



UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI SCIENZE AGRARIE E AMBIENTALI
PRODUZIONE, TERRITORIO, AGROENERGIA

Metodi Statistici per la Ricerca Ambientale

Marco Acutis

marco.acutis@unimi.it

www.acutis.it

a.a. 2019 - 2020

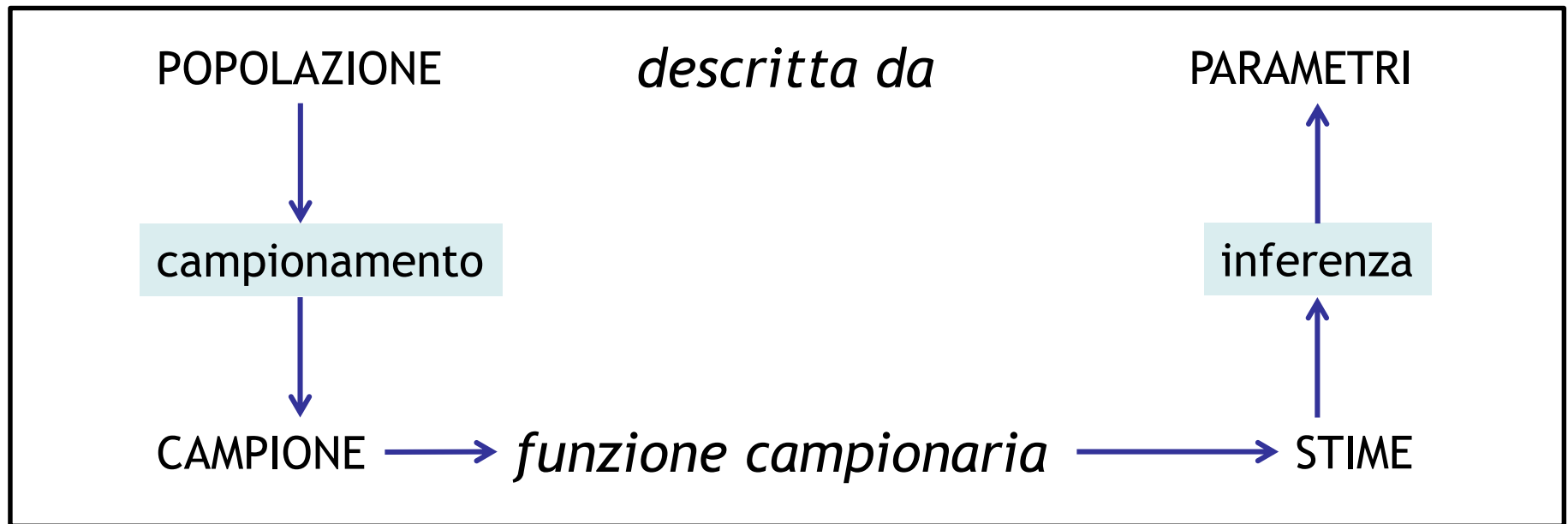
Lezione 02 - Sommario

- ❑ Statistica Inferenziale
 - Test delle Ipotesi
- ❑ Analisi della Varianza
 - ANOVA a 1 via



Obiettivi dell'Inferenza Statistica

1. Stima dei parametri della popolazione
2. Test delle ipotesi



Test delle Ipotesi - Definizione

Il test consiste nel formulare un'ipotesi (ipotesi nulla) e nel verificare se, con i dati a disposizione, è possibile rifiutarla o no.

L'ipotesi che viene formulata è detta **ipotesi nulla** (H_0) e rappresenta di solito lo stato di fatto.

Se il campione fornisce risultati fortemente in contrasto con H_0 , quest'ultima viene rifiutata a favore di una **ipotesi alternativa** (H_1).

Test delle Ipotesi - Procedimento

Usando una distribuzione campionaria, identifichiamo un range di valori che hanno bassa probabilità di accadere se l'ipotesi nulla è vera.

Questo range di valori costituisce la cosiddetta regione critica o **regione di rifiuto** dell'ipotesi nulla.

Dalla distribuzione campionaria della statistica posso conoscere le probabilità di ottenere determinati valori e, sulla base di queste, definire la regione di rifiuto.

Test della media di una popolazione (1/5)

Eseguiamo il test $H_0: \mu = \mu_0$

Ipotesi nulla H_0 : la media μ della popolazione da cui abbiamo estratto il campione è $= \mu_0$

Ipotesi nulla H_1 : la media μ della popolazione da cui abbiamo estratto il campione è $\neq \mu_0$

Verifichiamo quindi se la deviazione della media campionaria da μ_0 è compatibile con l'ipotesi nulla.

A questo scopo confrontiamo la media del campione con μ_0 .

Test della media di una popolazione (2/5)

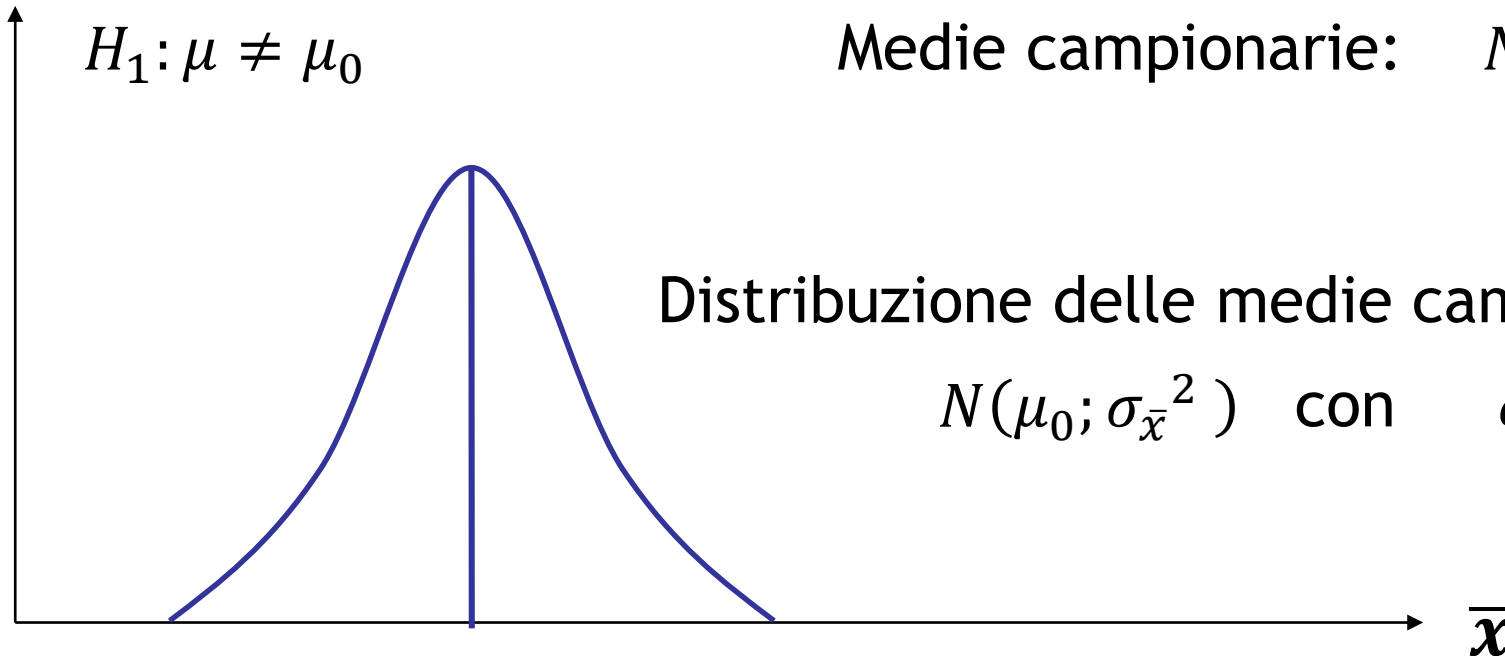
Situazione teorica, ipotizzando una popolazione «normale».

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

$$\text{Popolazione: } N(\mu_0; \sigma^2)$$

$$\text{Medie campionarie: } N(\mu_0; \sigma^2/n)$$



Test della media di una popolazione (3/5)

Verifico $H_0: \mu_0 = 10$ in una popolazione con $\sigma = 6$, estraendo a caso un campione di taglia $n = 9$.

$$H_0: \mu = 10$$

$$\text{Popolazione: } N(10; \sigma^2)$$

$$H_1: \mu \neq 10$$

$$\text{Medie campionarie: } N(10; \sigma_{\bar{x}}^2)$$

$$N(10; \sigma_{\bar{x}}^2)$$

Posto $\alpha = 0,05$, rifiuto H_0 se la media campionaria è al di fuori dei limiti $\mu_0 \pm 1,96(\sigma/\sqrt{n})$.

