Corso di Calcolo delle Probabilità e Statistica
II Parte - STATISTICA

X Lezione
Analisi della varianza
Esempi e esercizi

Argomenti della X Lezione

- **Tests per il confronto di più medie: ANOVA**

Utilità e impiego dei tests

Caso dell'analisi della varianza a una via

- **Esempi**

Confronto tra medie

- test per **l'uguaglianza tra medie di due popolazioni**

  (distribuzione normale; varianza nota o sconosciuta, ma uguale/varianza sconosciuta)

- test per **l'uguaglianza tra medie in campioni accoppiati $(X_1,\ldots,X_n)$, $(Y_1,\ldots,Y_n)$**

  si procede come per un test relativo al valore di una sola media, ma utilizzando la media e la varianza delle differenze campionarie corrispondenti $X_i - Y_i$, $i=1,\ldots,n$

  Si effettua p.es. nei casi in cui i dati campionari sono relativi a condizioni precedenti e successive a un certo trattamento o evento.

Per estendere questo tipo di analisi a più di due popolazioni normali di cui si vogliano confrontare le medie

ANALISI DELLA VARIANZA (ANOVA)

Metodo: la variazione totale nella risposta che viene misurata (a una certa sollecitazione o **trattamento**) è suddivisa in componenti che vengono attribuite a specifiche **cause di variabilità**

È inserito nel ramo della Statistica che si occupa di **disegno degli esperimenti (DOE)**

Tests per confrontare *p* medie

Vogliamo verificare l'ipotesi nulla che non ci sia differenza tra diverse popolazioni (p. es. 5 popolazioni)

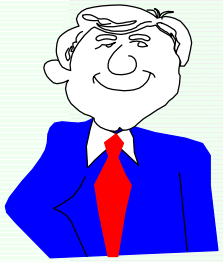
Si potrebbe effettuare il test tra due popolazioni per tutte le coppie possibili

Ci sono 10 coppie possibili su cui effettuare un test caratterizzato ogni volta da probabilità di errore di prima specie 0.05

Se i test sono tutti indipendenti la probabilità di trovare almeno un test che porti al rifiuto è $1-0.95^{10}=0.4013$ e quindi la probabilità di errore di I specie è **0.4013**

Troppo grande!

E' necessario introdurre una nuova metodologia per studiare questi casi: L'ANALISI DELLA VARIANZA



L'analisi della varianza può essere **a una o più vie**. Nel caso **a una via** si riconoscono solo le differenze tra diversi "trattamenti", nel caso **a più vie** si studiano più fattori che possono creare delle differenze tra le medie

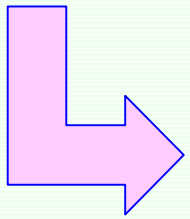
Considereremo questo caso

Analisi della varianza

Metodologia per verificare se due o più popolazioni sono caratterizzate dalla stessa media (o più medie sono estratte dalla stessa popolazione)

- **Analisi della varianza a una via**: si considera una sola causa di variazione nell'esito di ciascun esperimento

- **Fattore**: causa di variazione considerata.



Se si indaga sull'effetto di 5 tipi di deframmentatori sulla velocità di prestazioni di un certo PC si devono raccogliere 5 campioni, ciascuno relativo all'applicazione di un diverso deframmentatore. **Il fattore in questo caso è il deframmentatore.**

TERMINOLOGIA Disegno completamente casuale:

ciascun trattamento viene

assegnato in modo TOTALMENTE

casuale ai soggetti su cui si

effettuano le misure per determinare

l'effetto dei singoli trattamenti

Se voglio confrontare l'effetto di 4 antivirus e ho 100 terminali con uguali caratteristiche, divido casualmente i 100 terminali in gruppi di 25 e su ciascun gruppo installo un diverso antivirus.

Analisi della varianza: esperimenti completamente casualizzati

Quantità coinvolte:

x_{ij} = i-ma osservazione relativa al j-mo trattamento

$i = 1, 2, ..., n_j \qquad j = 1, ..., k$

Si definiscono:

