TECNICHE DI ANALISI DEI DATI

AA 2018/2019

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

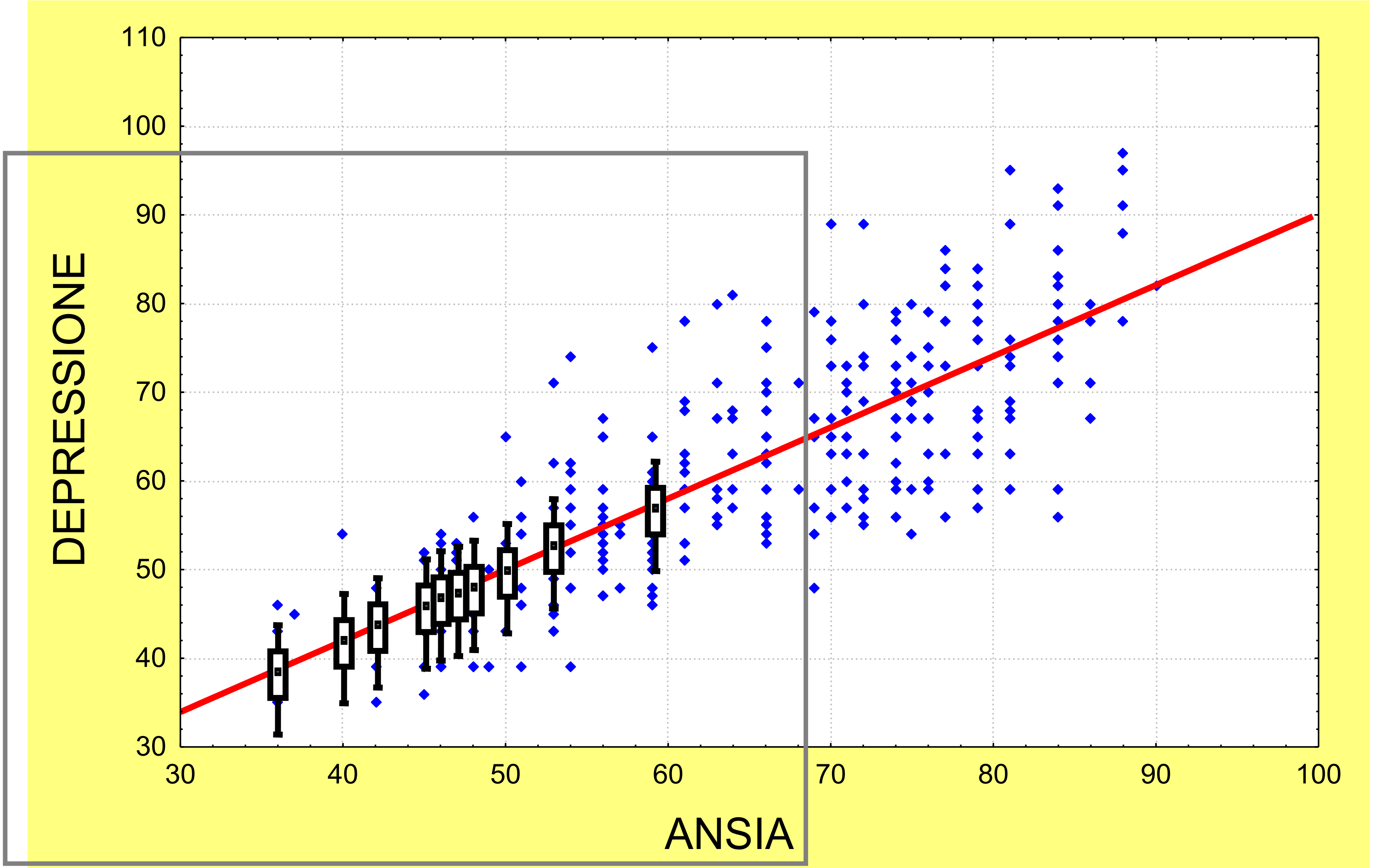
https://goo.gl/xY15fR

ANOVA

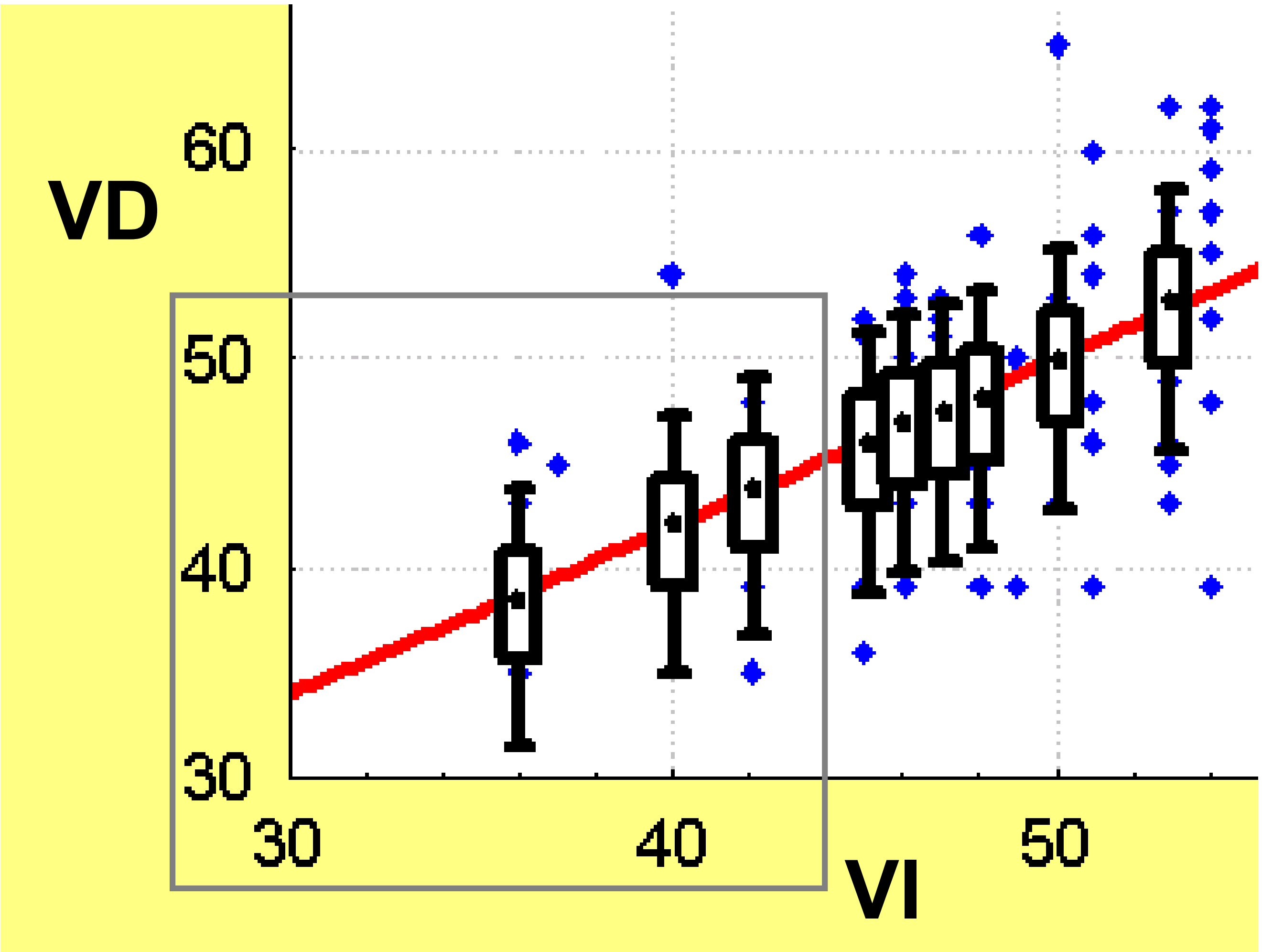
Quando la relazione causale chiama in causa **variabili indipendenti** (**VI**) di tipo **qualitativo** (**N** o **O**), mentre la **variabile dipendente** (**VD**) è di tipo **quantitativo** (**I** o **R**) l'analisi che può essere impiegata è l'**ANALISI DELLA VARIANZA** (**ANOVA**).

