



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

# Come raccogliere ed utilizzare i dati sperimentali

**Daniele Moro**

Piacenza, 10 marzo 2014 – La preparazione della tesi di Laurea Magistrale



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

- ma questa 'statistica' a che cosa serve?
- non vedo l'ora di cominciare a lavorare per la tesi....  
e dimenticarmi la statistica!!



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

il mio relatore mi ha proposto questo argomento di tesi

- il *Solanum nigrum* è un'infestante del pomodoro
- vi sono due erbicidi (*Metribuzin* e *Rimsulfuron*), ma poco efficaci
- lo studio del meccanismo di azione dei due erbicidi gli fa pensare che una miscela dei due potrebbe dare un effetto sinergico e dunque essere più efficace
- devo 'confermare' questa ipotesi



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

e adesso?



## lo schema logico della metodologia sperimentale





UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

## metodologia sperimentale e trattamento statistico dei dati

- osservazione della realtà
- raccolta di informazioni (indagine bibliografica)
- formulazione di un'ipotesi
- pianificazione ed esecuzione di un *esperimento*  
scientifico *riproducibile*
- *analisi dei dati raccolti*: misurazione e  
interpretazione del dato sperimentale



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

cos'è un *esperimento*?

processo investigativo con il quale, sulla base di un protocollo adeguato, si realizzano determinate circostanze che consentono di avere le informazioni necessarie per la verifica empirica

un esperimento viene organizzato a priori dal ricercatore: le modalità di organizzazione dell'esperimento costituiscono il *disegno sperimentale*

in genere le circostanze vengono realizzate dal ricercatore, imponendo condizioni (*trattamenti*) differenti su soggetti/individui selezionati, il più uniformi possibili in partenza

i trattamenti spesso sono confrontati con un trattamento di riferimento o *controllo* (nessun trattamento, placebo, pratica usuale)



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

## esempio

- lo studio del meccanismo di azione dei due erbicidi fa pensare che una miscela dei due potrebbe dare un effetto sinergico e dunque essere più efficace  
*(ipotesi scientifica)*
- pianificare un esperimento (*disegno sperimentale*)
- completamente **randomizzato** (attribuzione casuale dei trattamenti)  
trattamenti (Me, Ri, Me+Ri-MR) più un controllo non trattato (NT)





UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

esistono regole nel disegno sperimentale

la *randomizzazione?*

quante storie, e che sarà mai.....

che cosa potrà mai cambiare.....



un esercizio di simulazione

consideriamo tre trattamenti,

*uguali* tra loro, A-B-C, con repliche

realizzate su tre differenti substrati,

uno dei quali (substrato 1) *incida*

*in maniera significativa* sul

risultato dell'esperimento (valori

della variabile risposta)

A	B	C
1		
1		
1		
1		
1		
1		
1		
1		
1		
1		
1		
1		
1		
1		

esperimento non  
randomizzato

A	B	C
1		
1		1
1		1
1		
1	1	
		1
1	1	1
	1	1
	1	

esperimento  
randomizzato



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

usiamo la procedura ANOVA ad 1 VIA

$$H_0 : \mu_A = \mu_B = \mu_C$$

NB: **l'ipotesi nulla è vera** (dunque sarebbe *bene che il test ANOVA non la rifiutasse*, altrimenti commetterei un errore del I tipo)

*vengono simulati 40 esperimenti*

la procedura ANOVA ad 1 VIA viene applicata sui 40 esperimenti simulati (usando sia lo schema della randomizzazione che quello della non randomizzazione)



usiamo la procedura ANOVA ad 1 VIA

i risultati

	$\alpha$	# errori I tipo
esperimento non randomizzato	0.01	20
	0.05	31
	0.10	33
esperimento randomizzato	0.01	0
	0.05	1
	0.10	2

*NB: se sapessimo che abbiamo terreni diversi (blocchi) dovremmo applicare un esperimento a blocchi randomizzati e una ANOVA a 2 VIE*



