UNIVERSITÀ
DEGLI STUDI
DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 17: ANALISI DELLA VARIANZA (ANOVA)

A.A. 2024-2025

Prof. Martina Vettoretti

TEST STATISTICI PARAMETRICI PER IL CONFRONTO DELLE MEDIE TRA GRUPPI

- Test statistici parametrici per verificare se due campioni normali indipendenti hanno la stessa media:

Test	Assunzioni	Ipotesi nulla	Funzione Matlab
z test a due campioni	Campioni normali, varianze note	$H_0: \mu_1 = \mu_2$	Non disponibile
t test a due campioni	Campioni normali, varianze incognite ma uguali	$H_0: \mu_1 = \mu_2$	ttest2
t test di Welch	Campioni normali, varianze incognite	$H_0: \mu_1 = \mu_2$	ttest2 con l'opzione 'Vartype' 'unequal'

- E se i gruppi sono più di due come facciamo?

DEFINIZIONE DEL PROBLEMA

- Vogliamo confrontare N gruppi di osservazioni e determinare se sussiste una differenza significativa dal punto di vista statistico tra le medie degli N gruppi.

Esempio: vogliamo confrontare la pressione sanguigna in pazienti sottoposti a 3 trattamenti diversi (A, B, C) per determinare se il tipo di trattamento può avere un'influenza sulla pressione sanguigna.

- Selezioniamo 3 gruppi di pazienti sottoposti ai 3 trattamenti in esame (A, B, C)*.
- Calcoliamo la pressione media in ciascun gruppo.
- Domanda: c'è una differenza significativa tra la pressione media osservata nei 3 gruppi?

- Se avessimo solo 2 gruppi potremmo usare un t test per il confronto di due campioni indipendenti. Se i gruppi sono più di due possiamo usare il metodo ANOVA.

* La selezione deve essere fatta in modo che i pazienti nei 3 gruppi presentino caratteristiche simili, ovvero non ci siano altri fattori confondenti che possono influenzare il risultato (es. se un gruppo è formato da pazienti più vecchi rispetto agli altri, è normale attendersi una pressione mediamente più alta).

ANOVA

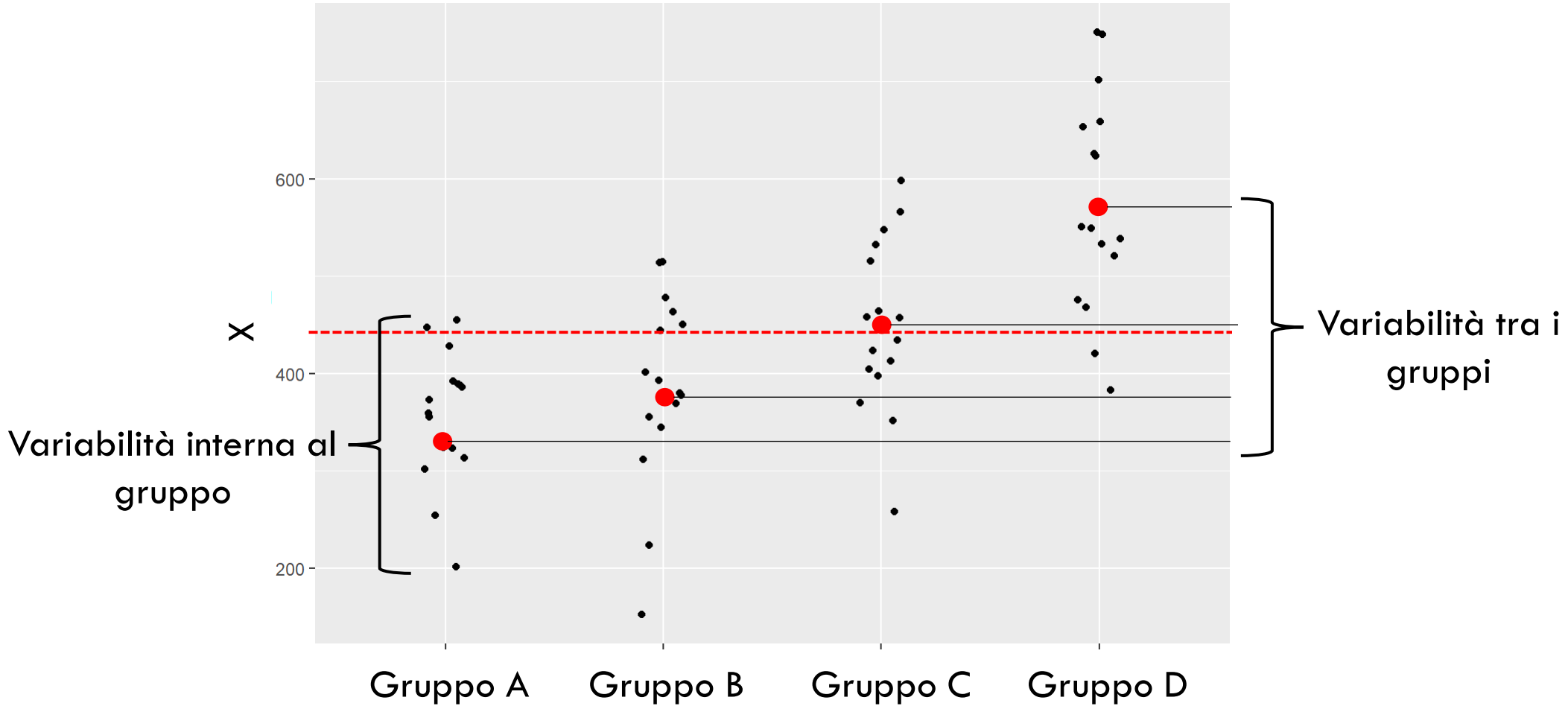
- **Analisi della varianza, o ANOVA**: tecnica statistica per confrontare se sussiste una differenza significativa tra le medie di N campioni (N gruppi).
- Assunzioni: N campioni indipendenti aventi distribuzione normale con uguale varianza pari a σ^2 e medie incognite: $\mu_1, \mu_2, \ldots, \mu_N$.
- Sistema di ipotesi:
  - $H_0: \mu_1 = \mu_2 = \ldots = \mu_N$ → le medie degli N campioni sono uguali
  - H_1: <u>almeno una</u> delle medie degli N campioni è diversa dalle altre