La regione di rifiuto ha probabilità α (livello di significatività), che rappresenta la probabilità di rifiutare H_0 quando H_0 è vera.

10

$\bar{x}$

Test della media di una popolazione (4/5)

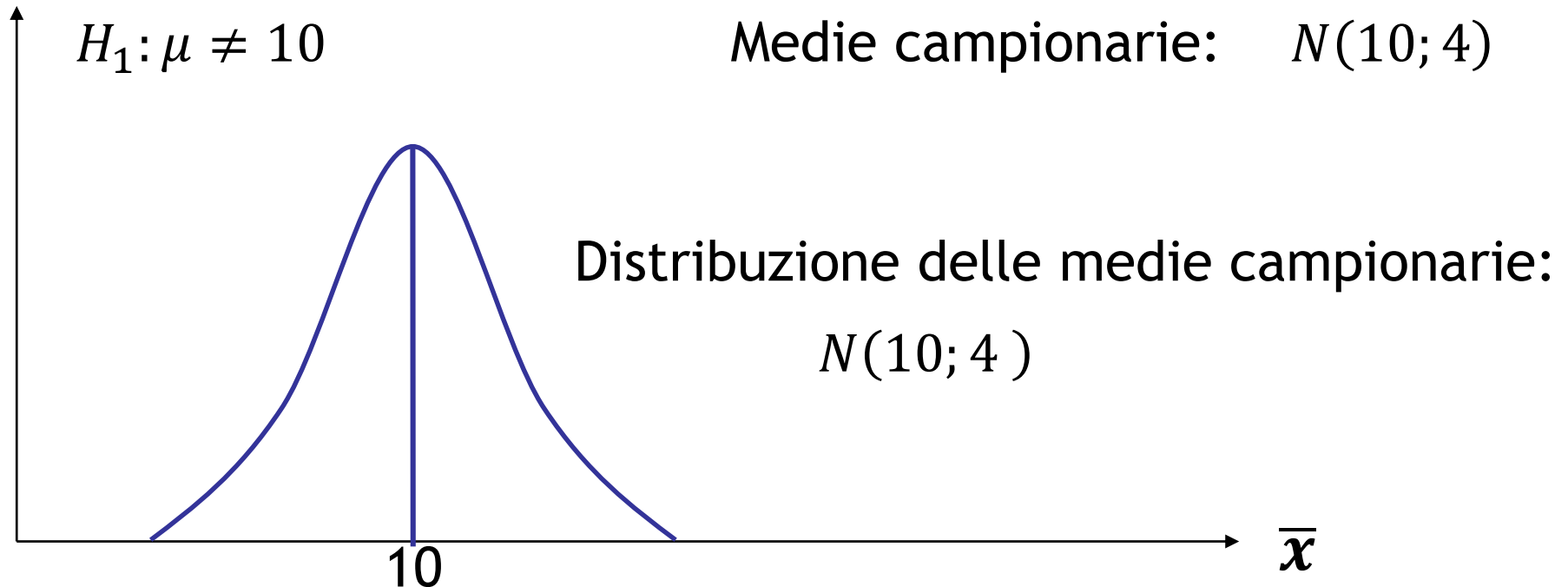
Questa è dunque in sintesi la situazione dell'esempio, essendo noti $\sigma = 6$ e $n = 9$.

$$H_0: \mu = 10$$

$$H_1: \mu \neq 10$$

Popolazione: $N(10; 36)$

Medie campionarie: $N(10; 4)$



Test della media di una popolazione (5/5)

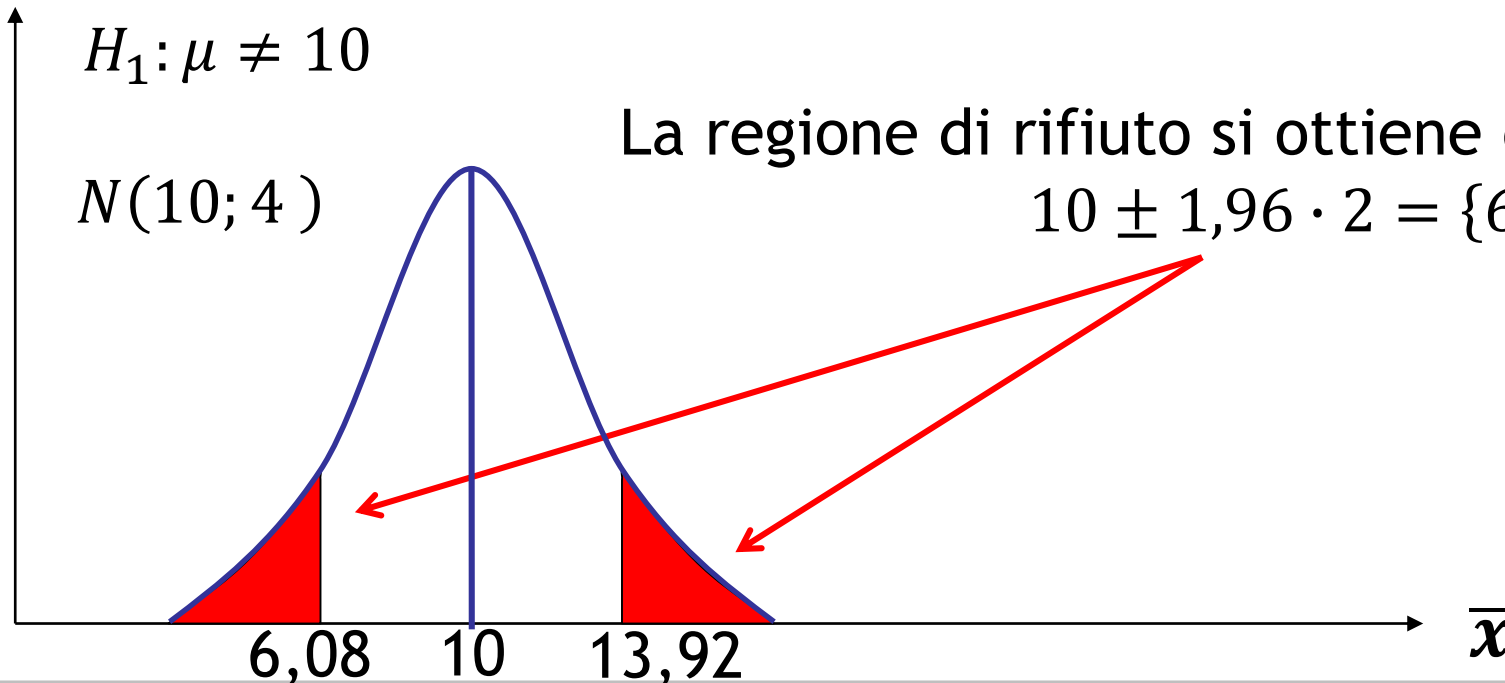
Posto quindi $\alpha = 0,05$, rifiuto H_0 se la media campionaria è al di fuori dei limiti $\mu_0 \pm 1,96(\sigma/\sqrt{n})$.

$$H_0: \mu = 10$$

$$H_1: \mu \neq 10$$

$$N(10; 4)$$

La regione di rifiuto si ottiene calcolando
 $10 \pm 1,96 \cdot 2 = \{6,08; 13,92\}$



Le fasi da seguire per un test delle ipotesi

1. Specificare H_0 , H_1 ed un livello α .
2. Definire una statistica per il test (statistica di cui sia definibile la distribuzione campionaria) e la zona di rifiuto per H_0 (cioè i valori della statistica di probabilità che risultano $< \alpha$ quando H_0 è vera).
3. Eseguire il campionamento (o l'esperimento) e calcolare la statistica.
4. Se la statistica calcolata cade nella zona di rifiuto decido di rifiutare H_0 , altrimenti decido di non rifiutare H_0 .

...tornando all'esempio precedente

1. $H_0: \mu = 10, H_1: \mu \neq 10, \alpha = 0,05$.
2. La statistica è Z . Dal momento che $P(|z| > 1,96) = 0,05$, la zona di rifiuto è $z < -1,96$ o $z > 1,96$ (test a due code).
3. Calcolo la media campionaria e la converto nella variabile standardizzata: $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$.
4. Rifiuto l'ipotesi nulla se $|z| > 1,96$; in questo modo si ha una probabilità di rifiuto pari a 0,05 quando H_0 è vera (e quindi una probabilità d'errore del 5%).

L'uso del p -value

In alternativa si può riportare direttamente il valore della probabilità p di commettere il cosiddetto «errore di I specie» (livello di significatività osservato).

