L'analisi della varianza

Introduzione e concetti generali

Giovanni Battista Flebus

Scopo dell'ANOVA

L'analisi della varianza (ANOVA, ANalysi Of VAriance) è una tecnica statistica che permette di valutare se le medie di due o più gruppi sono uguali fra loro.

Requisiti

1. La variabile Dipendente è misurata su una scala a intervalli
2. Ha una distribuzione normale
3. La classificazione è fatta in modo indipendente dai dati osservati (esiste in precedenza e non è influenzata dei valori osservati)
4. Le varianze all'interno dei gruppi sono omogenee (simili fra di loro)

La variabile indipendente (classificazione in più gruppi) è una misurazione a livello di scala nominale

Meccanismo

L' ANOVA si basa su due principi:

- (1) si può stimare la varianza della popolazione in **due modi diversi**, che tengano conto della suddivisione in gruppi
- (2) Si possono confrontare le due stime e verificare se sono estratte dalla stessa popolazione

Le ipotesi di ricerca

- Le due ipotesi di ricerca sono le seguenti
- H0 : le medie dei k gruppi sono uguali (a parte la variabilità stocastica)
- H1 : almeno una delle medie dei k gruppi è diversa dalle altre

Ulteriori esplorazioni

- Se il test statistico permette di concludere che c'è almeno un gruppo diverso dagli altri, si possono applicare altre tecniche per individuare i gruppi diversi

Esempio preliminare

- In un campione di studenti, si rileva il senso di benessere (un test, scala a intervalli) per vedere se le bocciature a scuola hanno influenza su tale tratto.
- Il benessere si rileva con un test (BeSco, Questionario di Benessere Scolastico)
- Le bocciature a scuola (nessuna, una o due), anche se sono una scala a rapporti, sono considerate qui come una classificazione e quindi come scala nominale.
- La frequenza dei tre gruppi è la seguente