Anche in questo caso (come nella regressione) l'obiettivo è quello di voler verificare se la **capacità di prevedere** i valori di una data variabile **Y**, **E(Y)**, aumenta conoscendo i valori assunti da una data variabile **X**; ovvero se nei diversi livelli della/e variabile/i indipendente la variabile dipendente si distribuisce in modo differente.

ANOVA


110
100
90
80
70
60
50
40
30
DEPRESSIONE
30
40
50
60
70
80
90
100
ANSIA

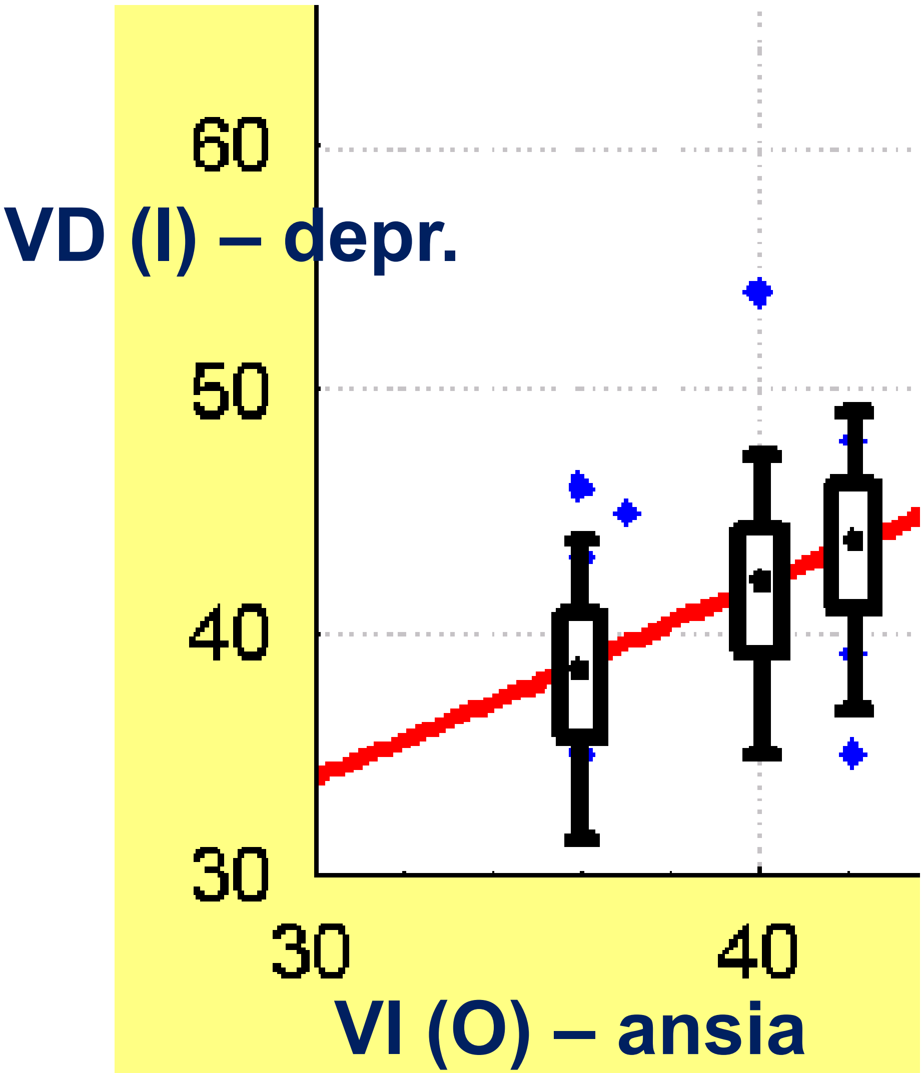
ANOVA



Trasformiamo la variabile indipendente e consideriamola come variabile misurata su scala ordinale (O)

Consideriamo di questa variabile solo un numero ristretto di valori (es. tre: 36, 40 e 42)

ANOVA



Questa situazione è quella dalla quale si parte quando abbiamo una VI di tipo N o O e una VD ad I o R e vogliamo confrontare la distribuzione della VD nei gruppi determinati dai livelli della VI.

Per effettuare il confronto dobbiamo scegliere il test statistico da applicare.

Es.