Me	Ri	M+R	NT
NT	NT	Me	M+R
Ri	Me	M+R	NT
M+R	Ri	Me	Ri

attribuzione casuale dei trattamenti alle parcelle

come **misuriamo** l'effetto dei trattamenti?

dopo un certo periodo dal trattamento, su ogni parcella si preleva la vegetazione infestante, per unità di superficie, si secca la biomassa, e si pesa (g/m<sup>2</sup>) → i **DATI** ottenuti dall'esperimento

attenzione: le '*confounding variables*':

**controllare tutti gli altri possibili fattori di incidenza**



una volta ottenuto il risultato dell'esperimento, cioè i dati?

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>NT</b>	24.62	30.94	24.02	27.51
<b>Me</b>	15.20	4.38	10.32	6.80
<b>MR</b>	6.14	1.95	7.27	5.15
<b>Ri</b>	10.50	20.70	20.74	15.50

organizzazione e descrizione dei dati

quale analisi statistiche effettuare e come  
presentare i risultati

l'analisi statistica deve essere funzionale agli  
obiettivi della ricerca e presentata in modo chiaro e  
sintetico



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

quali strumenti posso usare per l'organizzazione  
dei dati e l'analisi descrittiva e grafica

- **Excel** (foglio elettronico)
- **SPSS** (Statistical Package for Social Science)
- **R** (the R Project for Statistical Computing)
- **SAS** (Statistical Analysis System)
- **Stata**
- .....



che cosa possiamo quindi fare con un software statistico?

- **pulizia dei dati:** controllare errori di inserimento dei dati, presenza di dati mancanti, ricerca di *outlier* mediante l'analisi delle frequenze, ...
- **trasformazione dei dati:** ottenere nuove variabili effettuando operazioni o trasformazioni sulle variabili pre-esistenti
- **rappresentazione dei dati:** costruire grafici o tabelle
- **calcolo delle statistiche descrittive:** calcolare le statistiche descrittive quali: media, varianza, deviazione standard, mediana, moda...
- **verifica delle assunzioni:** se i dati si distribuiscono normalmente, se le distribuzioni siano simmetriche, se esista omoschedasticità , ...
- **verifica delle ipotesi di lavoro**





## 1. analisi descrittiva

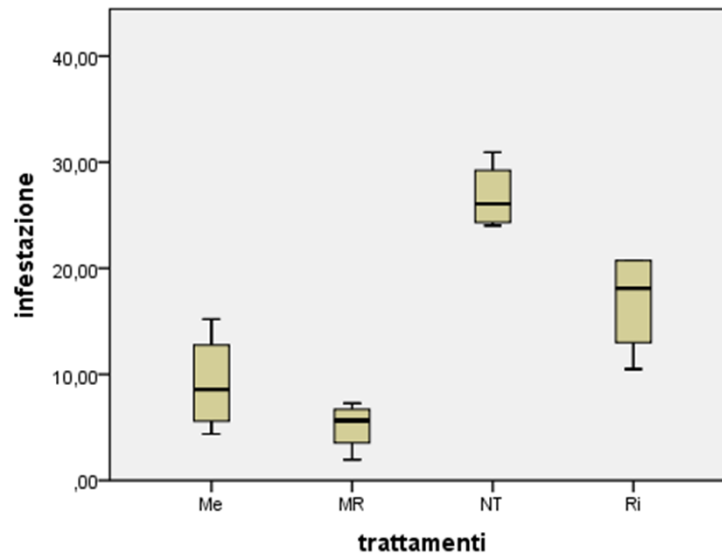


grafico 'a scatole' – box-plot  
(SPSS)

	n	Media	Deviazione standard
NT	4	26.77	3.17
Me	4	9.18	4.70
MR	4	5.13	2.29
Ri	4	16.86	4.90
Totale	16	14.48	9.21

statistiche descrittive  
(Excel)

sembra che i trattamenti siano diversi





## una breve parentesi

### la scelta dell'analisi statistica corretta

ad esempio

è stato realizzato un intervento per la pulizia delle acque di un fiume; per valutarne l'efficacia si sono prese sei località sul fiume e misurata la domanda biologica di ossigeno in tre occasioni: prima dell'intervento, dopo un mese dall'intervento, dopo un anno dall'intervento

