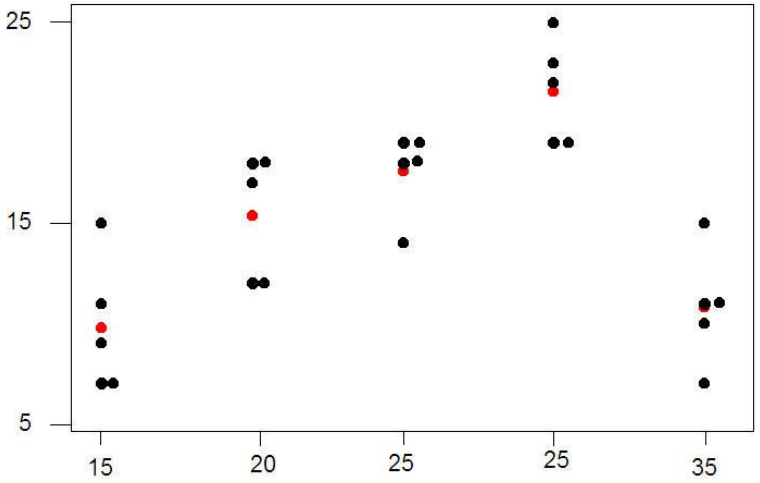
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

DIPARTIMENTO DI INGEGNERIA AEROSPAZIALE – D.I.A.S.

STATISTICA PER L'INNOVAZIONE

a.a. 2007/2008



ANOVA: ANALISI DELLA VARIANZA

Prof. Antonio Lanzotti

A cura di: Ing. Giovanna Matrone

giovanna.matrone@unina.it

Analisi della varianza

L'Analisi della Varianza (ANOVA – *Analysis of Variance*) permette di valutare gli effetti esercitati da determinati fattori (X_i), ognuno a più livelli, su una prestazione (Y) di nostro interesse.

L'ANOVA consiste nell'eseguire un test d'ipotesi sull'influenza del singolo fattore sulla *variabilità* globale; in questo modo è poi possibile stabilire la significatività dell'effetto del fattore sulla prestazione.

Effetto → Media → Varianza

1) Analisi della varianza ad una via

Un esempio: la resistenza di una fibra (Montgomery 2001)

Un ingegnere addetto allo sviluppo prodotto è interessato a studiare la resistenza a trazione (Y) di una nuova fibra sintetica per maglie maschili. Da precedenti esperienze egli sa che la resistenza a trazione è influenzata dalla percentuale di cotone (A) presente nella miscela di materiale usata per creare la fibra. Inoltre sospetta che, almeno inizialmente, incrementando la percentuale di cotone crescerà anche la resistenza della fibra. Altri vincoli ingegneristici suggeriscono di usare una percentuale di cotone nel range 10-40%. L'ingegnere decide quindi di esaminare dei campioni (numerosità 5) di fibre con percentuali di cotone rispettivamente del 15, 20, 25, 30 e 35%.

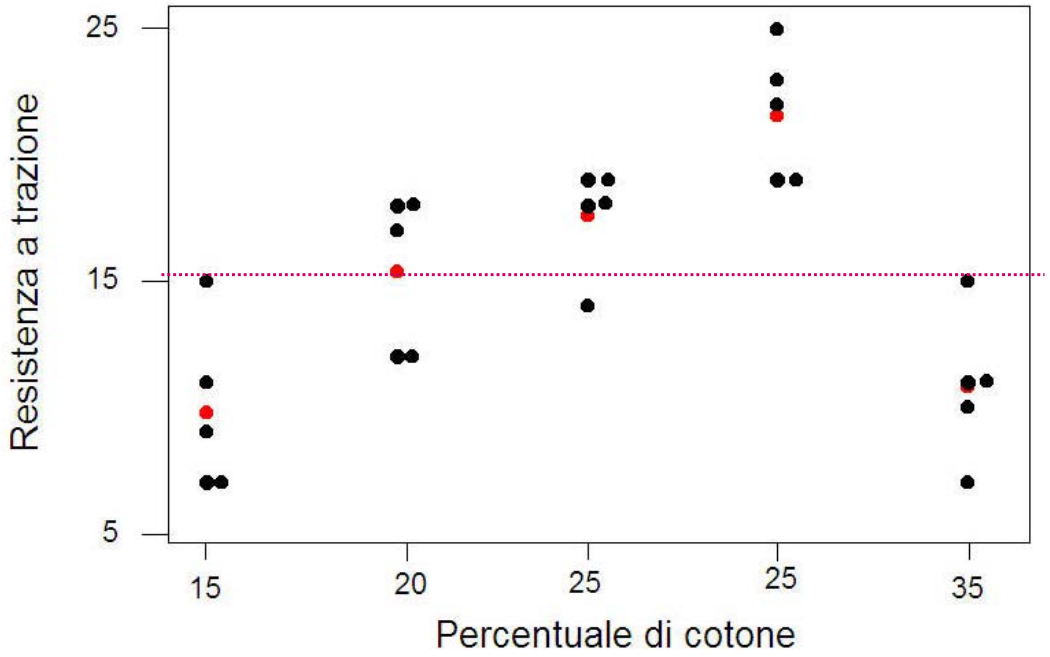
L'esperimento quindi consiste nell'eseguire 25 prove, opportunamente randomizzate, per testare la resistenza a trazione (lb/in^2) delle fibre scelte.