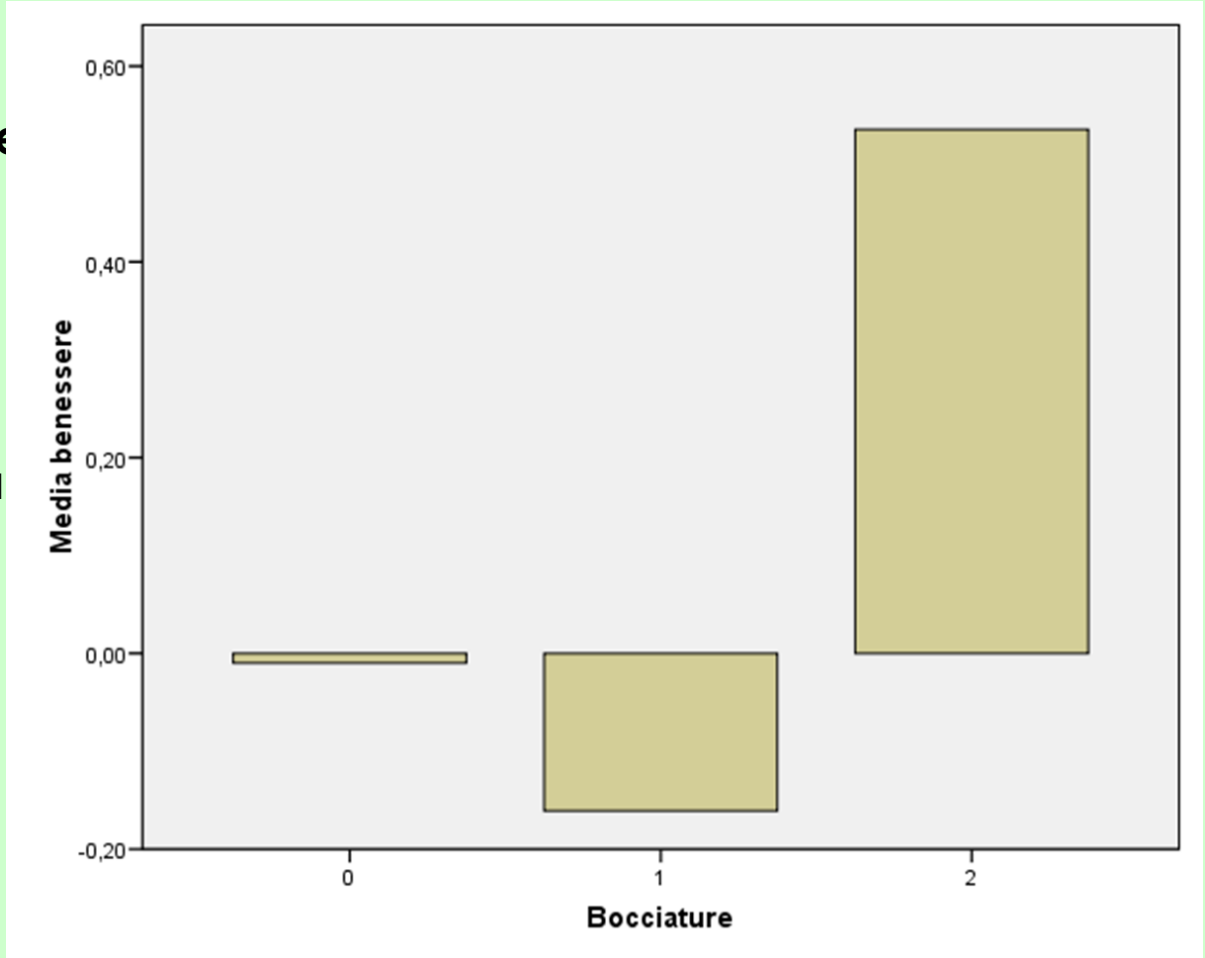
Ecco i dati del campione

Bocciature

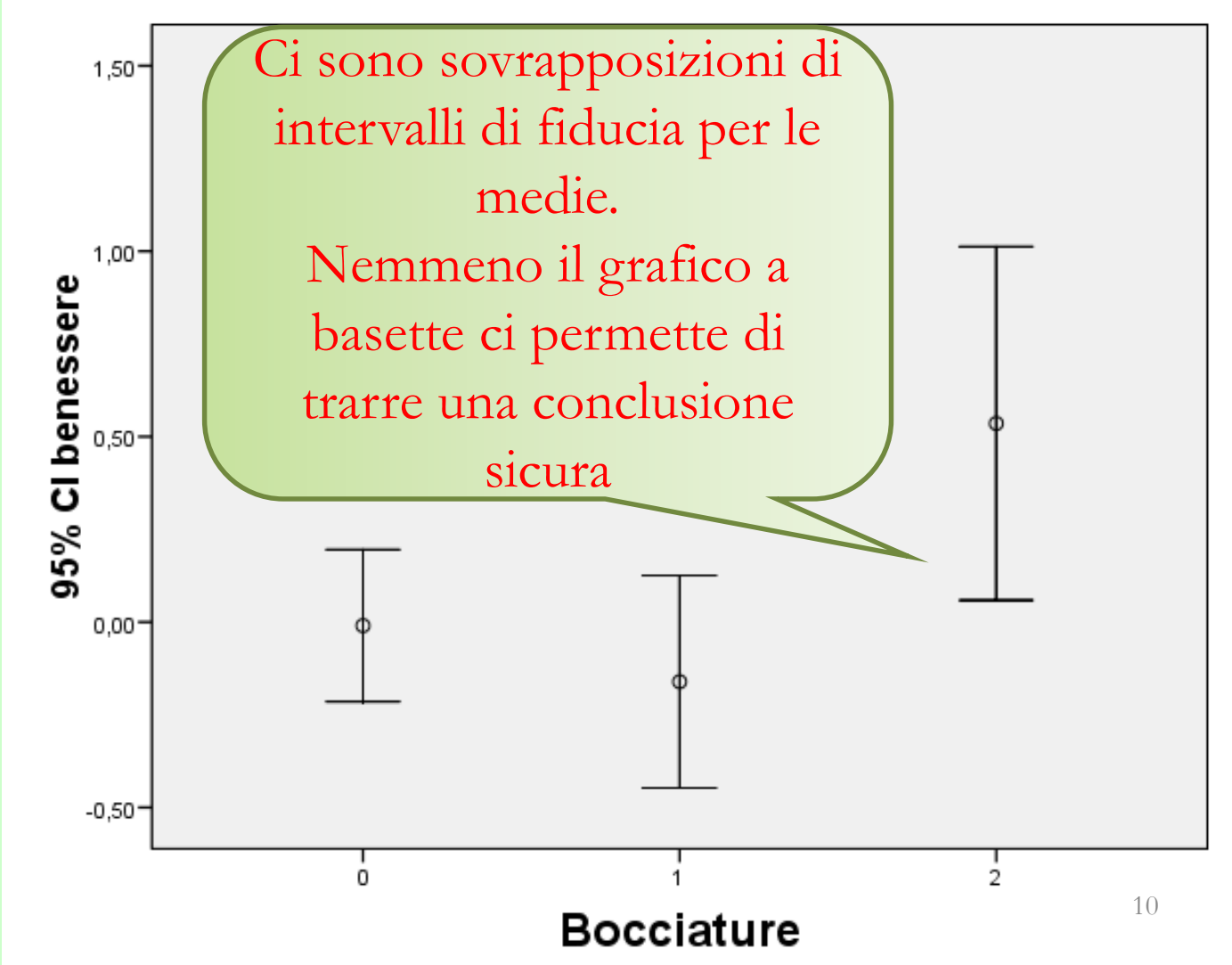
		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	0	87	55,4	55,4	55,4
	1	51	32,5	32,5	87,9
	2	19	12,1	12,1	100,0
	Totale	157	100,0	100,0	

Il punteggio di benessere nei tre gruppi pare diverso.

Ma le differenze sono attribuibili alla variabilità stocastica o sono veramente consistenti?



Esaminiamo il grafico a basette



Esaminiamo i risultati dell'ANOVA

ANOVA univariata

benessere

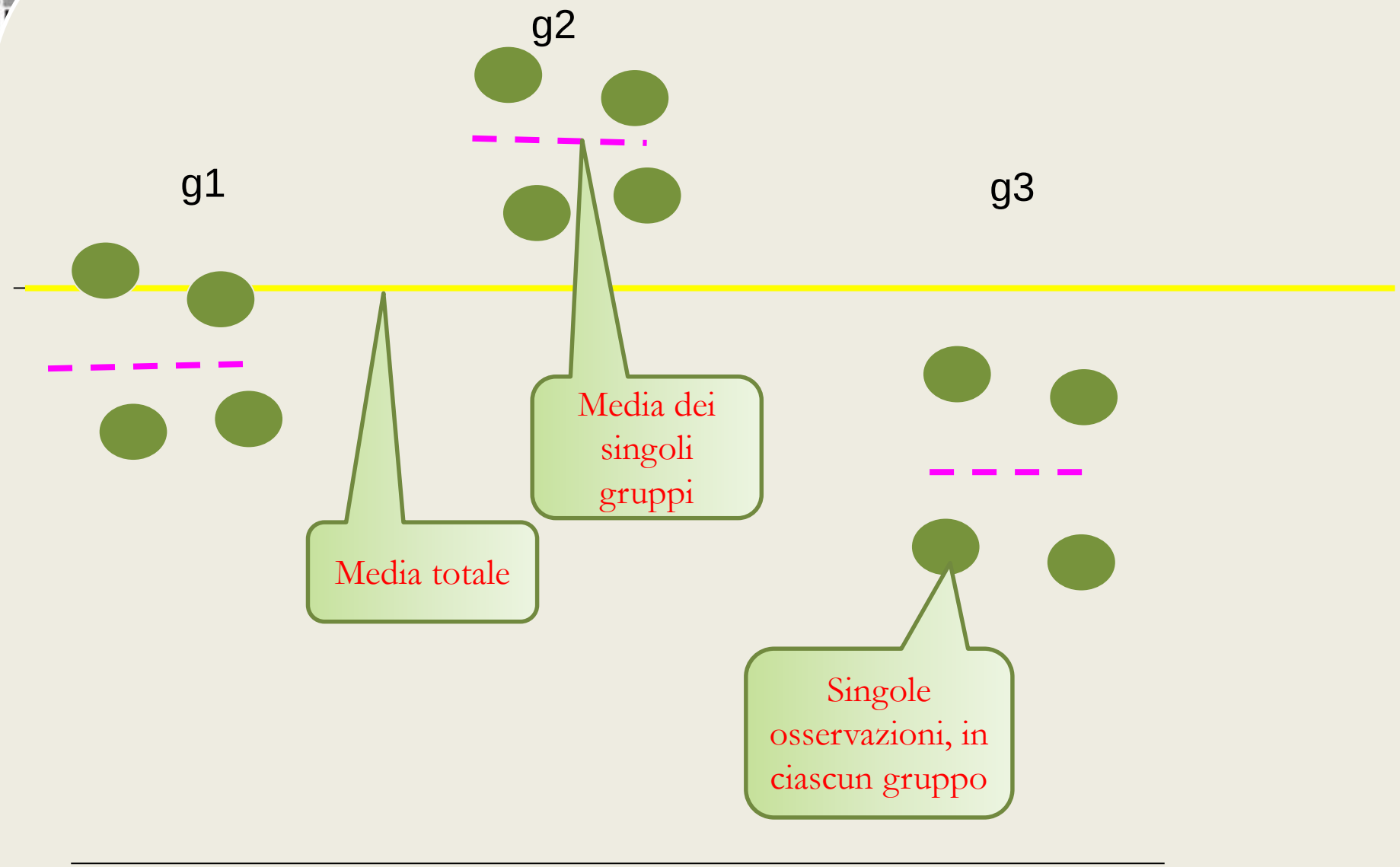
	Somma dei quadrati	df	Media dei quadrati	F	Sig.
Fra gruppi	6,767	2	3,384	3,495	,033
Entro gruppi	149,111	154	,968		
Totale	155,878	156			