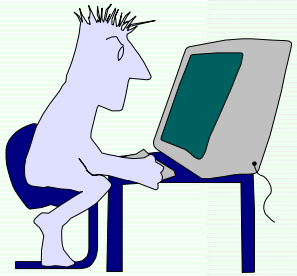
$$T_{.j} = \sum_{i=1}^{n_j} x_{ij}$$

$$\overline{x}_{.j} = \frac{T_{.j}}{n_j}$$

Per il trattamento j-simo

	Trattamenti				
	1	2	...	k	
	x_{11}	x_{12}	...	x_{1k}	
	x_{21}	x_{22}	...	x_{2k}	
	x_{31}	x_{32}	...	x_{3k}	
	⋮	⋮	⋮	⋮	
	$x_{n_1 1}$	$x_{n_2 2}$	...	$x_{n_k k}$	
Totale	$T_{.1}$	$T_{.2}$	...	$T_{.k}$	$T_{..}$
Media	$\overline{x}_{.1}$	$\overline{x}_{.2}$	...	$\overline{x}_{.k}$	$\overline{x}_{..}$

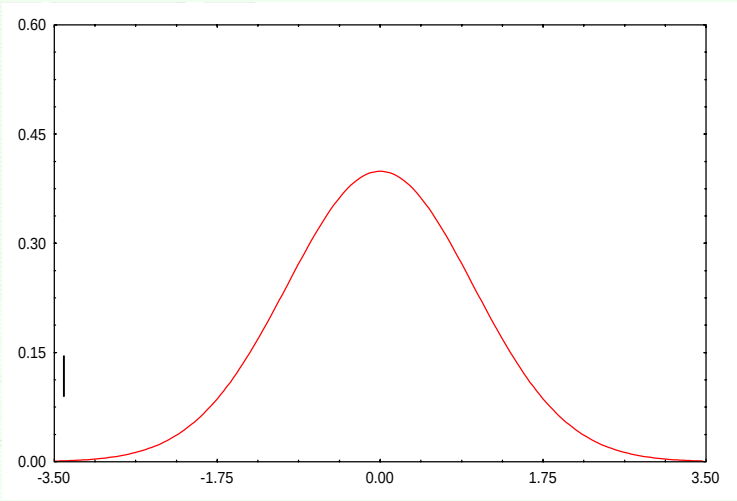
Problema:



Trovare un metodo che ci permetta di distinguere se le differenze che osserviamo tra le medie dei diversi campioni sono dovute a **reali differenze tra i trattamenti** o alla naturale differenza che si osserva campionando ripetutamente dalla stessa popolazione

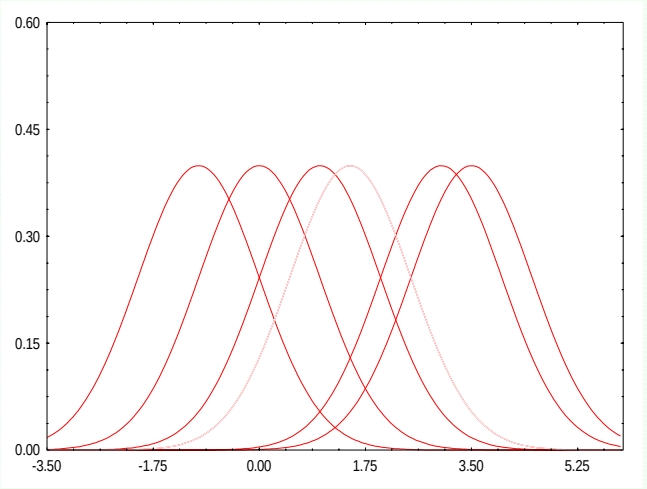
ESEMPIO

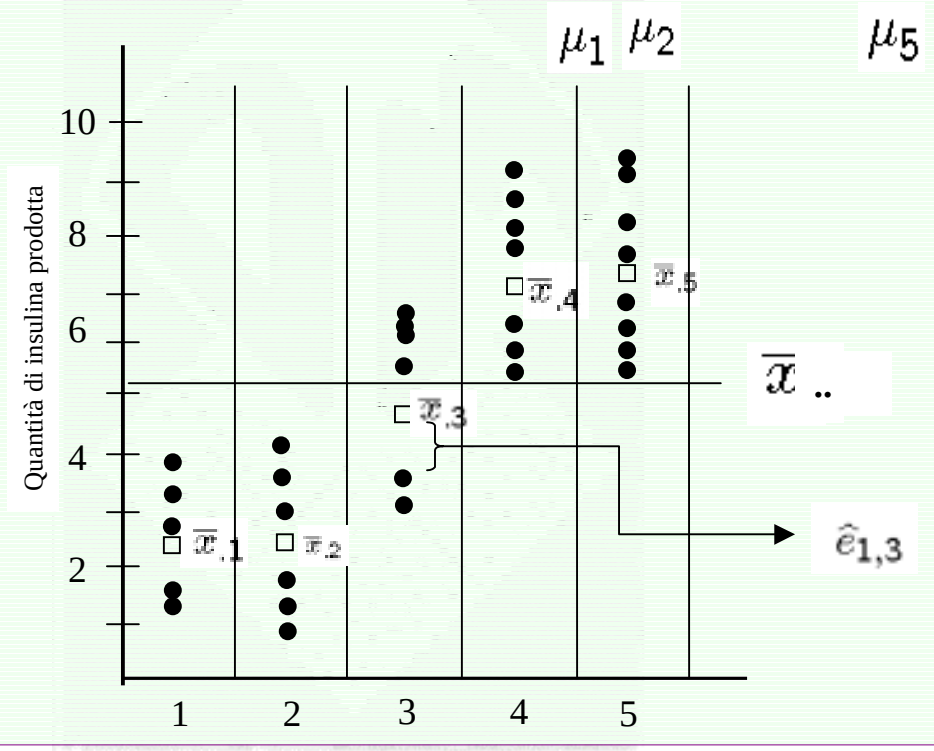
Se le medie sono tutte uguali ho:



