Metodi statistici per la ricerca sociale

Capitolo 12. Confronto fra gruppi: L'analisi della varianza (ANOVA)

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM–88 SOCIOLOGIA E RICERCA SOCIALE

Confronto fra gruppi: Analisi della Varianza (ANOVA)

- Confrontare g (sotto-)popolazioni
- Obiettivo: Stabilire se le g popolazioni sono identiche
- Confronto tra g campioni: I g campioni provengono dalla stessa popolazione o da popolazioni aventi un parametro caratteristico di uguale valore?
- Esiste evidenza (significatività statistica) che le osservazioni campionarie siano generate da g popolazioni diverse?
- Valutare se le variazioni osservate su una variabile risposta siano dovute alle differenti "situazioni sperimentali" riguardanti uno, due, o più fattori ritenuti influenti

Fattori

Si dice *fattore* una variabile esplicativa discreta con un certo numero finito g di modalità chiamate *livelli*.

- Un fattore può rappresentare un *trattamento*: variabile le cui modalità sono assegnate ad ogni individuo/unità secondo opportune regole di assegnazione
- Un fattore può rappresentare un carattere intrinseco come l'età o il sesso di un individuo
- Un fattore può rappresentare un raggruppamento arbitrario delle unità come, ad esempio, i blocchi di un esperimento o delle zone geografiche

Il livelli del fattore sono considerati su scala nominale, ma verranno per comodità indicati con numeri interi $i = 1, \ldots, g$

Esempi

- Sperimentazione agraria finalizzata a valutare come varia la produzione di mais a seconda del tipo di fertilizzante impiegato

Fertilizzante (A,B,C)	=	Fattore a 3 livelli
Produzione di mais (per unità di superficie)	=	Variabile risposta

- Una ditta che produce elettrodomestici ha condotto un esperimento su 5 prototipi di lavatrice, misurando il loro livello di rumorosità durante alcune prove di lavaggio: l'obiettivo è quello di individuare il prototipo a cui è associata una minore rumorosità

Prototipo di lavatrice	=	Fattore a 5 livelli
Rumorosità	=	Variabile risposta

- Un'eccessiva presenza di ozono nell'aria è indice di inquinamento atmosferico. In relazione a ciò, sono stati raccolti ed esaminati sei campioni di aria per quattro luoghi diversi.

Sito	=	Fattore a 4 livelli
Quantitativi di ozono rilevati (in parti per milione)	=	Variabile risposta

Ozono

Quantitativi di ozono rilevati (in parti per milione) in quattro diversi siti

Siti	Osservazioni					
1	0.08	0.10	0.09	0.07	0.09	0.06
2	0.15	0.09	0.11	0.10	0.08	0.13
3	0.13	0.10	0.15	0.09	0.09	0.17
4	0.05	0.11	0.07	0.09	0.11	0.08

- Le unità sperimentali sono costituite dalle 24 operazioni di misurazione dell'ozono
- Il fattore studiato è la localizzazione: i 4 siti in cui sono state effettuate le osservazioni
- In ogni sito si sono effettuate 6 misurazioni: campioni casuali della stessa ampiezza, pari a 6, da ogni popolazione (sito)

Capacità cognitive dei bambini

- I dati seguenti riguardano uno studio condotto per confrontare l'effetto di diversi programmi di assistenza all'infanzia (fattore a 3 livelli) sulle capacità cognitive di bambini tra 3 e 4 anni
- I bambini sono suddivisi in tre gruppi:
  1. No Programma: Bambini che non frequentano l'asilo;
  2. Programma Standard: Bambini che frequentano asili in cui è adottato il programma educativo standard;
  3. Programma innovativo: Bambini che frequentano asili in cui è adottato un programma educativo innovativo,
- Al termine dell'anno scolastico viene somministrato un test a tutti i bambini per rilevare le loro capacità cognitive

Capacità cognitive dei bambini (Dati simulati)

No Programma	Programma Standard	Programma Innovativo
62	103	96
85	88	120
96	92	100
74	101	118
105	74	124
81	95	91
86	109	105
95	83	97
	92	107
	128	135
	88	95
	101	
	113	
	98	

Schematizzazione dei dati negli studi a un solo fattore

Fattore	Osservazioni					
1	y_{11}	y_{12}	$\cdots$	y_{1i}	$\cdots$	y_{1n_1}
2	y_{21}	y_{22}	$\cdots$	y_{2i}	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
g	y_{g1}	y_{g2}	$\cdots$	y_{gi}	$\cdots$	y_{gn_g}

oppure

Livelli del fattore			
1	2	$\ldots$	g
y_{11}	y_{21}	$\cdots$	y_{g1}
y_{12}	y_{22}	$\cdots$	y_{g2}
$\vdots$	$\vdots$	$\ldots$	$\vdots$
y_{1i}	y_{2i}	$\cdots$	y_{gi}
$\vdots$	$\vdots$	$\ldots$	$\vdots$
y_{1n_1}	y_{2n_2}	$\cdots$	y_{gn_g}

Schematizzazione dei dati negli studi a un solo fattore

oppure

Unità u_i	Variabile risposta y_i	Livelli del fattore
1	y_1	1
2	y_2	1
$\vdots$	$\vdots$	$\vdots$
n_1	y_{n_1}	1
$n_1 + 1$	y_{n_1+1}	2
$n_1 + 2$	y_{n_1+2}	2
$\vdots$	$\vdots$	$\vdots$
$n_1 + n_2$	$y_{n_1+n_2}$	2
$\vdots$	$\vdots$	$\vdots$
i	y_i	ℓ
$\vdots$	$\vdots$	$\vdots$
n	y_n	g

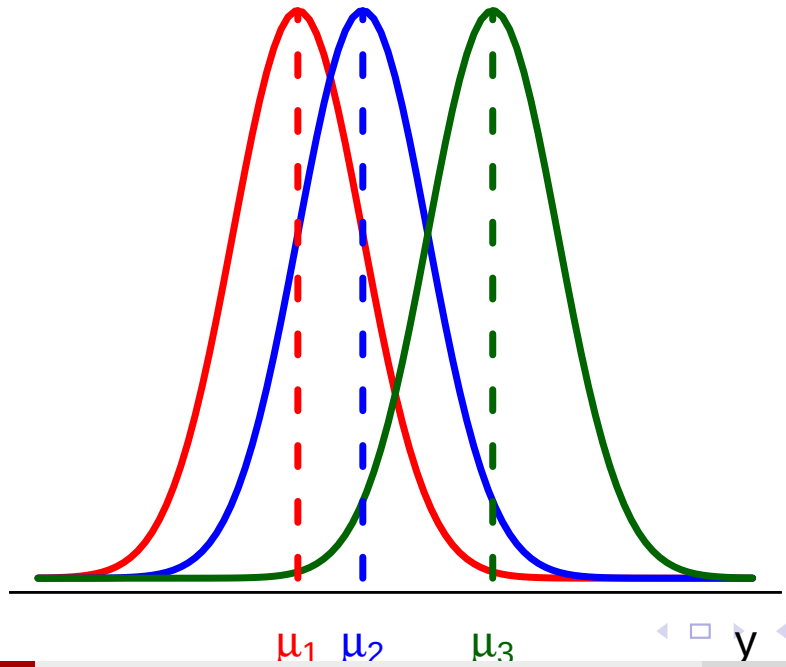
Assunzioni

- Sono disponibili g campioni casuali semplici indipendenti di numerosità $n_1, \ldots, n_g$

$$Y_{11}, \ldots, Y_{1n_1}; \quad Y_{21}, \ldots, Y_{2n_2}; \quad \ldots \quad Y_{1g}, \ldots, Y_{gn_g}$$

- Ogni campione è tratto da una distribuzione normale di media μ_ℓ ($\ell = 1, \ldots, g$) e varianza costante σ^2 (ipotesi di omoschedasticità)

$$Y_{\ell i} \sim N(\mu_\ell, \sigma^2) \quad i = 1, \ldots, n_\ell$$



Medie campionarie

- Medie entro gruppi

$$\begin{aligned}
\hat{\mu}_1 &= \overline{Y}_1 &&= \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} \\
&\vdots &&\vdots \\
\hat{\mu}_\ell &= \overline{Y}_\ell &&= \frac{1}{n_\ell}\sum_{i=1}^{n_\ell} Y_{\ell i} \\
&\vdots &&\vdots \\
\hat{\mu}_g &= \overline{Y}_g &&= \frac{1}{n_g}\sum_{i=1}^{n_g} Y_{gi}
\end{aligned}$$

- Media globale

$$\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{\ell=1}^{g}\sum_{i=1}^{n_\ell} Y_{\ell i} = \frac{1}{n}\sum_{\ell=1}^{g} \overline{Y}_\ell \cdot n_\ell$$

dove $n = n_1 + n_2 + \cdots + n_g$

Medie campionarie: Capacità cognitive dei bambini

	No Programma	Programma Standard	Programma Innovativo
	62	103	96
	85	88	120
	96	92	100
	74	101	118
	105	74	124
	81	95	91
	86	109	105
	95	83	97
		92	107
		128	135
		88	95
		101	
		113	
		98	
Totale	684	1365	1188
n_ℓ	8	14	11
Medie	85.5	97.5	108.0

Medie campionarie: Capacità cognitive dei bambini

$$n = 8 + 14 + 11 = 33$$

$$\hat{\mu} = \overline{Y} = \frac{1}{n} \sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} Y_{\ell i} = \frac{3237}{33} = 98.09$$

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{\ell=1}^{g} \overline{Y}_\ell \cdot n_\ell = \\ &= \frac{85.5 \cdot 8 + 97.5 \cdot 14 + 108.0 \cdot 11}{33} = \frac{684 + 1365 + 1188}{33} = \frac{3237}{33} = 98.09 \end{aligned}$$

Varianze campionarie

- Varianze entro gruppi

$$
\begin{aligned}
S_1^2 &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(Y_{1i} - \overline{Y}_1 \right)^2 \\
&\vdots \qquad\qquad\qquad \vdots \\
S_\ell^2 &= \frac{1}{n_\ell - 1} \sum_{i=1}^{n_\ell} \left(Y_{\ell i} - \overline{Y}_\ell \right)^2 \\
&\vdots \qquad\qquad\qquad \vdots \\
S_g^2 &= \frac{1}{n_g - 1} \sum_{i=1}^{n_g} \left(Y_{gi} - \overline{Y}_g \right)^2
\end{aligned}
$$

- Varianza totale (marginale)

$$
\frac{1}{n - 1} \sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} \left(Y_{\ell i} - \overline{Y} \right)^2
$$

Varianze entro gruppi: Capacità cognitive dei bambini

	No Programma	Programma Standard	Programma Innovativo
	62	103	96
	85	88	120
	96	92	100
	74	101	118
	105	74	124
	81	95	91
	86	109	105
	95	83	97
		92	107
		128	135
		88	95
		101	
		113	
		98	
n_ℓ	8	14	11
Medie	85.5	97.5	108.0
Varianze	183.7	182.1	202.6