	1	2	3	4	5	6
prima	17.4	15.7	12.9	9.8	13.4	19.6
1 mese	13.6	10.1	10.3	9.2	11.1	20.4
1 anno	13.2	9.8	9.7	9.0	10.7	19.6

ci viene suggerito di usare un test non-parametrico

test di Kruskal-Wallis (1 VIA)

test di Friedman (2 VIE) (ok)



una breve parentesi

la scelta dell'analisi statistica corretta

ranghi per $F_R$							
	1	2	3	4	5	6	somme
prima	3	3	3	3	3	1.5	16.5
1 mese	2	2	2	2	2	3	13.0
1 anno	1	1	1	1	1	1.5	6.5

ranghi pr KW							
	1	2	3	4	5	6	somme
prima	4	5	10	16	13	2.5	50.5
1 mese	6	8	11	17	14	1	57.0
1 anno	7	9	12	18	15	2.5	63.5

$F_R$  8.583

$H$  0.494

chiquad 5.991

intervento  
efficace

intervento  
non efficace



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

quali strumenti usare per l'analisi statistica dei dati?

- **Excel** (foglio elettronico)
- **SPSS** (Statistical Package for Social Science)
- **R** (the R Project for Statistical Computing)
- **SAS** (Statistical Analysis System)
- **Stata**
- .....

i software lavorano per noi ma  
siamo sempre noi a scegliere la  
strada!



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

## metodologia sperimentale e trattamento statistico dei dati

6 :

	dati	trat	var	var	var	var	var	var	var	var	var	var	var	var
1	24,62	NT												
2	30,94	NT												
3	24,02	NT												
4	27,51	NT												
5	15,20	Me												
6	4,38	Me												
7	10,32	Me												
8	6,80	Me												
9	6,14	MR												
10	1,95	MR												
11	7,27	MR												
12	5,15	MR												
13	10,50	Ri												
14	20,70	Ri												
15	20,74	Ri												
16	15,50	Ri												
17														
18														
19														
20														
21														
22														

Visualizzazione dati Visualizzazione variabili

IBM SPSS Statistics Il processore è pronto

ogni software ha il suo 'linguaggio'