Il p -value è una misura di quanto i dati sono in disaccordo con H_0 .

Si procede come segue:

1. $H_0: \mu = 10$, $H_1: \mu \neq 10$.
2. Calcolo la media campionaria e la converto nella variabile standardizzata: $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$.
3. Calcolo la probabilità p di ottenere il valore di z calcolato, vale a dire $P(Z > |z|)$ (test a due code).

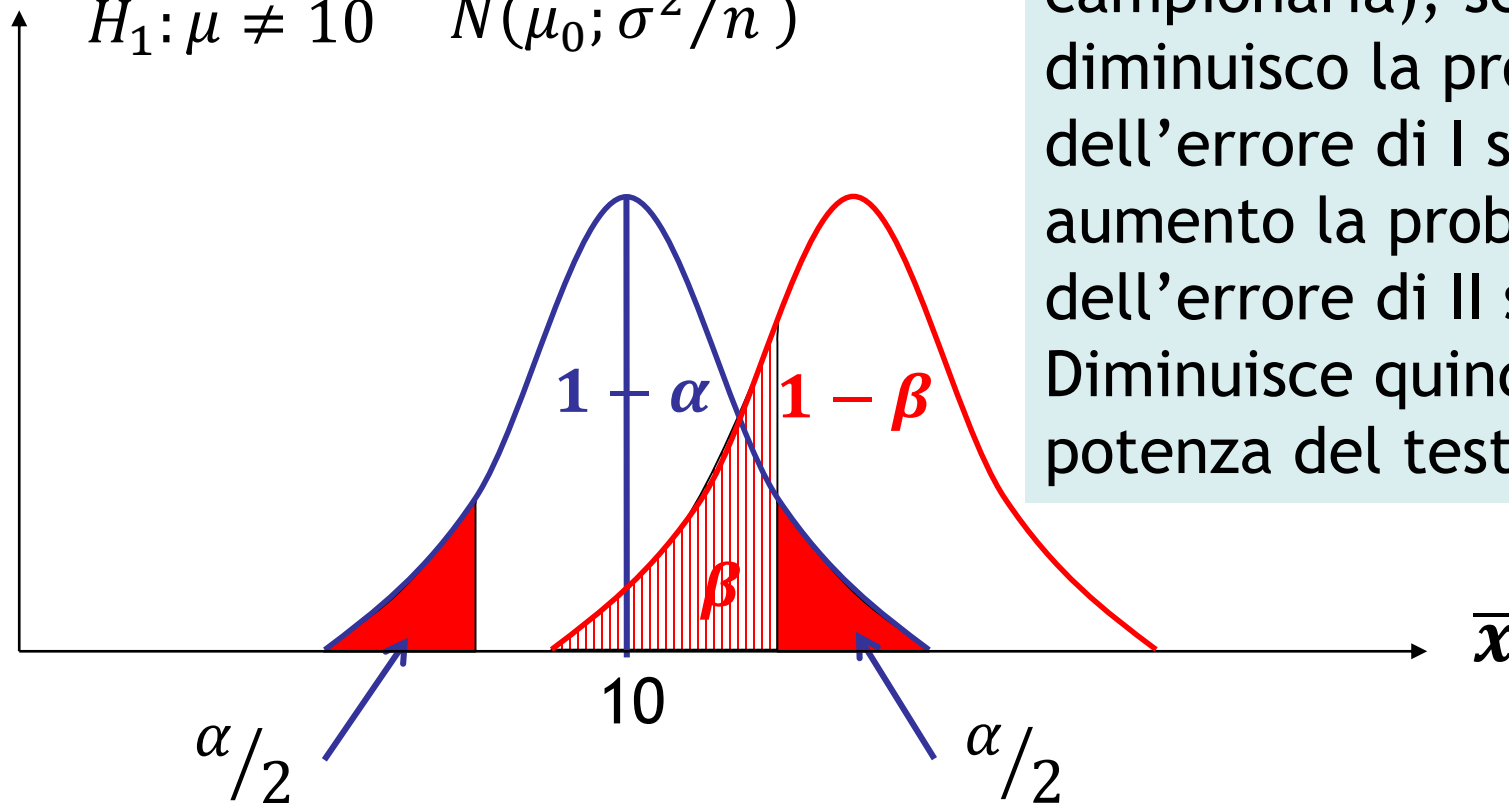
Significatività e potenza del test (1/2)

		verità	
		H_0 vera	H_0 falsa
conclusione	H_0 vera	GIUSTO $P = 1 - \alpha$ Livello di protezione	ERRORE II specie $P = \beta$
	H_0 falsa	ERRORE I specie $P = \alpha$ Significatività	GIUSTO $P = 1 - \beta$ Livello di potenza

Significatività e potenza del test (2/2)

$$H_0: \mu = 10$$

$$H_1: \mu \neq 10 \quad N(\mu_0; \sigma^2/n)$$

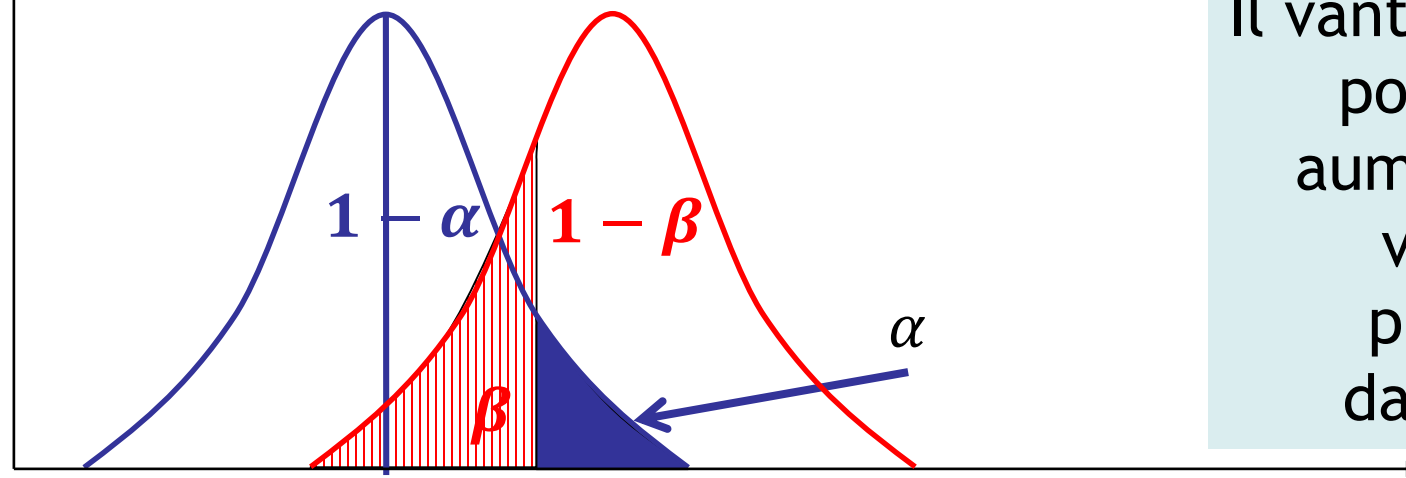


A parità di n (numerosità campionaria), se diminuisco la probabilità dell'errore di I specie (α), aumento la probabilità dell'errore di II specie (β). Diminuisce quindi la potenza del test ($1 - \beta$).

Test a una o due code

Se siamo interessati a rifiutare H_0 solo nel caso in cui la differenza è in un senso o nell'altro, allora eseguiamo il test a una coda, o test unilaterale.

L'ipotesi alternativa sarà $H_1: \mu > \mu_0$ oppure $H_1: \mu < \mu_0$.



Il vantaggio è che la potenza del test aumenta andando verso H_1 , ma è praticamente 0 dall'altra parte.

ANOVA - Definizione (1/2)

L'analisi della varianza (ANOVA) studia le relazioni tra variabili discrete, che definiscono delle categorie, e variabili continue.

Esempi

- confronti culti-varietali
- produzioni in funzione di tecniche colturali diverse
- effetti di andamenti climatici sulla produzione
- effetti di diversi trattamenti sperimentali
- ...

Si tratta dell'analisi maggiormente utilizzata nelle scienze agrarie e forestali .

ANOVA - Definizione (2/2)

Tipi di analisi della varianza:

- ✓ ANOVA a 1 via, quando si ha 1 sola variabile discreta;
- ✓ ANOVA a 2 o più vie, quando si hanno 2 o più variabili discrete e viene studiata anche l'interazione.

Principio Base:

il confronto tra due o più gruppi di dati si fonda sul rapporto tra la variabilità entro i gruppi e la variabilità tra i gruppi.