Analisi della varianza ad una via

Dati in lb/in^2 della resistenza a trazione delle fibre testate

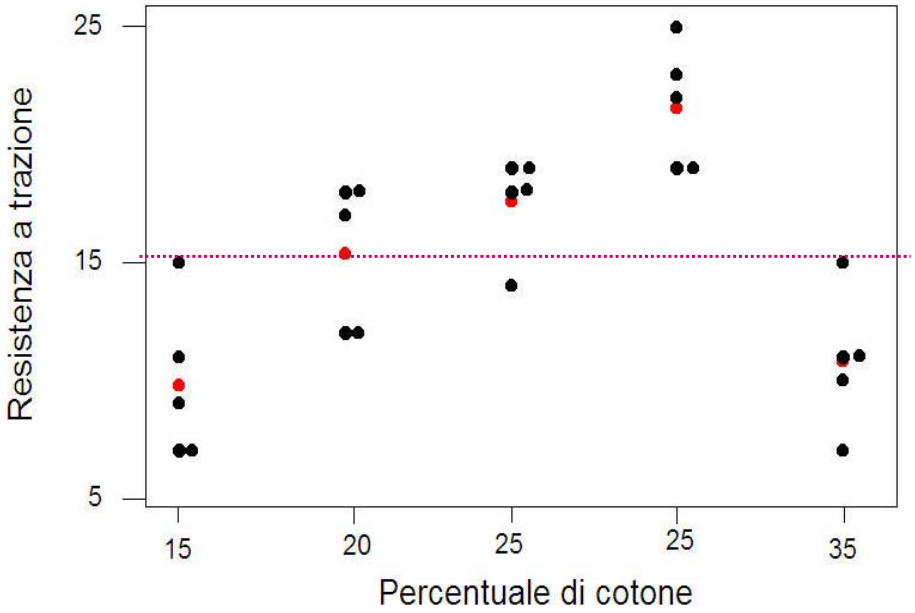
Fattore	A	Percentuale di cotone
Livelli	A_1	15
	A_2	20
	A_3	25
	A_4	30
	A_5	35

	Osservazioni							
A	1	2	3	4	5	Totale	Media	
A_1	7	7	15	11	9	49	9,8	
A_2	12	17	12	18	18	77	15,4	
A_3	14	18	18	19	19	88	17,6	
A_4	19	25	22	19	23	108	21,6	
A_5	7	10	11	15	11	54	10,8	
						376	15,04	TOT



Analisi della varianza ad una via

Una prima osservazione di questo grafico potrebbe permettere all'ingegnere di dire che la resistenza a trazione cresce al crescere della percentuale di cotone fino ad una percentuale del 30% e successivamente decrescere notevolmente.



Si può osservare inoltre che la variabilità della resistenza a trazione rispetto alle medie non dipende in maniera marcata dalla percentuale di cotone.

L'ingegnere addetto al processo di sviluppo del prodotto vuole ottenere dei risultati quantitativi che gli facciano prendere una posizione chiara riguardo alla percentuale di cotone da utilizzare.

Analisi della varianza ad una via: il modello

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

- μ media generale (parametro comune a tutti i livelli)
- α_i effetto imputabile al fattore in esame al livello i-esimo
- ε_{ij} componente casuale di errore

Test di ipotesi

$$H_0 = \{\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0\} \qquad H_1 = \{\alpha_i \neq 0 \text{ per almeno un } i\}$$

ASSUNZIONI DEL MODELLO

1) dipendenza lineare degli effetti
2) additività degli effetti
3) s-indipendenza delle osservazioni sperimentali
4) errori casuali distribuiti normalmente $\varepsilon_{ij} \sim N(0, \sigma^2)$
5) omogeneità delle varianze nelle varie condizioni sperimentali $x_{ij} \sim N(\mu + \alpha_i, \sigma^2)$

ANOVA ad una via: verifica delle ipotesi

Le violazioni delle assunzioni base di un modello di analisi della varianza possono essere esaminate attraverso l'analisi dei *residui*.

Definiamo *residuo* la devianza tra il valore osservato ed il valore 'stimato' con il modello: $e_{ij} = x_{ij} - \hat{x}_{ij}$ dove: $\hat{x}_{ij} = \hat{\mu} + \hat{\alpha}_i = \overline{x} + (\overline{x}_i - \overline{x}) = \overline{x}_i$