- Nota: si tratta di un test statistico sulle medie che però opera con una statistica basata sul calcolo di varianze. Per questo si chiama analisi della varianza.

SCOMPOSIZIONE DELLA VARIABILITA' TOTALE

- Chiamiamo $\bar{X}_i$ la media campionaria del gruppo i-esimo.
- Chiamiamo $\bar{X}$ la media campionaria globale (considerando le osservazioni di tutti i gruppi).
- La **variabilità totale** della variabile di interesse (nell'esempio precedente la pressione sanguigna), considerando le osservazioni di tutti i gruppi, può essere scomposta nella somma di due componenti:
  - **Variabilità interna ai gruppi** → quanto variano le osservazioni all'interno di un gruppo rispetto alla sua media
  - **Variabilità tra i gruppi** → quanto variano le medie dei diversi gruppi rispetto alla media globale $\bar{X}$

SCOMPOSIZIONE DELLA VARIABILITA' TOTALE



VARIABILITA' INTERNA AI GRUPPI

- La variabilità interna ai gruppi è quella componente di variabilità indipendente dal gruppo di appartenenza.
- Viene anche chiamata ***mean square error*** **(MSE).**
- E' calcolata come media pesata delle varianze campionarie dei singoli gruppi:

$$MSE = \frac{\sum_{i=1}^{N}(n_i - 1)S_i^2}{\sum_{i=1}^{N}(n_i - 1)} = \frac{\sum_{i=1}^{N}(n_i - 1)S_i^2}{(n_{tot} - N)}$$

  - S_i^2 → varianza campionaria del gruppo i-esimo
  - n_i → numero di osservazioni appartenenti al gruppo i-esimo
  - n_{tot} → numero totale di osservazioni

VARIABILITA' TRA I GRUPPI

- La variabilità tra i gruppi è la componente di variabilità influenzata dal gruppo di appartenenza.
- Essa è chiamata anche **mean square between groups (MSB).**
- E' calcolata come varianza delle medie dei singoli gruppi rispetto alla media globale:

$$MSB = \frac{\sum_{i=1}^{N} n_i (\bar{X}_i - \bar{X})^2}{N - 1}$$

STATISTICA DEL TEST

- Quanto più ci allontaniamo dall'ipotesi nulla, tanto più ci aspettiamo che la variabilità tra i gruppi (MSB) sia elevata e rappresenti un'importante frazione della variabilità totale.
- Statistica del test:

$$F = \frac{MSB}{MSE}$$

- Quando vale H_0, F è distribuita come una F di Fisher avente gradi di libertà N-1 e n_{tot}-N.
- Livello di significatività α.
- Regola decisionale:
  - Se $F > F_{\alpha,N-1,n_{tot}-N}$ → rifiuto H_0 → almeno uno dei gruppi ha media significativamente diversa dagli altri
  - Se $F \leq F_{\alpha,N-1,n_{tot}-N}$ → non possiamo rifiutare H_0 → non possiamo dire nulla.