$$\mu_1 = \mu_2 = \ldots = \mu_5$$

Se invece non vale l'ipotesi nulla le 5 popolazioni hanno medie diverse





$$x_{ij} = \mu_j + e_{ij}$$

Media relativa al j-mo trattamento

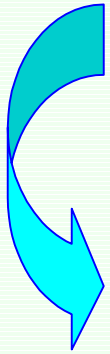
Errore commesso nella misura i-ma relativa al j-mo trattamento,
ovvero deviazione casuale dalla media dell'i-sima popolazione

$$\mu = \frac{\sum \mu_j}{k}$$

Media del campione totale, costituito dall'unione di tutti i campioni

$$\tau_j = \mu_j - \mu$$

Scarto tra la media relativa a un trattamento e quella del campione totale dovuto al fatto che l'elemento ha ricevuto il trattamento j



$$x_{ij} = \mu + \tau_j + e_{ij}$$

Somma degli scarti al quadrato per ogni gruppo di trattamenti

$$SST = \underbrace{\sum_{j=1}^{k} \overbrace{\sum_{i=1}^{n_j} \left(x_{ij} - \overline{x}_{\cdot\,\cdot}\right)^2}}_{\text{somma totale dei quadrati}}$$

$$SST = \underbrace{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{x}_{\cdot\,j}\right)^2}_{SSW} + \underbrace{\sum_{j=1}^{k} n_j \left(\overline{x}_{\cdot\,j} - \overline{x}_{\cdot\,\cdot}\right)^2}_{SSA}$$

$$SST = \underbrace{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{x}_{.\,j}\right)^2}_{SSW} + \underbrace{\sum_{j=1}^{k} n_j \left(\overline{x}_{.\,j} - \overline{x}_{.\,.}\right)^2}_{SSA}$$

SSW : somma dei quadrati degli scarti interni a ogni gruppo

SSA : somma dei quadrati degli scarti TRA gruppi

Statistica del test

$$VR = \frac{[SSA/(k-1)]}{\left[SSW/\left(\sum_{j=1}^{k}(n_j - 1)\right)\right]}$$

$$\left[SSW/\left(\sum_{j=1}^{k}(n_j - 1)\right)\right] = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{x}_{.\,j}\right)^2}{\sum_{j=1}^{k}(n_j - 1)} = \text{varianza interna ai gruppi pesata (o media quadratica nei gruppi)}$$

$$[SSA/(k-1)] = \frac{\sum_{j=1}^{k} n_j \left(\overline{x}_{.\,j} - \overline{x}_{.\,.}\right)^2}{k-1} = \text{media quadratica tra i gruppi}$$

Statistica del test

$$VR = \frac{[SSA/(k-1)]}{\left[SSW/\left(\sum_{j=1}^{k}(n_j - 1)\right)\right]}$$

$N-k$

Distribuzione della statistica:

Se i campioni possono venir considerati estratti dalla stessa popolazione VR dev'essere circa uguale a 1

VR ~ F (*k*-1, *N*-*k*)

F di Fisher

Riassumendo:

Se si vuole eseguire un test **sull'uguaglianza di più medie** estratte da popolazioni normali, si considera la statistica

$$VR = \frac{[SSA/(k-1)]}{\left[SSW/\ (\mathrm{N}-\mathrm{k})\right]}$$

e si esprime la regola di decisione del test, in dipendenza dall'ipotesi nulla e dal livello di significatività scelto, come:

