METODOLOGIA

DELLA

RICERCA PSICOLOGICA

ROBERTO BOLZANI

2000

Introduzione

Legge deterministica: corrispondenza univoca fra due eventi, causa ed effetto.

Legge probabilistica: corrispondenza fra un evento e un insieme di possibili eventi

Definizioni di Probabilità

Classica. Dato un insieme di eventi equiprobabili la probabilità di un evento è data da

$$\frac{\text{numero di eventi favorevoli}}{\text{numero di casi possibili}}$$

Assiomatica. La probabilità è definita dalle condizioni:

a) Ad ogni evento A corrisponde un valore p(A) maggiore o uguale a zero
b) La probabilità di tutti gli eventi possibili è uno
c) La probabilità che si verifichi A o B, essendo A e B mutuamente escludenti, è data dalla somma della probabilità di A e della probabilità di B

In formule:

a) $p(A) \geq 0$

b) $p(\Omega) = 1$

c) $p(A \text{ o } B) = p(A) + p(B)$ se $p(A\&B)=0$

Soggettiva. La probabilità di un evento E è la misura del grado di fiducia che un individuo *coerente* attribuisce, secondo le sue *informazioni*, all'avverarsi di E.

Frequentista. La probabilità di un evento è la frequenza con cui esso si presenta in un numero molto elevato di prove.

Proprietà della Probabilità

La probabilità di un **evento impossibile** è zero.

Non vale la proposizione inversa. Se la probabilità è zero l'evento non è necessariamente impossibile.

Es. La probabilità di ottenere 7 nel lancio di un dado a sei facce è zero. La probabilità di avere su infiniti lanci di una moneta nemmeno un risultato 'testa' è zero ma l'evento non è impossibile.

La probabilità di un **evento certo** è uno.

Non vale la proposizione inversa.

Es. La probabilità di ottenere un numero compreso fra uno e sei in un lancio di un dado è uno.
La probabilità di avere su infiniti lanci di una moneta almeno un risultato 'testa' è uno pur non essendo l'evento certo.

Probabilità condizionata: p(A|B) = probabilità che avvenga A essendo avvenuto B.
Es. probabilità di ottenere 12 in due lanci di un dado sapendo che nel primo lancio è risultato 6.

Eventi indipendenti: A e B sono indipendenti quando p(A|B) = p(A).
Es. la probabilità di avere testa nel primo lancio e croce nel secondo.

Eventi disgiunti: A e B sono eventi disgiunti se il verificarsi dell'uno esclude il verificarsi dell'altro.
Es. testa e croce.

Evento prodotto: Evento in cui si verifica sia A che B:
$p(A\&B) = p(A) \times p(B|A)$.
Se A e B sono indipendenti:
$p(A\&B) = p(A) \times p(B)$

Evento somma: Evento in cui si verifica A o B o, se non sono disgiunti, entrambi:
$p(A+B) = p(A) + p(B) - p(A\&B)$

Evento complementare: Evento in cui non si verifica A: p(Ã)=1 - p(A).
Es. il complementare del risultato 6 è il risultato 1 o 2 o 3 o 4 o 5.

PARAMETRI DESCRITTIVI

Variabili: qualitative
quantitative discrete
continue

Frequenza di un evento: Numero di volte in cui si verifica un evento diviso per il numero totale delle occorrenze.

Legge dei grandi numeri: Al crescere del numero delle Prove

$$P(|\, p_E - f_E \,| < \varepsilon) \rightarrow 1$$

dove p_E è la probabilità dell'evento E, f_E la sua frequenza, ε una costante qualsiasi > 0.

Media: somma di tutti i valori di una variabile divisa per il numero totale dei valori.

$$\bar{x} = \sum_{1}^{n} \frac{x_i}{n}$$

Varianza: somma dei quadrati degli scarti dei singoli valori dalla media divisa per i gradi di libertà.

$$s^2 = \sum_{1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Deviazione standard: radice quadrata della varianza.

Percentile: ordinando i casi secondo il valore di una variabile, l'*n*-esimo percentile è il limite al di sotto del quale si trova l'*n%* dei casi.

Mediana: punto che divide la popolazione in due parti di uguale numerosità. Corrisponde al 50° percentile.

Moda: valore per cui si ha un picco di frequenza. Caratterizza la distribuzione, che risulta unimodale, bimodale etc. a seconda dei picchi presenti.

Distribuzione di Probabilità di Variabile Aleatoria

Insieme dei valori di probabilità che competono a ciascun valore della variabile.

Funzione di distribuzione: funzione che rappresenta per ogni *x* la probabilità di ottenere un valore minore o uguale a *x*.

Se la *variabile è discreta* abbiamo una probabilità per ogni valore *x* discreto della variabile e la funzione di distribuzione si ottiene sommando le probabilità di tutti i casi aventi un valore inferiore ad *X*.

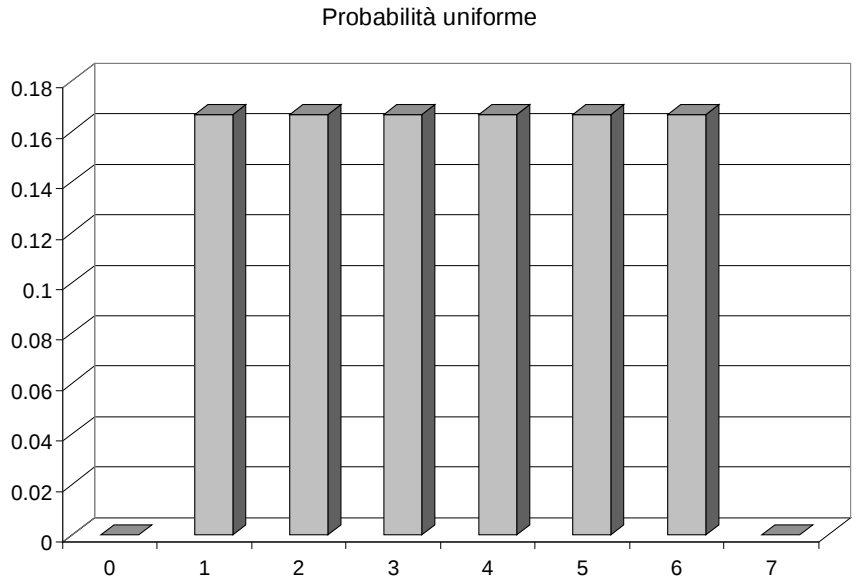
Se la *variabile è continua* la probabilità di un singolo valore della variabile è nulla essendo la probabilità di un valore su infiniti valori possibili. La funzione di distribuzione viene allora definita da

$$F(X) = p(x < X) = \int_{-\infty}^{X} f(x)dx$$

dove *f(x)* è la **densità di probabilità**. Essa può essere vista anche come quella curva che sottende (fra l'estremo sinistro e *X*) un'area pari alla funzione di distribuzione.