Questa tabella è prodotta dall'applicazione dell'ANOVA ai dati, che ci permette di passare alla conclusione…

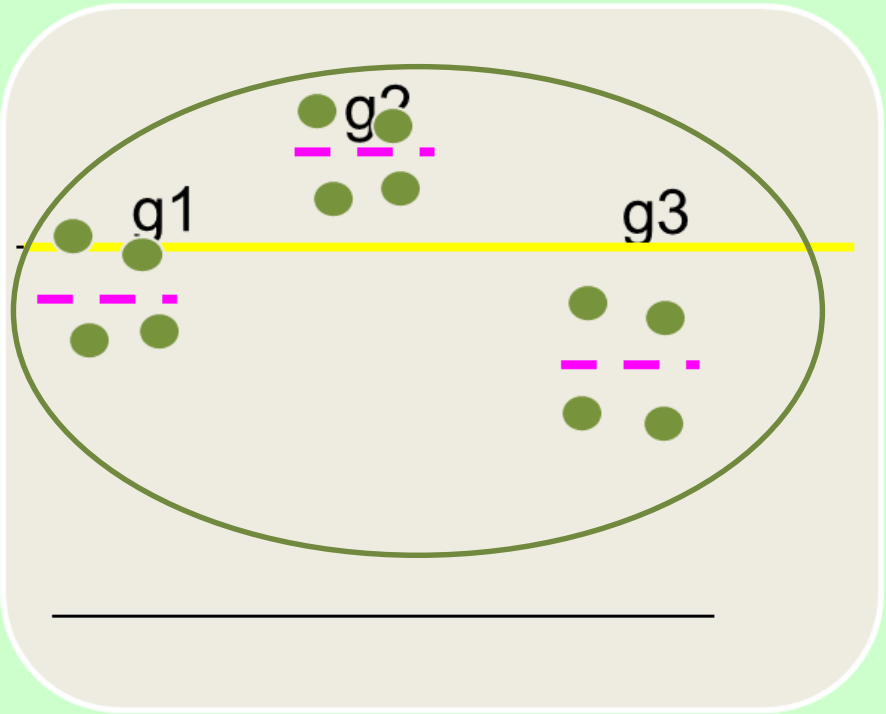
La significatività è il valore di probabilità dell'ipotesi nulla di uguaglianza delle medie: la probabilità bassi ci porta a escludere che sia vera. Perciò le medie dei tre gruppi non sono uguali.

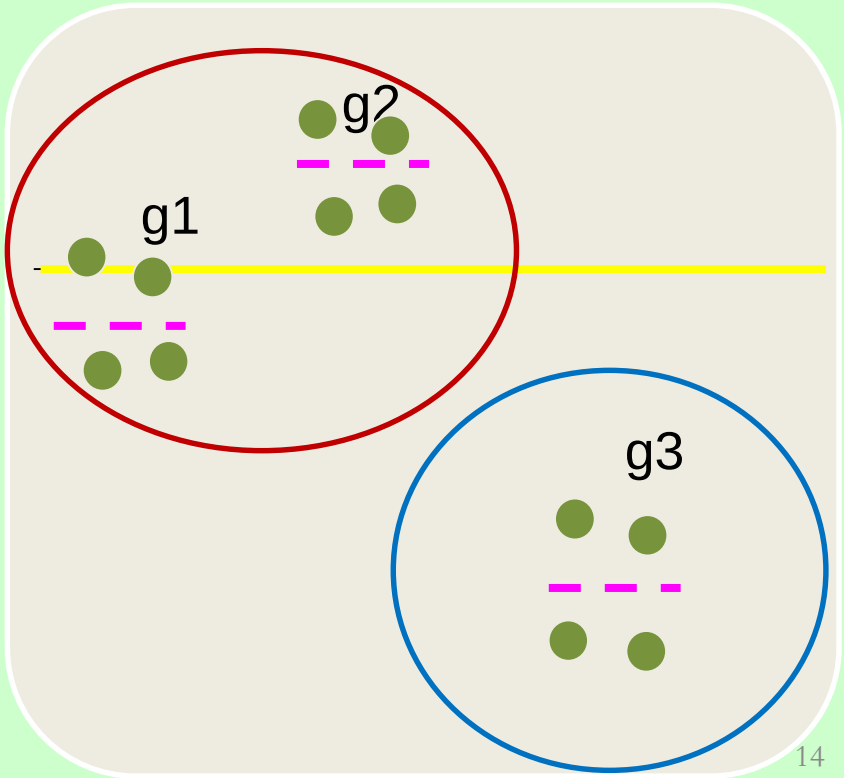
Principio dell'ANOVA

- Si può stimare la varianza della popolazione in due modi diversi e confrontare le due stime
- Primo metodo: calcolare la varianza delle k medie come se fossero k osservazioni
- Secondo metodo: calcolare la varianza media, usando tutte le osservazioni, eliminando però da ciascuna osservazione l'influenza del proprio gruppo.