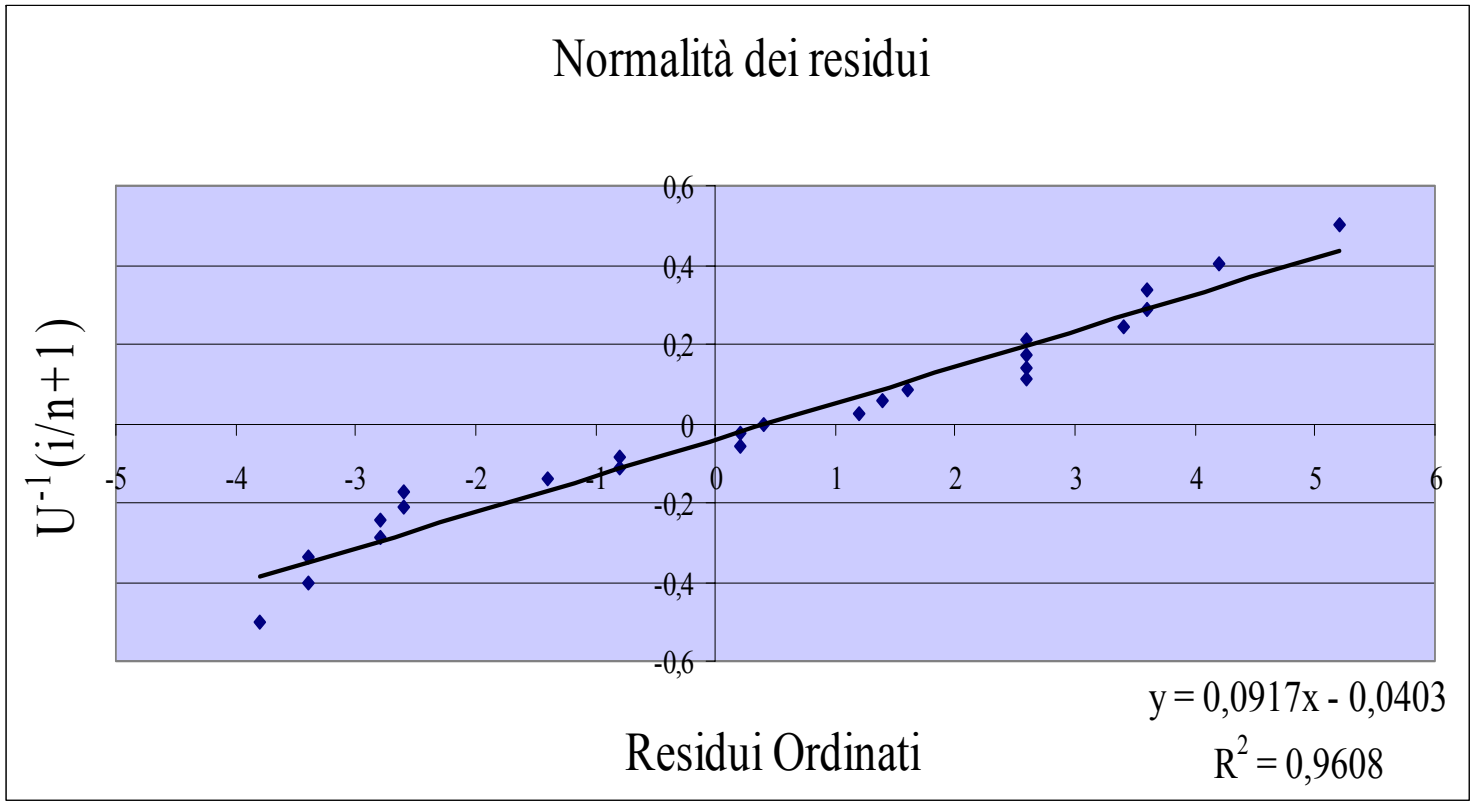
L'adeguatezza del modello si dimostra solo se i residui risultano essere **non strutturati**, cioè se non presentano andamenti particolari e riconoscibili.

La prima assunzione di normalità degli errori casuali può essere verificata attraverso l'istogramma dei residui e verificando che tale istogramma richiama la curva normale.

Poiché il test F soffre poco della divergenza dalla normalità (**test robusto**) è necessario che i residui siano "pressappoco" distribuiti normalmente.

ANOVA ad una via: verifica della normalità dei residui

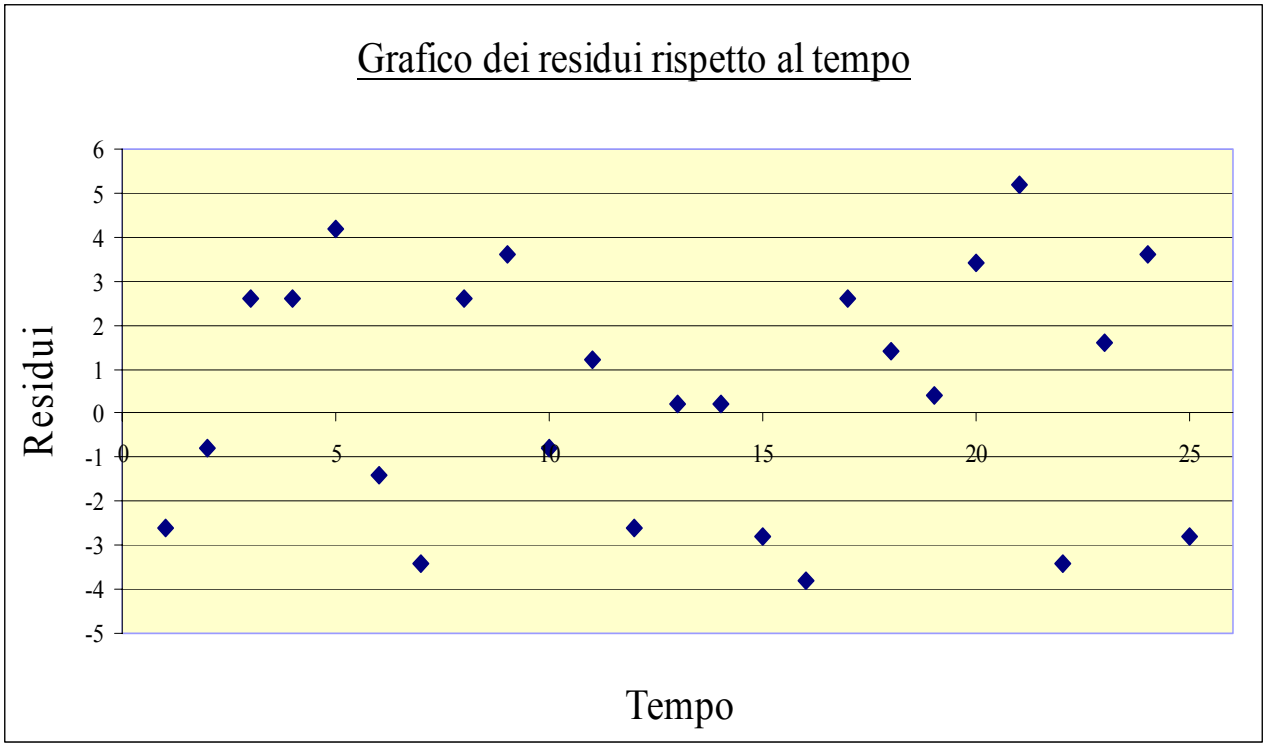
	Residui				
A	1	2	3	4	5
A_1	-2,8	-2,8	5,2	1,2	-0,8
A_2	-3,4	1,6	-3,4	2,6	2,6
A_3	-1,4	2,6	2,6	3,6	3,6
A_4	-2,6	3,4	0,4	-2,6	1,4
A_5	-3,8	-0,8	0,2	4,2	0,2