DISTRIBUZIONE UNIFORME

Distribuzione relativa ad una variabile discreta o continua avente uguale probabilità per ciascun suo valore.



Il grafico rappresenta una probabilità nulla per valori esterni all'intervallo 1-6 e uguale a 1/6 per i valori (discreti) interni a tale intervallo.

DISTRIBUZIONE BINOMIALE

Se il risultato di una prova può essere il successo S o l'insuccesso I con uguale probabilità p=q=1/2, i risultati possibili di due prove sono

SS SI IS II

ciascuno con probabilità 1/4.

L'evento 2 successi ha probabilità 1/4, 1 successo ha probabilità 1/2, zero successi probabilità 1/4.

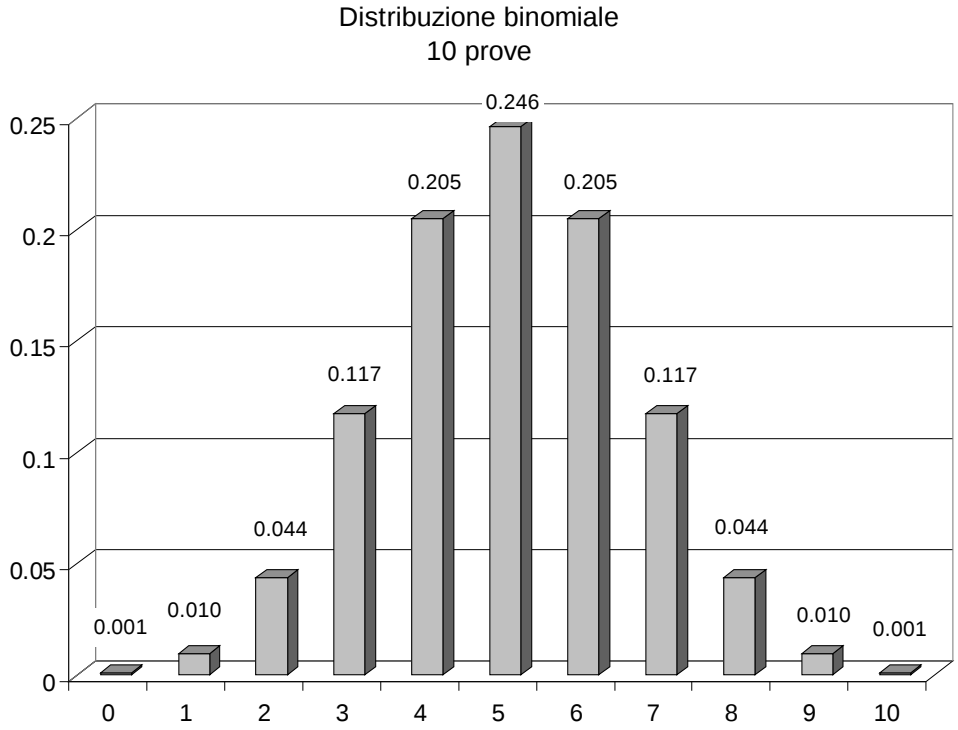
In generale su n prove la probabilità di s successi è data da:

$$p = \frac{1}{2^n} \binom{n}{s}$$

dove

$$\binom{n}{s} = \frac{n!}{s!(n-s)!}$$

Su 10 prove si ottiene ad esempio la seguente distribuzione:



la curva risulta simmetrica.

Se la probabilità di successo p è diversa dalla probabilità di insuccesso q=1-p allora la probabilità di s successi è data da

$$p \neq q$$

$$p = p^s q^i \binom{n}{s}$$

dove *i* rappresenta il numero di insuccessi.
La distribuzione risulta asimmetrica. Tale asimmetria è più evidente se il numero totale delle prove è piccolo.
Al tendere di n all'infinto la distribuzione binomiale tende alla distribuzione normale.

DISTRIBUZIONE NORMALE (GAUSSIANA)

La distribuzione normale riveste particolare importanza in statistica in quanto soddisfa ad alcuni requisiti molto generali.

Limite della distribuzione binomiale.
Al crescere di n la distribuzione binomiale tende ad una distribuzione normale con media np e varianza npq.