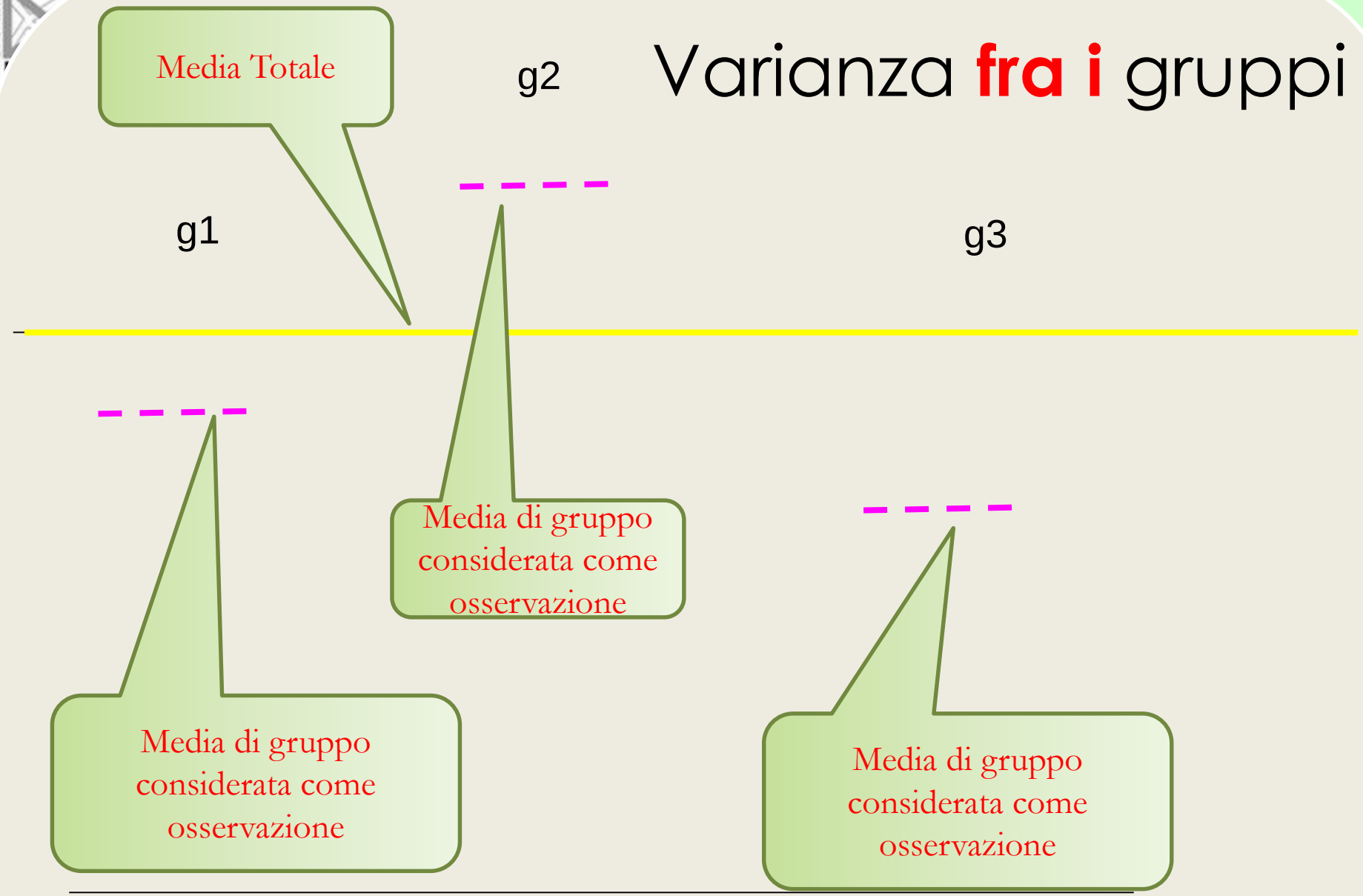

g2
g1
g3
Media dei singoli gruppi
Media totale
Singole osservazioni, in ciascun gruppo

H_0 (uguali) --- H_1 (diversi) ?

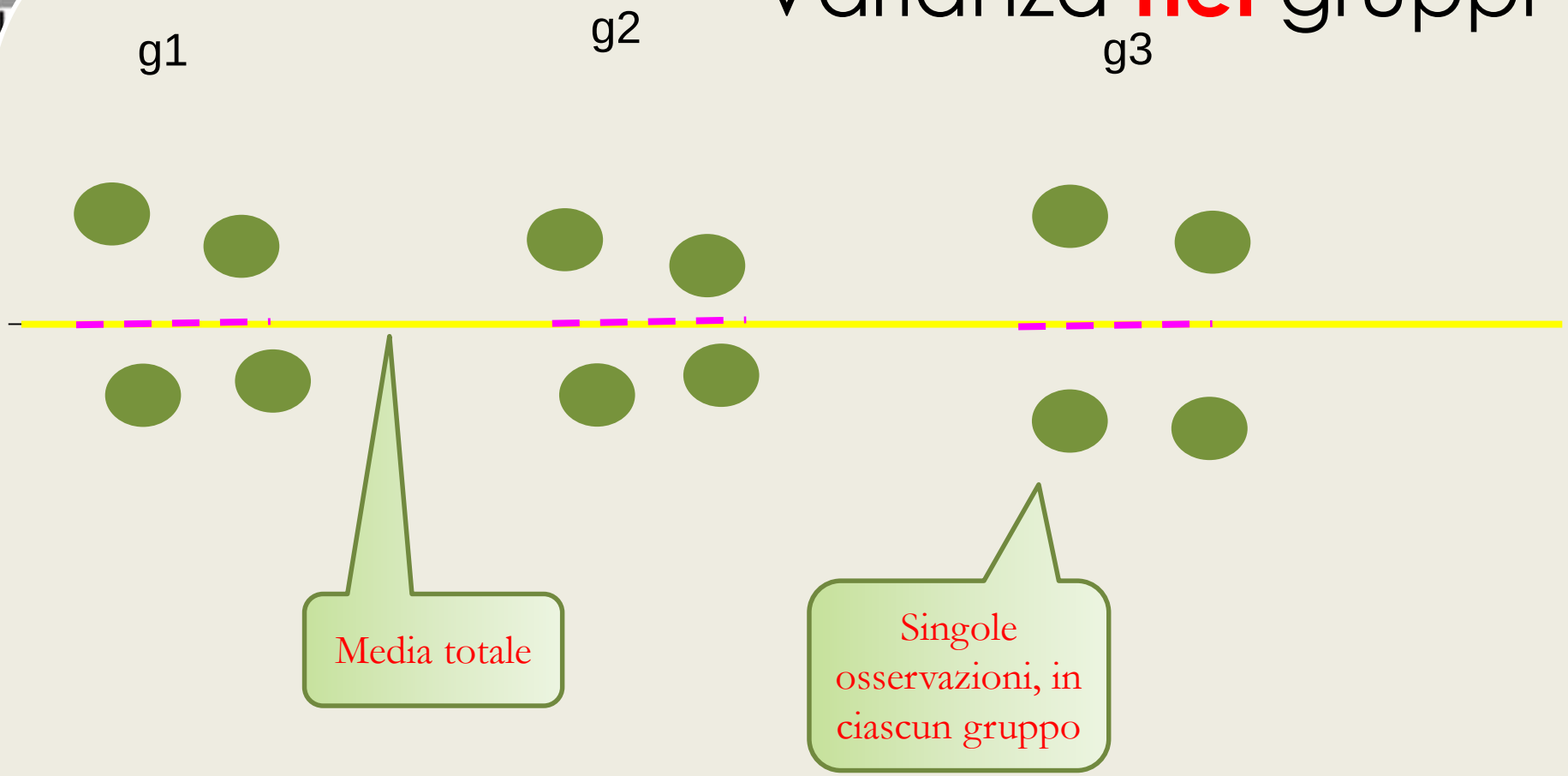




Varianza **fra i** gruppi



Varianza **nei** gruppi



Piccolo esempio numerico

- Un ricercatore pensa che il tempo passato a muoversi in città sia di detrimento per il rendimento accademico degli studenti universitari. Ha osservato il numero di esami di 12 studenti, suddivisi in tre gruppi secondo l'uso di trasporto per andare in facoltà:
- A) prendono i mezzi
- B) Hanno un loro mezzo (moto – auto)
- C) vivono in zona e quindi vanno a piedi

studente	gruppo	N_esami
s1	Mezzi pubblici	2
s2	Mezzi pubblici	4
s3	Mezzi pubblici	4
s4	Mezzi pubblici	6
media		4
s5	Mezzi propri	4
s6	Mezzi propri	5
s7	Mezzi propri	7
s8	Mezzi propri	8
Media		6
s9	Residenti	5
s10	Residenti	7
s11	Residenti	8
s12	Residenti	8
media		7
Media totale		5,7

	gruppo 1	gruppo 2	gruppo 3
	2	4	5
	4	5	7
	4	7	8
	6	8	8
media dei gruppi	**4**	**6**	**7**