Non ci sono evidenze di non-normalità nè valori outliers.

ANOVA ad una via: verifica dell'indipendenza dei residui

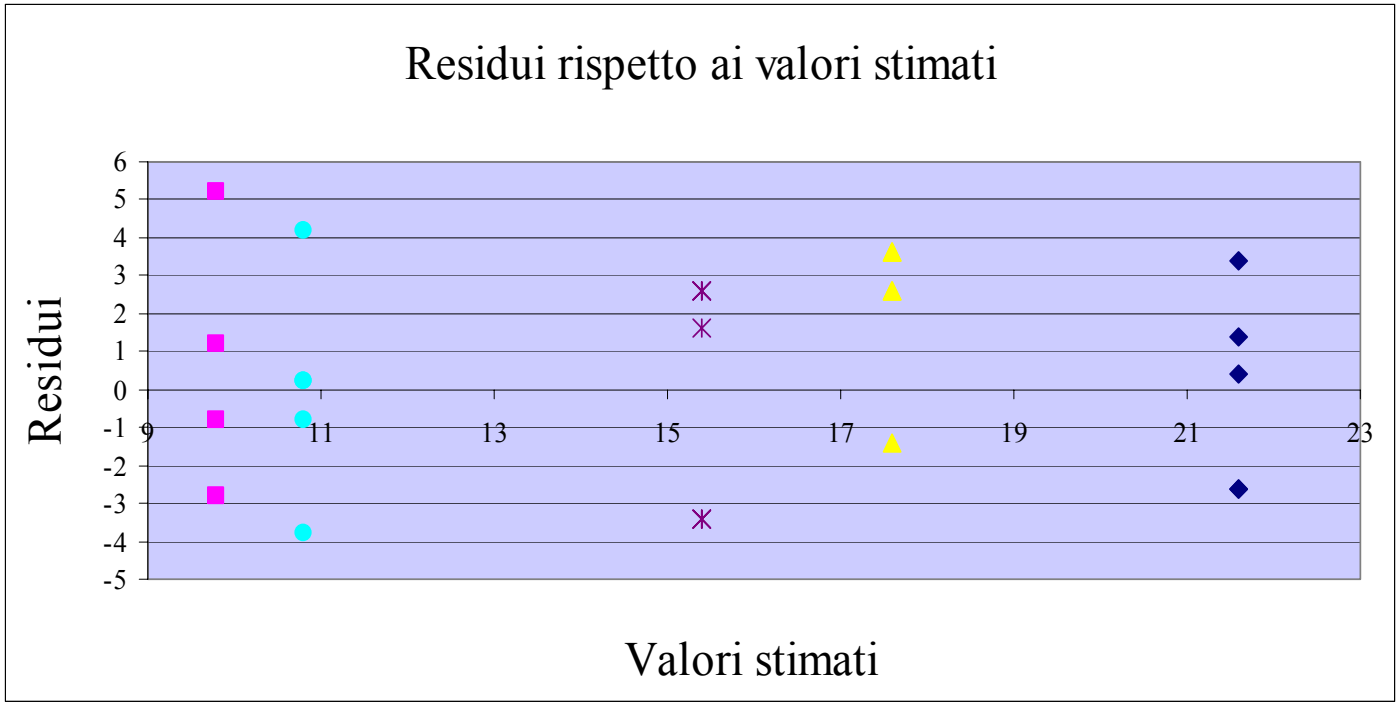
Un grafico dei residui rispetto all'ordine temporale di raccolta dei dati (tempo o numero dell'osservazione campionaria) è utile nel verificare la presenza di correlazioni nei residui e quindi l'assunzione di indipendenza degli stessi. Casualizzare opportunamente il processo di esecuzione delle prove spesso aiuta ad ottenere l'indipendenza.



Non si notano particolari aumenti di variabilità con il tempo: non c'è quindi motivo di sospettare della violazione dell'assunzione di indipendenza e dell'assunzione di varianza costante.

ANOVA ad una via: verifica dell'assunzione di residui non 'strutturati'

Se il modello è corretto e le assunzioni sono soddisfatte i residui non dovrebbero mostrare una particolare struttura. Inoltre non dovrebbero essere in relazione con nessuna altra variabile del modello e quindi anche con i valori stimati delle variabili $\hat{x}_{ij}$.



Non ci sono particolari strutture quindi ancora una volta le assunzioni base del modello sono rispettate.