- ***t test* ? *Chi-quadrato* ? *test W* ? *ANOVA* ?**

ANOVA

L'analisi della varianza **ANOVA** si basa sulla scomposizione della variabilità totale in due parti:

I PARTE) VAR. DOVUTA ALL'EFFETTO (VI)
varianza **tra-gruppi** (*between*) o **spiegata**

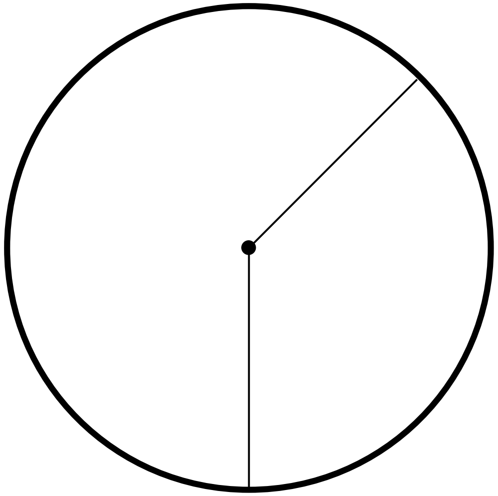
II PARTE)
VAR. DOVUTA ALLA DIVERSITÀ DEI SOGGETTI
varianza **entro i gruppi** (*within*) o
varianza **residua** o
varianza **casuale**

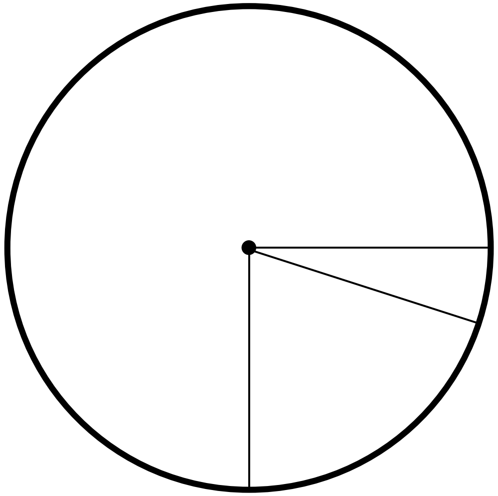
$$SS_{totale} = SS_{spiegata} + SS_{errore}$$

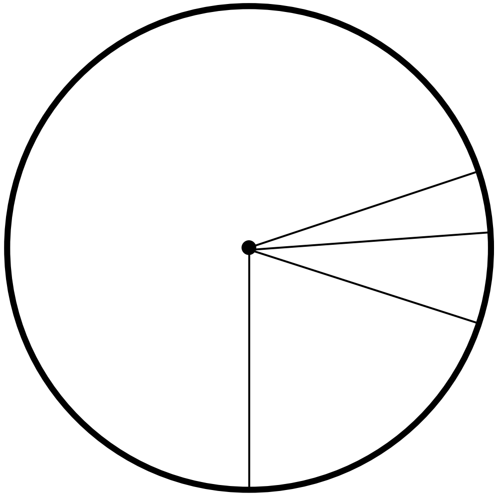
ANOVA

La scomposizione della variabilità **dipende** direttamente dal **disegno di ricerca**. Distinguiamo tra:

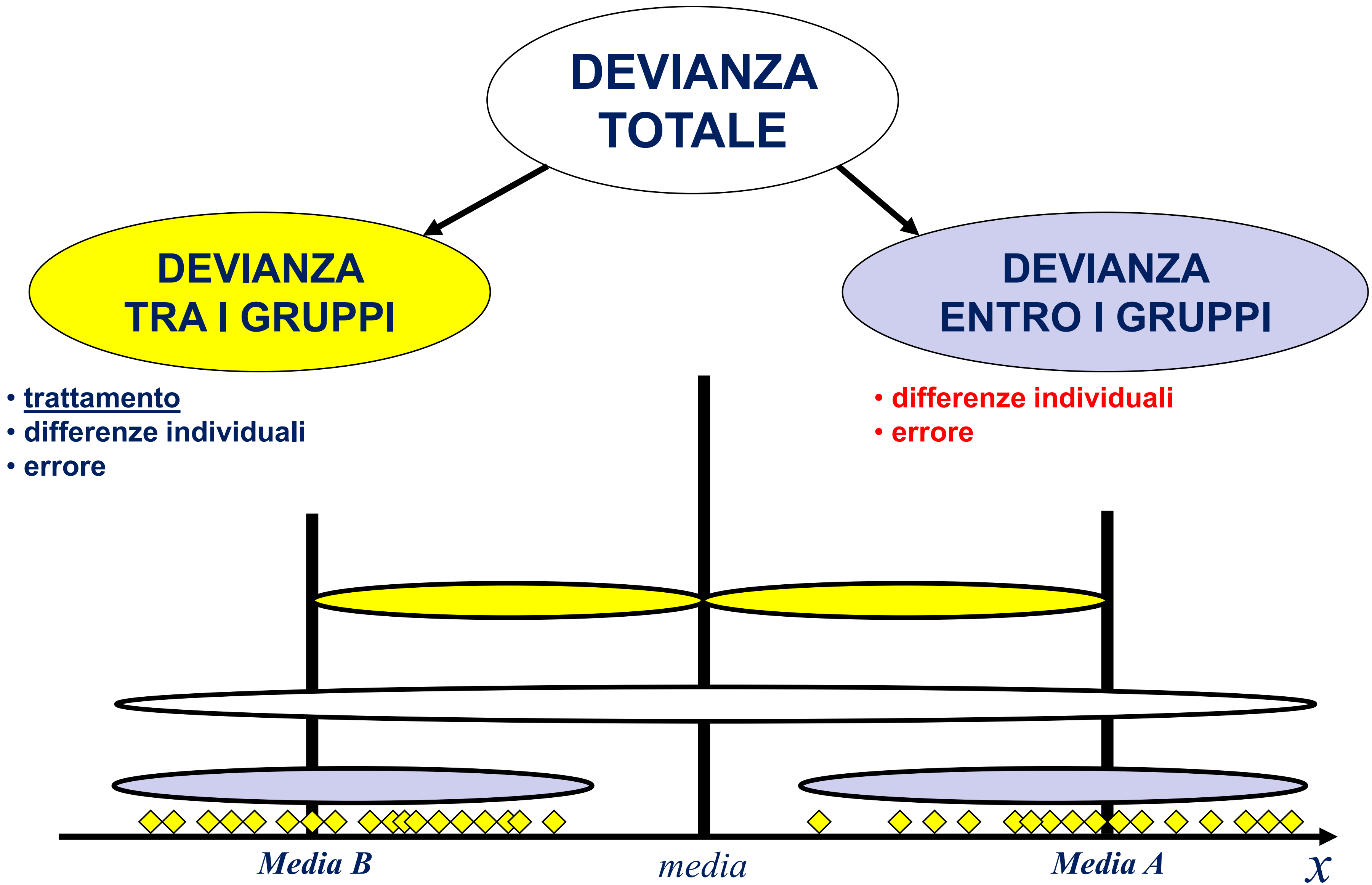
- disegni a **FATTORI *BETWEEN*** (misure/gruppi indipendenti);
- disegni a **FATTORI *WITHIN*** (misure/gruppi dipendenti);
- disegni **MISTI** (misure/gruppi sia dipendenti sia indipendenti).