Varianza entro il gruppo di bambini con programma innovativo

Gruppo 3: Programma Innovativo

Bambino	$y_{3,i}$	$(y_{3,i} - \overline{y}_3)^2$
1	96	144
2	120	144
3	100	64
4	118	100
5	124	256
6	91	289
7	105	9
8	97	121
9	107	1
10	135	729
11	95	169
Totale	1188	2026.0

$$\overline{y}_3 = 108.0$$

$$s_1^2 = \frac{2026}{11 - 1} = 202.6$$

Varianza campionaria: Capacità cognitive dei bambini

$$n = 33 \qquad \overline{Y} = \frac{3237}{33} = 98.09$$

$$\frac{1}{n-1} \sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} \left(Y_{\ell i} - \overline{Y} \right)^2 =$$

$$\frac{1}{33-1} \Big[(62 - 98.09)^2 + \cdots + (95 - 98.09)^2 +$$

$$(103 - 98.09)^2 + \cdots + (98 - 98.09)^2 + (96 - 98.09)^2 + \cdots + (95 - 98.09)^2 \Big] =$$

$$\frac{2554.248 + 2372.388 + 3106.091}{32} = \frac{8032.727}{32} = 251.02$$

Devianze

- Devianza totale (Somma dei quadrati totale)

$$D_T = \sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} \left(Y_{\ell i} - \overline{Y} \right)^2 = \sum_{i=1}^{n_1} \left(Y_{1i} - \overline{Y} \right)^2 + \cdots + \sum_{i=1}^{n_\ell} \left(Y_{gi} - \overline{Y} \right)^2$$

- Devianza entro gruppi (Somma dei quadrati entro gruppi): Somma delle devianze entro gruppi

$$D_W = \sum_{\ell=1}^{g} \sum_{i=1}^{n_\ell} \left(Y_{\ell i} - \overline{Y}_\ell \right)^2 = \sum_{\ell=1}^{g} D_\ell$$

dove $D_\ell = \sum_{i=1}^{n_\ell} \left(Y_{\ell i} - \overline{Y}_\ell \right)^2 = s_\ell^2 \cdot (n_\ell - 1)$

- Devianza tra gruppi (Somma dei quadrati tra gruppi): Devianza delle medie entro gruppi

$$D_B = \sum_{\ell=1}^{g} \left(\overline{Y}_\ell - \overline{Y} \right)^2 \cdot n_\ell = \left(\overline{Y}_1 - \overline{Y} \right)^2 \cdot n_1 + \cdots + \left(\overline{Y}_g - \overline{Y} \right)^2 \cdot n_g$$

Devianze: Capacità cognitive dei bambini

Gruppo	n_ℓ	Media	Devianza	s^2_ℓ
No programma	8	85.5	1286.0	183.714
Programma standard	14	97.5	2367.5	182.115
Programma innovativo	11	108.0	2026.0	202.600
Tutti	33	98.09	8032.727	251.023

$$D_T = 2554.248 + 2372.388 + 3106.091 = 8032.727$$

$$D_W = 1286.0 + 2367.5 + 2026.0 = 5679.5$$

$$\begin{aligned} D_B &= (85.5 - 98.09)^2 \cdot 8 + (97.5 - 98.09)^2 \cdot 14 + (108.0 - 98.09)^2 \cdot 11 \\ &= 1268.248 + 4.888 + 1080.091 = 2353.227 \end{aligned}$$

Scomposizione della devianza

$$\text{Devianza Totale} = \text{Devianza entro gruppi} + \text{Devianza tra gruppi}$$

$$\Updownarrow$$

$$\begin{array}{ccccc} \text{Somma dei} & & \text{Somma dei} & & \text{Somma dei} \\ \text{Quadrati Totale} & = & \text{Quadrati Entro Gruppi} & = & \text{Quadrati Tra Gruppi} \end{array}$$

$$\Updownarrow$$

$$D_T = D_W + D_B$$

- Esempio: Capacità cognitive dei bambini

$$D_T = 8032.727 = 5679.5 + 2353.227 = D_W + D_B$$

Varianza tra gruppi

$$\text{Varianza tra gruppi} = \frac{\text{Devianza tra gruppi}}{gdl_B} = \text{Media dei quadrati tra gruppi}$$

In formule

$$S_B^2 = \frac{D_B}{g-1}$$

- Esempio: Capacità cognitive dei bambini

$$s_B^2 = \frac{2353.227}{3-1} = 1176.613$$

Varianza entro i gruppi

$$\text{Varianza entro i gruppi} = \frac{\text{Devianza entro i gruppi}}{gdl_W} = \begin{array}{c}\text{Media dei quadrati}\\ \text{entro i gruppi}\end{array}$$

In formule

$$S^2 = \frac{D_W}{n-g}$$

- Varianza entro i gruppi = Media ponderata delle varianze entro i singoli gruppi

$$S^2 = \frac{S_1^2 \cdot (n_1 - 1) + S_2^2 \cdot (n_2 - 1) + \cdots + S_g^2 \cdot (n_g - 1)}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_g - 1)}$$

dove

$$(n_1 - 1) + (n_2 - 1) + \cdots + (n_g - 1) = (n_1 + n_2 + \cdots + n_g) - g = n - g$$

- La varianza entro i gruppi, S^2, è uno stimatore della varianza σ^2 comune (pooled) delle popolazioni da cui sono estratti i g campioni

Varianza entro i gruppi – Esempio: Capacità cognitive dei bambini

$$
\begin{aligned}
s^2 &= \frac{5679.5}{33-3} \\
&= \frac{1}{30}\Big[1286.0 \cdot (8-1) + 2367.5 \cdot (14-1) + 2026.0 \cdot (11-1)\Big] \\
&= 189.317
\end{aligned}
$$

dove

$$30 = 33 - 3 = (8-1) + (14-1) + (11-1)$$

Scomposizione della devianza: sintesi

Fonte di Variabilità	Somma dei Quadrati	GdL	Media dei Quadrati
Tra Gruppi	D_B	$g-1$	$D_B/(g-1)$
Entro i Gruppi	D_W	$n-g$	$D_W/(n-g)$
Totale	D_T	$n-1$	$D_T/(n-1)$

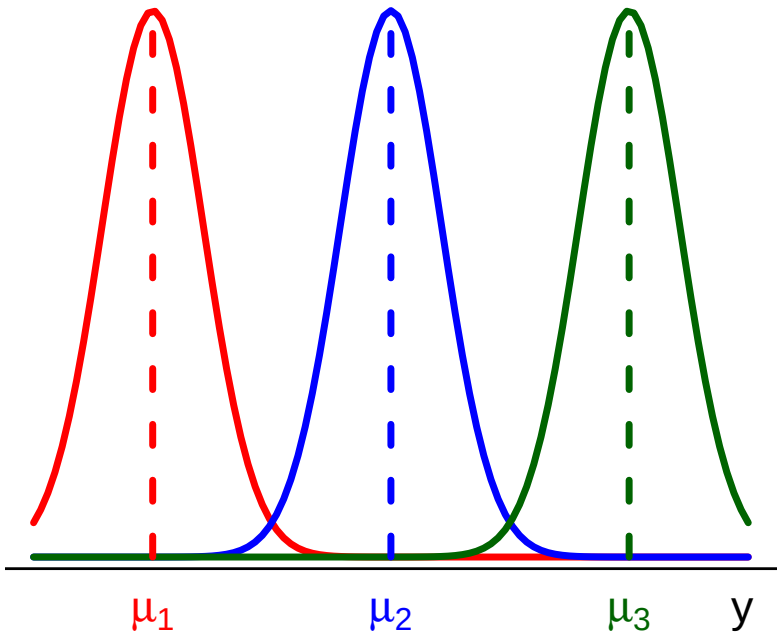
- Esempio: Capacità cognitive dei bambini

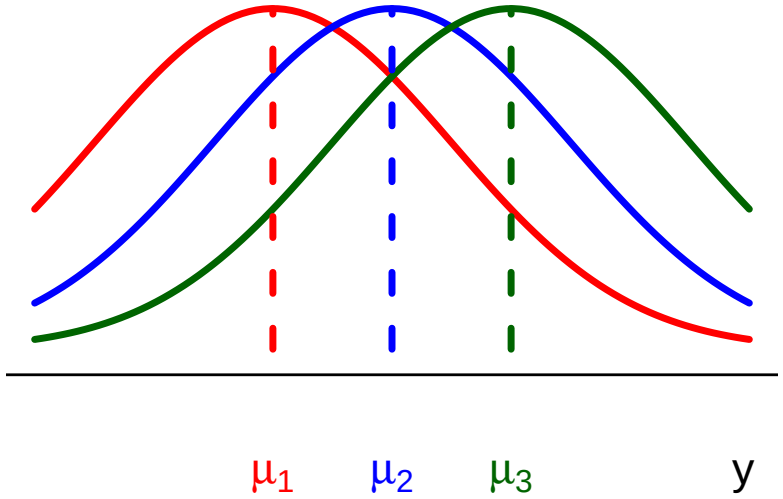
Fonte di Variabilità	Somma dei Quadrati	GdL	Media dei Quadrati
Tra Gruppi	$D_B = 2353.227$	$g-1 = 3-1 = 2$	1176.614
Entro i Gruppi	$D_W = 5679.500$	$n-g = 33-3 = 30$	189.317
Totale	$D_T = 8032.727$	$n-1 = 33-1 = 32$	251.023

Varianza spiegata e varianza residua

- La varianza tra gruppi, varianza delle medie entro i gruppi, viene anche detta **varianza spiegata**
- La varianza spiegata rappresenta la parte di variabilità totale riprodotta dalle medie condizionate
- La varianza entro gruppi, media ponderata delle varianze entro i gruppi, viene anche detta **varianza residua**
- La varianza residua rappresenta la parte di variabilità totale non spiegata dai gruppi
- Misura la variabilità interna alle distribuzioni condizionate rispetto alle proprie medie

Varianza spiegata e varianza residua





Verifica di ipotesi per l'uguaglianza tra le medie

- Sistema di ipotesi

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$$
$$\textit{versus}$$
$$H_a : \text{ Almeno un'uguaglianza in } H_0 \text{ è falsa}$$

- Statistica test

$$F = \frac{\text{Varianza tra gruppi}}{\text{Varianza entro i gruppi}} = \frac{D_B/(g-1)}{D_W/(n-g)}$$

  Sotto H_0 $F \sim F_{(g-1),(n-g)}$
- Regione di rifiuto al livello di significatività α

$$RC_\alpha : \quad F \geq f_{(g-1),(n-g)}(\alpha)$$

  dove $f_{(g-1),(n-g)}(\alpha) : Pr(F_{(g-1),(n-g)} \geq f_{(g-1),(n-g)}(\alpha)) = \alpha$
- $p - value = Pr(F_{(g-1),(n-g)} \geq F_{oss}; H_0)$

Tavola ANOVA

Fonte di Variabilità	Somma dei Quadrati	GdL	Media dei Quadrati	F-value	p–value
Tra Gruppi	D_B	$g-1$	$s^2_B = D_B/(g-1)$	s^2_B/s^2	$p-value$
Entro i Gruppi	D_W	$n-g$	$s^2 = D_W/(n-g)$		
Totale	D_T	$n-1$	$D_T/(n-1)$		

Verifica di ipotesi per l'uguaglianza tra le medie
Esempio: Capacità cognitive dei bambini

- Sistema di ipotesi

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \textit{versus} \quad H_a : \text{ Almeno un'uguaglianza in } H_0 \text{ è falsa}$$
$$(\mu_1 \neq \mu_2 \text{ oppure } \mu_1 \neq \mu_3 \text{ oppure } \mu_2 \neq \mu_3)$$

- Statistica test

$$F_{oss} = \frac{2353.227/(3-1)}{5679.5/(33-3)} = \frac{1176.614}{189.317} = 6.215$$