ANOVA ad una via: la tabella

- n livelli del fattore sotto studio
- m numero di osservazioni sperimentali per ciascun livello del fattore

Origine della varianza	Somma dei quadrati degli scarti	Gradi di libertà	Scarto quadratico medio (s.q.m.)	Valore atteso dello s.q.m.
Fattore (var. interclassi)	$SS_{fattori} = m\sum_{i=1}^{n}(\bar{x}_i - \bar{x})^2$	$n-1$	$MS_{fattori} = \frac{m\sum_{i=1}^{n}(\bar{x}_i - \bar{x})^2}{n-1}$	$\sigma^2 + \frac{m}{(n-1)}\sum_{i=1}^{n}\alpha_i^2$
Errore (var. intraclassi)	$SS_E = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2$	$n(m-1)$	$MS_{errore} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2}{n(m-1)}$	σ^2
Totale	$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x})^2$	$n \cdot m - 1$	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x})^2}{n \cdot m - 1}$	

ANOVA ad una via: la tabella

Poiché $x_{ij} \sim N(\mu + \alpha_i, \sigma^2)$ la SS_T è una somma di variabili normali al quadrato e quindi $\frac{SS_T}{\sigma^2} \sim \chi_{nm-1}$ ed analogamente $\frac{SS_E}{\sigma^2} \sim \chi_{n(m-1)}$

Sotto l'ipotesi nulla inoltre: $\frac{SS_{fattori}}{\sigma^2} \sim \chi_{n-1}$

Per il teorema di Cochran le due chi-quadrato sono indipendenti e quindi sotto l'ipotesi nulla: $F_0 = \frac{SS_{fattori} / n-1}{SS_E / n(m-1)} = \frac{MS_{fattori}}{MS_E} \sim F_{n-1,n(m-1)}$

ANOVA ad una via: analisi

Fattore	A	Percentuale di cotone
Livelli	A_1	15
	A_2	20
	A_3	25
	A_4	30
	A_5	35

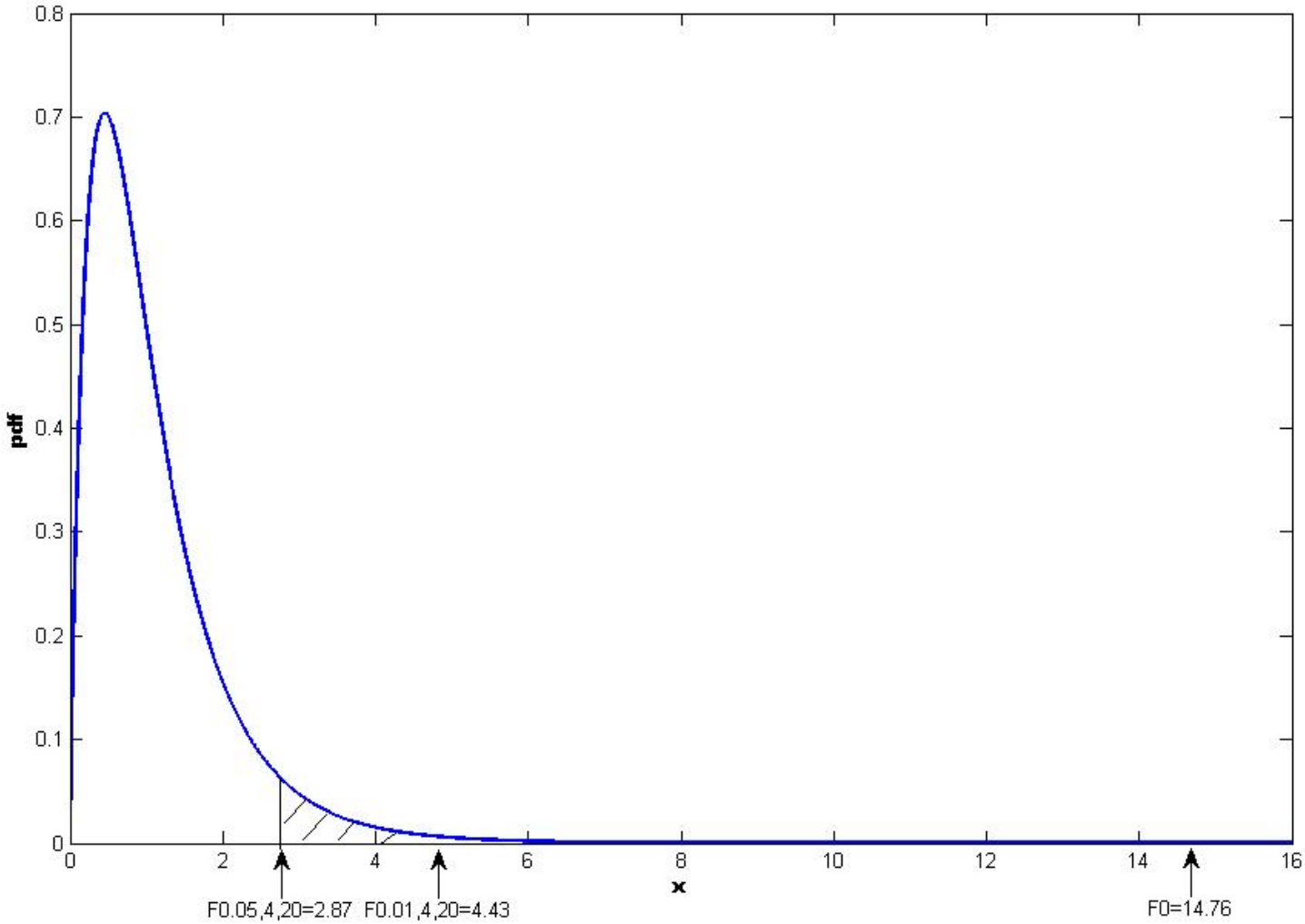
	Osservazioni						
A	1	2	3	4	5	Totale	Media
A_1	7	7	15	11	9	49	9,8
A_2	12	17	12	18	18	77	15,4
A_3	14	18	18	19	19	88	17,6
A_4	19	25	22	19	23	108	21,6
A_5	7	10	11	15	11	54	10,8
						376	15,04	TOT

ANOVA

						$\alpha = 0,05$
Origine della varianza	SS	g.d.l	MS	F0	Ftest	p-value
Fattori	475,76	4	118,94	14,76	2,87	0,00001
Errore	161,2	20	8,06			
Totale	636,96	24				

Poiché $F_0 > F_{test}$ (p-value < α) possiamo rigettare l'ipotesi nulla e concludere che le medie dei trattamenti sono differenti, ovvero l'effetto del fattore risulta essere significativo e il modello assunto per descrivere il modello è corretto.