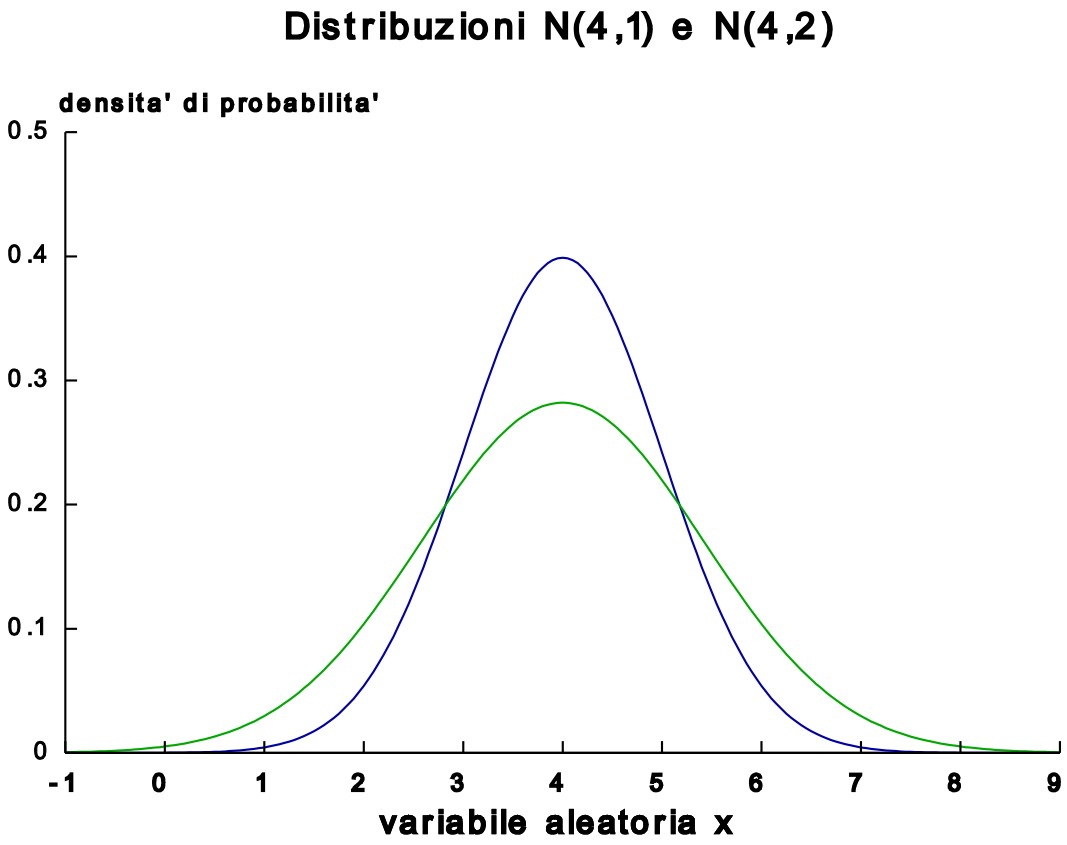
Curva degli errori. Sotto le condizioni che

- un errore sia la somma di molte componenti di uguale ampiezza
- le diverse componenti siano fra loro indipendenti
- ciascuna componente possa essere positiva o negativa con uguale probabilità

allora l'ampiezza dell'errore ha una distribuzione normale.

Distribuzione a massima entropia.

La distribuzione normale è la distribuzione di probabilità con la massima entropia per una variabile compresa fra $-\infty$ e $+\infty$ ed avente un data media e varianza. È quindi la distribuzione meno *strutturata*, la più casuale.



Una **generica variabile normale** con media μ e varianza σ^2 è indicata con $\mathbf{N(\mu, \sigma^2)}$ e la sua densità di probabilità è

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

È simmetrica con massimo in corrispondenza del valor medio. La sua forma è tanto più allargata quanto maggiore è la varianza.

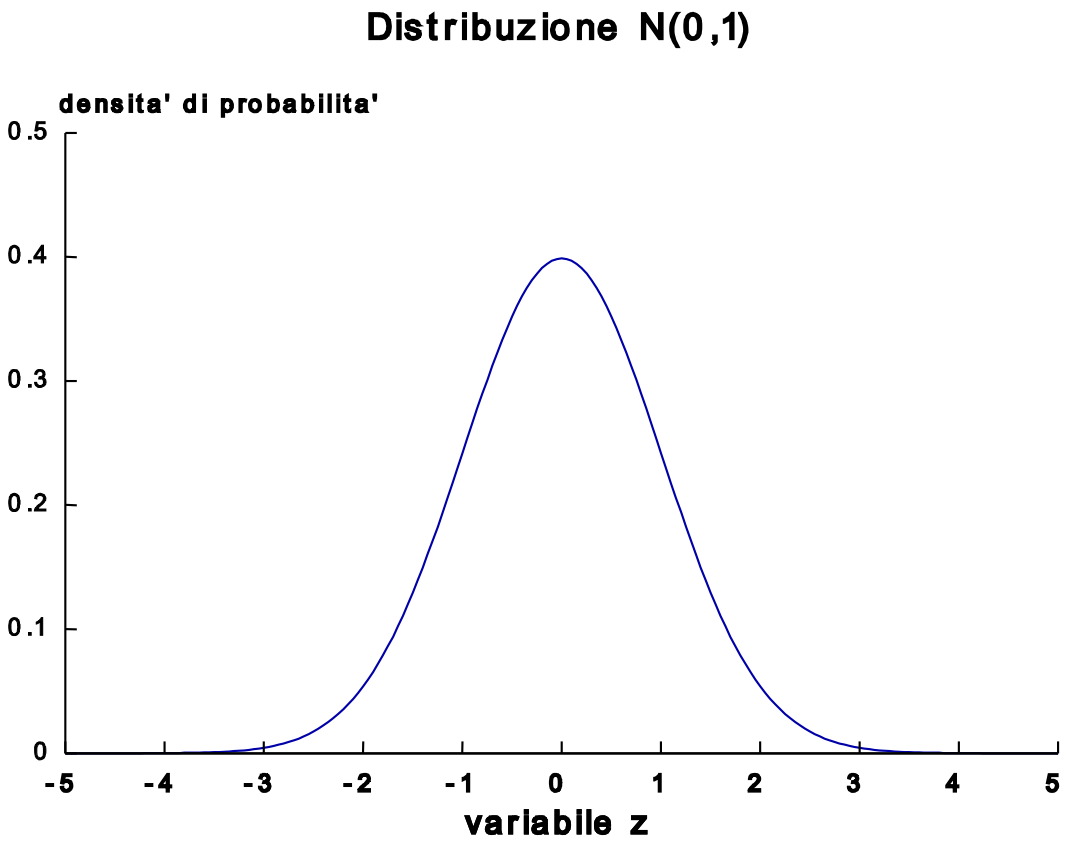
Una variabile normale a media zero e varianza unitaria è detta **variabile z** o **standard**, si indica con N(0,1) e la sua densità di probabilità è data da