Varianza di errore

	gruppo 1	gruppo 2	gruppo 3
	2	4	5
	4	5	7
	4	7	8
	6	8	8
media dei gruppi	**4**	**6**	**7**

Varianza fra i campioni

Le medie e varianze dei tre gruppi

Report

num_esami

gruppo	Media	N	Varianza
1 mezzi pubblici	4,00	4	2,667
2 auto	6,00	4	3,333
3 residenti	7,00	4	2,000
Totale	5,67	12	3,879

Consideriamo gli elementi utili

Report

num_esami

gruppo	Media	N	Varianza
1 mezzi pubblici	4,00	4	2,667
2 auto	6,00	4	3,333
3 residenti	7,00	4	2,000
Totale	5,67	12	3,879

1 Le medie dei gruppi

2 La media totale

3 Le varianze dei gruppi

Calcoliamo la varianza **fra** i gruppi

Report

num_esami

gruppo	Media	N	Varianza
1 mezzi pubblici	4,00	4	2,667
2 auto	6,00	4	3,333
3 residenti	7,00	4	2,000
Totale	5,67	12	3,879

1 Le medie dei gruppi

2 La numerosità dei gruppi è 3

3 La media totale

Calcoliamo la varianza delle medie dei gruppi (varianza **fra** i k gruppi $\Sigma(X_i-M)^2/(n-1)$

Report

num_esami

gruppo	Media	N	Varianza
1 mezzi pubblici	4,00	4	2,667
2 auto	6,00	4	3,333
3 residenti	7,00	4	2,000
Totale	5,67	12	3,879

Varianza fra i gruppi =
$[(4-5{,}67)^2+(6-5{,}67)^2+(7-5{,}67)^2]\ /\ 2 =$
$(2{,}7889+0{,}1089+1{,}7689)/2 = \ 2{,}3335$

Varianza della popolazione o varianza della distribuzione campionaria delle medie?

La varianza delle k medie **(s^2)** è però la varianza della distribuzione campionaria delle medie:

s^2 / n

A noi serve la varianza della popolazione: **s^2**
Perciò dobbiamo moltiplicare il valore per **n** (numerosità nei gruppi):

Calcoliamo la varianza della popolazione con la stima della varianza fra i gruppi

Report

num_esami

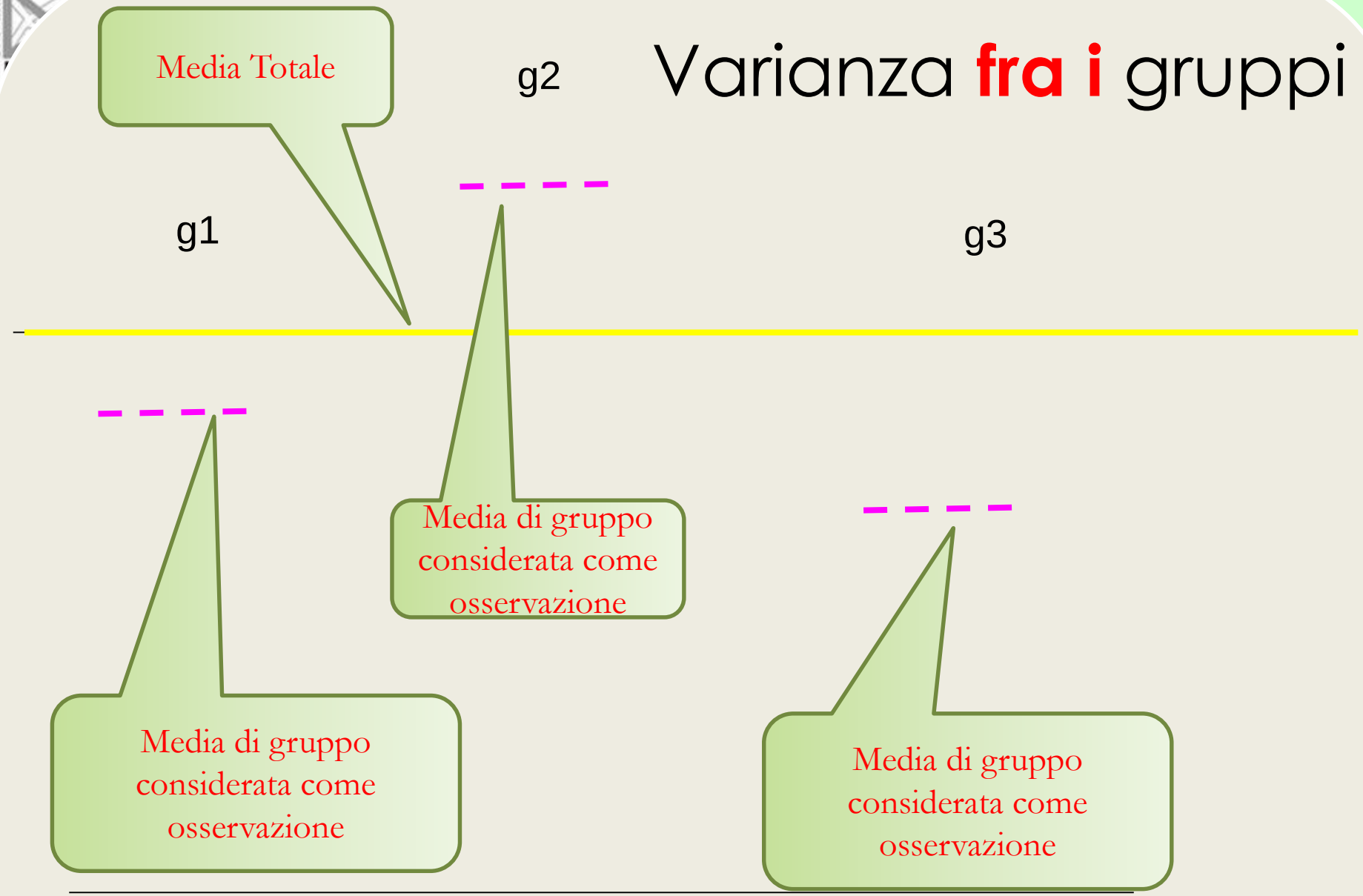
gruppo	Media	N	Varianza
1 mezzi pubblici	4,00	4	2,667
2 auto	6,00	4	3,333
3 residenti	7,00	4	2,000
Totale	5,67	12	3,879

Varianza fra i gruppi= $[(4-5{,}67)^2+(6-5{,}67)^2+(7-5{,}67)^2]$ / 2 =
(2,7889+0,1089+1,7689)/2= 2,3335 ossia