Analisi della Varianza - Il caso a 1 via

Definizione

Nel caso dell'ANOVA a 1 via, si considera 1 sola fonte o causa delle variazioni dei dati, che viene definita fattore sperimentale o trattamento.

Un trattamento può prevedere:

- più livelli quantitativi, come le dosi crescenti di un farmaco,
oppure
- diverse modalità qualitative, come la somministrazione di farmaci differenti.

Ogni unità sperimentale si chiama replica e, TEORICAMENTE, devono essercene almeno 2 per almeno un livello o modalità del trattamento. In altre parole il numero minimo di repliche è pari al numero dei livelli + 1.



Formalizzazione del Test d'Ipotesi

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

Le k popolazioni da cui sono tratti i campioni hanno la stessa media.

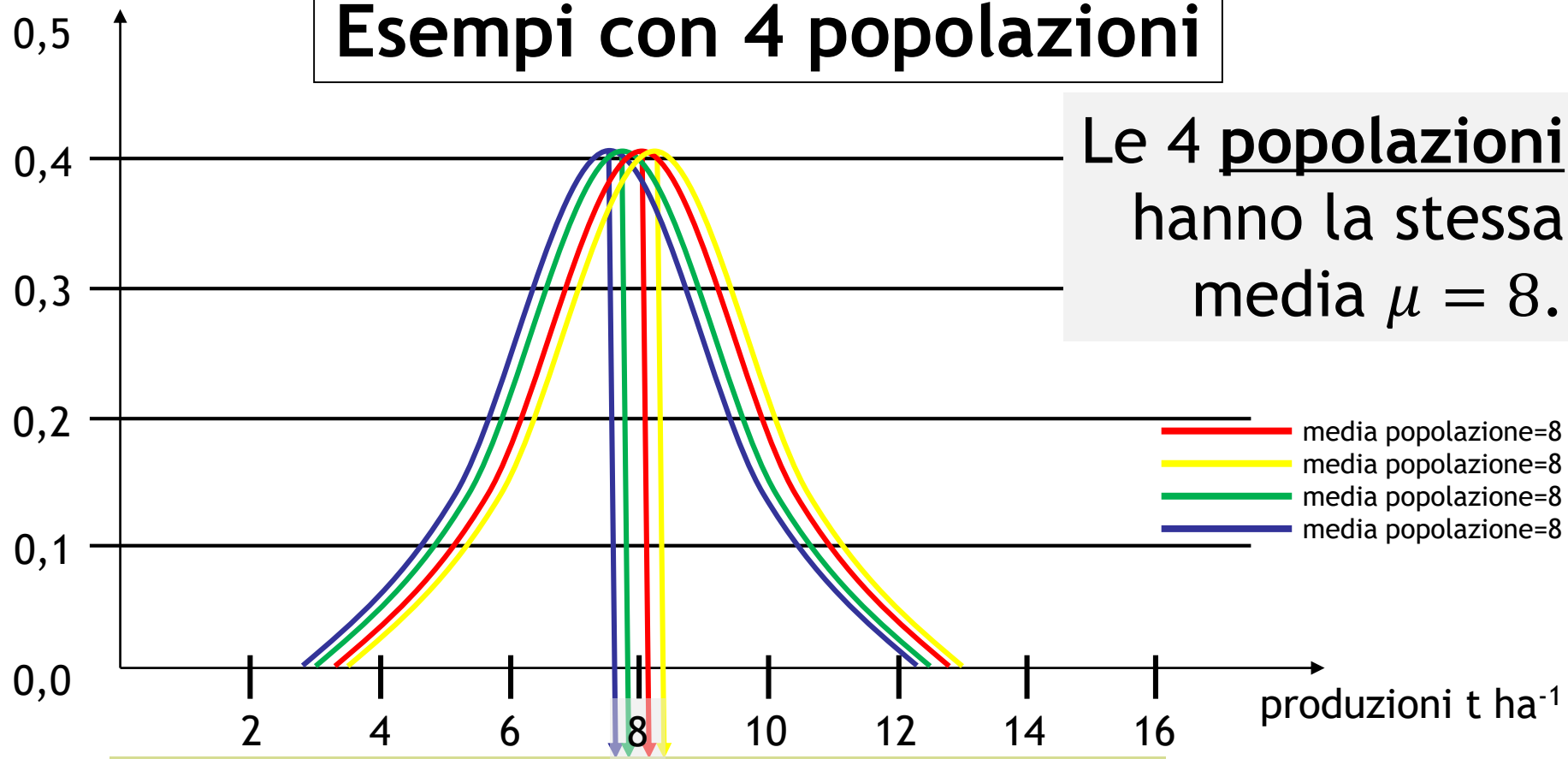
$$H_1: \exists i \text{ tale per cui } \mu_i \neq \mu_j, \text{ con } j = 1, \dots, k \text{ e } j \neq i$$

Almeno una delle medie delle popolazioni da cui sono tratti i campioni differisce dalle altre.

N.B. m rappresenta il numero di livelli del trattamento!

Analisi della Varianza - Il caso a 1 via

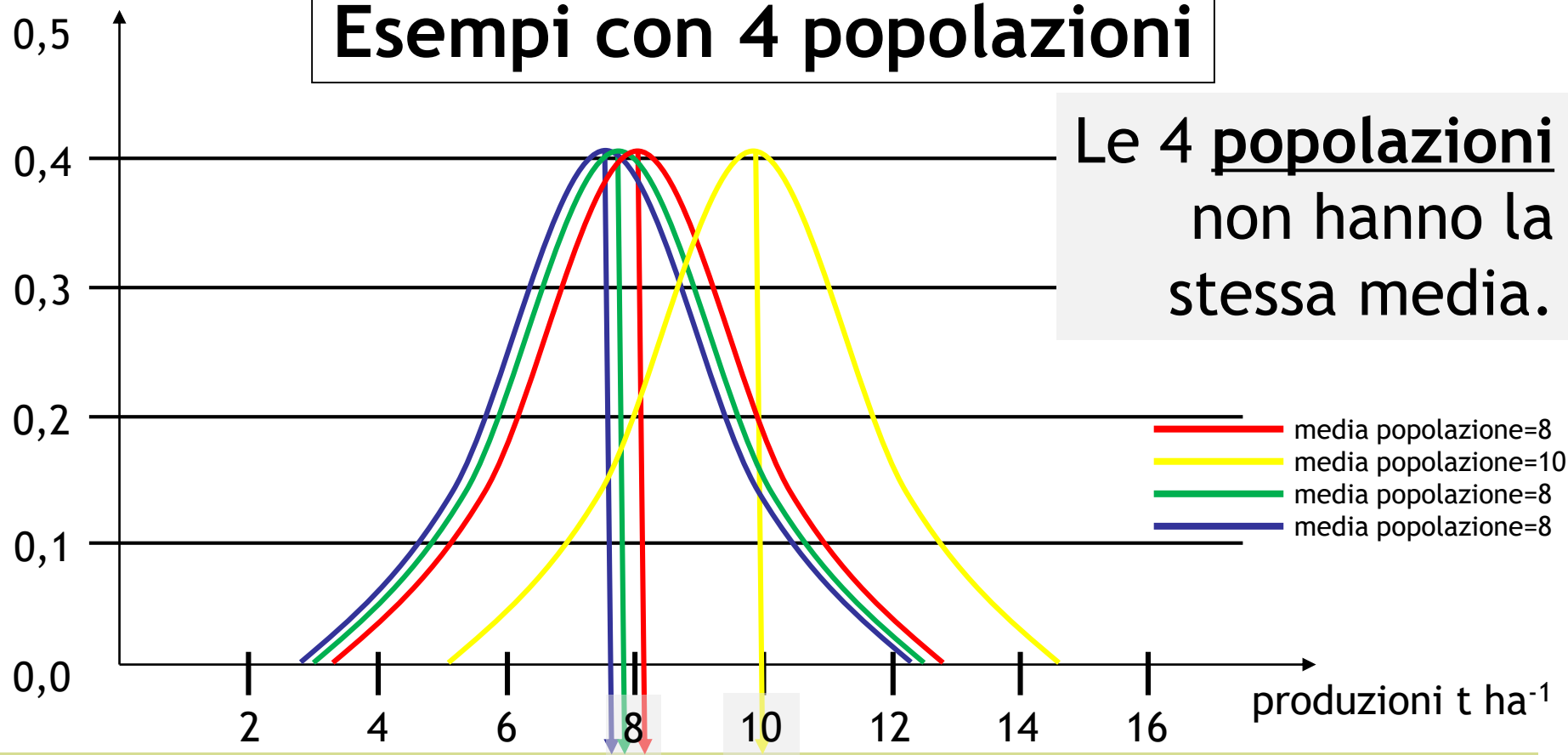
Esempi con 4 popolazioni



I valori medi dei 4 campioni differiscono tra loro solo per fluttuazioni casuali!