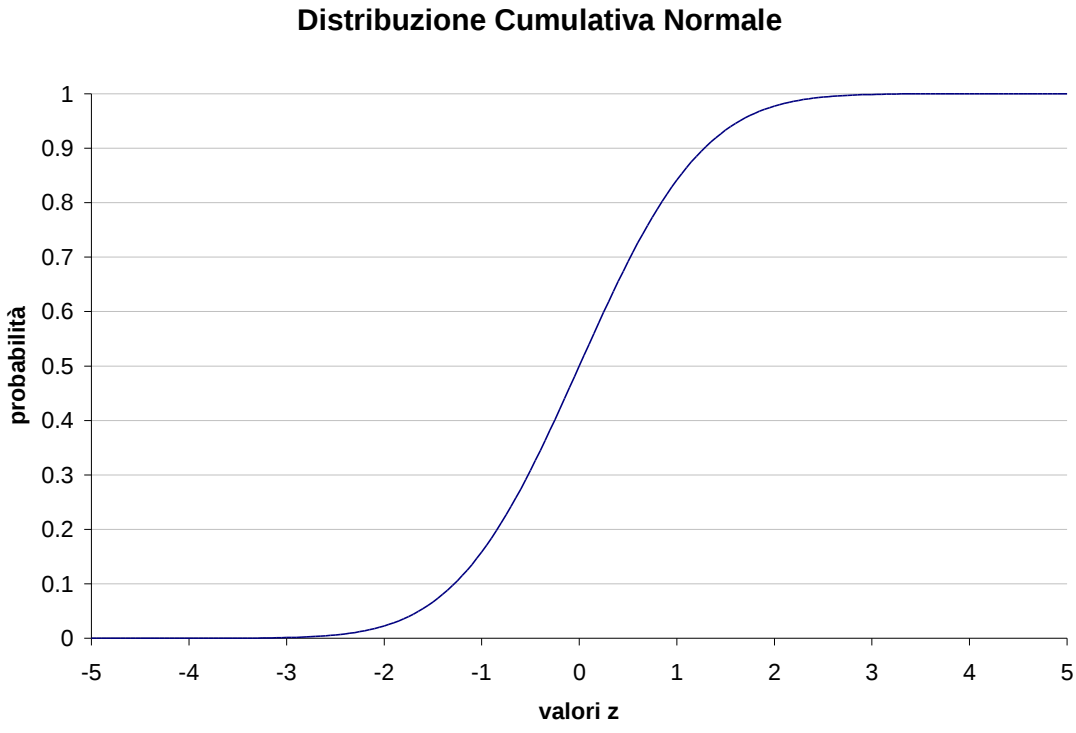
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

dove **z** e **x** sono legati dalla relazione

$$z = \frac{x - \mu}{\sigma}; \quad x = z \cdot \sigma + \mu$$

Essendo la distribuzione di una variabile continua il suo valore per un dato x corrisponde alla densità di probabilità per quel valore. Trova frequente applicazione la sua funzione di distribuzione che viene data in forma tabulata su molti manuali.


Distribuzione N(0,1)
densita' di probabilita'
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
variabile z


Distribuzione Cumulativa Normale
probabilità
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
valori z

La **somma di variabili normali** è anche essa normale.Il prodotto di variabili normali invece non mantiene la normalità della distribuzione.

DISTRIBUZIONE χ^2

È la distribuzione della variabile somma di variabili N(0,1) elevate al quadrato

$$\chi^2 = z_1^2 + z_2^2 + z_3^2 + ... + z_n^2$$

z:N(0,1)

Il numero n delle componenti definisce i gradi di libertà della distribuzione.

Ha espressione

$$f(\chi^2) = \frac{2^{-\frac{n}{2}}(\chi^2)^{n-1} e^{-\frac{1}{2}\chi^2}}{\Gamma(\frac{n}{2})}$$

e si trova tabulata essendo utilizzata nei test sulle frequenze e in test sulle matrici.
È asimmetrica e tende alla distribuzione normale al crescere dei gradi di libertà.

DISTRIBUZIONE *t*

Distribuzione di una variabile rapporto fra una variabile N(0,1) e la radice quadrata di una variabile χ^2 divisa per i gradi di libertà.È simmetrica e tende alla normale.
Ha espressione

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})} \frac{1}{(1+\frac{t^2}{n})^{\frac{n+1}{2}}}$$

DISTRIBUZIONE *F*

Distribuzione di una variabile rapporto di due variabili χ^2 divise per i rispettivi gradi di libertà.

Ha espressione

$$f(F) = \frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\ \Gamma(\frac{n_2}{2})} (\frac{n_1}{n_2})^{\frac{n_1}{2}}\ F^{\frac{n_1-2}{2}}\ (1+\frac{n_1}{n_2}F)^{\frac{-n_1+n_2}{2}}$$

con gradi di libertà n_1 e n_2. Si utilizza in molti test riguardanti l'analisi della varianza e della regressione sia univariata che multivariata.

Statistica descrittiva: insieme di tecniche idonee alla rappresentazione sintetica dei diversi valori relativi ai soggetti di un determinato gruppo.

- media
- frequenza
- percentuale etc.

Riguarda esclusivamente i soggetti esaminati.

Es. Rappresentazione dei dati elettorali (punteggi totali, percentuali sui votanti, variazioni etc.)

Statistica inferenziale: insieme di procedure atte a:

-saggiare l'influenza di alcuni fattori sui parametri
-classificare soggetti in vari gruppi
-prevedere l'andamento di certi parametri.

Riguarda concetti generali e quindi tutti i possibili soggetti che rispondono a certe caratteristiche.

Es. Influenza dell'età sull'apprendimento.

TEST STATISTICI

idea scientifica
verifica sperimentale
falsificazione o conferma

Verifica sperimentale deterministica:

verifica di una pura relazione causa-effetto. Es. rotazione terrestre -----> giorno/notte

Verifica sperimentale probabilistica:

verifica sperimentale di un legame generico fra due eventi. Es.
rotazione terrestre -----> temperatura al suolo
grado di istruzione -----> reddito

STATISTICA DESCRITTIVA

raccolta dati
descrizione sintetica di *quei dati*
idea generale

STATISTICA INDUTTIVA (INFERENZA)

idea generale
evento che può falsificare l'idea
scelta di un *campione* opportuno e
raccolta dati
test statistico
conferma dell'idea o non conferma

CAMPIONE. Idoneo a confermare l'idea. Rappresentativo dell'intera popolazione (casuale, sufficientemente ampio). Conforme alle richieste del test che si intende utilizzare (distribuzione,indipendenza).