ANOVA *BETWEEN*

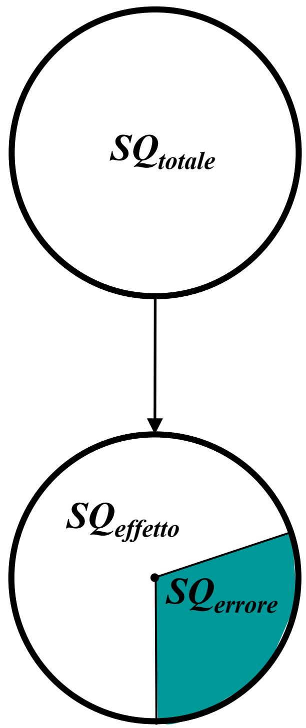


ANOVA *BETWEEN*

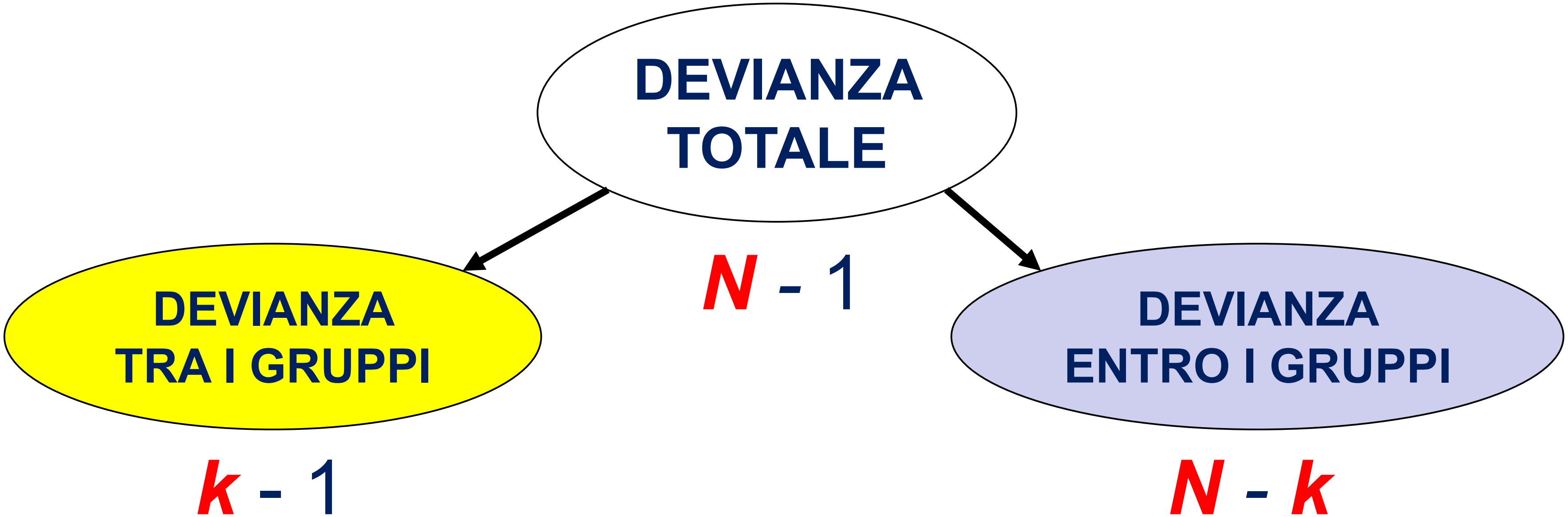
$$SS_{totale} = SS_{spiegata} + SS_{errore}$$

SQ_{totale}

$SQ_{effetto}$ + SQ_{errore}



ANOVA *BETWEEN*



N = numero di osservazioni

k = numero di gruppi (livelli della **VI**)

ANOVA *BETWEEN*

Es. Un fattore *between* 3 livelli:

A (alti) – B (medi) – C (bassi)

DEVIANZA TOTALE

$$\sum\left(y_{ij}-\bar{y}_{..}\right)^2$$

DEVIANZA ENTRO I GRUPPI

$$\sum\left(y_{ij}-\bar{y}_{i.}\right)^2$$

DEVIANZA TRA I GRUPPI

$$\sum\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2$$

Effetto VI

$y_{ij} = punteggio\ ss_i$

$\bar{y}_{i.} = media\ gruppo$

$\bar{y}_{..} = media\ totale$

ANOVA *BETWEEN*

Per **confrontare la due varianze** e verificare se quella spiegata dall'effetto (**VI**) è maggiore di quella residua, si calcola la statistica **F**. La **varianza spiegata** (dal modello) va al numeratore, quella **residua** al denominatore.

$$\mathrm{F} = \frac{\text{var spiegata}}{\text{var errore}}$$

$$F = \frac{Var_{tra\_gr}}{Var_{entro\_gr}} = \frac{\dfrac{Dev_{tra\_gr}}{gdl}}{\dfrac{Dev_{entro\_gr}}{gdl}}$$

H_0 : la varianza spiegata è uguale a quella residua (casuale) $\rightarrow F = 1$

$\mathrm{H}_1 : F > 1$

$$\mathrm{gdl}_{\mathrm{F}} = \frac{k-1}{n-k}$$

H_0 : tutte le medie dei k gruppi sono uguali

H_1 : esiste almeno un gruppo con una media diversa dalle altre

ANOVA *BETWEEN*

Se si rifiuta $\mathbf{H_0}$, ovvero si osserva che la **variabilità *between*** è **significativamente maggiore** di quella ***within*** ($\mathbf{H_1}$), possiamo affermare che:

(1) la variabilità osservata nella **variabile dipendente** è riconducibile alla **variabile indipendente** che ha generato i gruppi;

(2) esiste **almeno una differenza** tra le medie dei gruppi riconducibile alla variabile indipendente, ovvero esiste **almeno un gruppo** in cui la variabile dipendente si distribuisce in modo diverso dagli altri gruppi.