Varianza delle distribuzione campionaria delle medie (s^2/n)

Varianza della popolazione = n S^2 → 2,3335 x 4 = 9,3334

Varianza **fra i** gruppi



Calcoliamo la varianza della popolazione con la stima della varianza dentro i gruppi

Report

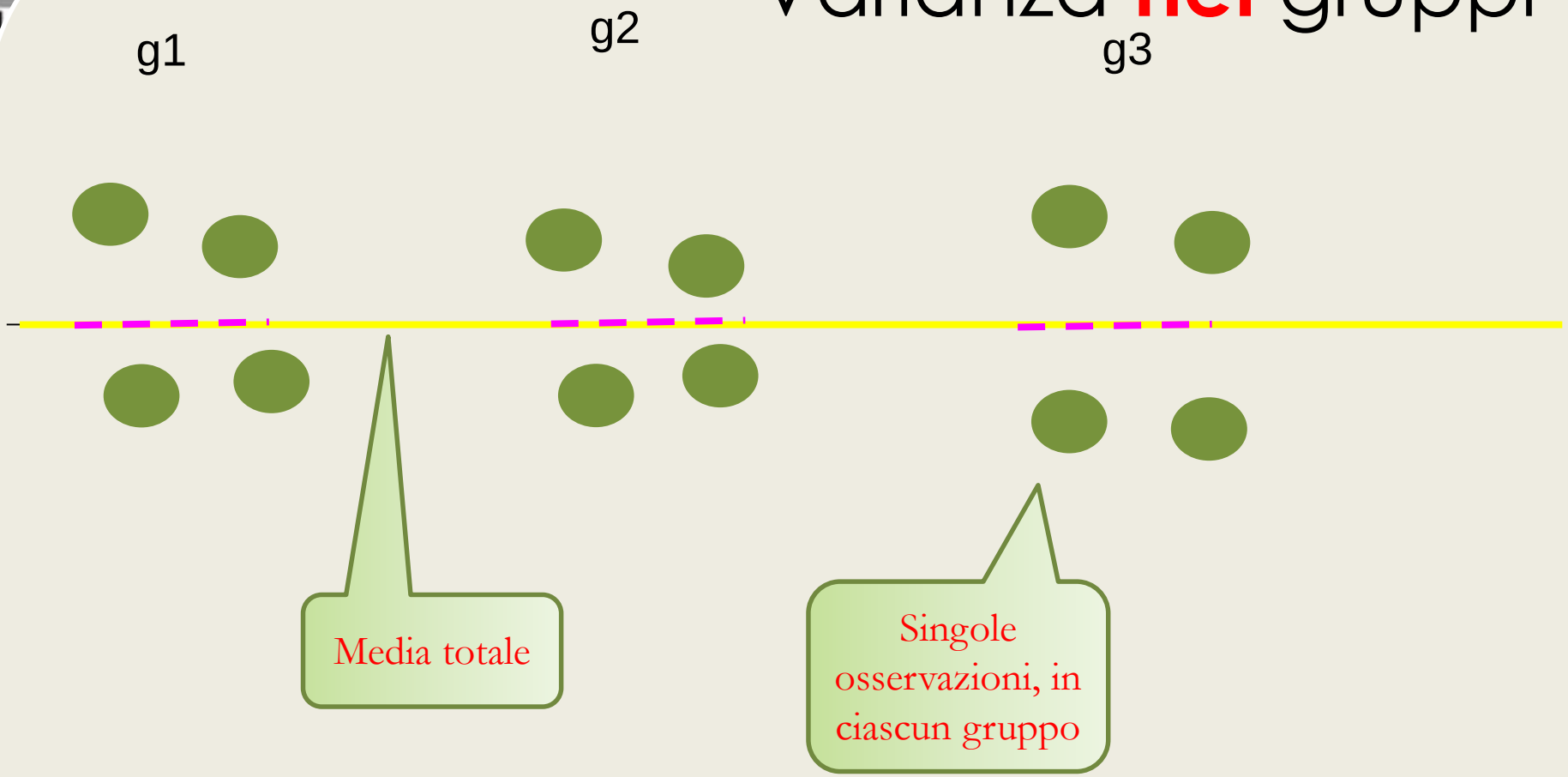
num_esami

gruppo	Media	N	Varianza
1 mezzi pubblici	4,00	4	2,667
2 auto	6,00	4	3,333
3 residenti	7,00	4	2,000
Totale	5,67	12	3,879

Calcoliamo la media delle varianze nei gruppi: 2,667+3,333+2,000=8,00

Media della varianza nei gruppi 8,00/3= 2,667

Varianza **nei** gruppi



I gradi di libertà

- I gradi di libertà sono dati da
- (1) Numero di gruppi -1 per la varianza fra i gruppi
- (2) Numero di osservazioni meno i gruppi, per la varianza nei gruppi.
- Nel nostro caso, 3-1= 2 gl per la varianza fra i gruppi
- 12-3 = 9 gl per la varianza nei gruppi

Otteniamo il valore di F

- Il rapporto fra le due stime della varianza della popolazione (una nei gruppi e l'altra fra i gruppi) ha una distribuzione descritta dalla variabile casuale F di Fisher Snedecor con gl1 e gl2 gradi di libertà.

Nel nostro caso otteniamo
F= 9,334/ 2,666 = 3,500 con 2 e 9 gradi di libertà.

Le tavole di F ci dicono che il valore 3,500 ricade al di sotto della zona critica e perciò accettiamo l'ipotesi nulla di uguaglianza delle medie dei tre gruppi

$\alpha = 0.05$

NUMERATORE

DEN.	1	2	3	4	5	6	7
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14

Grafico di F con 2 e 9 g.l.