TEST. Creati per essere applicati in modo indipendente. Richiedono che i dati sperimentali abbiano determinate distribuzioni teoriche (continuità, normalità ..). In grado di falsificare alcuni tipi determinati di ipotesi nulle.

IPOTESI NULLA. Ipotesi la cui accettazione renderebbe falsa l'idea da verificare. Viene in genere indicata con H_0.

Es. x - y H_0: x = y

SIGNIFICATIVITÀ. Probabilità di respingere l'ipotesi nulla pur essendo questa vera (α , errore del I tipo).

INDIPENDENZA. Lo stesso campione non può essere utilizzato per condurre test separati. La probabilità di falsi positivi cresce col numero dei test ed in ragione della loro dipendenza.

È possibile scomporre un test in più test fra loro indipendenti.

LIVELLO DI SIGNIFICATIVITÀ. L'essenza del test statistico consiste nel valutare se le differenze riscontrate in un campione casuale si possano attribuire a fattori diversi dalla oscillazione casuale.

Poiché sperimentalmente non è possibile raggiungere tale conclusione con assoluta certezza, si stabilisce a priori quale

probabilità di errore consideriamo accettabile per la verifica (*livello di significatività* normalmente 0.05 o 0.01).

Scegliere un livello di significatività basso (0.005 0.001) significa proteggersi da errori del I tipo ma esporsi maggiormente ad errori del II tipo: non respingere cioè l'ipotesi nulla H_0 quando questa è falsa.

	H_0 vera H_1 falsa	H_0 falsa H_1 vera
Respingo H_0	errore I tipo α	corretto
non respingo H_0	corretto	errore II tipo β

POTENZA DI UN TEST. Probabilità di respingere H_0 quando H_0 è falsa. È dato da $1-\beta$.
La potenza di un test dipende sia da H_0 che da H_1.

FALSIFICAZIONE DELL'IPOTESI NULLA. Una volta scelto il livello di significatività è possibile determinare i limiti entro cui potrebbero
cadere i dati per effetto di spostamenti casuali dall'ipotesi H_0, pur essendo questa vera.
Se il campione è esterno a tali limiti si respinge l'ipotesi nulla H_0 con un livello di significatività
α e si accetta l'ipotesi alternativa H_1 di cui volevamo dimostrare la validità.
Se il campione è interno a tali limiti non si respinge l'ipotesi nulla e quindi non si accetta l'ipotesi H_1.

INTERVALLO DI CONFIDENZA. Un modo alternativo per misurare la variabilità dell'ipotesi nulla è quello di costruire

l'*intervallo di confidenza*. Esso rappresenta la zona, attorno al parametro stimato sperimentalmente, in cui potrebbe cadere il valore vero del parametro con una probabilità $1-\alpha$. Ha la stessa estensione dell'intervallo attorno all'ipotesi nulla. Se nell'intervallo di confidenza cade il valore di H_0 non si può escludere che la differenza trovata sia dovuta al caso. Non posso quindi respingere l'ipotesi nulla.

SITUAZIONI IN CUI NON VIENE FALSIFICATA L'IPOTESI NULLA

La conferma in statistica viene fatta dimostrando che l'ipotesi opposta (ipotesi nulla) non è sostenibile, è altamente improbabile.

Ne segue che la non conferma non indica necessariamente che l'idea è falsa che è cioè vera l'ipotesi nulla ma solamente che l'ipotesi nulla non è altamente improbabile.

La non falsificazione dell'ipotesi nulla si può avere perchè:

- l'ipotesi nulla è vera
- scarsa potenza del test:
  - il campione ha varianza elevata
  - scarsa numerosità del campione
  - il campione non soddisfa le condizioni relative alla distribuzione
  - il campione non è rappresentativo dell'intera popolazione
  - non sufficiente separazione fra H_0 e H_1

CONFRONTO FRA FREQUENZE

Si confrontano i valori sperimentali E_i con quelli teorici T_i, provenienti dall'ipotesi sulla distribuzione, e si calcola la variabile

$$\chi^2 = \sum_i \frac{(T_i - E_i)^2}{T_i}$$

Analoga procedura può essere utilizzata per eseguire test sulle proporzioni.

Esempio
In dieci lanci di una moneta abbiamo ottenuto T=8 e C=2. L'ipotesi da dimostrare è che la moneta sia truccata. La corrispondente ipotesi nulla è

H_0: p(T)=p(C)=.5

Calcoliamo

$$\chi^2 = (8-5)^2/5 + (2-5)^2/5 = 3.6$$

poichè $\chi^2_{.05,1}$ (per una significatività di 0.05 e 1 grado di libertà) vale 3.84 > 3.6, non è possibile respingere l'ipotesi nulla.

TAVOLE DI CONTINGENZA

Se abbiamo una tabella di frequenza l'ipotesi nulla è che le frequenze nelle varie caselle siano proporzionali ai totali di riga e colonna.

Esempio
I risultati di una terapia applicata in due gruppi è presentata in tabella

	Risultato Negativo	risultato positivo	totale
gruppo A	41 x_{11}	216 x_{12}	257 $x_{1.}$
gruppo B	64 x_{21}	180 x_{22}	244 $x_{2.}$
totale	105 $x_{.1}$	396 $x_{.2}$	501 $x_{..}$

La tabella teorica che si ottiene ammettendo che le frequenze nelle caselle siano proporzionali ai totali marginali si ottengono dalla

$$T_{ij} = \frac{x_{i.}\, x_{.j}}{x_{..}}$$

	risultato negativo	risultato positivo	totale
gruppo A	41 53.9	216 203.1	257
gruppo B	64 51.1	180 192.9	244
totale	105	396	501

si calcola con la solita formula

$\chi^2 = 7.978$

con (r-1)×(c-1) gradi di libertà che risulta sign al 0.5%.