Post hoc ► Se $k > 2$ e non pianificati ► Varianti del ***t*-test** per confronti fra coppie di campioni con correzione del valore di probabilità ($\Sigma p = \alpha$) per verificare quali gruppi sono diversi.

ANOVA *POST-HOC*

Per la valutazione dei **confronti non pianificati** bisogna fare attenzione a contenere l'errore di **I tipo**. A tal scopo sono stati proposti differenti procedure utili alla correzione:

Test di Tukey (Tukey, 1953): consente di controllare l'errore di I tipo nel caso in cui vengano eseguiti solo confronti a coppie.

Test di Scheffé (Sheffé, 1953): consente di eseguire **tutti i confronti possibili**, sia tra coppie di medie sia tra medie complesse, impedendo all'errore di I tipo di eccedere il livello prescelto.

Correzione di Bonferroni (Bonferroni, 1936): consente di eseguire tutti i confronti possibili correggendo il livello di probabilità (ad esempio $\alpha = .05$) di ciascun confronto per il numero (n) di confronti effettuati: α/n.

Sono stati proposti **modelli alternativi** per correggere i livelli di probabilità dei confronti multipli che tengano in considerazione sia una riduzione dell'errore di **I tipo** sia dell'errore di **II tipo** (es., Pastore, Nucci e Galfano, 2005).

ANOVA *BETWEEN*

EFFECT SIZE

Eta quadrato

$$\eta^2 = \frac{\text{devianza spiegata}}{\text{devianza errore}} = \frac{SS_{between}}{SS_{totale}}$$

Effect size	η^2
small	.06
medium	.14
large	.20

Omega quadrato

$$\omega^2 = \frac{(k-1)\cdot(F-1)}{(k-1)\cdot(F-1)+n}$$

ASSUNTI ANOVA *BETWEEN*

(1) i ***k*** campioni sono **indipendenti**;

(2) le popolazioni da cui provengono i ***k*** campioni sono distribuite in modo **normale**;

(3) le varianze di tali popolazioni sono omogenee (**omoschedasticità**);

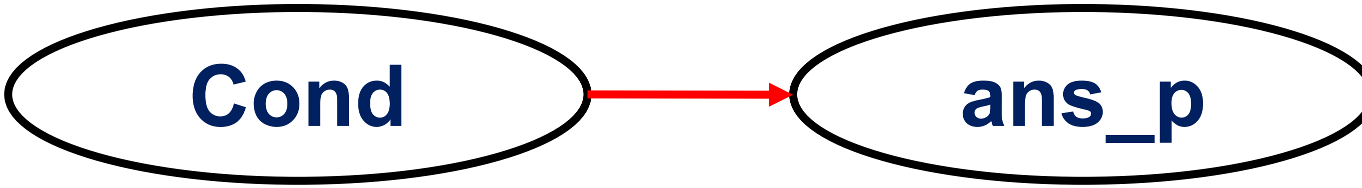
(4) la variabile indipendente che ha ***k*** livelli (2 o più) è misurata su scala **qualitativa** o non parametrica (N o O);

(5) la variabile dipendente è misurata su scala **metrica** (I o R)

(6) gli effetti delle VI sulla VD sono **additivi**.

ESEMPIO #3

La condizione (cond) sperimentale (facile, media, difficile, impossibile) influenza il livello di ansia **pre**-test (ans_p)?



ANOVA between a **un fattore** a **4 livelli** con la **Condizione** come variabile indipendente (Cond; VI-O) e il livello di **ansia pre-test** come variabile dipendente (ans_p; VD-I).

$$\mathrm{H}_0 \Rightarrow \mu_{G1} = \mu_{G2} = \mu_{G3} = \mu_{G4}$$

$$\mathrm{H}_1 \Rightarrow \mu_{G1} \neq (\mu_{G2} \text{ o } \mu_{G3} \text{ o } \mu_{G4})$$

$$\mathrm{H}_1 \Rightarrow \mu_{G2} \neq (\mu_{G3} \text{ o } \mu_{G4})$$

$$\mathrm{H}_1 \Rightarrow \mu_{G3} \neq (\mu_{G4})$$

$$\alpha = .05$$

ESEMPIO #3

ANOVA - ans_p

Cases	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
Cond	20.84	3	6.947	0.164	0.921	0.005	0.000
Residual	4075.52	96	42.453				

Note. Type III Sum of Squares

Questo risultato ci porta ad accettare l'ipotesi nulla.

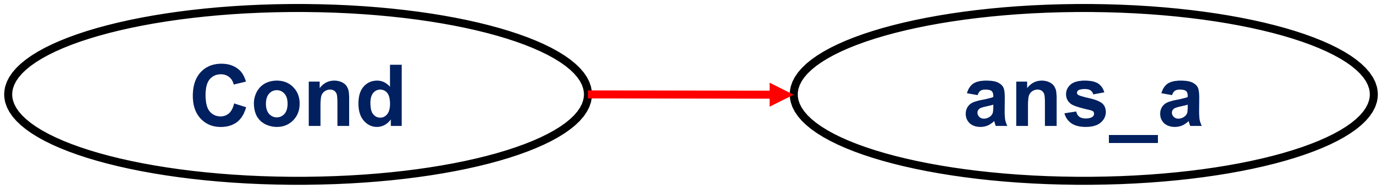
$$H_0 \Rightarrow \mu_{G1} = \mu_{G2} = \mu_{G3} = \mu_{G4}$$

Il livello di ansia prima della prova non viene influenzato dalla condizione sperimentale, ***F*(3,96) = 0.164, *p* = .921, η^2 = .005**.

N.B. Se si calcola il ***power*** di questo test il risultato è: ***power* = .518**. In queste condizioni di analisi per avere un *power* di almeno .80 è necessario un campione di **180 ss**.

ESEMPIO #4

La condizione (cond) sperimentale (facile, media, difficile, impossibile) influenza il livello di ansia **post**-test (ans_a)?