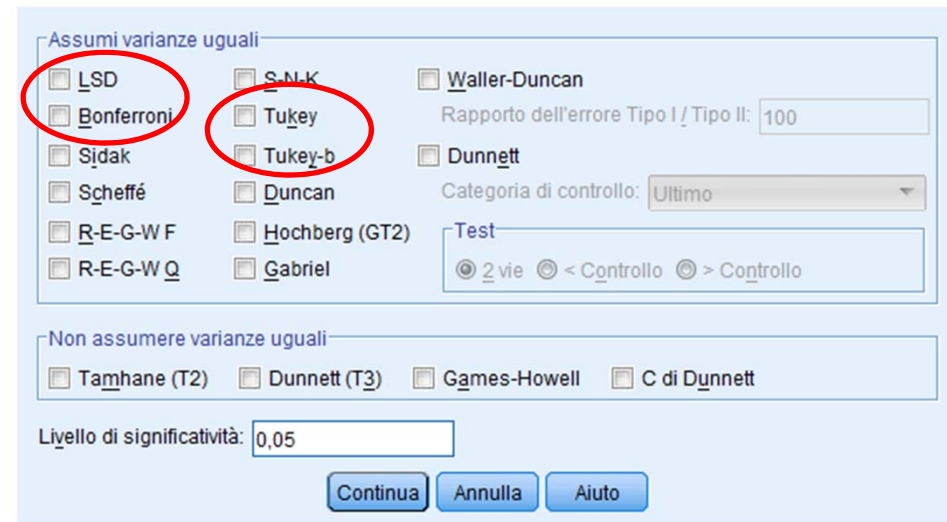
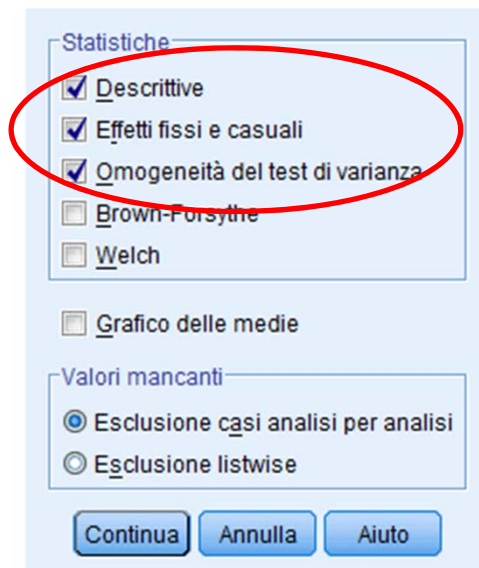
**importante:** la possibilità di trasferire dati da un pacchetto all'altro



### 3. 'validazione' dell'ipotesi di lavoro e presentazione dei risultati



- scegliere la procedura
- verificare le assunzioni
- valutare le opzioni





Descrittivi									
dati									
	n	Media	Deviazione std.	Errore std.	Intervallo di confidenza 95% per la media		Minimo	Massimo	Varianza tra componenti
					Limite inferiore	Limite superiore			
<b>1 - NT</b>	4	<b>26.773</b>	3.169	1.584	21.730	31.815	24.020	30.940	
<b>2- Me</b>	4	<b>9.175</b>	4.699	2.350	1.698	16.652	4.380	15.200	
<b>3 - MR</b>	4	<b>5.128</b>	2.289	1.144	1.486	8.769	1.950	7.270	
<b>4 - Ri</b>	4	<b>16.860</b>	4.902	2.451	9.059	24.661	10.500	20.740	
Totale	16	<b>14.484</b>	9.215	2.304	9.573	19.394	1.950	30.940	
<b>Modello</b>	<b>Effetti fissi</b>		3.918	0.979	12.350	16.618			
	<b>Effetti casuali</b>			4.764	-0.678	29.646			86.957

Test di omogeneità delle varianze ( $H_0$ )			
dati			
Statistica di Levene	df1	df2	Sig.
1.356	3	12	<b>.303</b>





ANOVA univariata					
	Somma dei quadrati	gdl	Media dei quadrati	F	p-value
Fra gruppi	1089.529	3	363.176	<b>23.663</b>	<b>.000</b>
Entro gruppi	184.177	12	15.348		
Totale	1273.706	15			

- esiste una differenza tra i trattamenti

...ma...abbiamo risposto alla 'domanda di ricerca'?



- la combinazione MR è migliore delle altre?

sembra che MR sia migliore di Ri (e NT),  
**ma non di Me**

Confronti multipli						
HSD di Tukey						
(I) tratt		Differenza fra medie (I-J)	Errore std.	p-value	Intervallo di	
					Limite inferiore	Limite superiore
<b>1 - NT</b>	2	17.598	2.770	0.000	9.373	25.822
	3	21.645	2.770	0.000	13.421	29.869
	4	9.913	2.770	0.017	1.688	18.137
<b>2 - Me</b>	1	-17.598	2.770	0.000	-25.822	-9.373
	3	4.048	2.770	0.488	-4.177	12.272
	4	-7.685	2.770	0.070	-15.909	0.539
<b>3 - MR</b>	1	-21.645	2.770	<b>0.000</b>	-29.869	-13.421
	2	-4.048	2.770	<b>0.488</b>	-12.272	4.177
	4	-11.733	2.770	<b>0.006</b>	-19.957	-3.508
<b>4 - Ri</b>	1	-9.913	2.770	0.017	-18.137	-1.688
	2	7.685	2.770	0.070	-0.539	15.909
	3	11.733	2.770	0.006	3.508	19.957