Il test del χ^2 è approssimato. Per tabelle 2×2 è abbastanza semplice calcolare un test esatto basato sulla distribuzione binomiale. Questo calcolo esatto è consigliabile quando vi sono frequenze teoriche < 5.
Nel caso che alcune caselle abbiano valori piccoli ma >5, è consigliabile utilizzare la correzione di Yates.

CONFRONTO FRA MEDIE E VARIANZE

IPOTESI H_0	CONDIZIONI	DISTRIBUZ.
$\mu = \mu_o$	σ^2 nota σ^2 ignota	N t
$\sigma^2 = \sigma^2_o$		χ^2
$\mu_1 = \mu_2$	σ_1^2 e σ_2^2 note $\sigma_1^2 = \sigma_2^2$ ignote $\sigma_1^2 \neq \sigma_2^2$ ignote	N t ?
$\sigma_1^2 = \sigma_2^2$		F

MEDIA e VARIANZA teoriche sono COSTANTI, non cambiano al variare del campione…sono quelle della popolazione globale e non sono soggette ad oscillazioni.

Per avere una media e una varianza teoriche occorre far rilevamenti su una popolazione ampia sia numericamente sia geograficamente, oppure quando la M e la V sono valori CONVENZIONALI fissati!

La VARIABILE Z [$z = \chi - \mu / \sigma$] non viene quasi mai usata, tranne in test non-parametrici perché non conosciamo sempre la V.

La DISTRIBUZIONE t dà lo stesso rapporto di z, ma con M e V stimate sul campione, ovvero quando non sono stimate M e V e ho solo M e V calcolate sul mio campione. La distrib t è relativa alla stessa formula della distrib z dove al posto di M e V teoriche ho M e V calcolate!

Le differenze non vengono valutate in assoluto, ma in rapporto all'ooscillazione casuale.

La distrib z e il t-test e l'analisi della Var e Regres non fanno altro che rapportare la differenza in studio con l'oscillazione casuale e poi nel valutare questo rapporto con l'opportuna distribuzione.

Se non abbiamo M e V teoriche usiamo la stessa formula, ma questa formula non dà luogo ad una var Gaussiana.

Var G/ var χ^2 è una distribuzione t e la devo applicare in tutti i test in cui non conosco la M e V teoriche (ciò succede il 99% delle volte)….questi sono i t-test in cui posso confrontare delle medie!!

TEST SULLE MEDIE

t-test

Nella maggioranza dei casi NON abbiamo la V teorica, allora faccio una stima della var sul campione, ma se questa stima la metto in una formula, essa NON è Gaussiana, ma ha una distrib χ^2 e non è una costante! Se noi calcoliamo una formula analoga a quella z otteniamo una distrib chiamata t, simmetrica!!

Di distrib t non ce n'è una sola, ma cambia la variare dei gradi di libertà (n-1)!

1. H_0: $\mu = \mu_o$

σ^2 nota

viene utilizzata la distribuzione normale $N(\mu_o, \sigma^2/n)$

σ^2 ignota

si utilizza

$$t = \frac{\bar{x} - \mu_o}{\sqrt{\frac{s^2}{n}}}$$

Se prendo gruppetti di 10 e calcolo l'altezza media dei gruppi io ho varie altezze nedie che non saranno uguali tra loro, ho quindi delle oscillazioni! La VARIANZA TRA LE MEDIE di questi campioni è l'ennesima parte della varianza che c'è tra gli individui.

+ sono grandi i num del campione,

- ci saranno oscillazioni!!

La formula del t-test è sostanzialmente la stessa della variabile z in cui xò non è nota la varianza teorica e da costante passa a variabile casuale!!

La DIFFERENZA TRA LE MEDIE la valuto pesandola inversamente alla probabilità:

+ i miei dati sono variabili,

- importante sarà la differenza tra le medie

Mettere al den la VARIANZA esprime che non è la differenza tra i parametri che contano, ma le DIFFERENZE rapportate alla var casuale!!!

Nessuna differenza in sé x sé è significativa se non viene rapportata all'oscillazione casuale!

t = differenza tra medie/ √varianza delle medie

2. H_0: $\mu_1 = \mu_2$

σ_1^2, σ_2^2 note

si utilizza la var z

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

$\sigma_1^2 = \sigma_2^2$ ignote

si utilizza il t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

dove S^2_p = gli scarti sono calcolati rispetto alla media del proprio gruppo…VARIANZA ACCORPATA

$$s_p^2 = \frac{\sum_1^{n_1}(x_{1j} - \bar{x}_1)^2 + \sum_1^{n_2}(x_{2j} - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{SSQ_1 + SSQ_2}{n_1 + n_2 - 2}$$

$\sigma_1^2 \neq \sigma_2^2$ ignote si utilizza

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

correggendo per i gradi di libertà

3. OSSERVAZIONI CORRELATE

utilizzare lo stesso gruppo x vedere x es. la pressione al mattino e alla sera mi avvantaggia xkè so le 2 pressioni sullo stesso soggetto e posso confrontare le 2 pressioni togliendo così la variabilità casuale dovuta ai diversi valori basali dei 2 gruppi (se ne prendevo 2 anziché lo stesso)!

Questo test viene preferito xkè toglie quindi la variabilità casuale individuale!

Così il t-test diventa:

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}}$$

dove

-il termine a numeratore d rappresenta la media delle differenze fra due osservazioni appaiate

-s^2_d rappresenta la varianza delle differenze

-il denominatore la varianza delle medie delle differenze

Anziché utilizzare i dati misurati, uso le differenze tra le misurazioni

4. TEST T^2 MULTIVARIATO

$$T^2 = \frac{n_1 \cdot n_2}{n_1 + n_2}(\overline{X}_1 - \overline{X}_2)\, S^{-1}(\overline{X}_1 - \overline{X}_2)$$

$$T^2 \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}$$

ha distribuzione F(p,n_1+n_2-p-1)

Nel confronto tra le medie se conosco la varianza utilizzo una distrib normale, se non conosco la varianza uso il t-test!!!