Analisi della Varianza - Il caso a 1 via

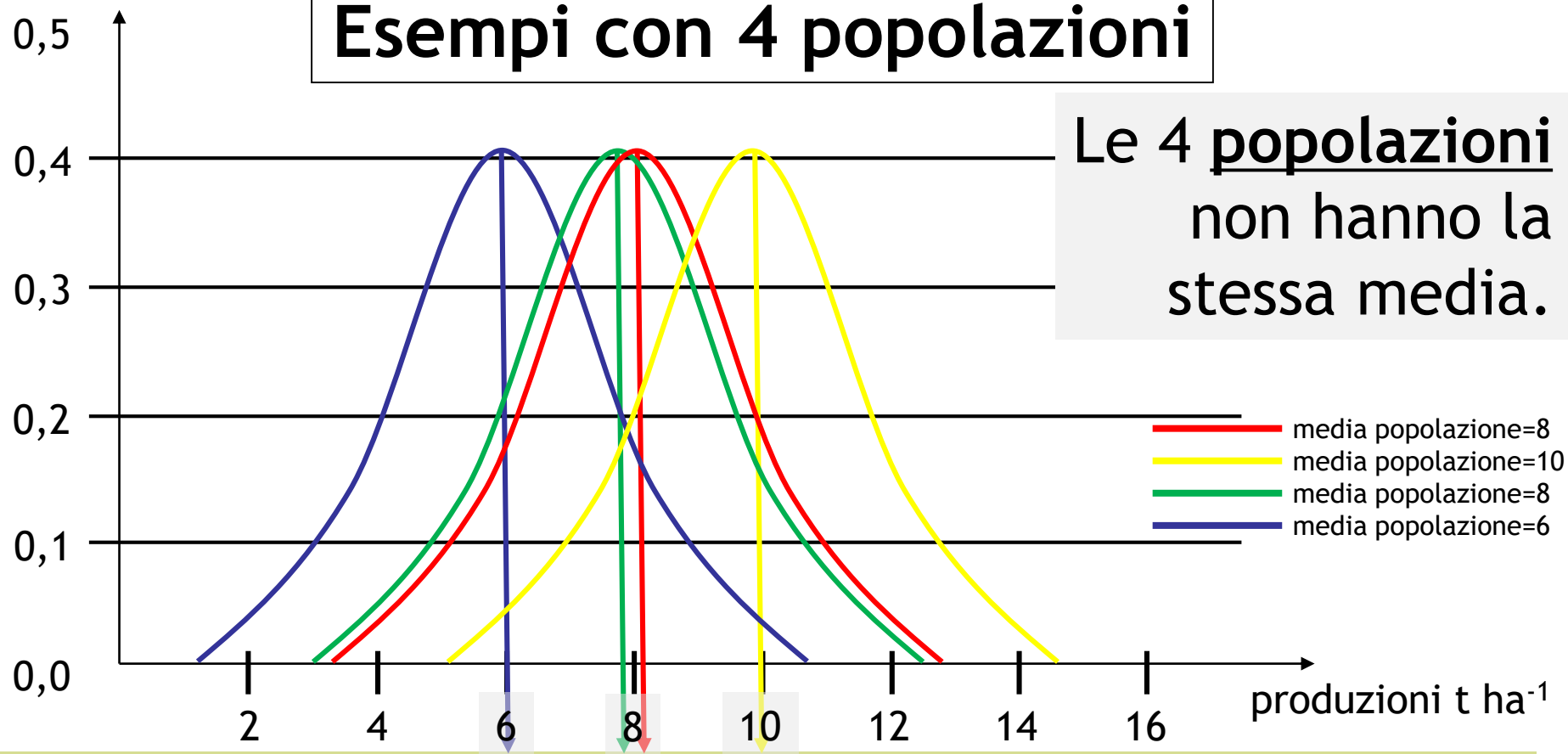
Esempi con 4 popolazioni



Le differenze tra le medie dei campioni sono da attribuirsi a differenze di medie tra le popolazioni da cui sono stati tratti.

Analisi della Varianza - Il caso a 1 via

Esempi con 4 popolazioni



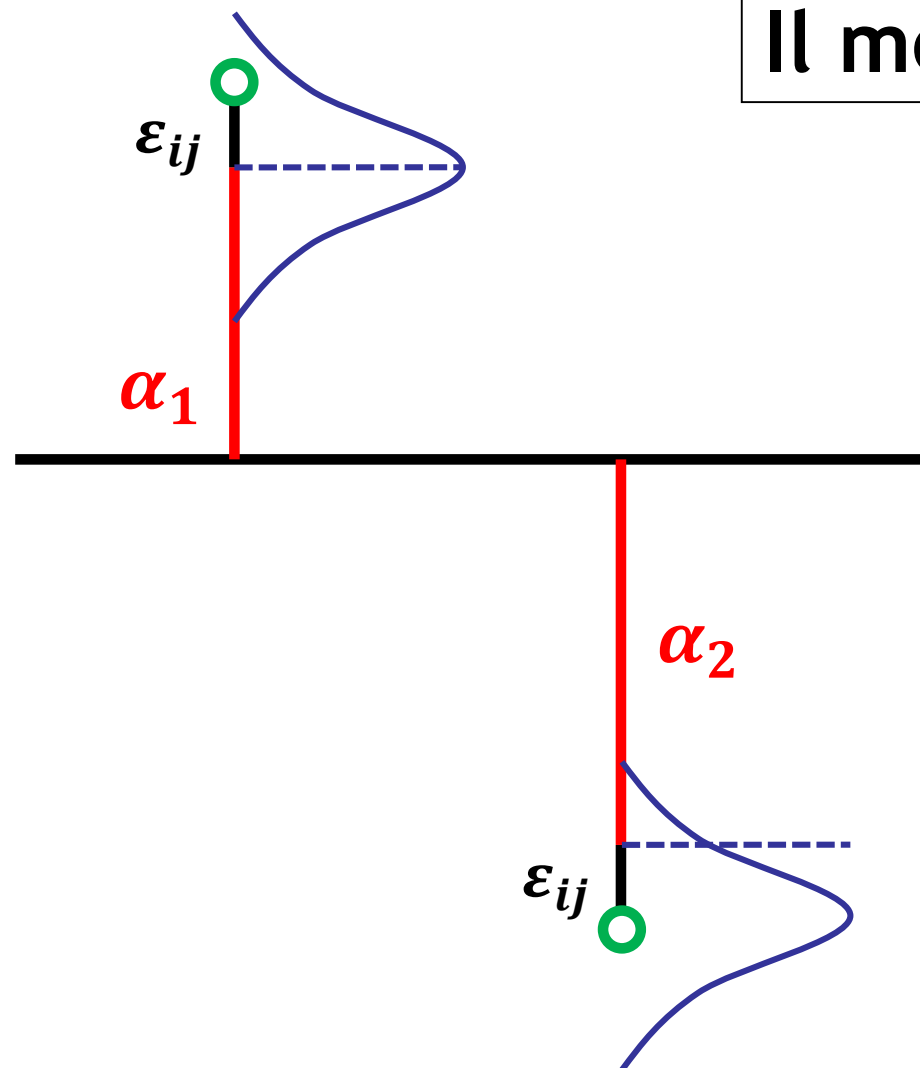
Le differenze tra le medie dei campioni sono da attribuirsi a differenze di medie tra le popolazioni da cui sono stati tratti.

Analisi della Varianza - Il caso a 1 via

Il modello

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Il valore di un dato Y_{ij} è la somma dell'effetto di uno specifico livello del fattore (o trattamento) α_i e di una componente accidentale (residuo o errore sperimentale) ε_{ij} , tenuto conto della media generale μ .



La Randomizzazione Completa

Si tratta del modello più semplice di ANOVA e prevede un campionamento in cui:

- ciascun trattamento viene assegnato in modo totalmente casuale ai soggetti (omogenei) su cui si effettuano le misure, per determinare l'effetto dei singoli trattamenti;
- le repliche(osservazioni) sono assegnate casualmente ai vari livelli del trattamento, i quali possono non avere lo stesso numero di osservazioni.