Sotto H_0, $F \sim F_{(3-1),(33-3)}$

- Regione di rifiuto al livello di significatività $\alpha = 0.05$

$$RC_{0.05} : \quad F \geq f_{(3-1),(33-3)}(0.05) = 3.316$$

- $F_{oss} = 6.215 > 3.316 = f_{(3-1),(33-3)}(0.05) \Longrightarrow$ I dati mostrano evidenza contro l'ipotesi nulla al livello di significatività del 5%
- $p - value = Pr(F_{(3-1),(33-3)} \geq 6.215; H_0) = 0.00009$

Tavola ANOVA – Esempio: Capacità cognitive dei bambini

Fonte di Variabilità	Somma dei Quadrati	GdL	Media dei Quadrati	F-value	p–value
Tra Gruppi	2353.227	2	1176.614	6.215	0.00009
Entro i Gruppi	5679.5	30	189.317		
Totale	8032.727	32	251.023		

Test per la differenza tra le medie
Popolazioni Normali e varianze ignote ma uguali

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti, σ_1^2 e σ_2^2 ignote ma $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{1n_1})\, i.i.d. \quad (Y_{21}, \ldots, Y_{2n_2})\, i.i.d. \quad \text{indipendenti}$$

- Sistema di ipotesi

$$H_0 : \mu_1 = \mu_2 \qquad \textit{versus} \qquad H_a : \mu_1 \neq \mu_2$$

- Stimatore congiunto (pooled) della varianza:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(Y_{1i} - \overline{Y}_1\right)^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(Y_{2i} - \overline{Y}_2\right)^2$$

$$\Longrightarrow S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Test per la differenza tra le medie
Popolazioni Normali e varianze ignote ma uguali

- Statistica test

$$\overline{Y}_1 - \overline{Y}_2 \quad \longmapsto \quad T = \frac{(\overline{Y}_1 - \overline{Y}_2) - 0}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Sotto l'ipotesi nulla $T \sim t_{n_1+n_2-2}$
- Livello di significatività del test = α
- Regione Critica (Regione di Rifiuto)

$$RC(\alpha) : T \leq -t_{(n_1+n_2-2),\alpha/2} \text{ oppure } T \geq t_{(n_1+n_2-2),\alpha/2}$$

- $p - value = 2 \cdot [1 - P(T_{n_1+n_2-2} \leq t_{oss}; H_0)] = P(F_{1,n_1+n_2-2} \geq F_{oss}; H_0)$

Test per la differenza tra le medie
Popolazioni Normali e varianze ignote ma uguali

- Tavola ANOVA

Devianza	SQ	GdL	MQ	F-value	p–value
Tra Gruppi	D_B	$2-1$	$s_B^2 = D_B/(2-1)$	s_B^2/s^2	$p-value$
Entro i Gruppi	D_W	$n-2$	$s^2 = D_W/(n-2)$		
Totale	D_T	$n-1$	$D_T/(n-1)$		

dove $n = n_1 + n_2$

$$s^2 = \frac{D_W}{n-2} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2}{n_1+n_2-2}$$

$$D_B = (\overline{y}_1 - \overline{y})^2 \cdot n_1 + (\overline{y}_2 - \overline{y})^2 \cdot n_2$$

$$\overline{y} = \frac{(y_{11}+\cdots+y_{1n_1}) + (y_{21}+\cdots+y_{2n_2})}{n_1+n_2}$$

- Sotto H_0, $F \sim F_{1,n_1+n_2-2} = t^2_{n_1+n_2-2}$

Test per la differenza tra le medie
Popolazioni Normali e varianze ignote ma uguali: Esempio

- Obiettivo: Confrontare il voto medio alla fine del primo anno di studenti di istituti professionali in classi in cui si sono adottati metodi di insegnamento standard (popolazione 1) e in classi in cui si sono adottati metodi di insegnamento interattivi (popolazione 2)

Metodo di insegnamento standard	Metodo di insegnamento interattivo
6.6	6.5
6.2	7.2
5.7	6.7
6.1	
5.4	

- Ipotesi

$$H_0 : \mu_1 = \mu_2 \quad vs \quad H_a : \mu_1 \neq \mu_2$$

- Stima della differenza tra le medie

$$\overline{y}_2 - \overline{y}_1 = \frac{20.4}{3} - \frac{30}{5} = 6.8 - 6 = 0.8$$

- Stima della varianza pooled

$$s_1^2 = \frac{0.86}{5-1} = 0.215 \qquad \text{e} \qquad s_2^2 = \frac{0.26}{3-1} = 0.13$$

Quindi

$$s_p^2 = \frac{4 \cdot 0.215 + 2 \cdot 0.13}{5+3-2} = \frac{0.86 + 0.26}{6} = \frac{1.12}{6} = 0.187$$

Test per la differenza tra le medie
Popolazioni Normali e varianze ignote ma uguali: Esempio

- Regione di rifiuto al livello di significatività $\alpha = 0.01$

$$RC_{0.01} = T \leq -3.707 \text{ oppure } T \geq 3.707$$

perché sotto H_0 $T \sim t_6$

- Valore osservato della statistica test

$$t_{oss} = \frac{\overline{y}_2 - \overline{y}_1}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8}{\sqrt{0.187\left(\frac{1}{5} + \frac{1}{3}\right)}} = \frac{0.8}{\sqrt{0.0996}} = \frac{0.8}{0.3155} = 2.535$$

- Decisione: Il valore osservato della statistica test NON appartiene alla regione di rifiuto: $-t_{0.005,6} = -3.707 < t_{oss} = 2.535 < 3.707 = t_{0.005,6}$. I dati NON mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 1%
- $p - value = 0.04435$

Test per la differenza tra le medie
Popolazioni Normali e varianze ignote ma uguali: Esempio

- Tavola ANOVA

Gruppo	n_ℓ	Media	Devianza	s^2_ℓ
1	5	6	0.86	0.216
2	3	6.8	0.26	0.130
1 + 2	8	6.3	2.32	

$$
\begin{aligned}
D_T &= 1.31 + 1.01 = 2.32 \\
D_W &= 0.86 + 0.26 = 1.12 \\
D_B &= (6 - 6.3)^2 \cdot 5 + (6.8 - 6.3)^2 \cdot 3 = 0.45 + 0.75 = 1.2
\end{aligned}
$$

Devianza	SQ	GdL	MQ	F-value	$p-$value
Tra Gruppi	1.20	1	1.2	6.43	0.04435
Entro i Gruppi	1.12	6	0.187		
Totale	2.32	7	0.331		

$$RC(0.01) : F \geq f_{1,6}(0.01) = 13.74 = 3.707^2$$

- Decisione: I dati NON mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 1%: $F_{oss} = 6.43 = (T_{oss})^2 = (2.535)^2 < 13.74 = f_{1,6}(0.01)$

Test F versus test t

- Per $g > 2$ si possono considerare confronti a coppie
- $g \cdot (g-1)/2$ possibili confronti
- Multipli test t: la probabilità di commettere l'errore di prima specie è relativo al singolo confronto ma non si applica a tutti i confronti congiuntamente considerati
- Fissato il livello di significatività del test t, α, α è la probabilità di commettere l'errore di prima specie in ciascuno dei $g \cdot (g-1)/2$ possibili confronti, ossia la probabilità di rifiutare l'ipotesi che due medie siano uguali quando le due popolazioni hanno uguale media
- Il test F permette di controllare la probabilità di commettere l'errore di I specie nei confronti multipli
- Fissato il livello di significatività del test F, α, α è la probabilità di commettere l'errore di prima specie, ossia di rifiutare l'ipotesi che le g medie siano uguali quando le g popolazioni hanno *tutte* la stessa media

Confronti a coppie: Intervalli di confidenza per la differenza tra due medie

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma^2), \ldots, Y_g \sim N(\mu_g, \sigma^2)$ indipendenti,
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{1n_1})\ i.i.d. \quad \ldots \quad (Y_{g1}, \ldots, Y_{gn_g})\ i.i.d. \quad \text{indipendenti}$$

- Stimatore della differenza tra le medie

$$\overline{Y}_h - \overline{Y}_k \sim N\left(\mu_h - \mu_k, \frac{\sigma^2}{n_h} + \frac{\sigma^2}{n_k}\right)$$

- Stimatore congiunto (pooled) della varianza:

$$S^2 = \frac{D_W}{n-g} = \frac{S_1^2 \cdot (n_1 - 1) + S_2^2 \cdot (n_2 - 1) + \cdots + S_g^2 \cdot (n_g - 1)}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_g - 1)}$$

Confronti a coppie: Intervalli di confidenza per la differenza tra due medie

- Statistica

$$T = \frac{(\overline{Y}_h - \overline{Y}_k) - (\mu_h - \mu_k)}{\sqrt{S_p^2 \left(\frac{1}{n_h} + \frac{1}{n_k} \right)}} \sim t_{n-g}$$

- Intervallo di confidenza al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\mu_h - \mu_k) = \left[(\overline{y}_h - \overline{y}_k) - t_{(n-g),\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_h} + \frac{1}{n_k} \right)}; \right.$$
$$\left. (\overline{y}_h - \overline{y}_k) + t_{(n-g),\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_h} + \frac{1}{n_k} \right)} \right]$$

Il livello di confidenza nei confronti multipli

- Il livello di confidenza $1-\alpha$ di ogni intervallo di confidenza non può essere interpretato come il livello di fiducia che ciascun intervallo di confididenza contenga la vera differenza tra le medie
- Confronti multipli: Fissare il livello di confidenza congiunto
- Nel campionamento ripetuto, il livello di confidenza congiunto $(1-\alpha)$ rappresenta la probabilità che tutti i $g \cdot (g-1)/2$ intervalli di confidenza contengano la vera differenza tra le medie
- $\alpha\%$ dei campioni porta a almeno un intervallo di confidenza tra i $g \cdot (g-1)/2$ intervalli di confidenza che non contiene la vera differenza tra le medie
- Confronto multiplo di Bonferroni: Per ottenere intervalli di confidenza di livello di confidenza congiunto $(1-\alpha)$, ciascun intervallo di confidenza è costruito utilizzando un livello di confidenza

$$1-\alpha_{Bonferroni} = 1 - \frac{\alpha}{\text{Numero di confronti}} = 1 - \frac{\alpha}{g \cdot (g-1)/2}$$

- Il livello di confidenza di Bonferroni garantisce che il livello di confidenza complessivo sia *almeno* pari a $1-\alpha$ (approccio conservativo)

Confronti multipli: Capacità cognitive dei bambini

- $g = 3 \Longrightarrow g \cdot (g-1)/2 = 3 \cdot (3-1)/2 = 3$ possibili confronti
- Stima pooled della varianza

$$s^2 = \frac{5679.5}{33-3} = 189.317$$

- Livello di confidenza: $1 - \alpha = 0.95 \Longrightarrow t_{30,0.025} = 2.042$
- Livello di confidenza di Bonferroni:

$$1 - \alpha_{Bonferroni} = 1 - \frac{0.05}{3} = 1 - 0.017 = 0.98 \Longrightarrow t_{30,0.017/2} = 2.536$$

- Confronti

Gruppi	$\overline{y}_h - \overline{y}_k$	$IC_{0.95}$	$IC_{0.95}$ Bonferroni
(P.I, No P.)	22.5	(9.44; 35.56)	(6.29; 38.71)
(P.I.,P.S.)	10.5	(−0.82; 21.82)	(−3.56; 24.56)
(P.S.,No.P)	12.0	(−0.45; 24.45)	(−3.46; 27.46)