Nel confronto tra 2 varianza uso la distribu F che riguarda l'Analisi della Varianza..che serve quando confronto medie di gruppo se ho + di 2 gruppi!!!

CONFRONTO FRA LE MEDIE DI PIÙ GRUPPI.

ANALISI DELLA VARIANZA

Quando i gruppi sono più di due non è più possibile applicare il t-test per il confronto fra due medie. Perché se ho 5 gruppi non posso applicare il t-test tra il 1 e il 2, poi tra il 1 e il 3…ecc non è questa la procedura: viene meno la condizione dell'indipendenza e casualità del campione, inoltre aumenta l'errore α!!

Io devo poter confrontare tutti i gruppi che ho x ottenere un unico risultato!

Bisogna allora ricorrere all'analisi della varianza (estensione del t-test, che a sua volta è estensione del test z). Il suo presupposto fondamentale è che, se è vera l'ipotesi nulla che non vi sia differenza fra i gruppi, la variabilità all'interno dei gruppi è uguale alla variabilità fra i gruppi.

Si tratta quindi di un confronto di varianze che può essere saggiato con la distribuzione F.

Se io calcolo la var tra le medie dei gruppi, questa var mi dice quanto queste medie variano tra loro, ovvero la variabilità delle medie…questa vari la confronto con la var casuale, cioè quella interna al gruppo.

Prendo la var xkè la medie posso confrontarla solo x 2 gruppi!

Per ciascun soggetto *i* del gruppo *j* lo scarto dalla media generale può essere scomposto in uno scarto dalla media di gruppo più uno scarto della media di gruppo dalla media generale, vale cioè la relazione

$$x_{ij} - x.. = (x_{ij} - x._j) + (x._j - x..)$$

La stessa scomposizione può essere fatta anche sulle somme degli scarti al quadrato (SSQ)

$$SSQ_{tot} = SSQ_{intgr} + SSQ_{tragr}.$$

La somma dei quadrati degli scarti totali è calcolata sui valori di tutti i soggetti del campione rispetto la media generale.

La somma dei quadrati degli scarti tra i gruppi si ottiene attribuendo a ciascun soggetto il valore medio del suo gruppo e calcolando gli scarti dei valori così modificati dalla media generale.

La somma dei quadrati degli scarti all'interno dei gruppi si ottiene per differenza.

Le relative varianze si ottengono dividendo le somme dei quadrati degli scarti per i rispettivi gradi di libertà. La varianza all'interno dei gruppi è nota anche come *varianza residua*.

La variabile statistica su cui viene effettuato il test è rappresentata dal rapporto

$$F = \frac{\text{varianza tra gruppi}}{\text{varianza interno gruppi}}$$

VAR TRA I GRUPPI=

-la presenza delle differenze che io ipotizzo

-var dovuta agli effetti che sto studiando

VAR INTERNO AI GRUPPI=

-l'oscillazione casuale che uso come termine di confronto

-il rumore di fondo che deve essere piccolo

Se la differenza tra i due gruppi è casuale allora deve essere dello stesso ordine di quello all'interno dello stesso gruppo!

Se non è così allora significa che quelle differenze sono SIGNIFICATIVA quindi non è solo dovuta all'oscillazione casuale!!

Il rapporto F è tra 2 varianze xcui ha una distribuzione F (rapporto tra 2 variabili χ^2)

Se la varianza tra i gruppi è dovuta solo all'oscillazione casuale e non agli effetti che sto studiando (vera Ho) allora la var tra i gruppi sarà della stessa entità della var interna ai gruppi.

Se Ho è vera, le medie avranno differenze dovute al caso!

I vari test statistici rapportano la variabilità di un effetto con la variabilità casuale!!

Noi valutiamo gli effetti misurando le differenze!!

L'Analisi della Varianza è un tets GENERALE xkè contiene in sé il t-test…se ho infatti solo 2 gruppi questo test mi fa ciò che fa il t-test!!

INFATTI $\mathbf{F = t^2}$

gruppo1	gruppo2
16	12
14	14
17	11
13	13
14	14
16	12
15	15
17	12
14	11
15	13

	Mean	Std Dev	Cases
Tot	13.9000	1.8035	20
Group 1	15.1000	1.3703	10
Group 2	12.7000	1.3375	10

t= 3.96 d.f.= 18 p= .001

Source	SS	DF	MS	F	Sig
WITHIN	33.00	18	1.83		
CONST	3864.20	1	3864.20	2107.75	.000
GROUP	28.80	1	28.80	15.71	.001

L'AdV fa ciò che fa il t-test anche quando c'è un gruppo solo e un valore di riferimento! In questo caso non abbiamo l'effetto-gruppo xcui rimane la costante: con essa l'Adv confronta la media del gruppo con un valore zero… questo confronto corrisponde a chiedersi se i valori di partenza sono significativamente diversi da 100 (nel caso del QI).

La costante ha la caratteristica di essere sempre confrontata con zero!

Se oltre ad una divisione in gruppi abbiamo anche una suddivisione in sottogruppi (+ suddivido in gruppi e + è facile che polverizzo le differenze…quindi è meglio ridurre al minimo le classificazioni), si hanno più criteri di classificazione. In questo caso, oltre agli effetti dovuti ai gruppi, si hanno anche gli effetti dovuti alla interazione fra i vari criteri.

L'**interazione** non è rappresentato negli effetti principali, non è semplicemente dovuto alla somma degli effetti principali, ma rappresenta l'effetto di particolari combinazioni degli effetti principali..è la MANCANZA di ADDITIVITÀ degli effetti principali!

È l'effetto che esiste tra i fattori, rappresenta il non parallelismo tra i fattori principali e rende conto di particolari risultati dovuti alla COMBINAZIONE degli effetti principali!

Quando i fattori sono più di 1 in tutte le procedure c'è la possibilità dell'interazione…è un 'ingenuità non tener conto dell'interazione!