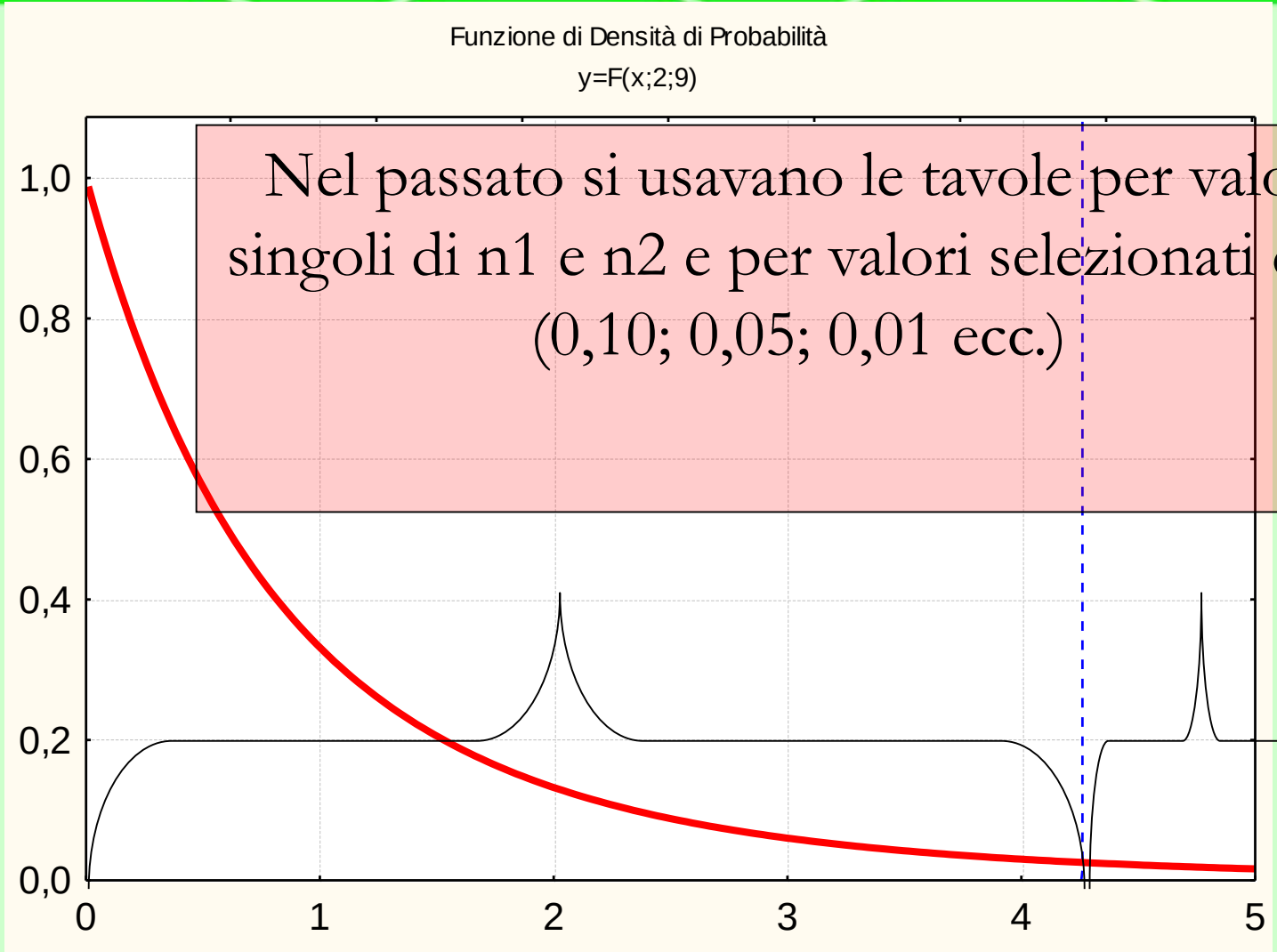
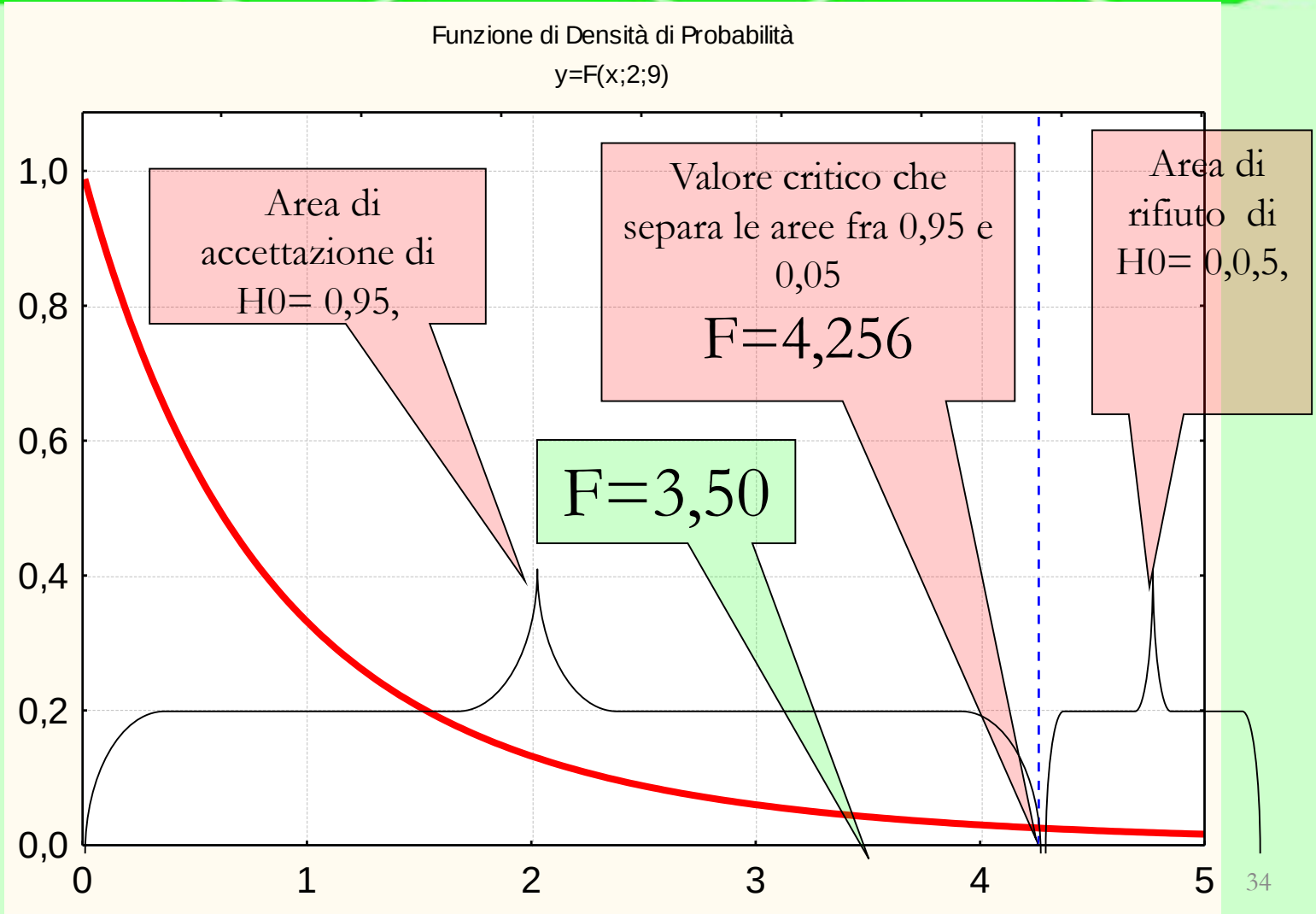


Grafico di F con 2 e 9 g.l.