Confronti multipli: Capacità cognitive dei bambini

- Livello di confidenza: $1 - \alpha = 0.90 \Longrightarrow t_{30,0.05} = 1.607$
- Livello di confidenza di Bonferroni:

$$1 - \alpha_{Bonferroni} = 1 - \frac{0.10}{3} = 1 - 0.033 = 0.967 \Longrightarrow t_{30,0.033/3} = 2.231$$

- Confronti

Gruppi	$\overline{y}_h - \overline{y}_k$	$IC_{0.90}$	$IC_{0.90}$Bonferroni
(P.I, No P.)	22.5	(11.65; 33.35)	(8.24; 36.76)
(P.I.,P.S.)	10.5	(1.09; 19.91)	(–1.87; 22.87)
(P.S.,No.P)	12.0	(1.65; 22.35)	(–1.60; 25.60)

Variabili indicatrici

Dato un fattore A con livelli $\ell = 1, \ldots, g$ si definiscono le variabili

$$A_{i\ell} = \begin{cases} 1 & \text{se l'unità } i \text{ appartiene al gruppo } \ell \\ 0 & \text{Altrimenti} \end{cases}$$

Le variabili $A_{i\ell}$ si dicono anche **variabili dummy** o variabili indicatrici associate ai livelli del fattore A e indicano la presenza o l'assenza di quel livello, ossia l'appartenenza dell'unità i a un dato gruppo ℓ

Variabili indicatrici

u_i	y_i	$A_{i,1}$	$A_{i,2}$	$A_{i,3}$	u_i	y_i	$A_{i,1}$	$A_{i,2}$	$A_{i,3}$
1	62	1	0	0	18	128	0	1	0
2	85	1	0	0	19	88	0	1	0
3	96	1	0	0	20	101	0	1	0
4	74	1	0	0	21	113	0	1	0
5	105	1	0	0	22	98	0	1	0
6	81	1	0	0	23	96	0	0	1
7	86	1	0	0	24	120	0	0	1
8	95	1	0	0	25	100	0	0	1
9	103	0	1	0	26	118	0	0	1
10	88	0	1	0	27	124	0	0	1
11	92	0	1	0	28	91	0	0	1
12	101	0	1	0	29	105	0	0	1
13	74	0	1	0	30	97	0	0	1
14	95	0	1	0	31	107	0	0	1
15	109	0	1	0	32	135	0	0	1
16	83	0	1	0	33	95	0	0	1
17	92	0	1	0					

Il modello di analisi della varianza

$$Y_{\ell i} = \mu_\ell + \epsilon_{\ell i} \qquad \epsilon_{\ell i} \sim N(0, \sigma^2) \text{ indipendenti}$$

$$(Y_{\ell i} \sim N(\mu_\ell, \sigma^2) \text{ indipendenti})$$

Oppure, equivalentemente,

$$Y_i = \mu_1 \cdot A_{i1} + \mu_2 \cdot A_{i2} + \cdots + \mu_g \cdot A_{ig} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \text{ indipendenti}$$

$$(Y_i \sim N(\mu_1 \cdot A_{i1} + \cdots + \mu_g \cdot A_{ig}, \sigma^2) \text{ indipendenti})$$

$$Y_{1,i} = \mu_1 + \epsilon_{1,i} \qquad Y_{2,i} = \mu_2 + \epsilon_{2,i} \qquad Y_{3,i} = \mu_3 + \epsilon_{3,i}$$

$$\epsilon_{\ell,i} \sim N(0, \sigma^2) \text{ indipendenti}$$

$\ell = 1, 2, 3 =$ No programma, Programma standard, Programma innovativo

Oppure, equivalentemente,

$$Y_i = \mu_1 \cdot A_{i,1} + \mu_2 \cdot A_{i,2} + \mu_3 \cdot A_{i,3} + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

Il modello di analisi della varianza con vincolo baseline

$$Y_i = \beta_0 + \beta_2 \cdot A_{i2} + \cdots + \beta_g \cdot A_{ig} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \text{ indipendenti}$$

$$\Downarrow$$

$$Y_i \sim N(\beta_0 + \beta_2 \cdot A_{i2} + \cdots + \beta_g \cdot A_{ig}, \sigma^2) \quad \text{indipendenti}$$

- $\beta_0 = \mu_1$
- $\beta_\ell = \mu_\ell - \mu_1 \Longrightarrow \mu_\ell = \beta_\ell + \beta_0 \qquad \ell = 1, \ldots, g$
- Avremmo potuto includere anche la prima variabile indicatrice, A_{1i} ma in tal caso avremmo incluso $g + 1$ parametri, $\beta_0, \beta_1, \ldots, \beta_g$ contro g medie $\mu_1, \ldots, \mu_k$, strettamente necessarie.
- Dal punto di vista della stima, il modello con $g + 1$ parametri ha infinite soluzioni per le stime dei minimi quadrati
- Per convenzione come gruppo di riferimento (baseline) si considera il gruppo 1 ma la scelta è arbitraria
- Le stime dei parametri β_ℓ sono differenze o *contrasti* tra la media nel livello/gruppo ℓ e la media del gruppo considerato come baseline

Il modello di analisi della varianza con vincolo baseline
Vincolo baseline

$$\widehat{\beta}_0 = \overline{Y}_1 \qquad \widehat{\beta}_1 = 0 \qquad \widehat{\beta}_\ell = \overline{Y}_\ell - \overline{Y}_1 \qquad \ell = 2, \ldots, g$$

- Stima dell'intercetta = Media campionaria per il gruppo baseline
- Le stime dei parametri β_ℓ sono differenze o *contrasti* tra la media nel livello/gruppo ℓ e la media del gruppo considerato come baseline
- I contrasti β_ℓ misurano l'effetto del fattore

Il modello di analisi della varianza con vincolo baseline
Esempio: Capacità cognitive dei bambini

$$Y_i = \beta_0 + \beta_2 \cdot A_{i,2} + \beta_3 \cdot A_{i,3} + \epsilon_i$$

con

$$\epsilon_i \sim N(0, \sigma^2) \text{ indipendenti}$$

e gruppo 1 ("No programma") come gruppo di riferimento

	Stima	*SE*	$t - value$	$p - value$
Costante	85.5	4.86	17.58	0.0000
Programma standard	12.0	6.10	1.97	0.0584
Programma Innovativo	22.5	6.39	3.52	0.0014

Il test F

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$$

$$\Longleftrightarrow$$

$$\mu_2 - \mu_1 = 0 \qquad \mu_3 - \mu_1 = 0 \quad \cdots \quad \mu_g - \mu_1 = 0$$

$$\Longleftrightarrow$$

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_g = 0$$

$$\Downarrow$$

$$F = \frac{\text{Varianza tra gruppi}}{\text{Varianza entro i gruppi}} = \frac{\text{Media dei quadrati di regressione}}{\text{Media dei quadrati dei residui}}$$

$$\| \qquad\qquad \|$$

$$\frac{D_B/(g-1)}{D_W/(n-g)} = \frac{RSS/k}{SSE/(n-k-1)}$$

con $k = g - 1 \Longrightarrow n - k - 1 = n - (g-1) - 1 = n - g$.

Sotto l'ipotesi nulla $F \sim F_{k,(n-k-1)} \equiv F_{(g-1),(n-g)}$, quindi

$$\begin{aligned} RC(\alpha) &= F \geq f_{k,(n-k-1)}(\alpha) \equiv f_{(g-1),(n-g)}(\alpha) \\ p - value &= Pr(F_{k,(n-k-1)} \geq F_{oss}; H_0) \equiv Pr(F_{(g-1),(n-g)} \geq F_{oss}; H_0) \end{aligned}$$

Tavola ANOVA

Fonte di Variabilità	Somma dei Quadrati	GdL	Media dei Quadrati	F-value	p–value
Tra Gruppi	D_B	$g-1$	$s^2_B = D_B/(g-1)$	s^2_B/s^2	$p-value$
Entro i Gruppi	D_W	$n-g$	$s^2 = D_W/(n-g)$		
Totale	D_T	$n-1$	$D_T/(n-1)$		

$\Longleftrightarrow$

Fonte di Variabilità	Somma dei Quadrati	GdL	Media dei Quadrati	F-value	p–value
Regressione	RSS	k	RSS/k	RMS/MSE	$p-value$
Residua	SSE	$n-k-1$	$SSE/(n-k-1)$		
Totale	TSS	$n-1$	$TSS/(n-1)$		

Verifica di ipotesi per l'uguaglianza tra le medie
Esempio: Capacità cognitive dei bambini

- Ipotesi

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \textit{versus} \quad H_a : \mu_1 \neq \mu_2 \text{ o } \mu_1 \neq \mu_3 \text{ o } \mu_2 \neq \mu_3$$

oppure, equivalentemente,

$$H_0 : \beta_2 = \beta_3 = 0 \quad \textit{versus} \quad H_a : \beta_2 \neq 0 \text{ o } \beta_3 \neq 0$$

- Tavola ANOVA

Fonte di Variabilità	Somma dei Quadrati	GdL	Media dei Quadrati	F-value	p–value
Tra Gruppi (Regressione)	2353.227	2	1176.614	6.215	0.00009
Entro i Gruppi (Residua)	5679.5	30	189.317		
Totale	8032.727	32	251.023		

Verifica di ipotesi per l'uguaglianza tra le medie
Esempio: Capacità cognitive dei bambini

- $k = g - 1 = 3 - 1 = 2 \Longrightarrow F \sim F_{2,(33-2-1)} \equiv F_{(3-1),(33-3)}$ sotto H_0
- Regione di rifiuto al livello di significatività $\alpha = 0.05$

$$RC_{0.05}: \quad F \geq f_{(3-1),(33-3)}(0.05) = 3.316$$

- $F_{oss} = 6.215 > 3.316 = f_{2,(33-2-1)}(0.05) \equiv f_{(3-1),(33-3)}(0.05) \Longrightarrow$ I dati mostrano evidenza contro l'ipotesi nulla al livello di significatività del 5%
- p-value

$$p-value = Pr(F_{2,(33-2-1)} \geq 6.215; H_0) = Pr(F_{(3-1),(33-3)} \geq 6.215; H_0) = 0.00009$$

Analisi della varianza con due fattori

- Due fattori (variabili categoriche) A e B possono esercitare un'influenza sulla variabile risposta Y
- Esempio I: Il direttore di una società ha raccolto le entrate (in migliaia di dollari) per 5 anni e in base al mese

  Anno = Fattore a 5 livelli (Anno 1, Anno 2,... Anno 5)
  Mese = Fattore a 12 livelli (Gennaio, ..., Dicembre)
  Entrate = Variabile risposta
- Esempio II: Valutare l'effetto di tre differenti trattamenti dietetici per soggetti obesi classificati per sesso.