\*. La differenza media è significativa al livello 0.05



sembra che MR sia migliore di Ri (e NT), ma non di Me

non possiamo dire che la combinazione sinergica *Metribuzin* e *Rimsulfuron* produca effetti migliori del solo *Metribuzin*

HSD di Tukey<sup>a</sup>

tratt	N	Sottoinsieme per alfa = 0.05		
		1	2	3
3 - MR	4	5.128		
2 - Me	4	9.175	9.175	
4 - Ri	4		16.860	
1 - NT	4			26.773
Sig.		0.488	0.070	1.000



UNIVERSITÀ  
CATTOLICA  
del Sacro Cuore

metodologia sperimentale e trattamento statistico dei dati

attenzione

i risultati delle procedure inferenziali della statistica sono di carattere **probabilistico**

... e dipendono anche dalle condizioni dell'esperimento

... cercare di essere sempre **'critici'** sulla validità e sul significato dei propri risultati

... e sul perché

...quali sono i limiti della nostra analisi? che cosa avremmo potuto fare di meglio?

*pensarci in anticipo*



ad esempio:  
tenendo le *stesse differenze tra le medie* dei campioni ma  
effettuando un esperimento con dodici replicazioni.....

tratt	N	Sottoinsieme per alfa = 0.05			
		1	2	3	4
3 - MR	12	5.128			
2 - Me	12		9.175		
4 - Ri	12			16.860	
1 - NT	12				26.773
Sig.		1.000	1.000	1.000	1.000

l'effetto sinergico risulterebbe  
significativo!



## 4. presentazione dei risultati e discussione della tesi

descrivere il disegno sperimentale e le modalità di conduzione dell'esperimento  
presentare in maniera sintetica i dati sperimentali  
descrivere e motivare qualsiasi trasformazione o pulizia dei dati  
non 'piegare' i dati ai propri desideri  
descrivere le metodologie statistiche usate nell'analisi e il significato dei test condotti  
presentare i risultati e la loro interpretazione  
non forzare l'interpretazione dei risultati  
riportare la parte essenziale della propria analisi statistica in sede di discussione finale di tesi

	Somma dei quadrati	df	Media dei quadrati	F	Sig.
Fra gruppi	1089.529	3	363.176	<b>23.663</b>	<b>.000</b>
Entro gruppi	184.177	12	15.348		
Totale	1273.706	15			

HSD di Tukey						
(I) tratt		Differenza fra medie (I-J)	Errore std.	Sig.	Intervallo di	
					Limite inferiore	Limite superiore
1 - NT	2	17.598	2.770	0.000	9.373	25.822
	3	21.645	2.770	0.000	13.421	29.869
	4	9.913	2.770	0.017	1.688	18.137
2 - Me	1	-17.598	2.770	0.000	-25.822	-9.373
	3	4.048	2.770	0.488	-4.177	12.272
	4	-7.685	2.770	0.070	-15.909	0.539
3 - MR	1	-21.645	2.770	<b>0.000</b>	-29.869	-13.421
	2	-4.048	2.770	<b>0.488</b>	-12.272	4.177
	4	-11.733	2.770	<b>0.006</b>	-19.957	-3.508
4 - Ri	1	-9.913	2.770	0.017	-18.137	-1.688
	2	7.685	2.770	0.070	-0.539	15.909
	3	11.733	2.770	0.006	3.508	19.957

\*. La differenza media è significativa al livello 0.05