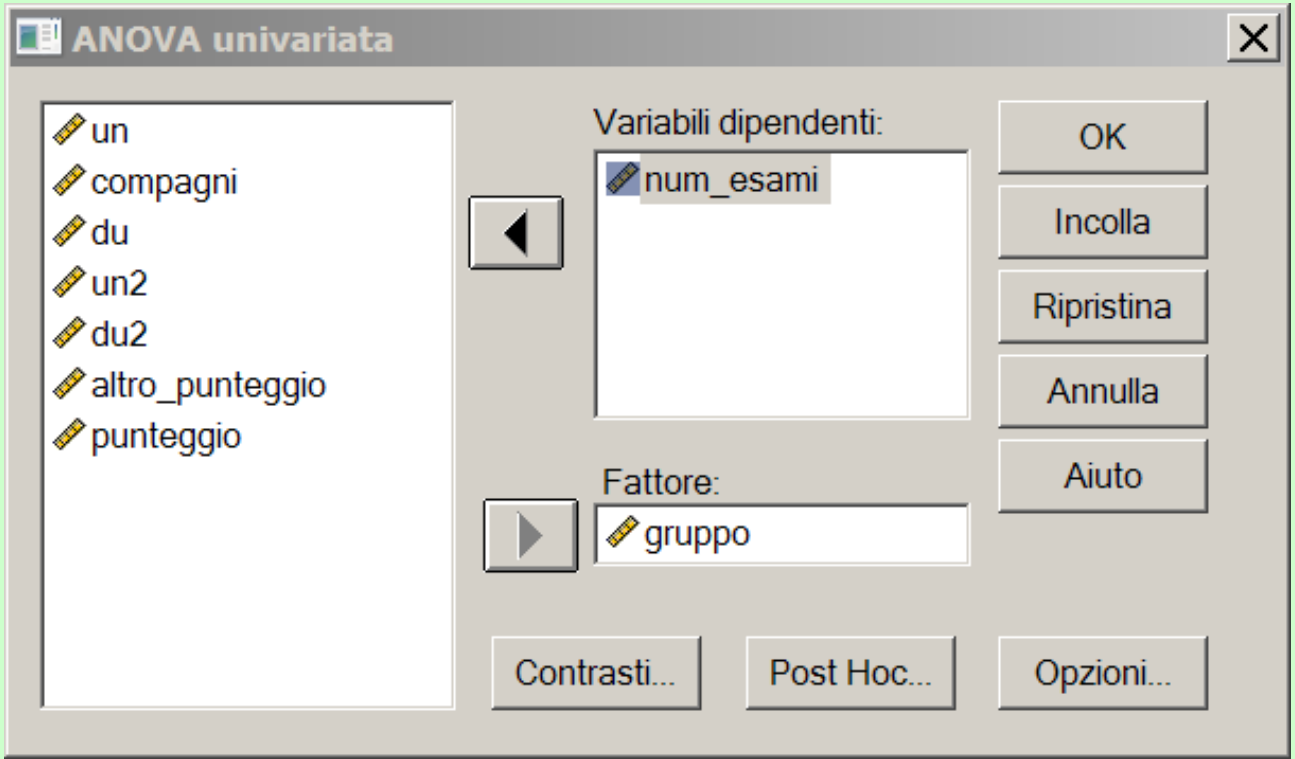
Per il calcolo con SPSS

Le due varianze sono però calcolate in modo diverso da quello che è stato presentato: si parte dalla somma dei quadrati (distanza dell'osservazione dalla media) (devianza in italiano, Sum of squares in inglese) dentro e fra i gruppi, divisi per i rispettivi gradi di libertà.

Il rapporto F è sempre stampato usando la devianza nei e fra i gruppi. La loro somma è uguale alla devianza totale

Passiamo a SPSS

- Selezioniamo il menu Analizza->Confronta Medie-> ANOVA univariata. Compare questo finestra. Inseriamo la variabile Gruppo come fattore, e il numero di esami come variabile dipendente



Output di SPSS per l'ANOVA

Guardiamo solo una parte della tabella

ANOVA univariata

	df	Media dei quadrati	F	Sig.
	2	9,333	3,500	,075
	9	2,667		
	11			

Significatività di F

Valore F calcolato

Gradi di liberta FRA e DENTRO i gruppi, quelli totali

Le due varianze calcolate nei due modi diversi

Output di SPSS per l'ANOVA

ANOVA univariata

	df	Media dei quadrati	F	Sig.
	2	9,333	3,500	,075
	9	2,667		
	11			

- Dati che abbiamo calcolato in precedenza:
- F= 9,334/ 2,666 = 3,500 con 2 e 9 gradi di libertà.

Significatività di F

Valore F calcolato

Gradi di liberta FRA e DENTRO i gruppi, quelli totali

Le due varianze calcolate nei due modi diversi

Il metodo di calcolo seguito è diverso

- Le due varianze appena confrontate sono di solito concepite come un rapporto di scarti quadrati, divisi per i rispettivi gradi di libertà, per produrre delle stime delle varianze
- Per rendere questo metodo di calcolo utilizzabile con gruppi di diversa numerosità, si procede ricordando il concetto di devianza totale, suddivisa in devianza fra i gruppi e devianza nei gruppi

Scomposizione della variabilità totale

La variabilità totale è descritta da SQT, ovvero Devianza totale:

$$SQT = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y} \right)^2$$

Scomposizione della variabilità totale

La variabilità fra i gruppi è descritta con la formula seguente

Devianza fra i gruppi:

$$SQF = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

Scomposizione della variabilità totale

La variabilità nei (o dentro i) gruppi è descritta dalla SQE detta anche variabilità dell'errore:

Devianza dentro i gruppi:

$$SQE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i \right)^2$$

Dalle devianze alle due varianze

- Le due varianze (dentro e fra i gruppi) sono quindi calcolate come rapporti fra due somme di quadrati, divise dai rispettivi gradi di libertà.

Test F per ANOVA

I risultati del test F per la ANOVA sono generalmente presentati in una tabella come questa:

Output di SPSS per l'ANOVA

ANOVA univariata

num_esami

	Somma dei quadrati	df	Media dei quadrati	F	Sig.
Fra gruppi	18,667	2	9,333	3,500	,075
Entro gruppi	24,000	9	2,667		
Totale	42,667	11			

Significatività di F

Valore F calcolato

Gradi di liberta FRA e DENTRO i gruppi, quelli totali

Le due varianze calcolate nei due modi diversi

Calcolare la media dei quadrati fra i gruppi

- Media dei quadrati $= \frac{devianza}{gradi\ di\ libertà}$

- $MQF = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2}{gl}$

Calcolare la media dei quadrati nei gruppi

- Media dei quadrati = $\frac{devianza}{gradi\ di\ libertà}$

- $MQE = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_j} (y_{ij} - \bar{y})^2}{gl}$

La devianza

Si usa il termine devianza per indicare

la somma dei quadrati delle distanze dalla media.

In inglese *Sum of Squares*

- La varianza stimata della popolazione si ottiene dividendo la devianza per il numero dei gradi di libertà
  - Si usano i termini inglesi within (W) per indicare la devianza nei gruppi e between (B) per indicare la devianza fra i gruppi