ANOVA between a **un fattore** a **4 livelli** con la **Condizione** come variabile indipendente (Cond; VI-O) e il livello di **Ansia post-test** come variabile dipendente (ans_a; VD-I).

$$H_0 \Rightarrow \mu_{G1} = \mu_{G2} = \mu_{G3} = \mu_{G4}$$

$$H_1 \Rightarrow \mu_{G1} \neq (\mu_{G2} \text{ o } \mu_{G3} \text{ o } \mu_{G4})$$

$$H_1 \Rightarrow \mu_{G2} \neq (\mu_{G3} \text{ o } \mu_{G4})$$

$$H_1 \Rightarrow \mu_{G3} \neq (\mu_{G4})$$

$$\alpha = .05$$

ESEMPIO #4

ANOVA - ans_a

Cases	Sum of Squares	df	Mean Square	F	p	η²	ω²
Cond	197.8	3	65.95	3.082	0.031	0.088	0.059
Residual	2054.2	96	21.40				

Note. Type III Sum of Squares

Questo risultato ci porta a respingere l'ipotesi nulla e a supportare l'ipotesi alternativa.

$$H_0 \Rightarrow \mu_{G1} = \mu_{G2} = \mu_{G3} = \mu_{G4}$$

$$H_1 \Rightarrow \mu_{G1} \neq (\mu_{G2} \text{ o } \mu_{G3} \text{ o } \mu_{G4})$$

$$H_1 \Rightarrow \mu_{G2} \neq (\mu_{G3} \text{ o } \mu_{G4})$$

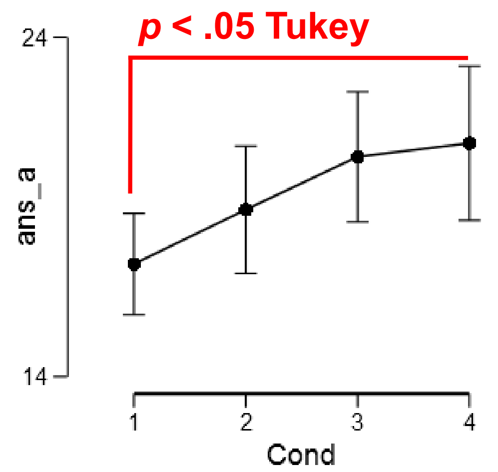
$$H_1 \Rightarrow \mu_{G3} \neq (\mu_{G4})$$

ESEMPIO #4

Cond
1- facile
2- media
3- difficile
4- impossibile

Descriptives - ans_a

Cond	Mean	SD	N
1	17.32	3.614	25
2	18.92	4.545	25
3	20.48	4.647	25
4	20.88	5.502	25



Post Hoc Comparisons - Cond

		Mean Difference	SE	t	p_{tukey}	$p_{scheffe}$	p_{bonf}	p_{holm}
1	2	−1.600	1.308	−1.223	0.614	0.684	1.000	0.673
	3	−3.160	1.308	−2.415	0.081	0.128	0.106	0.088
	4	−3.560	1.308	−2.721	0.038	0.067	0.046	0.046
2	3	−1.560	1.308	−1.192	0.633	0.701	1.000	0.673
	4	−1.960	1.308	−1.498	0.443	0.526	0.824	0.550
3	4	−0.400	1.308	−0.306	0.990	0.993	1.000	0.760

ESEMPIO #4

L'**ANOVA** ha messo in evidenza che la condizione sperimentale influenza significativamente il livello di ansia riportato dopo il test di statistica, ***F*(3,96) = 3.08, *p* = .031, η^2 = .088**. In particolare, i test ***post-hoc*** condotti con la correzione di Tukey evidenziano che gli studenti posti nella condizione impossibile (***M* = 20.9**) dichiarano di avere un livello di ansia significativamente superiore agli studenti posti nella condizione facile (***M* = 17.3**). Non significative risultano le differenze tra gli altri gruppi confrontati.

N.B. Se si calcola il ***power*** di questo test il risultato è: ***power* = .518**. In queste condizioni di analisi per avere un *power* di almeno .80 è necessario un campione di **180 ss**.

TECNICHE DI ANALISI DEI DATI

AA 2018/2019

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:

https://goo.gl/xY15fR

ANOVA *WITHIN*

DEVIANZA TOTALE

DEVIANZA TRA LE MISURE

- trattamento
- errore

N.B. Non ci sono più le differenze individuali.