ANOVA ad una via: la tabella



Zona di rigetto

2) Analisi della varianza ad una via

Durante uno studio sulla durata (in ore) di un componente meccanico, si notò un comportamento particolare al variare delle condizioni ambientali. Si decise pertanto di verificare questa ipotesi attraverso una sperimentazione programmata. A causa di motivi tecnici, però, non fu possibile eseguire lo stesso numero di esperimenti per ciascuno dei livelli fissati.

Fattore	A
Livelli	A_1
	A_2
	A_3

	Osservazioni									
A	1	2	3	4	5	6	7	8	Media	
A_1	82	114	90	80	88	93	80	105	91,5	$m1$ =8
A_2	128	90	130	110	133	130	104		117,86	$m1$ =7
A_3	156	128	151	140					143,75	$m1$ =4
									117,7	TOT

Analisi della varianza ad una via

$$SS_f = m\sum_{i=1}^{n}\left(\bar{x}_i - \bar{x}\right)^2 \qquad SS_e = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_i\right)^2$$

$$SS_f = \sum_{i=1}^{n} m_i\left(\bar{x}_i - \bar{x}\right)^2 \qquad SS_e = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(x_{ij} - \bar{x}_i\right)^2$$

3) Analisi della varianza a due vie

Un esempio: la resa di soia

L'azienda agricola *Agrilab* si era posta l'obiettivo di massimizzare la resa della soia. Attraverso una serie di studi su dati storici, si era delineato il sospetto che la resa della soia (Y) fosse influenzata da due fattori:

- Fattore A – grado di umidità del terreno
- Fattore B – tipo di concime utilizzato

Si decise, quindi, di effettuare una sperimentazione fattoriale completa per comprendere gli effetti dei fattori sulla resa della soia.

In particolare, gli esperimenti di semina della soia sono stati effettuati in 10 appezzamenti di terreno diversi per umidità, utilizzando 4 diversi tipi di concime, rispettivamente a base di stallatico, Na, K ed N.

Analisi della varianza a due vie: il modello

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad \text{con } i = 1,\ldots n \text{ e } j = 1,\ldots, b$$

- μ media generale (parametro comune a tutti i livelli)
- α_i effetto imputabile al primo fattore in esame al livello i-esimo
- β_j effetto imputabile al secondo fattore in esame al livello j-esimo
- ε_{ij} componente casuale di errore

Test di ipotesi

$$H_0 = \{\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0\} \qquad H_1 = \{\alpha_i \neq 0 \text{ per almeno un } i\}$$

$$H_0 = \{\beta_1 = \beta_2 = \ldots = \beta_b = 0\} \qquad H_1 = \{\beta_j \neq 0 \text{ per almeno un } j\}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$x_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2)$$

$$\hat{\varepsilon}_{ij} = x_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j = x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}$$

Analisi della varianza a due vie: la tabella

Origine della varianza	Somma dei quadrati degli scarti	gdl	Scarto quadratico medio	Valore atteso dello s.q.m.
Fattore principale	$b\sum_{i=1}^{n}\left(\bar{x}_{i\bullet}-\bar{x}\right)^2$	*n-1*	$\frac{b\sum_{f}}{n-1}$	$\sigma^2+\frac{b\sum_{i=1}^{n}\alpha_i^2}{(n-1)}$
Fattore secondario	$n\sum_{j=1}^{b}\left(\bar{x}_{\bullet j}-\bar{x}\right)^2$	*b-1*	$\frac{n\sum_{b}}{b-1}$	$\sigma^2+\frac{b\sum_{i=1}^{b}\beta_j^2}{(b-1)}$
Errore	$\sum_{i=1}^{n}\sum_{j=1}^{b}\left(x_{ij}-\bar{x}_{i\bullet}-\bar{x}_{\bullet j}+\bar{x}\right)^2$	*(n-1)(b-1)*	$\frac{\sum_{e}}{(n-1)(b-1)}$	σ^2
Totale	$\sum_{i=1}^{n}\sum_{j=1}^{b}\left(x_{ij}-\bar{x}\right)^2$	*nb-1*	$\frac{\sum_{tot}}{nb-1}$	