  Trattamento dietetico = Fattore a 3 livelli
  Sesso = Fattore a 2 livelli (Maschio, Femmina)
  Peso = Variabile risposta

Analisi della varianza con due fattori

Una impresa di servizi informatici ha condotto un'indagine per valutare i fattori che determinano le differenze di salario

Livello di istruzione (Fattore A)	=	Fattore a 3 livelli
(1 = Diploma, 2 = Laurea, 3 = Master/Dottorato)		
Ruolo nell'azienda (Fattore B)	=	Fattore a 2 livelli
(0 = Impiegato, 1 = Funzionario)		
Salario	=	Variabile risposta

Salari degli informatici

u_i	y_i	A_i	B_i	u_i	y_i	A_i	B_i	u_i	y_i	A_i	B_i
1	13876	1	1	16	13231	3	0	31	15942	2	0
2	11608	3	0	17	12884	2	0	32	23174	3	1
3	18701	3	1	18	13245	2	0	33	23780	2	1
4	11283	2	0	19	13677	3	0	34	25410	2	1
5	11767	3	0	20	15965	1	1	35	14861	1	0
6	20872	2	1	21	12336	1	0	36	16882	2	0
7	11772	2	0	22	21352	3	1	37	24170	3	1
8	10535	1	0	23	13839	2	0	38	15990	1	0
9	12195	3	0	24	22884	2	1	39	26330	2	1
10	12313	2	0	25	16978	1	1	40	17949	2	0
11	14975	1	1	26	14803	2	0	41	25685	3	1
12	21371	2	1	27	17404	1	1	42	27837	2	1
13	19800	3	1	28	22184	3	1	43	18838	2	0
14	11417	1	0	29	13548	1	0	44	17483	1	0
15	20263	3	1	30	14467	1	0	45	19207	2	0
								46	19346	1	0

Analisi della varianza con due fattori: Medie di gruppo

- Siano A e B due fattori con H e K livelli rispettivamente
- Totale gruppi: $g = H \times K$
- Si possono calcolare $H \times K$ medie di gruppo

Fattore A	Fattore B				
	1	$\dots$	k	$\dots$	K
1	μ_{11}	$\dots$	μ_{1k}	$\dots$	μ_{1K}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h	μ_{h1}	$\dots$	μ_{hk}	$\dots$	μ_{hK}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
H	μ_{H1}	$\dots$	μ_{Hk}	$\dots$	μ_{HK}

- Confrontare le medie di Y per i livelli di A controllando per B
- Confrontare le medie di Y per i livelli di B controllando per A
- Confrontare le medie di cella

Analisi della varianza con due fattori: Esempi teorici

- Esempio I: Il salario medio non cambia rispetto al ruolo nell'azienda, controllando per il livello di istruzione

Livello di Istruzione	Ruolo nell'azienda	
	Impiegato	Funzionario
Diploma	15000	15000
Laurea	18500	18500
Master/Dottorato	18000	18000

- Esempio II: Il salario medio non varia con il livello di istruzione, controllando per il ruolo nell'azienda

Livello di Istruzione	Ruolo nell'azienda	
	Impiegato	Funzionario
Diploma	15000	20000
Laurea	15000	20000
Master/Dottorato	15000	20000

Analisi della varianza con due fattori: Esempi

- Esempio III: Istruzione e ruolo nell'azienda non hanno alcun effetto sul salario

Livello di Istruzione	Ruolo nell'azienda	
	Impiegato	Funzionario
Diploma	17000	17000
Laurea	17000	17000
Master/Dottorato	17000	17000

Analisi della varianza con due fattori: Medie di gruppo
Esempio: Salari degli informatici

- A e B hanno rispettivamente $H = 3$ e $K = 2$ livelli
- Totale gruppi: $g = H \times K = 3 \times 2 = 6$
- Si possono calcolare $H \times K = 3 \times 2 = 6$ medie di gruppo

Livello di Istruzione	Ruolo nell'azienda: Impiegato	Ruolo nell'azienda: Funzionario	Totale
Diploma	14442.56	15839.60	14941.50
	$(n_{11} = 9)$	$(n_{12} = 5)$	$(n_{1.} = 14)$
Laurea	14913.08	24069.14	18286.37
	$(n_{21} = 12)$	$(n_{22} = 7)$	$(n_{2.} = 19)$
Master/Dottorato	12495.60	21916.12	18292.85
	$(n_{31} = 5)$	$(n_{32} = 8)$	$(n_{3.} = 13)$
Totale	14285.31	21150.55	17270.2
	$(n_{.1} = 26)$	$(n_{.2} = 20)$	$(n = 46)$

Variabili indicatrici

- Variabili indicatrici per il fattore A

$$A_{ih} = \begin{cases} 1 & \text{se l'unità } i \text{ appartiene al gruppo } h \\ 0 & \text{Altrimenti} \end{cases} \qquad h = 1, \ldots, H$$

- Variabili indicatrici per il fattore B

$$B_{ik} = \begin{cases} 1 & \text{se l'unità } i \text{ appartiene al gruppo } k \\ 0 & \text{Altrimenti} \end{cases} \qquad k = 1, \ldots, K$$

Variabili indicatrici

Esempio: Salari degli informatici

u_i	y_i	A_{i1}	A_{i2}	A_{i3}	B_{i1}	B_{i2}
1	13876	1	0	0	0	1
2	11608	0	0	1	1	0
3	18701	0	0	1	0	1
4	11283	0	1	0	1	0
5	11767	0	0	1	1	0
6	20872	0	1	0	0	1
7	11772	0	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	17949	0	1	0	1	0
41	25685	0	0	1	0	1
42	27837	0	1	0	0	1
43	18838	0	1	0	1	0
44	17483	1	0	0	1	0
45	19207	0	1	0	1	0
46	19346	1	0	0	1	0

Modello di analisi della varianza con due fattori senza interazione

Vincolo Baseline

$$Y_i = \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

- Nel modello di analisi della varianza con due fattori A e B (senza interazione) aventi H e K livelli si costruiscono $H-1$ variabili indicatrici per A e $K-1$ variabili indicatrici per B
- Numero parametri: $1 + (H-1) + (K-1)$
- GdL: $n - 1 - (H-1) - (K-1) = n - H - K + 1$

Modello di analisi della varianza con due fattori senza interazione

Vincolo Baseline - Interpretazione dei parametri

- $\beta_0 = \mu_{11}$: Valore medio di Y nel gruppo con $A = 1$ e $B = 1$
- $\beta_h^A = \mu_{hk} - \mu_{1k}$: Differenza tra la media di Y nel gruppo con $A = h$ e $B = k$ e la media di Y nel gruppo con $A = 1$ e $B = k$, quale che sia $k = 1, \ldots, K$
- $\beta_k^B = \mu_{hk} - \mu_{h1}$: Differenza tra la media di Y nel gruppo con $A = h$ e $B = k$ e la media di Y nel gruppo con $A = h$ e $B = 1$, quale che sia $h = 1, \ldots, H$

$$\Downarrow$$

$$\mu_{11} = \beta_0 \qquad \mu_{h1} = \beta_0 + \beta_h^A \qquad \mu_{1k} = \beta_0 + \beta_k^B \qquad \mu_{hk} = \beta_0 + \beta_h^A + \beta_k^B$$

Modello di analisi della varianza con due fattori senza interazione

Vincolo Baseline: Salari degli informatici

$$Y_i = \beta_0 + \beta^A_{Laurea} A_{i,Laurea} + \beta^A_{Master/PhD} A_{i,Master/PhD} + \beta^B_{Funzionario} B_{i,Funzionario} + \epsilon_i$$

- β_0: Salario medio degli impiegati con diploma
- β^A_{Laurea}: Differenza tra il salario medio dei laureati e il salario medio dei diplomati fissato il ruolo nell'azienda

$$\begin{aligned} \beta^A_{Laurea} &= \mu_{Laurea,Impiegato} - \mu_{Diploma,Impiegato} \\ &= \mu_{Laurea,Funzionario} - \mu_{Diploma,Funzionario} \end{aligned}$$

- $\beta^A_{Master/PhD}$: Differenza tra il salario medio di dipendenti con master o dottorato e il salario medio dei diplomati fissato il ruolo nell'azienda

$$\begin{aligned} \beta^A_{Master/PhD} &= \mu_{Master/PhD,Impiegato} - \mu_{Diploma,Impiegato} \\ &= \mu_{Master/PhD,Funzionario} - \mu_{Diplona,Funzionario} \end{aligned}$$

- $\beta^B_{Funzionario}$: Differenza tra il salario medio di funzionari e il salario medio di impiegati fissato il livello di istruzione

$$\begin{aligned} \beta^B_{Funzionario} &= \mu_{Diploma,Funzionario} - \mu_{Diploma,Impiegato} \\ &= \mu_{Laurea,Funzionario} - \mu_{Laurea,Impiegato} \\ &= \mu_{Master/PhD,Funzionario} - \mu_{Master/PhD,Impiegato} \end{aligned}$$

Modello di analisi della varianza con due fattori senza interazione

Vincolo Baseline: Salari degli informatici

Variabile	*Stima*	*ES*	*tvalue*	$p - value$
Costante	12475.8291	869.8989	14.34	0.0000
Istruzione (Baseline: Diploma)				
Laurea	3267.0051	1061.4278	3.08	0.0037
Master/PhD	1568.4764	1184.7484	1.32	0.1927
Ruolo (Baseline: Impiegato)				
Funzionario	6903.8785	920.6890	7.50	0.0000

- $\widehat{\beta}_0 = 12475.8291$: Salario medio degli impiegati con diploma
- $\widehat{\beta}^A_{Laurea} = 3267.0051$: Differenza tra il salario medio dei laureati e il salario medio dei diplomati fissato il ruolo nell'azienda
- $\widehat{\beta}^A_{Master/PhD} = 1568.4764$: Differenza tra il salario medio di dipendenti con master o dottorato e il salario medio dei diplomati fissato il ruolo nell'azienda
- $\widehat{\beta}^B_{Funzionario} = 6903.8785$: Differenza tra il salario medio di funzionari e il salario medio di impiegati fissato il livello di istruzione

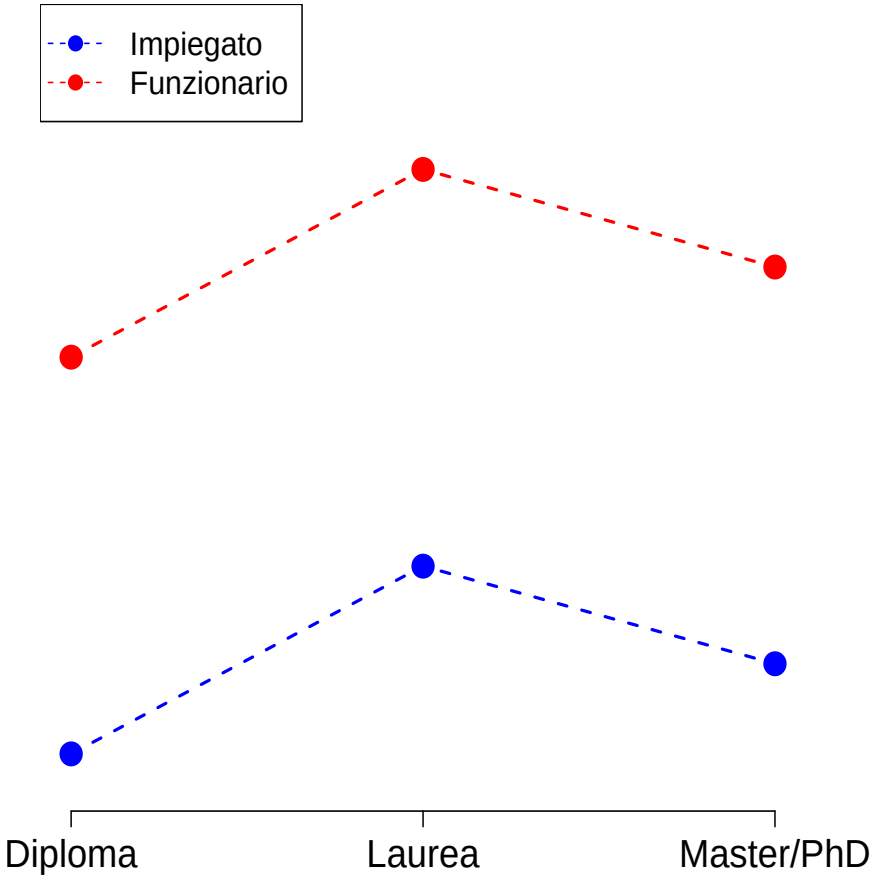
Modello di analisi della varianza con due fattori senza interazione