DEVIANZA ENTRO I GRUPPI

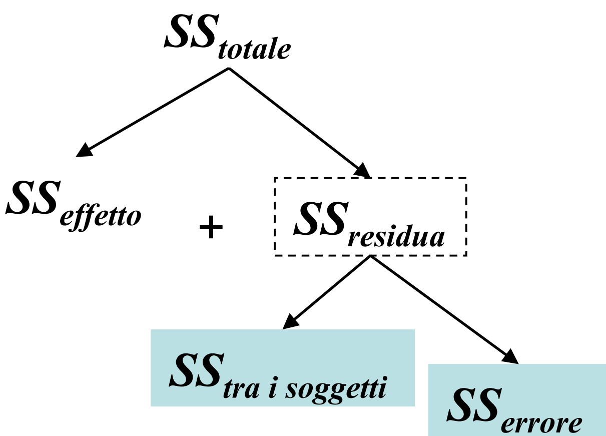
- differenze individuali
- errore

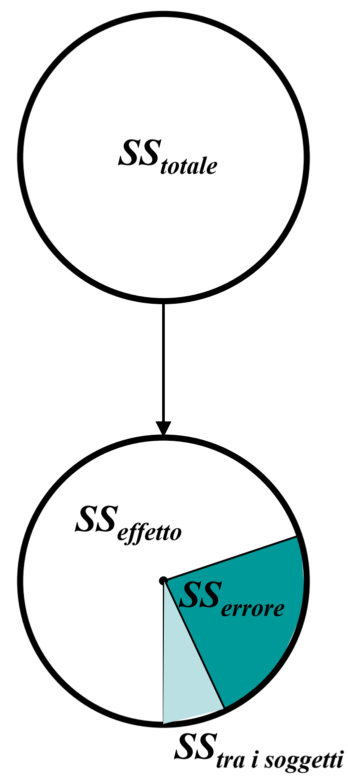
DEVIANZA TRA SOGGETTI

DEVIANZA ERRORE

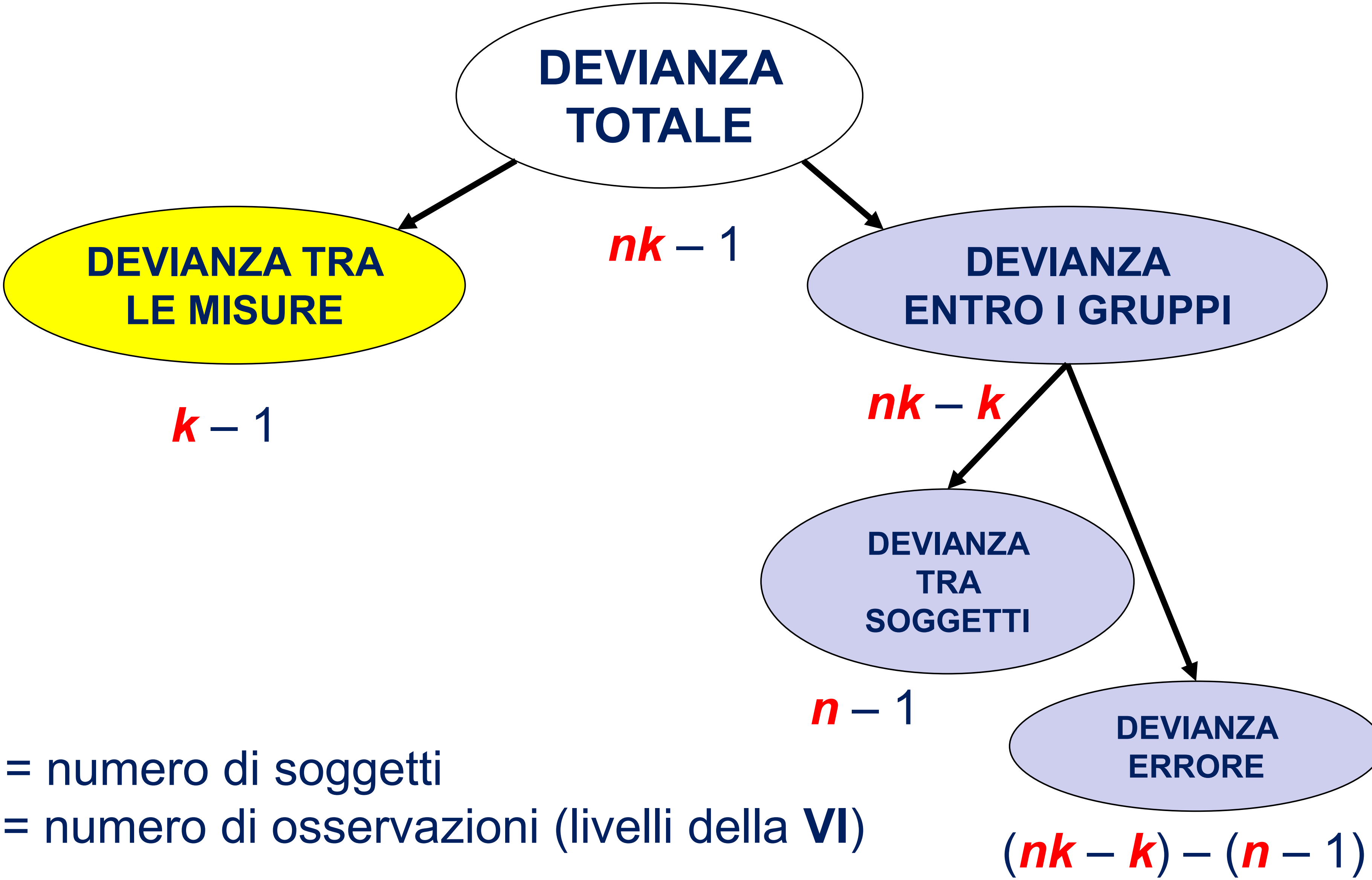
ANOVA *BETWEEN*

$$SS_{totale} = SS_{spiegata} + SS_{errore}$$





ANOVA *WITHIN*



n = numero di soggetti

k = numero di osservazioni (livelli della **VI**)

ANOVA *WITHIN*

$$\sum\left(y_{ij}-\bar{y}_{..}\right)^2$$

DEVIANZA TOTALE

$$\sum\left(y_{ij}-\bar{y}_{.j}\right)^2$$

DEVIANZA ENTRO I GRUPPI

$$n\sum\left(y_{.j}-\bar{y}_{..}\right)^2$$

DEVIANZA TRA LE MISURE

DEVIANZA TRA I SOGGETTI

$$k\sum\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2$$

DEVIANZA ERRORE

Dev. entro i gr. – Dev. tra i sogg.

$y_{ij} = punteggio\ ss_i$

$\bar{y}_{i.} = media\ ss$

$\bar{y}_{.j} = media\ misura\ k$

$\bar{y}_{..} = media\ totale$

$k = osservazioni$

ANOVA *WITHIN*

Per **confrontare la due varianze** e verificare se quella spiegata dall'effetto (**VI**) è maggiore di quella residua, si calcola la statistica **F**. La **varianza spiegata** (dal modello) va al numeratore, quella **residua** al denominatore.

$$F = \frac{\text{var spiegata}}{\text{var errore}}$$

$$F = \frac{Var_{spiegata}}{Var_{errore}} = \frac{\dfrac{Dev_{spiegata}}{gdl}}{\dfrac{Dev_{errore}}{gdl}}$$

H_0 : la varianza spiegata è uguale a quella residua (casuale) $\rightarrow F = 1$

$\mathrm{H}_1 : F > 1$

$$\mathrm{gdl}_{\mathrm{F}} = \frac{k-1}{(nk-k)-(n-1)}$$

H_0 : tutte le medie dei k gruppi sono uguali

H_1 : esiste almeno un gruppo con una media diversa dalle altre

ANOVA *WITHIN*

Se si rifiuta **H_0**, ovvero si osserva che la **variabilità *between*** **(spiegata)** è **significativamente maggiore** di quella ***within*** **(errore)** (**H_1**), possiamo affermare che:

(1) la variabilità osservata nella **variabile dipendente** è riconducibile alla **variabile indipendente** che ha influenzato le misure;

(2) esiste **almeno una differenza** tra le ***k*** misure riconducibile alla variabile indipendente, ovvero esiste **almeno una misura** in cui la variabile dipendente si distribuisce in modo diverso dalle altre.