Analisi della varianza a due vie

Fattore	A
Livelli	A_1
	A_2
	A_3
	A_4
	A_5
	A_6
	A_7
	A_8
	A_9
	A_{10}

Fattore	B
Livelli	B_1
	B_2
	B_3
	B_4
	B_5

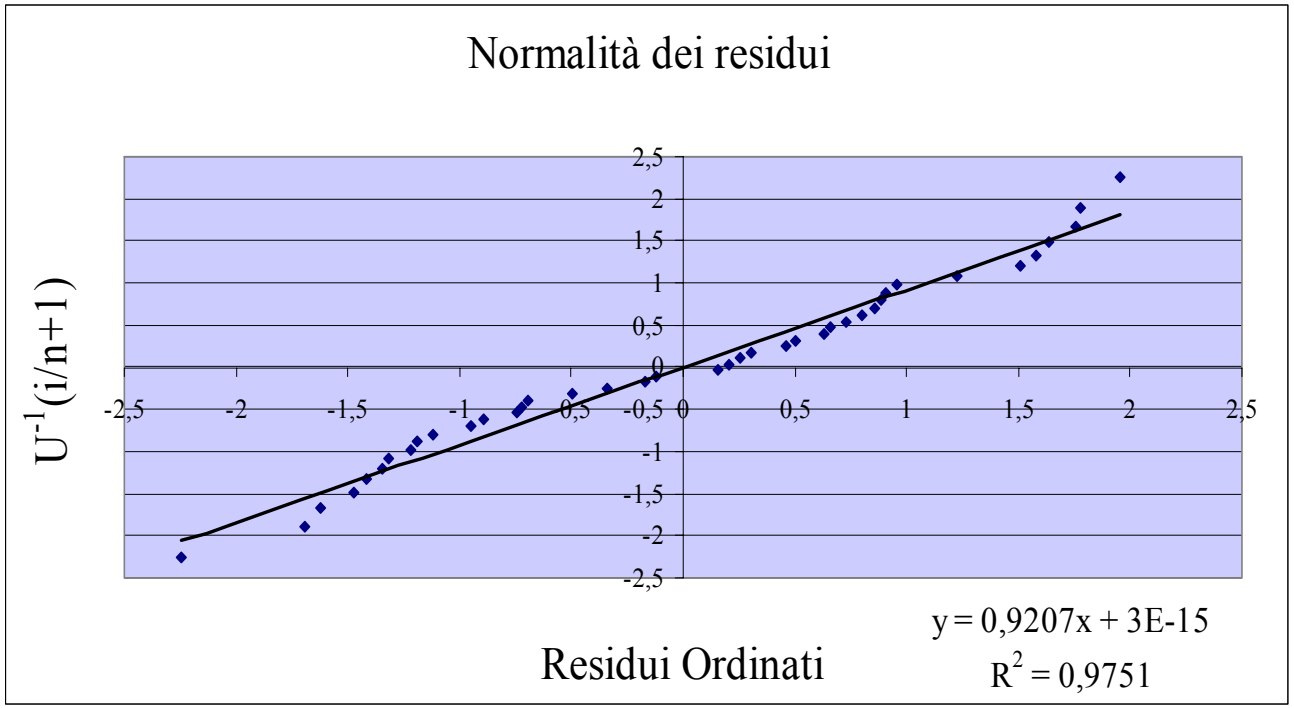
	Osservazioni				
	B_1	B_2	B_3	B_4	Media
A_1	71,2	74	71,9	70,1	71,8
A_2	73,4	73,3	74,3	73,4	73,6
A_3	68,2	72,3	70,3	70,4	70,3
A_4	75,9	75,8	74,5	73,3	74,875
A_5	72	69,8	69,6	72,6	71
A_1	73	76,2	75,3	73,7	74,55
A_2	75,6	72,9	74,4	72,9	73,95
A_3	76,2	73,9	72,8	74,8	74,425
A_4	74,6	73,5	73,1	75,6	74,2
A_5	72,4	73,6	72,8	70,7	72,375
Media	73,25	73,53	72,9	72,75	73,108