Analisi della Varianza - Il caso a 1 via

La Rappresentazione

	<i>k</i> Livelli del Trattamento			
	T_1	T_2	...	T_k
$n_1 + n_2 + \dots + n_k = N$ Osservazioni	x_{11}	x_{12}	...	x_{1k}
	x_{21}	x_{22}	...	x_{2k}
	...	...	...	...
	$x_{n_1 1}$	$x_{n_2 2}$	...	$x_{n_k k}$
Numero di repliche	n_1	n_2	...	n_k
<i>k</i> Medie dei trattamenti	$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_k$
Media Generale	$\bar{\bar{X}}$			

Analisi della Varianza - Il caso a 1 via

Le Devianze

La metodologia dell'ANOVA si basa sul fatto che la variazione totale nella risposta misurata a una certa sollecitazione o trattamento si può suddividere in componenti che vengono attribuite a specifiche cause di variabilità.



occorre calcolare:

Media generale $\bar{\bar{X}} = \frac{1}{k} \sum_{j=1}^k \bar{X}_j$

Media dei trattamenti $\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$

Devianza totale $dev_{TOT} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{X}})^2$

Devianza TRA trattamenti $dev_{TRA} = \sum_{j=1}^k n_j (\bar{X}_j - \bar{\bar{X}})^2$

Devianza ENTRO trattamenti $dev_{ENTRO} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_j)^2$

Principio di scomposizione della varianza:

$$dev_{TOT} = dev_{TRA} + dev_{ENTRO}$$

Analisi della Varianza - Il caso a 1 via

I Gradi di Libertà (1/3)

Alcune stime si basano su un numero di informazioni più alto rispetto ad altre. Ad esempio, una stima della varianza a partire da un campione di taglia 100 si basa su un numero maggiore di informazioni rispetto al caso in cui la taglia del campione sia pari a 5.



I Gradi di Libertà rappresentano il numero di informazioni indipendenti su cui si basa una stima.

Esempio 1) Supponiamo che la media (μ) di una popolazione sia nota.

Se, «per assurdo», volessimo stimare la media della popolazione, di quante informazioni avremmo bisogno?

Ovviamente zero perché μ la conosciamo già.

Di conseguenza, in questo caso (paradossale), i gradi di libertà della stima sono pari a zero.

Esempio 2) Supponiamo che la media di una popolazione sia ignota.

Se vogliamo stimare la media della popolazione, di quante informazioni abbiamo bisogno?

Sappiamo che le osservazioni (cioè i campionamenti dalla popolazione) devono essere indipendenti e perciò tale è l'informazione che contengono. In questo caso quindi i gradi di libertà coincidono con il numero delle osservazioni stesse.

D'altra parte la media campionaria si definisce proprio come la somma delle singole osservazioni diviso il loro numero...



Analisi della Varianza - Il caso a 1 via

I Gradi di Libertà (2/3)

Esempio 3) Supponiamo che la media di una popolazione sia nota (pari a 10).

Se vogliamo stimare la varianza della popolazione, di quante informazioni abbiamo bisogno?

Immaginiamo di fare un primo campionamento e di ottenere 8. Dato che la varianza è definita come la media dei quadrati degli scarti dalla media della popolazione, sappiamo che la nostra prima stima è pari a $\frac{(8-10)^2}{1} = 4$. Questa stima è basata su una singola informazione, quindi i suoi gradi di libertà sono pari a 1.

Ora ipotizziamo di fare un secondo campionamento e di ottenere 9. La stima della varianza ora è pari a $\frac{(8-10)^2 + (9-10)^2}{2} = 2,5$ ed è basata su due informazioni indipendenti. Di conseguenza i gradi di libertà sono 2.

Ripetendo la procedura, si può concludere che i gradi di libertà, in questo caso, coincidono con il numero di osservazioni e infatti $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} \dots$



Analisi della Varianza - Il caso a 1 via

I Gradi di Libertà (3/3)

Esempio 4) Supponiamo che la media di una popolazione sia ignota.

Se vogliamo stimare la varianza della popolazione, di quante informazioni abbiamo bisogno?

Immaginiamo di fare 2 campionamenti e di ottenere 8 e 6. Dal momento che non la conosciamo, dobbiamo stimare la media della popolazione, prima di poter stimare la varianza. Abbiamo quindi $\bar{x} = (8 + 6)/2 = 7$ e 2 stime: $stima1 = (8 - 7)^2 = 1$ e $stima2 = (6 - 7)^2 = 1$.

Il punto chiave è che queste 2 informazioni non sono indipendenti, dal momento che ciascuna osservazione contribuisce al calcolo di $\bar{x}$. 8 (prima osservazione) influenza $\bar{x}$ e conseguentemente influenza anche $stima2$. Infatti se la prima osservazione fosse stata pari a 9, allora avremmo avuto $\bar{x} = (9 + 6)/2 = 7,5$ e $stima2 = (6 - 7,5)^2 = 2,25$. Analogo discorso vale per la seconda osservazione e perciò i gradi di libertà non possono essere 2.

Generalizziamo e immaginiamo di aver fatto 4 campionamenti.

Sappiamo che la procedura di stima della varianza richiede PRELIMINARMENTE il calcolo di $\bar{x}$.

A questo punto calcoliamo, per ognuno dei 4 campionamenti, lo scarto dalla media $(x_i - \bar{x})$.

Ricordiamo però che $(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + (x_4 - \bar{x}) = 0$.

Da questa formula segue che il quarto scarto è univocamente determinato dai primi tre, infatti si ha $(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) = -(x_4 - \bar{x})$, e perciò quest'ultima informazione DIPENDE dalle prime 3.

In conclusione quindi i gradi di libertà della stima della varianza, in questo caso, sono 3 (numero di osservazioni - 1) perché 3 sono le informazioni indipendenti su cui si basa e infatti $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \dots$



Analisi della Varianza - Il caso a 1 via

Le Varianze

sfruttando la

regola di calcolo dei gdl: il numero di gradi di libertà si ricava dalla differenza tra il numero di dati e il numero di parametri (o informazioni preliminari) che è necessario calcolare per ottenere la stima voluta

otteniamo

$$dev_{TOT} \longrightarrow N - 1 \text{ gradi di libertà}^{(1)} \longrightarrow s_{TOT}^2 = \frac{dev_{TOT}}{N - 1} = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{X}})^2}{N - 1}$$

$$dev_{TRA} \longrightarrow k - 1 \text{ gradi di libertà}^{(2)} \longrightarrow s_{TRA}^2 = \frac{dev_{TRA}}{k - 1} = \frac{\sum_{j=1}^k n_j (\bar{X}_j - \bar{\bar{X}})^2}{k - 1}$$

$$dev_{ENTRO} \longrightarrow N - k \text{ gradi di libertà}^{(3)} \longrightarrow s_{ENTRO}^2 = \frac{dev_{ENTRO}}{N - k} = \frac{\sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_j)^2}{N - k}$$

⁽¹⁾ N dati $- 1$ media generale

⁽²⁾ k dati (medie dei trattamenti) $- 1$ media generale

⁽³⁾ N dati $- k$ medie dei trattamenti oppure, in modo equivalente, $(n_1 - 1) + \dots + (n_k - 1) = (n_1 + \dots + n_k) - \overbrace{(1 + \dots + 1)}^{k \text{ volte}} = N - k$



Analisi della Varianza - Il caso a 1 via

Test F di Fisher (1/2)

Come detto in precedenza, lo scopo dell'ANOVA a 1 via è quello di verificare l'ipotesi nulla $H_0: \mu_1 = \mu_2 = \dots = \mu_k$.



Occorre quindi definire un metodo che ci permetta di distinguere se le differenze che osserviamo tra le medie dei diversi campioni sono dovute a reali differenze TRA i trattamenti o alla naturale differenza che si osserva campionando ripetutamente ENTRO la stessa popolazione.



Sappiamo che “Varianza TRA” e “Varianza ENTRO” dipendono dalla variabilità esistente nei dati ed essendo due misure della stessa variabilità, dovrebbero avere lo stesso valore.



Indice dell'uguaglianza tra queste due componenti della varianza è il rapporto:



$$F = \frac{S_{TRA}^2}{S_{ENTRO}^2}$$

Test F di Fisher (2/2)

- ✓ Il rapporto F è la statistica del test F di Fisher
- ✓ F si distribuisce approssimativamente come una distribuzione di Fisher con $k - 1$ e $N - k$ gradi di libertà: $F \sim F(k - 1, N - k)$



Intuitivamente:

- se è vera H_0 , dovrebbe risultare $F = 1$
- se è vera H_1 , dovrebbe risultare $F > 1$

Rigorosamente (fissato un livello di significatività α):

con un numero **finito** di trattamenti e repliche

- se $F < F^{-1}_{(k-1, N-k)}(\alpha)$, allora non si può rifiutare H_0
- se $F \geq F^{-1}_{(k-1, N-k)}(\alpha)$, allora si rifiuta H_0

con un numero **infinito** di trattamenti e repliche

- se è vera H_0 , allora $F = 1$
- se è vera H_1 , allora $F > 1$

N.B. $F^{-1}_{(k-1, N-k)}(\alpha)$
è un valore tabulato

Esempio - Il calcolo «a mano» (1/8)

Supponiamo di avere il seguente dataset:

	4 Livelli del Trattamento			
	T_1	T_2	T_3	T_4
3 + 5 + 4 + 4 = 16 Osservazioni	10,0	8,5	12,0	10,0
	9,5	8,0	9,5	9,5
	11,0	9,0	11,0	11,0
		9,0	11,5	10,0
		10,0		
Numero di repliche	3	5	4	4
4 Medie dei trattamenti	10,2	8,9	11,0	10,1
Media Generale	10,0			

e riportiamo i dati in un foglio di lavoro Excel...