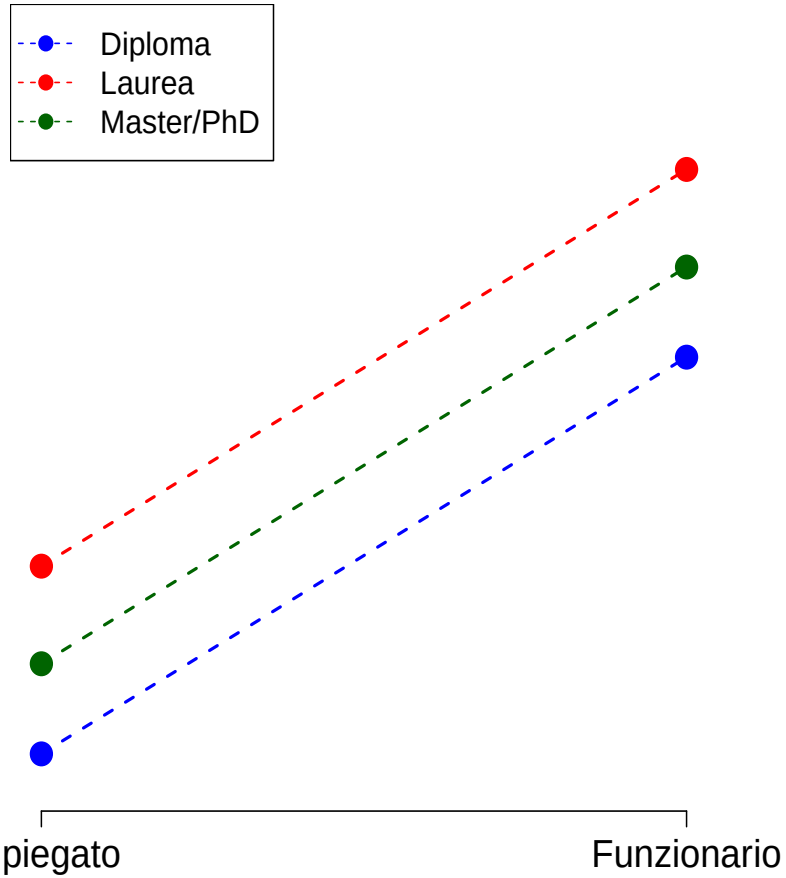
Valori teorici

Livello di istruzione	Ruolo nell'azienda: Impiegato	Funzionario
Diploma	$\widehat{\mu}_{D,I} = \widehat{\beta}_0$	$\widehat{\mu}_{D,F} = \widehat{\beta}_0 + \widehat{\beta}^B_{Funzionario}$
Laurea	$\widehat{\mu}_{L,I} = \widehat{\beta}_0 + \widehat{\beta}^A_{Laurea}$	$\widehat{\mu}_{L,F} = \widehat{\beta}_0 + \widehat{\beta}^A_{Laurea} + \widehat{\beta}^B_{Funzionario}$
Master/PhD	$\widehat{\mu}_{M/PhD,I} = \widehat{\beta}_0 + \widehat{\beta}^A_{Master/PhD}$	$\widehat{\mu}_{M/PhD,F} = \widehat{\beta}_0 + \widehat{\beta}^A_{Master/PhD} + \widehat{\beta}^B_{Funzionario}$

Diploma,Impiegato	$\widehat{\mu}_{D,I}$	=	12475.83
Diploma,Funzionario	$\widehat{\mu}_{D,F}$	=	12475.83 + 6903.8785 = 19379.71
Laurea,Impiegato	$\widehat{\mu}_{L,I}$	=	12475.83 + 3267.00 = 15742.83
Laurea,Funzionario	$\widehat{\mu}_{L,F}$	=	12475.83 + 3267.00 + 6903.88 = 22646.71
Master/PhD,Impiegato	$\widehat{\mu}_{M/PhD,I}$	=	12475.83 + 1568.48 = 14044.31
Master/PhD,Funzionario	$\widehat{\mu}_{M/PhD,F}$	=	12475.83 + 1568.48 + 6903.88 = 20948.19

Modello di analisi della varianza con due fattori senza interazione





Il test F

$$Y_i = \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$$

- Sistema di ipotesi

$$H_0 : \beta_2^A = \cdots = \beta_H^A = \beta_2^B = \cdots = \beta_K^B = 0$$

versus

$$H_a : \text{Almeno un coefficiente è diverso da zero}$$

- Statistica test

$$F = \frac{RSS/(H-1+K-1)}{SSE/(n-H-K+1)} = \frac{RMS}{MSE} \sim F_{(H-1+K-1),(n-H-K+1)} \quad \text{sotto } H_0$$

- Regione critica di livello di significatività α

$$RC_\alpha : \quad F \geq f_{(H-1+K-1),(n-H-K+1)}(\alpha)$$

dove $f_{(H-1+K-1),(n-H-K+1)}(\alpha)$:

$$Pr(F_{(H-1+K-1),(n-H-K+1)} \geq f_{(H-1+K-1),(n-H-K+1)}(\alpha)) = \alpha$$

Tavola di Analisi della Varianza (Tavola ANOVA)

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p-value$
Regressione	RSS	$(H-1)+(K-1)$	RMS	RMS/MSE	$p-value$
Residua	SSE	$n-H-K+1$	MSE		
Totale	TSS	$n-1$	MQT		

dove $p-value = Pr(F_{(H-1+K-1),(n-H-K+1)} \geq F_{oss}; H_0)$

Il test F: Salari degli informatici

$$Y_i = \beta_0 + \beta^A_{Laurea} A_{i,Laurea} + \beta^A_{Master/PhD} A_{i,Master/PhD} + \beta^B_{Funzionario} B_{i,Funzionario} + \epsilon_i$$

- Ipotesi

$$H_0 : \beta^A_{Laurea} = \beta^A_{Master/PhD} = \beta^B_{Funzionario} = 0$$

versus

H_a : Almeno un coefficiente è diverso da zero

- Tavola ANOVA

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	619718936	3	206572979	22.749	0.0000
Residua	381378642	42	9080444		
Totale	1001097577	45	22246613		

Il test F: Salari degli informatici

- Statistica test

$$F_{oss} = \frac{619718936/3}{381378642/42} = \frac{206572979}{9080444} = 22.749$$

- Regione critica di livello di significatività $\alpha = 0.05$

$$RC_{0.05}: \quad F \geq f_{3,42}(0.05) = 2.827$$

- p–value

$$p - value = Pr(F_{3,42} \geq 22.749; H_0) = 0.0000$$

- Forte evidenza contro l'ipotesi nulla

Modelli a confronto

Modello esteso :

$$Y_i = \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$$

versus

Modell ridotto : $Y_i = \beta_0 + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$

$$\Longleftrightarrow$$

$$H_0 : \beta_2^A = \beta_3^A = \cdots = \beta_H^A = 0$$

versus

H_a : Almeno un'uguaglianza in H_0 è falsa

Esempio: Salari degli informatici

$$M_e : Y_i = \beta_0 + \beta^A_{Laurea} A_{i,Laurea} + \beta^A_{Master/PhD} A_{i,Master/PhD} + \beta^B_{Funzionario} B_{i,Funzionario} + \epsilon_i$$

versus

$$M_r : Y_i = \beta_0 + \beta^B_{Funzionario} B_{i,Funzionario} + \epsilon_i$$

$$\Longleftrightarrow$$

$$H_0 : \beta^A_{Laurea} = \beta^A_{Master/PhD} = 0 \quad \text{versus} \quad H_a : \beta^A_{Laurea} \neq 0 \text{ oppure } \beta^A_{Master/PhD} \neq 0$$

Modelli a confronto

- Valori teorici

$$\widehat{y}_{ie} = \widehat{\beta}_0 + \widehat{\beta}_2^A \cdot A_{i2} + \cdots + \widehat{\beta}_H^A \cdot A_{iH} + \widehat{\beta}_2^B \cdot B_{i2} + \cdots + \widehat{\beta}_K^B \cdot B_{iK}$$
$$\widehat{y}_{ir} = \widehat{\beta}_0 + \widehat{\beta}_2^B \cdot B_{i2} + \cdots + \widehat{\beta}_K^B \cdot B_{iK}$$

- Somma dei quadrati degli errori

$$\text{Modello Esteso} \quad SSE_e = \sum_{i=1}^{n} (y_i - \widehat{y}_{ie})^2 \quad \text{con } gdl_e = n - H - K + 1$$
$$\text{Modello Ridotto} \quad SSE_r = \sum_{i=1}^{n} (y_i - \widehat{y}_{ir})^2 \quad \text{con } gdl_r = n - (K - 1) - 1$$

Modelli di regressione a confronto

- Statistica Test

$$F = \frac{(SSE_r - SSE_e)/(gdl_r - gdl_e)}{SSE_e/gdl_e} = \frac{(SSE_r - SSE_e)/(H-1)}{SSE_e/(n-H-K+1)}$$

  dove $gdl_r - gdl_e = H - 1$ = numero di termini aggiuntivi presenti nel modello esteso
- Sotto l'ipotesi nulla $F \sim F_{(H-1),n-H-K+1}$
- Regione di rifiuto al livello di significatività α:

$$RC_\alpha : \quad F \geq f_{H-1,n-H-K+1}(\alpha)$$

- $p - value = Pr(F_{H-1,n-H-K+1} \geq F_{oss}; H_0)$

Modelli a confronto – Esempio: Salari degli informatici

Modello esteso :

$$Y_i = \beta_0 + \beta^A_{Laurea} A_{i,Laurea} + \beta^A_{Master/PhD} A_{i,Master/PhD} + \beta^B_{Funzionario} B_{i,Funzionario} + \epsilon_i$$

versus

Modell ridotto : $Y_i = \beta_0 + \beta^B_{Funzionario} B_{i,Funzionario} + \epsilon_i$

$$\Longleftrightarrow$$

$$H_0 : \beta^A_{Laurea} = \beta^A_{Master/PhD} = 0$$

versus

$$H_a : \beta^A_{Laurea} \neq 0 \text{ oppure } \beta^A_{Master/PhD} \neq 0$$

Modelli a confronto – Esempio: Salari degli informatici

- Tavola ANOVA modello esteso

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	619718936	3	206572979	22.749	0.0000
Residua	381378642	42	9080444		
Totale	1001097577	45	22246613		

- Tavola ANOVA modello ridotto

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	532791457	1	532791457	50.059	0.0000
Residua	468306120	44	10643321		
Totale	1001097577	45	22246613		

Il test F: Salari degli informatici

- Statistica test

$$F_{oss} = \frac{(468306120 - 381378642)/(44 - 42)}{381378642/42} = \frac{43463739}{9080444} = 4.7865$$

- Regione critica di livello di significatività $\alpha = 0.05, 0.01$

$RC_{0.05}$: $F \geq f_{2,42}(0.05) = 3.22$ $RC_{0.01}$: $F \geq f_{2,42}(0.01) = 5.149$