$n = 10$

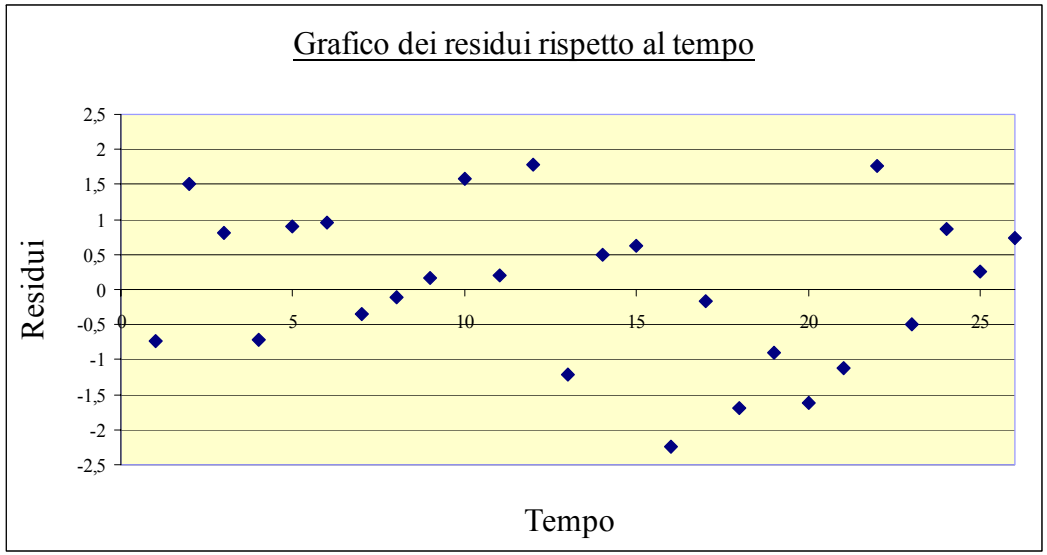
$b = 4$

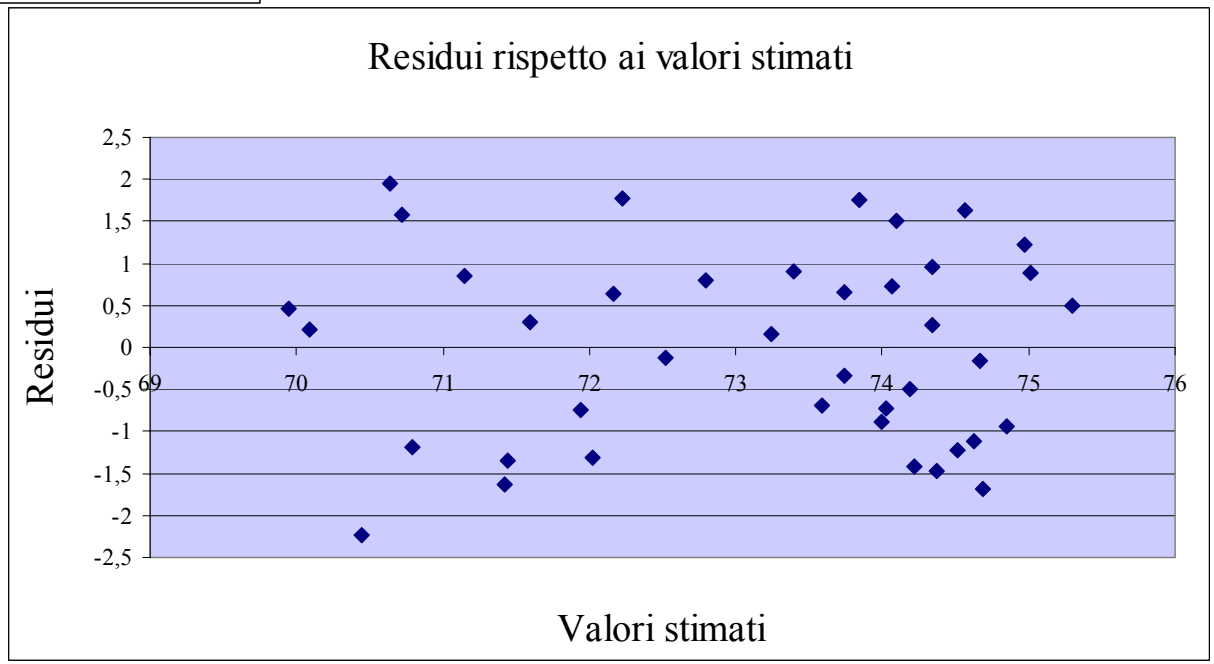
ANOVA a due vie: verifica della normalità dei residui

	Residui			
	B_1	B_2	B_3	B_4
A_1	-0,74	1,78	0,31	-1,34
A_2	-0,34	-0,72	0,91	0,16
A_3	-2,24	1,58	0,21	0,46
A_4	0,88	0,50	-0,17	-1,22
A_5	0,86	-1,62	-1,19	1,96
A_1	-1,69	1,23	0,96	-0,49
A_2	1,51	-1,47	0,66	-0,69
A_3	1,63	-0,95	-1,42	0,73
A_4	0,26	-1,12	-0,89	1,76
A_5	-0,12	0,80	0,63	-1,32



ANOVA a due vie: verifica dell'indipendenza dei residui e della loro 'destrutturazione'





Analisi della varianza a due vie: analisi

Fattore	A
Livelli	A_1
	A_2
	A_3
	A_4
	A_5
	A_6
	A_7
	A_8
	A_9
	A_{10}

Fattore	B
Livelli	B_1
	B_2
	B_3
	B_4
	B_5

	Osservazioni				
	B_1	B_2	B_3	B_4	Media
A_1	71,2	74	71,9	70,1	71,8
A_2	73,4	73,3	74,3	73,4	73,6
A_3	68,2	72,3	70,3	70,4	70,3
A_4	75,9	75,8	74,5	73,3	74,875
A_5	72	69,8	69,6	72,6	71
A_1	73	76,2	75,3	73,7	74,55
A_2	75,6	72,9	74,4	72,9	73,95
A_3	76,2	73,9	72,8	74,8	74,425
A_4	74,6	73,5	73,1	75,6	74,2
A_5	72,4	73,6	72,8	70,7	72,375
Media	73,25	73,53	72,9	72,75	73,108

$n = 10$

$b = 4$

ANOVA

$\alpha = 0{,}005$

Origine della varianza	SS	g.d.l	MS	F0	Ftest	p-value	Esito	Effetto del Fattore
Fattore Principale	94,625	9	10,514	5,56	3,56	0,00024	Rigetto H0	Significativo
Fattore Secondario	3,6968	3	1,2323	0,65	5,36	0,58860	Accetto H0	Non significativo
Errore	51,03	27	1,8898					
Totale	149,35	39						

Analisi della varianza a due vie: il modello

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad \text{con } i = 1,\ldots n \text{ e } j = 1,\ldots, b$$

- μ media generale (parametro comune a tutti i livelli)
- α_i effetto imputabile al primo fattore in esame al livello i-esimo
- β_j effetto imputabile al secondo fattore in esame al livello j-esimo
- ε_{ij} componente casuale di errore

- α_i effetto imputabile al primo fattore: significativo
- β_j effetto imputabile al secondo fattore: non significativo

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

Riferimenti sul Libro

Pasquale Erto
“Probabilità e statistica per le scienze e l’ingegneria”
***McGraw Hill* – seconda edizione**

→ **Capitolo 11**

- **Analisi della varianza ad una via pag. 252**
- **Analisi della varianza a due vie pag. 259**

Orario di ricevimento:

Giovedì ore 16:30-18:30
P.le Tecchio X piano
Stanza Dottorandi DIAS