H_0	H_1	Regione critica
$\mu_1 = \mu_2$ $\ldots = \mu_k$	$\exists i:$ $\mu_i \neq \mu_j$	$VR < F_{\frac{\alpha}{2}}(k-1, N-k)$ $VR > F_{1-\frac{\alpha}{2}}(k-1, N-k)$

Esempio: nello studio dell'effetto del glucosio nella produzione di insulina, si considerano campioni di tessuto pancreatico prelevati da cavie trattate con 5 differenti stimolanti. I ricercatori vogliono scoprire se c'è una differenza tra i 5 trattamenti. (campionamento completamente casuale)

x_{i1}: i-mo dato relativo a cavie soggette al primo trattamento

$\overline{x}_{.1}$: effetto medio del primo trattamento

	Stimolante					
	1	2	3	4	5	
	1.53	3.15	3.89	8.18	5..86	
	1.61	3.96	3.68	5.64	5.46	
	3.75	3.59	5.70	7.36	5.69	
	2.89	1.89	5.62	5.33	6.49	
	3.26	1.45	5.79	8.82	7.81	
		1.56	5.33	5.26	9.03	
				7.10	7.49	
					8.98	
Totale	13.04	15.60	30.01	47.69	56.81	163.15
Media	2.61	2.60	5.00	6.81	7.10	5.10

Dall'esempio:

SST = 162.54282 *SSW* = 41.35739 *SSA* = 121.18543

$$MSW = SSW/(N-k) = 1.5317552$$
$$MSA = SSA/(k-1) = 30.296358$$

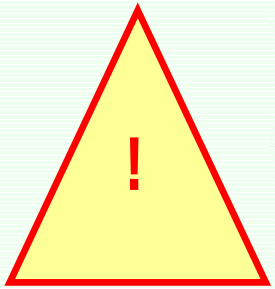
Causa variabilità	Somma dei quadrati	Gradi libertà	Media Quadratica	Rapporto delle varianze
Tra campioni	121.18543	4	30.296358	19.78
Interna ai campioni	41.35739	27	1.5317552	
Totale	162.54282	31		

Valore critico di F=2.73 < 19.78

Rifiuto l'ipotesi nulla: i diversi trattamenti hanno effetti diversi

Tabella ANOVA

Causa variabilità	Somma dei quadrati	Gradi libertà	Media Quadratica	Rapporto delle varianze
Tra campioni	$SSA = \sum_{j=1}^{k} n_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) = \sum_{j=1}^{k} \frac{T_{\cdot j}^2}{n_j} - \frac{T^2}{N}$	$k-1$	$MSA = SSA$ /(k-1)	$VR = \frac{MSA}{MSW}$
Interna ai campioni	$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\cdot j})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{k} \frac{T_{\cdot j}^2}{n_j}$	$N-k$	$MSW = SSW/(N-k)$	
Totale	$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\cdot\cdot})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}^2 - \frac{T^2}{N}$	$N-1$		



Attenzione: per usare il disegno completamente casuale occorre essere certi che gli individui che ricevono il trattamento siano omogenei.

Esempio: se si devono confrontare gli effetti di 4 farmaci antiinfluenzali, si devono selezionare un gruppo di individui con età, stato salute, … omogenei. Poi occorre selezionare il gruppo in 4 sottogruppi cui si distribuiranno i 4 farmaci.

Si potrebbe agire nello stesso modo se ci fossero, per esempio, individui di diverse età. Tuttavia in questo caso si dovrebbero attribuire gli individui tenendo conto di queste differenze, per esempio attribuendo un individuo anziano a caso a ciascun sottogruppo (**disegno completamente casuale a blocchi**).

ESEMPIO

Si domanda se le medie dei punteggi ottenuti in un certo test da studenti provenienti da 3 diverse scuole superiori siano diverse tra loro. Dato il seguente campione di 5 studenti per ogni scuola, al 5% di significatività si rifiuta o si accetta l'ipotesi che le tre scuole siano equivalenti?

Scuola 1	220	251	226	246	260
Scuola 2	244	235	232	242	225
Scuola 3	252	272	250	238	256

MSW = SSW/(N-k) = 165.9667

MSA = SSA / (K-1) = 431.6667

Valore della statistica di test : 2.6009

p-value : 0.1124 → Non si può rifiutare l'ipotesi nulla ai livelli < 0.1124