- p–value

$$p - value = Pr(F_{2,42} \geq 4.7865; H_0) = 0.01341$$

- Evidenza contro l'ipotesi nulla al livello di significatività del 5% ma i dati non mostrano sufficiente evidenza contraria all'ipotesi nulla al livello di significatività del 1%

Modelli a confronto – Esempio: Salari degli informatici

Modello	SSE	GdL	$F - value$	$p - value$
Modello esteso	381378642	42		
Istruzione: $H_0 : \beta^A_{Laurea} = \beta^A_{Master/PhD} = 0$	468306120	44	4.7865	0.0134
Ruolo nell'azienda: $H_0 : \beta^B_{Funzionario} = 0$	891962932	43	56.229	0.0000
Modello nullo: $H_0 : \beta^A_{Laurea} = \beta^A_{Master/PhD} = 0$ $\beta^B_{Funzionario} = 0$	1001097577	45	22.749	0.0000

Modello di analisi della varianza con due fattori con interazione

$$\begin{aligned} Y_i &= \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} \\ &+ \beta_{22}^{AB} A_{i2} \cdot B_{i2} + \ldots + \beta_{2K}^{AB} A_{i2} \cdot B_{iK} + \ldots \\ &+ \beta_{H2}^{AB} A_{iH} \cdot B_{i2} + \ldots + \beta_{HK}^{AB} A_{iH} \cdot B_{iK} + \epsilon_i \end{aligned}$$

con $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

- Nel modello di analisi della varianza con due fattori A e B (con interazione) aventi H e K livelli si costruiscono
  - ✓ $H-1$ variabili indicatrici per A;
  - ✓ $K-1$ variabili indicatrici per B;
  - ✓ $(H-1)\cdot(K-1)$ prodotti di variabili indicatrici per i parametri di interazione

  Non sono incluse le variabili indicatrici A_1 e B_1 e i prodotti di tali due variabili indicatrici con ogni altra variabile indicatrice $(A_1 \cdot B_1, \ldots, A_1 \cdot B_K,\ A_2 \cdot B_1, \ldots, A_H \cdot B_1)$
- Numero parametri: $1 + (H-1) + (K-1) + (H-1)\cdot(K-1)$
- GdL: $n - [1 + (H-1) + (K-1) + (H-1)\cdot(K-1)] = n - H \cdot K$

Modello di analisi della varianza con due fattori con interazione

Interpretazione dei parametri

$$\mu_{11} = \beta_0 \qquad \mu_{h1} = \beta_0 + \beta_h^A \qquad \mu_{1k} = \beta_0 + \beta_k^B \qquad \mu_{hk} = \beta_0 + \beta_h^A + \beta_k^B + \beta_{hk}^{AB}$$

- $\beta_0 = \mu_{11}$: Valore medio di Y nel gruppo con $A_{i1} = 1$ e $B_{i1} = 1$
- $\beta_h^A = \mu_{h1} - \mu_{11}$: Differenza tra la media di Y nel gruppo con $A_{ih} = 1$ e $B_{i1} = 1$ la media di Y nel gruppo con $A_{i1} = 1$ e $B_{i1} = 1$
- $\beta_k^B = \mu_{1k} - \mu_{11}$: Differenza tra la media di Y nel gruppo con $A_{i1} = 1$ e $B_{ik} = 1$ la media di Y nel gruppo con $A_{i1} = 1$ e $B_{i1} = 1$
- β_{hk}^{AB}: Variazione dell'effetto della variabile dummy A_h nel gruppo con $B_{ik} = 1$ rispetto al gruppo con $B_{i1} = 1 \iff$ Variazione dell'effetto della variabile dummy B_k nel gruppo con $A_{ih} = 1$ rispetto al gruppo con $A_{i1} = 1$
  - ✓ Il termine di interazione rappresenta una differenza di differenze tra medie:

$$\begin{aligned}
&[\text{Media per } A_{ih} = 1, B_{ik} = 1 - \text{Media per } A_{ih} = 1, B_{i1} = 1] - \\
&\quad [\text{Media per } A_{i1} = 1, B_{ik} = 1 - \text{Media per } A_{i1} = 1, B_{i1} = 1] \\
&[\text{Media per } A_{ih} = 1, B_{ik} = 1 - \text{Media per } A_{i1} = 1, B_{ik} = 1] - \\
&\quad [\text{Media per } A_{ih} = 1, B_{i1} = 1 - \text{Media per } A_{i1} = 1, B_{i1} = 1]
\end{aligned}$$

Modello di analisi della varianza con due fattori con interazione

Esempio: Salari degli informatici

$$
\begin{aligned}
Y_i \quad = \quad & \beta_0 + \beta^A_{Laurea} A_{i,Laurea} + \beta^A_{Master/PhD} A_{i,Master/PhD} + \beta^B_{Funzionario} B_{i,Funzionario} + \\
& \beta^{AB}_{L,F} A_{i,Laurea} \cdot B_{i,Funzionario} + \beta^{AB}_{Master/PhD,F} A_{i,Master/PhD} \cdot B_{i,Funzionario} + \epsilon_i
\end{aligned}
$$

- Numero parametri interazione

$$H \cdot K - (H + K - 1) = (H - 1) \cdot (K - 1) = (3 - 1) \cdot (2 - 1) = 2$$

- GdL

$$n - H \cdot K = 46 - 3 \cdot 2 = 40$$

Modello di analisi della varianza con due fattori con interazione

Esempio: Salari degli informatici

- β_0: Salario medio degli impiegati con diploma
- β^A_{Laurea}: Differenza tra il salario medio dei laureati *impiegati* e il salario medio dei diplomati *impiegati*

$$\beta^A_{Laurea} = \mu_{Laurea,Impiegato} - \mu_{Diploma,Impiegato}$$

- $\beta^A_{Master/PhD}$: Differenza tra il salario medio di dipendenti con master o dottorato *impiegati* e il salario medio dei diplomati *impiegati*

$$\beta^A_{Master/PhD} = \mu_{Master/PhD,Impiegato} - \mu_{Diploma,Impiegato}$$

- $\beta^B_{Funzionario}$: Differenza tra il salario medio di funzionari *diplomati* e il salario medio di impiegati *diplomati*

$$\beta^B_{Funzionario} = \mu_{Diploma,Funzionario} - \mu_{Diploma,Impiegato}$$

Modello di analisi della varianza con due fattori con interazione

Esempio: Salari degli informatici

- $\beta^{AB}_{L,F}$: Variazione dell'effetto della laurea per i funzionari rispetto agli impiegati $\Longleftrightarrow$ Variazione dell'effetto del ruolo nell'azienda tra i laureati rispetto ai diplomati

$$\begin{aligned}\beta^{AB}_{L,F} &= \left[\mu_{Laurea,Funzionario} - \mu_{Laurea,Impiegato}\right] - \left[\mu_{Diploma,Funzionario} - \mu_{Diploma,Impiegato}\right] \\ &= \left[\mu_{Laurea,Funzionario} - \mu_{Diploma,Funzionario}\right] - \left[\mu_{Laurea,Impigato} - \mu_{Diploma,Impiegato}\right]\end{aligned}$$

- $\beta^{AB}_{Master/PhD,F}$: Variazione dell'effetto del master/dottorato per i funzionari rispetto agli impiegati $\Longleftrightarrow$ Variazione dell'effetto del ruolo nell'azienda tra coloro che hanno il master o il dottorato rispetto a coloro che hanno solo il diploma

$$\begin{aligned}\beta^{AB}_{Master/PhD,F} = \\ &= \left[\mu_{Master/PhD,Funzionario} - \mu_{Master/PhD,Impiegato}\right] - \left[\mu_{Diploma,Funzionario} - \mu_{Diploma,Impiegato}\right] \\ &= \left[\mu_{Master/PhD,Funzionario} - \mu_{Diploma,Funzionario}\right] - \left[\mu_{Master/PhD,Impigato} - \mu_{Diploma,Impiegato}\right]\end{aligned}$$

Modello di analisi della varianza con due fattori con interazione

Esempio: Salari degli informatici

Variabile	*Stima*	*ES*	$t - value$	$p - value$
Costante	14442.5556	819.8780	17.62	0.0000
Istruzione (Baseline: Diploma)				
Laurea	470.5278	1084.5967	0.43	0.6667
Master/PhD	−1946.9556	1371.9184	−1.42	0.1636
Ruolo (Baseline: Impiegato)				
Funzionario	1397.0444	1371.9184	1.02	0.3146
Interazione				
Laurea · Funzionario	7759.0151	1802.9329	4.30	0.0001
Master/PhD · Funzionario	8023.4806	1961.7199	4.09	0.0002

Modello di analisi della varianza con due fattori con interazione

Esempio: Salari degli informatici

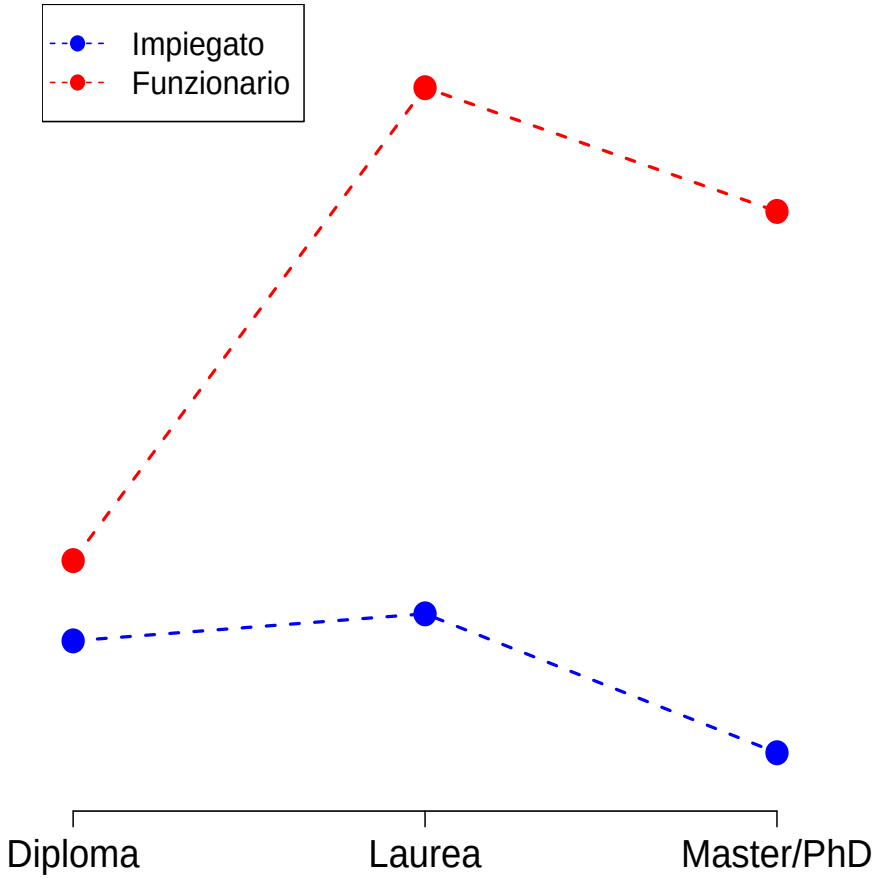
- $\widehat{\beta}_0 = 14442.5556$: Salario medio degli impiegati con diploma
- $\widehat{\beta}^A_{Laurea} = 470.5278$: Differenza tra il salario medio dei laureati *impiegati* e il salario medio dei diplomati *impiegati*
- $\widehat{\beta}^A_{Master/PhD} = -1946.9556$: Differenza tra il salario medio di dipendenti *impiegati* con master o dottorato e il salario medio dei dipendenti *impiegati* con diploma
- $\widehat{\beta}^B_{Funzionario} = 1397.0444$: Differenza tra il salario medio di funzionari *con diploma* e il salario medio di impiegati *con diploma*
- $\beta^{AB}_{L,F} = 7759.0151$: L'effetto della laurea sul salario è maggiore di 7759.0151 Euro per i funzionari rispetto agli impiegati $\Longleftrightarrow$ L'effetto del ruolo di funzionario sul salario è maggiore di 7759.0151 Euro per i laureati rispetto ai diplomati
- $\beta^{AB}_{Master/PhD,F} = 8023.4806$: L'effetto del Master/dottorato sul salario è maggiore di 8023.4806 Euro per i funzionari rispetto agli impiegati $\Longleftrightarrow$ L'effetto del ruolo di funzionario sul salario è maggiore di 8023.4806 Euro per i dipendenti con master/dottorato rispetto ai dipendenti con diploma