Analisi della Varianza - Il caso a 1 via

Esempio - Il calcolo «a mano» (2/8)

The screenshot shows an Excel spreadsheet titled 'Anova_a_mano.xlsx'. The formula bar for cell G7 displays `=DEV.Q(B3:E7)`, which is circled in red. A red arrow points from this formula to the text 'Attraverso la funzione DEV.Q, calcoliamo la devianza totale e riportiamo i relativi gdl.' Another red arrow points from the text to the cell containing the value 18.234375, which is labeled 'Devianza Totale'.

Dataset					
repliche * trattamenti	T1	T2	T3	T4	
r1	10.0	8.5	12.0	10.0	
r2	9.5	8.0	9.5	9.5	
r3	11.0	9.0	11.0	11.0	
r4		9.0	11.5	10.0	
r5		10.0			

Summary statistics:

Devianza Totale	18.234375
gdl Totale	15

Analisi della Varianza - Il caso a 1 via

Esempio - Il calcolo «a mano» (3/8)

Anova_a_mano.xlsx - Microsoft Excel

File Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Sviluppo

Garamond 14 A A

Incolla Appunti Carattere Allineamento Numeri Formattazione condizionale Formatta come tabella Stili cella Stili Inserisci Elimina Formato Ordina e filtra Trova e seleziona Modifica

G15 =DEV.Q(B11:E15)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Dataset														
2	repliche * trattamenti	T1	T2	T3	T4										
3	r1	10.0	8.5	12.0	10.0										
4	r2	9.5	8.0	9.5	9.5										
5	r3	11.0	9.0	11.0	11.0										
6	r4		9.0	11.5	10.0										
7	r5		10.0			Devianza Totale	18.234375								
8						gdl Totale	15								
9	Dataset "in assenza del caso"														
10	repliche * trattamenti	T1	T2	T3	T4										
11	r1	10.2	8.9	11.0	10.1										
12	r2	10.2	8.9	11.0	10.1										
13	r3	10.2	8.9	11.0	10.1										
14	r4		8.9	11.0	10.1										
15	r5		8.9			Devianza TRA	10.18020833								
16						gdl TRA	3								
17															
18															
19															
20															
21															

Se il caso non esistesse, allora non ci sarebbe variabilità nel campionamento e quindi ogni osservazione sarebbe uguale alla media del relativo trattamento.

Questo dataset «ideale» contiene esclusivamente la variabilità dovuta ai trattamenti, perciò possiamo usarlo per calcolare la devianza TRA i trattamenti esattamente come fatto prima. Riportiamo anche i relativi gdl.

Analisi della Varianza - Il caso a 1 via

Esempio - Il calcolo «a mano» (4/8)

Anova_a_mano.xlsx - Microsoft Excel

File Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Sviluppo

Garamond 14 A A

Incolla Appunti Carattere Allineamento Numeri Stili Formattazione condizionale Formatta come tabella Stili cella Inserisci Elimina Formato Ordina e filtra Trova e seleziona Modifica

I15 fx =G7-G15

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Dataset														
2	repliche * trattamenti	T1	T2	T3	T4										
3	r1	10.0	8.5	12.0	10.0										
4	r2	9.5	8.0	9.5	9.5										
5	r3	11.0	9.0	11.0	11.0										
6	r4		9.0	11.5	10.0										
7	r5		10.0			Devianza Totale	18.234375								
8						gdl Totale	15								
9	Dataset "in assenza del caso"														
10	repliche * trattamenti	T1	T2	T3	T4										
11	r1	10.2	8.9	11.0	10.1										
12	r2	10.2	8.9	11.0	10.1										
13	r3	10.2	8.9	11.0	10.1										
14	r4		8.9	11.0	10.1										
15	r5		8.9			Devianza TRA	10.18020833	Devianza ENTRO	8.05417						
16						gdl TRA	3	gdl ENTRO	12						
17						Varianza TRA	3.393402778	Varianza ENTRO	0.67118						
18															
19						Rapporto F	5.055871702	p-value	0.01186						
20								F tabulato	3.23887						
21															

Calcoliamo la Devianza Entro (=Devianza Totale- Devianza TRA).

Analisi della Varianza - Il caso a 1 via

Esempio - Il calcolo «a mano» (5/8)

Anova_a_mano.xlsx - Microsoft Excel

File Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Sviluppo

Garamond 14 A A

Incolla Appunti Carattere Allineamento Numeri Formattazione condizionale Formatta come tabella Stili cella Inserisci Elimina Formato Ordina e filtra Trova e seleziona Modifica

I16 fx =G8-G16

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Dataset														
2	repliche * trattamenti	T1	T2	T3	T4										
3	r1	10.0	8.5	12.0	10.0										
4	r2	9.5	8.0	9.5	9.5										
5	r3	11.0	9.0	11.0	11.0										
6	r4		9.0	11.5	10.0										
7	r5		10.0			Devianza Totale	18.234375								
8						gdl Totale	15								
9	Dataset "in assenza del caso"														
10	repliche * trattamenti	T1	T2	T3	T4										
11	r1	10.2	8.9	11.0	10.1										
12	r2	10.2	8.9	11.0	10.1										
13	r3	10.2	8.9	11.0	10.1										
14	r4		8.9	11.0	10.1										
15	r5		8.9			Devianza TRA	10.18020833	Devianza ENTRO	8.05417						
16						gdl TRA	3	gdl ENTRO	12						
17						Varianza TRA	3.393402778	Varianza ENTRO	0.67118						
18															
19						Rapporto F	5.055871702	p-value	0.01186						
20								F tabulato	3.23887						
21															

Calcoliamo i gdl Entro (=gdl Totale-gdl TRA).

Pronto

dati dati_2 Foglio3

90%

11:10 26/09/2014

Analisi della Varianza - Il caso a 1 via

Esempio - Il calcolo «a mano» (6/8)

Anova_a_mano.xlsx - Microsoft Excel

File Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Sviluppo

Garamond 14 A A

Incolla Appunti Carattere Allineamento Numeri Stili

Formattazione condizionale Formatta come tabella Stili cella

Inserisci Elimina Formato Ordina e filtra Trova e seleziona Modifica

G19 fx =G17/I17

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Dataset														
2	repliche * trattamenti	T1	T2	T3	T4										
3	r1	10.0	8.5	12.0	10.0										
4	r2	9.5	8.0	9.5	9.5										
5	r3	11.0	9.0	11.0	11.0										
6	r4		9.0	11.5	10.0										
7	r5		10.0			Devianza Totale	18.234375								
8						gdl Totale	15								
9	Dataset "in assenza del caso"														
10	repliche * trattamenti	T1	T2	T3	T4										
11	r1	10.2	8.9	11.0	10.1										
12	r2	10.2	8.9	11.0	10.1										
13	r3	10.2	8.9	11.0	10.1										
14	r4		8.9	11.0	10.1										
15	r5		8.9			Devianza TRA	10.18020833	Devianza ENTRO	8.05417						
16						gdl TRA	3	gdl ENTRO	12						
17						Varianza TRA	3.393402778	Varianza ENTRO	0.67118						
18															
19						Rapporto F	5.055871702	p-value	0.01186						
20								F tabulato	3.23887						
21															

dati dati_2 Foglio3

Pronto

90%

11:11 26/09/2014

A questo punto abbiamo tutte le informazioni per calcolare le due varianze e di conseguenza il valore del loro rapporto F .

Analisi della Varianza - Il caso a 1 via

Esempio - Il calcolo «a mano» (7/8)

Attraverso la funzione $INV.F.DS$ otteniamo il valore tabulato $F^{-1}_{(3,16)}(0,05)$. Rifiutiamo quindi l'ipotesi nulla con un livello di significatività pari al 5%.

