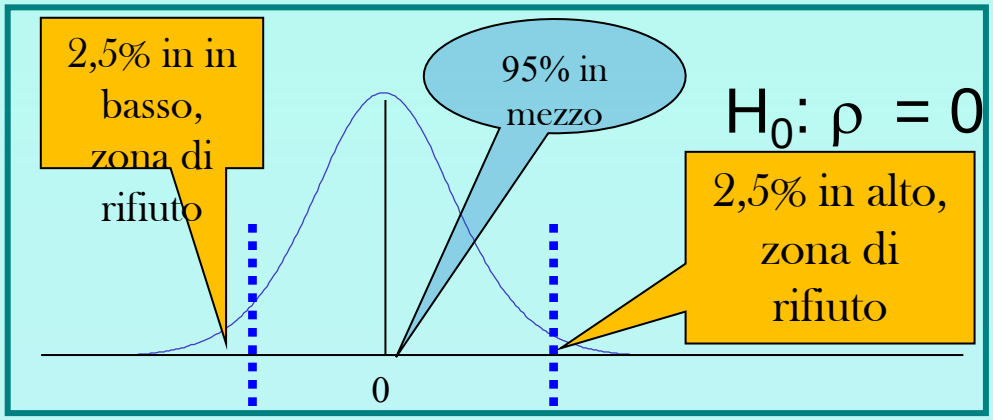

2,5% in in basso, zona di rifiuto
H_0: ρ = 0
2,5% in alto, zona di rifiuto
0
90% in mezzo

- Con tali



Ipotesi mono e bidirezionali


2,5% in in basso, zona di rifiuto
95% in mezzo
H_0: ρ = 0
2,5% in alto, zona di rifiuto
0


2,5% in in basso, zona di rifiuto
95% in mezzo
H_0: ρ = 0
2,5% in alto, zona di rifiuto
0