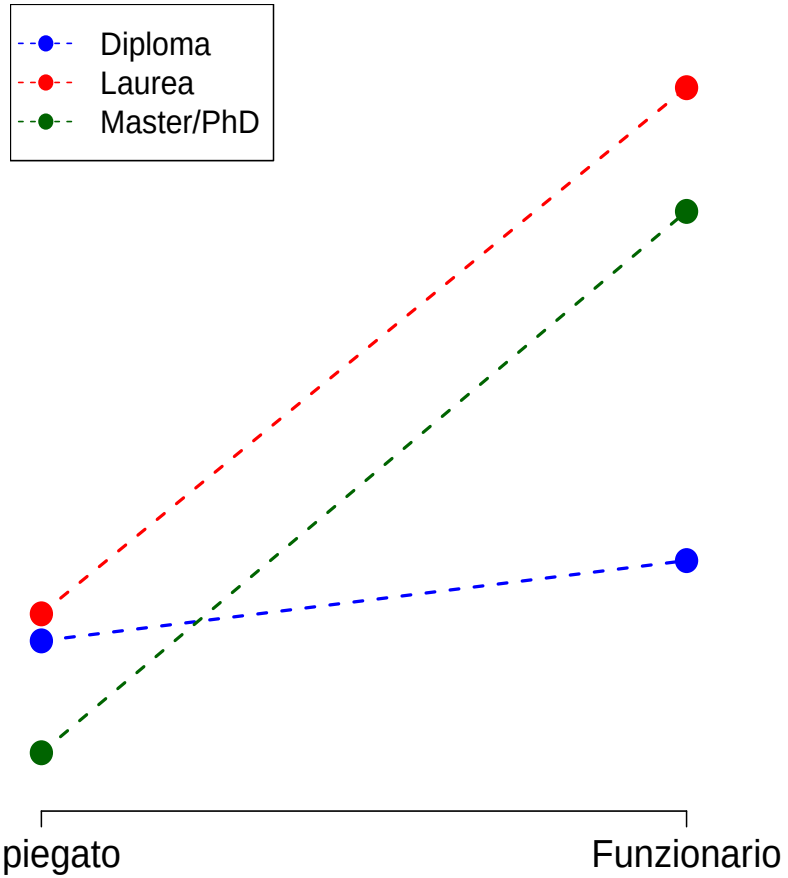
Analisi della varianza con due fattori con interazione
Valori teorici

Livello di istruzione	Ruolo nell'azienda: Impiegato	Funzionario
Diploma	$\widehat{\mu}_{D,I} = \widehat{\beta}_0$	$\widehat{\mu}_{D,F} = \widehat{\beta}_0 + \widehat{\beta}^B_{Funzionario}$
Laurea	$\widehat{\mu}_{L,I} = \widehat{\beta}_0 + \widehat{\beta}^A_{Laurea}$	$\widehat{\mu}_{L,F} = \widehat{\beta}_0 + \widehat{\beta}^A_{Laurea} + \widehat{\beta}^B_{Funzionario} + \beta^{AB}_{L,F}$
Master/PhD	$\widehat{\mu}_{M/PhD,I} = \widehat{\beta}_0 + \widehat{\beta}^A_{Master/PhD}$	$\widehat{\mu}_{M/PhD,F} = \widehat{\beta}_0 + \widehat{\beta}^A_{Master/PhD} + \widehat{\beta}^B_{Funzionario} + \beta^{AB}_{Master/PhD,F}$

Diploma,Impiegato	$\widehat{\mu}_{D,I}$	$= 14442.56$
Diploma,Funzionario	$\widehat{\mu}_{D,F}$	$= 14442.56 + 1397.04 = 15839.6$
Laurea,Impiegato	$\widehat{\mu}_{L,I}$	$= 14442.56 + 470.53 = 14913.09$
Laurea,Funzionario	$\widehat{\mu}_{L,F}$	$= 14442.56 + 470.53 + 1397.04 + 7759.01 = 24069.14$
Master/PhD,Impiegato	$\widehat{\mu}_{M/PhD,I}$	$= 14442.56 - 1946.96 = 12495.6$
Master/PhD,Funzionario	$\widehat{\mu}_{M/PhD,F}$	$= 14442.56 - 1946.96 + 1397.04 + 8023.48 = 21916.12$

Modello di analisi della varianza con due fattori con interazione





Valutare la significatività dell'interazione

Modello esteso:

$$\begin{aligned} Y_i &= \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} \\ &+ \beta_{22}^{AB} A_{i2} \cdot B_{i2} + \ldots + \beta_{2K}^{AB} A_{i2} \cdot B_{iK} + \ldots \\ &+ \beta_{H2}^{AB} A_{iH} \cdot B_{i2} + \ldots + \beta_{HK}^{AB} A_{iH} \cdot B_{iK} + \epsilon_i \end{aligned}$$

versus

Modello ridotto:

$$Y_i = \beta_0 + \beta_2^A A_{i2} + \ldots + \beta_H^A A_{iH} + \beta_2^B B_{i2} + \ldots + \beta_K^B B_{iK} + \epsilon_i$$

$$\Updownarrow$$

$$H_0 : \beta_{22}^{AB} = \cdots = \beta_{2K}^{AB} = \cdots = \beta_{H2}^{AB} = \cdots = \beta_{HK}^{AB} = 0$$

versus

H_a : Almeno un'uguaglianza in H_0 è falsa

Valutare la significatività dell'interazione
Esempio: Salari degli informatici

$$
\begin{aligned}
M_e: \quad Y_i \quad = \quad & \beta_0 + \beta^A_{Laurea} A_{i,Laurea} + \beta^A_{Master/PhD} A_{i,Master/PhD} + \\
& \beta^B_{Funzionario} B_{i,Funzionario} + \beta^{AB}_{L,F} A_{i,Laurea} \cdot B_{i,Funzionario} + \\
& \beta^{AB}_{Master/PhD,F} A_{i,Master/PhD} \cdot B_{i,Funzionario} + \epsilon_i
\end{aligned}
$$

versus

$$
\begin{aligned}
M_r: \quad Y_i \quad = \quad & \beta_0 + \beta^A_{Laurea} A_{i,Laurea} + \beta^A_{Master/PhD} A_{i,Master/PhD} + \\
& \beta^B_{Funzionario} B_{i,Funzionario} + \epsilon_i
\end{aligned}
$$

$$\Longleftrightarrow$$

$$H_0 : \beta^{AB}_{L,F} = \beta^{AB}_{Master/PhD,F} = 0 \text{ versus } H_a : \beta^{AB}_{L,F} \neq 0 \text{ oppure } \beta^{AB}_{Master/PhD,F} \neq 0$$

Valutare la significatività dell'interazione

- Valori teorici

$$\begin{aligned}\widehat{y}_{ie} &= \widehat{\beta}_0 + \widehat{\beta}_2^A \cdot A_{i2} + \cdots + \widehat{\beta}_h^A \cdot A_{ih} + \widehat{\beta}_2^B \cdot B_{i2} \cdots + \widehat{\beta}_K^B \cdot B_{iK} + \\ &\quad \widehat{\beta}_{22}^{AB} \cdot A_{i2} \cdot B_{i2} + \cdots + \widehat{\beta}_{HK}^{AB} \cdot A_{iH} \cdot B_{iK} \\ \widehat{y}_{ir} &= \widehat{\beta}_0 + \widehat{\beta}_2^A \cdot A_{i2} + \cdots + \widehat{\beta}_h^A \cdot A_{ih} + \widehat{\beta}_2^B \cdot B_{i2} \cdots + \widehat{\beta}_K^B \cdot B_{iK}\end{aligned}$$

- Somma dei quadrati degli errori

$$\text{Modello Esteso} \quad SSE_e = \sum_{i=1}^{n} (y_i - \widehat{y}_{ie})^2 \quad \text{con } gdl_e = n - H \cdot K$$

$$\text{Modello Ridotto} \quad SSE_r = \sum_{i=1}^{n} (y_i - \widehat{y}_{ir})^2 \quad \text{con } gdl_r = n - H - K + 1$$

Valutare la significatività dell'interazione

- Statistica Test

$$F = \frac{(SSE_r - SSE_e)/(gdl_r - gdl_e)}{SSE_e/gdl_e}$$
$$= \frac{(SSE_r - SSE_e)/(H \cdot K - H - K + 1)}{SSE_e/(n - H \cdot K)}$$

  dove $gdl_r - gdl_e = H \cdot K - H - K + 1$ = numero di termini aggiuntivi presenti nel modello esteso

- Sotto l'ipotesi nulla $F \sim F_{(H \cdot K - H - K + 1), n - H \cdot K}$
- Regione di rifiuto al livello di significatività α:

$$RC_\alpha : \quad F \geq f_{(H \cdot K - H - K + 1), n - H \cdot K}(\alpha)$$

- $p - value = Pr(F_{(H \cdot K - H - K + 1), n - H \cdot K} \geq F_{oss}; H_0)$

Valutare la significatività dell'interazione
Esempio: Salari degli informatici

$H_0 : \beta^{AB}_{L,F} = \beta^{AB}_{Master/PhD,F} = 0$ versus $H_a : \beta^{AB}_{L,F} \neq 0$ oppure $\beta^{AB}_{Master/PhD,F} \neq 0$

- Tavola ANOVA modello esteso (modello con interazione)

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	759105574	5	151821115	25.095	0.0000
Residua	241992003	40	6049800		
Totale	1001097577	45	22246613		

- Tavola ANOVA modello ridotto (modello senza interazione)

Fonte di Variabilità	Somma dei quadrati	GdL	Media dei quadrati	Statistica F	$p - value$
Regressione	619718936	3	206572979	22.749	0.0000
Residua	381378642	42	9080444		
Totale	1001097577	45	22246613		

Valutare la significatività dell'interazione
Esempio: Salari degli informatici

- Statistica test

$$F_{oss} = \frac{(381378642 - 241992003)/(6-3-2+1)}{241992003/40}$$
$$= \frac{139386639/2}{6049800} = 11.520$$

- Regione critica di livello di significatività $\alpha = 0.05$

$$RC_{0.05}: \quad F \geq f_{2,40}(0.05) = 3.232$$

- p–value

$$p - value = Pr(F_{2,40} \geq 11.52; H_0) = 0.00011192$$

- Forte evidenza contro l'ipotesi nulla