Dataset				
repliche * trattamenti	T1	T2	T3	T4
r1	10.0	8.5	12.0	10.0
r2	9.5	8.0	9.5	9.5
r3	11.0	9.0	11.0	11.0
r4		9.0	11.5	10.0
r5		10.0		

Dataset "in assenza del caso"				
repliche * trattamenti	T1	T2	T3	T4
r1	10.2	8.9	11.0	10.1
r2	10.2	8.9	11.0	10.1
r3	10.2	8.9	11.0	10.1
r4		8.9	11.0	10.1
r5		8.9		

	Devianza Totale	gdl Totale
	18.234375	15

	Devianza TRA	gdl TRA	Devianza ENTRO	gdl ENTRO
	10.18020833	3	8.05417	12

	Varianza TRA	Varianza ENTRO
	3.393402778	0.67118

	Rapporto F	p-value
	5.055871702	0.01186

	F tabulato
	3.23887

Analisi della Varianza - Il caso a 1 via

Esempio - Il calcolo «a mano» (8/8)

Infine, attraverso la funzione DISTRIB.F calcoliamo il valore della probabilità p di commettere il cosiddetto «errore di I specie» (detto anche p -value), che rappresenta il livello di significatività osservato).

Dataset				
repliche * trattamenti	T1	T2	T3	T4
r1	10.0	8.5	12.0	10.0
r2	9.5	8.0	9.5	9.5
r3	11.0	9.0	11.0	11.0
r4		9.0	11.5	10.0
r5		10.0		

Dataset "in assenza del caso"				
repliche * trattamenti	T1	T2	T3	T4
r1	10.2	8.9	11.0	10.1
r2	10.2	8.9	11.0	10.1
r3	10.2	8.9	11.0	10.1
r4		8.9	11.0	10.1
r5		8.9		

	Devianza Totale	Devianza TRA	Devianza ENTRO
gdl Totale	18.234375	10.18020833	8.05417
gdl TRA	15	3	12
Varianza TRA		3.393402778	0.67118
Rapporto F		5.055871702	
p -value			0.01186
F tabulato			3.23887

Esempio - L'errore di prima specie (1/3)

Chiariamo ora il concetto di errore di prima specie con un esempio basato sull'analisi della varianza a 1 via.

- ✓ Sappiamo che l'errore di prima specie rappresenta la probabilità di rifiutare H_0 , quando questa invece è vera. Impostare un livello di significatività $\alpha = 0,05$ vuol dire quindi accettare di avere falsi positivi nel 5% dei casi.
- ✓ Detto in altre parole, fissare $\alpha = 0,05$ significa che (TEORICAMENTE), ripetendo 100 volte lo stesso esperimento e verificando 100 volte se è possibile rifiutare o no H_0 , 5 volte accadrà di giungere ad una conclusione errata.

Esempio - L'errore di prima specie (2/3)

Illustriamo con un esempio pratico le affermazioni precedenti:

- 1) creiamo, attraverso una procedura di generazione casuale (ad esempio utilizzando opportune funzioni di SPSS), 100'000 dataset con 12 valori ciascuno, tutti estratti da una popolazione $\sim N(10, 1)$;
- 2) immaginiamo che ciascun dataset rappresenti l'esito di un esperimento, in cui sono stati messi a confronto 3 livelli di un trattamento, effettuando 4 osservazioni (o repliche) per ciascun livello;
- 3) applichiamo l'ANOVA a 1 via su tutti i 100'000 dataset per verificare $H_0: \mu_1 = \mu_2 = \mu_3$;
- 4) attraverso i passaggi visti prima, calcoliamo i 100'000 p-value di ciascun esperimento;
- 5) sappiamo a priori che H_0 è vera, dal momento i dati sono stati tutti estratti dalla medesima popolazione;
- 6) fissato $\alpha = 0,05$, ci aspettiamo di trarre conclusioni errate dall'ANOVA nel 5% dei casi e, in effetti, è esattamente quello che accade (ordinando le colonne per p-value crescente) ...

Analisi della Varianza - Il caso a 1 via

Esempio - L'errore di prima specie (3/3)

100000anova.xlsx - Microsoft Excel

FileHomeInserisciLayout di paginaFormuleDatiRevisioneVisualizzaSviluppo

Incolla

Appunti

Garamond12A⁺A⁻

GCS

Carattere

Unisci e allinea al centro

Allineamento

Numero

%000

0,00

0,00

Numeri

Formattazione condizionale

Formatta come tabella

Stili cella

Stili

Inserisci

Elimina

Formato

Celle

Ordina e filtra

Trova e seleziona

Modifica

S4994fx=DISTRIB.F(R4994;2;9)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	T1_r1	T1_r2	T1_r3	T1_r4	T2_r1	T2_r2	T2_r3	T2_r4	T3_r1	T3_r2	T3_r3	T3_r4	dev_tot	dev_tra	dev_entro	var_tra	var_entro	F	p_value	
4985	9.96	9.63	10.19	8.06	9.87	12.26	11.05	11.77	9.92	10.17	10.79	10.87	12.98	6.31	6.67	3.16	0.74	4.260227	0.049904	
4986	8.49	8.16	9.97	9.55	9.91	11.38	10.83	10.06	11.72	11.41	11.75	9.04	16.97	8.25	8.72	4.13	0.97	4.259274	0.049929	
4987	9.65	10.72	10.15	9.12	8.16	8.53	8.54	9.07	11.75	8.27	11.45	11.36	19.10	9.28	8.81	4.64	1.09	4.258703	0.049943	
4988	9.96	10.38	9.21	8.65	10.01	11.81	9.52	11.88	9.05	9.07	9.59	7.49	16.94	8.24	8.71	4.12	0.97	4.258371	0.049952	
4989	9.16	9.81	9.36	9.99	9.26	9.44	9.93	9.50	10.66	9.95	9.85	10.23	2.11	1.02	1.08	0.51	0.12	4.258350	0.049952	
4990	11.96	11.12	11.09	11.02	10.39	9.33	9.64	10.18	10.26	11.11	10.42	8.52	9.62	4.68	4.94	2.34	0.55	4.258169	0.049957	
4991	10.45	10.28	8.99	10.35	8.76	9.09	8.38	9.90	9.40	11.74	10.49	10.90	10.71	5.21	5.50	2.60	0.61	4.258082	0.049959	
4992	9.21	10.38	10.06	11.38	11.70	9.63	11.35	11.11	8.72	9.77	9.34	9.53	10.70	5.20	5.50	2.60	0.61	4.257399	0.049961	
4993	8.91	10.44	8.98	7.98	10.27	12.12	11.11	10.69	10.17	12.08	9.49	12.19	20.55	9.99	10.56	4.99	1.17	4.257136	0.049984	
4994	10.49	11.15	9.42	10.88	8.51	9.92	8.92	8.11	10.33	8.77	10.60	9.50	10.89	5.29	5.59	2.65	0.62	4.256397	0.050002	
4995	8.24	8.52	9.51	9.19	10.83	10.59	10.51	9.58	11.00	8.57	10.28	9.77	9.88	4.80	5.08	2.40	0.56	4.254882	0.050041	
4996	9.79	11.28	10.62	10.92	10.32	9.35	8.57	10.43	9.10	9.80	9.54	8.66	8.31	4.04	4.27	2.02	0.47	4.253333	0.050081	
4997	10.25	10.40	9.93	12.32	9.04	9.23	7.83	9.53	8.96	9.36	10.55	10.40	13.61	6.61	7.00	3.31	0.78	4.253292	0.050082	
4998	8.97	9.43	9.54	9.89	11.61	10.02	10.75	12.55	10.32	10.65	8.69	10.78	13.22	6.43	6.80	3.21	0.76	4.252924	0.050092	
4999	8.99	9.70	8.07	10.04	10.34	11.18	11.81	10.26	9.48	9.93	11.58	10.42	12.37	6.01	6.36	3.01	0.71	4.252869	0.050093	
5000	9.95	11.00	8.90	9.64	10.27	10.52	10.66	9.74	10.68	11.11	11.53	10.99	6.15	2.99	3.16	1.49	0.35	4.252812	0.050095	
5001	10.32	10.23	10.64	11.13	9.64	8.50	9.19	10.32	10.03	9.93	10.98	9.96	5.90	2.87	3.03	1.43	0.34	4.252076	0.050114	
5002	10.43	9.61	10.31	11.38	10.98	9.59	8.92	9.23	8.29	8.41	10.02	8.03	12.65	6.15	6.51	3.07	0.72	4.251724	0.050123	
5003	10.80	10.58	8.84	10.85	10.74	10.83	9.07	10.68	7.02	10.18	9.23	8.51	12.00	5.83	6.17	2.92	0.60	4.251606	0.050123	

Pronto

15:3226/09/2014