OSSERVAZIONI

- L'ANOVA mi consente di dire se almeno uno dei gruppi presenta una media diversa, ma non mi consente di dire quale gruppo (o quali gruppi) presentano media diversa dagli altri.
- Con 2 gruppi, ANOVA è equivalente a un t test per il confronto di due campioni aventi la stessa varianza.
- L'ANOVA è equivalente all'F test di una regressione lineare avente la variabile di interesse come outcome e delle variabili dummy rappresentanti il gruppo di appartenenza come variabili esplicative.

EQUIVALENZA TRA ANOVA E REGRESSIONE LINEARE

- L'ANOVA è equivalente ad un particolare tipo di regressione lineare.
- Stiamo confrontando una variabile di interesse X misurata in N gruppi.
- Possiamo usare una regressione lineare per valutare l'impatto del gruppo di appartenenza sul valore medio di X:

$$X = \beta_0 + \beta_1 D_1 + \cdots + \beta_{N-1} D_{N-1} + \varepsilon$$

La variabile di outcome è la variabile X misurata su tutti i gruppi

Le variabili esplicative sono le variabili dummy che codificano il gruppo di appartenenza (variabile qualitativa a N livelli)

EQUIVALENZA TRA ANOVA E REGRESSIONE LINEARE

- L'F test che valuta se il modello di regressione è diverso da quello nullo è equivalente all'F test dell'ANOVA (stesso identico p-value).
- Questo non dovrebbe sorprenderci!
- L'F test della regressione lineare di fatto valuta se almeno una delle variabili esplicative considerate ha un coefficiente β diverso da 0, ovvero ha un impatto non nullo sul valor medio dell'outcome.
- Poiché in questa regressione le variabili esplicative codificano la variabile qualitativa che rappresenta il gruppo di appartenenza, l'F test della regressione ci dice se il gruppo di appartenenza ha un impatto significativo sul valor medio di X, ovvero se almeno uno dei gruppi ha media significativamente diversa dagli altri, proprio come ANOVA.

ESEMPIO

Dati:

X	Gruppo
120	A
110	A
130	A
145	A
150	B
105	B
115	B
137	C
135	C
123	C

X	D_1 (livello B)	D_2 (livello C)
120	0	0
110	0	0
130	0	0
145	0	0
150	1	0
105	1	0
115	1	0
137	0	1
135	0	1
123	0	1

$$X = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \varepsilon$$

F test per valutare se il modello è significativamente diverso da quello nullo.

- $H_0: \beta_1 = \beta_2 = 0$
- H_1: almeno uno tra β_1 e β_2 è diverso da 0.

ANOVA A DUE VIE

- ANOVA a due vie (two-way ANOVA): estensione di ANOVA al caso in cui i gruppi sono definiti sulla base di due fattori.
  - Esempio: Vogliamo confrontare la pressione sanguigna per diversi gruppi di pazienti definiti in base al trattamento e al sesso.

Numero di pazienti nei gruppi

		Trattamento		
		A	B	C
Sesso	Maschio	20	25	15
	Femmina	15	10	15

ANOVA A DUE VIE

- Nell'ANOVA a due vie vengono eseguiti 2 F test, uno per ciascun fattore:
  - Un primo F test valuta se il primo fattore ha un impatto significativo sulla media della variabile di interesse.
  - Un secondo F test valuta se il secondo fattore ha un impatto significativo sulla media della variabile di interesse.

ANOVA A DUE VIE CON LA REGRESSIONE LINEARE

- L'ANOVA a due vie è equivalente ad eseguire una regressione lineare in cui le variabili esplicative sono le variabili dummy che codificano i due fattori che definiscono i sottogruppi.
- Considerando l'esempio di prima con due fattori, ovvero il trattamento (A, B, C) e il sesso (maschio, femmina), l'ANOVA a due vie si può realizzare con la regressione lineare di equazione:

$$X = \beta_0 + \beta_{t,B} D_{t,B} + \beta_{t,C} D_{t,C} + \cdots + \beta_{sesso,m} D_{sesso,m} + \varepsilon$$

Variabili dummy che codificano il trattamento

Variabile dummy che codifica il sesso

ANOVA A DUE VIE CON LA REGRESSIONE LINEARE

- F test sul fattore trattamento: F test che confronta il modello completo vs il modello senza le variabili $D_{t,B}$ e $D_{t,C}$.
  - $H_0: \beta_{t,B} = \beta_{t,C} = 0$
  - H_1: almeno uno tra $\beta_{t,B}$ e $\beta_{t,C}$ è diverso da 0.
- F test sul fattore sesso: F test che confronta il modello completo vs. il modello senza la variabile $D_{sesso,m}$.
  - $H_0: \beta_{sesso,m} = 0$
  - H_1: $\beta_{sesso,m} \neq 0$

Quando il fattore ha solo due livelli come in questo caso, l'F test è equivalente al t test che valuta se il coefficiente associato al fattore è significativamente diverso da 0.