Post hoc ► Se ***k* > 2** e non pianificati ► Varianti del ***t*-test** per confronti fra coppie di campioni con correzione del valore di probabilità (**$\Sigma p = \alpha$**) per verificare quali misure sono diverse.

ANOVA *WITHIN*
EFFECT SIZE

Effect size	η^2
small	.06
medium	.14
large	.20

Eta quadrato

$$\eta^2 = \frac{\text{devianza spiegata}}{\text{devianza errore}} = \frac{SS_{spiegata}}{SS_{totale}}$$

Omega quadrato

$$\omega^2 = \frac{SS_{spiegata} - (k-1)\cdot MS_{errore}}{SS_{totale} + MS_{soggetti} + n \cdot MS_{errore}}$$

ASSUNTI ANOVA *WITHIN*

(1) Gli **errori** (ε_{ij}) devono essere indipendenti (i soggetti non si devono influenzare reciprocamente);

(2) gli **errori** (ε_{ij}) devono essere distribuiti normalmente con una media uguale a **0**;

(3) le **varianze** delle differenze tra tutte le coppie di medie devono essere uguali (**sfericità o circolarità** → **test di Mauchley**).

Nel caso in cui ci sia una **violazione dell'assunto di sfericità** si procede correggendo i gradi di libertà della statistica **F** mediante il parametro (ε). Tre sono le possibili correzioni: →

- Greenhouse-Geisser
- Huynh-Feldt
- Lower-bound

GG è conservativa

ESEMPIO #5

Le fasi sperimentali influenzano (baseline, info sul test, test) il livello di ansia (ans)?

$$\text{Fase} \longrightarrow \text{ans}$$

ANOVA within a **un fattore** a **3 livelli** con la **Fase** (baseline; info sul test; test) come variabile **indipendente** (Fase; VI-O) e il livello di **Ansia** come variabile **dipendente** (ans; VD-I).

$$\mathrm{H}_0 \Rightarrow \mu_{F1} = \mu_{F2} = \mu_{F3}$$

$$\mathrm{H}_1 \Rightarrow \mu_{F1} \neq (\mu_{F2} \text{ o } \mu_{F3})$$

$$\mathrm{H}_1 \Rightarrow \mu_{F2} \neq (\mu_{F3})$$

$$\alpha = .05$$

ESEMPIO #5

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η²	ω²
Fase	None	76.53[a]	2.000[a]	38.26[a]	3.744[a]	0.025[a]	0.036	0.027
	Greenhouse-Geisser	76.53[a]	1.643[a]	46.59[a]	3.744[a]	0.034[a]	0.036	0.027
	Huynh-Feldt	76.53[a]	1.667[a]	45.92[a]	3.744[a]	0.033[a]	0.036	0.027
Residual	None	2023.47	198.000	10.22				
	Greenhouse-Geisser	2023.47	162.608	12.44				
	Huynh-Feldt	2023.47	164.988	12.26				

Note. Type III Sum of Squares
[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated (p < .05).

Between Subjects Effects ▼

	Sum of Squares	df	Mean Square	F	p	η²	ω²
Residual	2023	198	10.22				

Note. Type III Sum of Squares

Test of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Fase	0.782	< .001	0.821	0.833

Questo risultato ci porta a respingere l'ipotesi nulla e a supportare l'ipotesi alternativa.

$$H_0 \Rightarrow \mu_{F1} = \mu_{F2} = \mu_{F3}$$

$$H_1 \Rightarrow \mu_{F1} \neq (\mu_{F2} \text{ o } \mu_{F3})$$

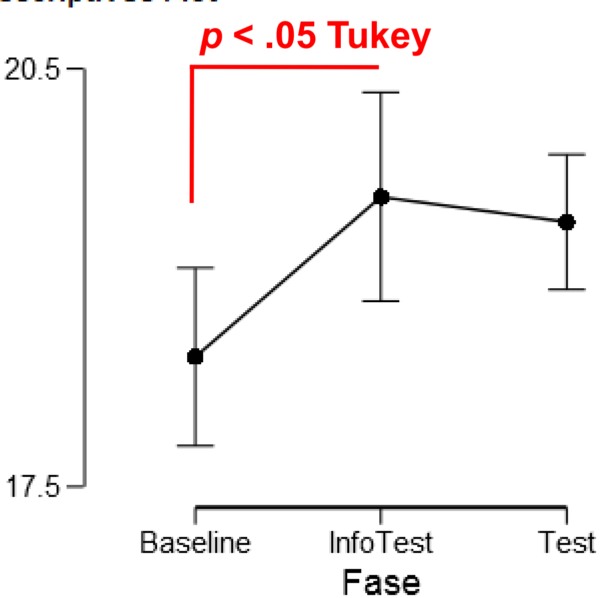
$$H_1 \Rightarrow \mu_{F2} \neq (\mu_{F3})$$

ESEMPIO #5

Post Hoc Comparisons - Fase

		Mean Difference	SE	t	p_{tukey}	$p_{scheffe}$	p_{bonf}	p_{holm}
Baseline	InfoTest	−1.150	0.452	−2.544	0.031	0.041	0.035	0.035
	Test	−0.970	0.452	−2.146	0.083	0.103	0.099	0.066
InfoTest	Test	0.180	0.452	0.398	0.916	0.924	1.000	0.691

Descriptives Plot



Descriptives

Fase	Mean	SD	N
Baseline	18.43	4.330	100
InfoTest	19.58	6.433	100
Test	19.40	4.769	100

ESEMPIO #5

L'**ANOVA** ha messo in evidenza che la fase sperimentale influenza significativamente il livello di ansia riportato, ***F*(2,198) = 3.74, *p* = .025, η^2 = .036**. In particolare, i test ***post-hoc*** condotti con la correzione di Tukey evidenziano che il livello di ansia degli studenti nella fase *baseline* (***M* = 18.4**) è significativamente inferiore a quello degli studenti nella fase prima del test (***M* = 19.6**). Non significativa è la differenza nel livello di ansia tra prima e dopo il test (***M* = 19.4**).

N.B. Se si calcola il ***power*** di questo test il risultato è: ***power* = .999**.