In base ai metodi utilizzati dall'analisi della varianza risulta che la sua utilizzazione richiede le seguenti condizioni:

- le variabili indipendenti devono essere variabili qualitative, di gruppo.
- la variabile dipendente deve essere una variabile quantitativa a distribuzione gaussiana.

DISEGNO SPERIMENTALE

Definisce il modo di dividere in gruppi e sottogruppi il campione sperimentale. Si intendono variabili considerate e criteri utilizzati.

Esempio:

due gruppi

trattati	controlli

Se prendiamo in considerazione anche il sesso i gruppi diventano

trattati M	controlli M
trattati F	controlli F

	gruppo 1	gruppo 2
M	31 35 34 32 36	39 41 43 38 40
F	36 37 38 33 38	41 36 35 41 38

Gruppi	Mean	Std Dev	Cases

gruppo 1			
M	33.600	2.074	5
F	36.400	2.074	5
gruppo 2			
M	40.200	1.924	5
F	38.200	2.775	5
tot	**37.100**	**3.227**	**20**

Source	SS	DF	MS	F	Sig
within	80.00	16	5.00		
constant	27528.20	1	27528.2	5505.64	.000
group	88.20	1	88.20	17.64	.001
sex	.80	1	.80	.16	.694
gr x sex	28.80	1	28.80	5.76	.029

Quando il disegno sperimentale comprende tutte le possibili combinazioni viene detto **disegno fattoriale**. È il tipo più frequente di disegno sperimentale.

Esistono altri tipi di **disegni incompleti** che si utilizzano quando non è possibile, o comunque non conveniente, esaminare tutte le possibili combinazioni dei fattori in esame. Il modello matematico utilizzato per l'analisi statistica è determinato dal tipo di disegno sperimentale.

A seconda del tipo di disegno sperimentale utilizzato possono essere analizzati diversi tipi di interazione.

MODELLO LINEARE

L'analisi della varianza può anche essere vista come la ricerca di una stima dei parametri di una
equazione, relativa al *k*-esimo soggetto appartenente al gruppo *i*, sottogruppo *j*, del tipo

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + e_{ijk}$$

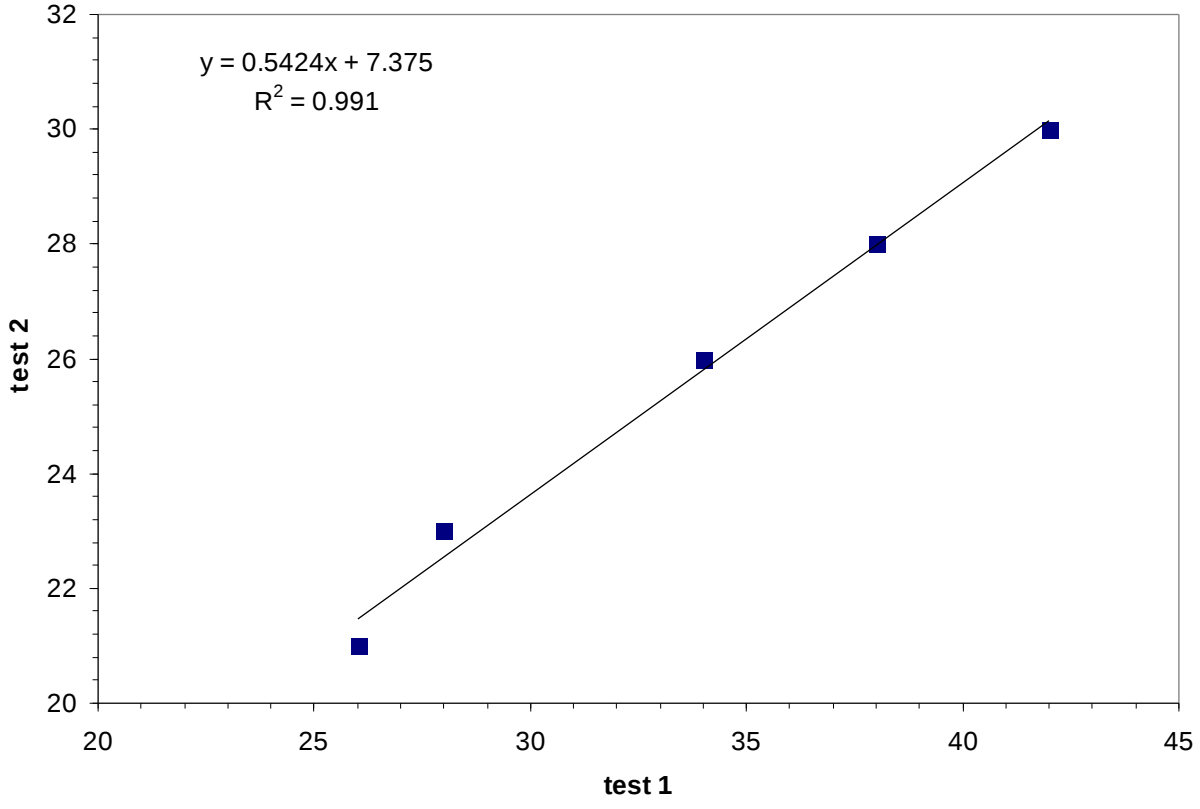
dove y_{ijk} rappresenta la variabile dipendente misurata e α e β rappresentano i parametri (variabili indipendenti) relativi agli effetti che influenzano la variabile dipendente. Il parametro τ rappresenta l'effetto dovuto all'interazione delle due variabili indipendenti.Il parametro *e* rappresenta il termine errore dovuto alla variazione casuale dei dati.

La formalizzazione presentata si presta ad un' analisi della varianza basata su metodi algebrici tipici dell'analisi della regressione.

ANALISI DELLA REGRESSIONE

Quando sia la variabile dipendente che le variabili indipendenti sono di tipo quantitativo a distribuzione gaussiana la loro relazione può essere studiata con l'analisi della regressione.

Soggetti	test 1	test 2
1	34	26
2	38	28
3	26	21
4	42	30
5	28	23



Nel caso di una sola variabile indipendente la relazione può essere rappresentata da una retta

o da una equazione del tipo

$$y = a + bx$$

La determinazione dei parametri *a* e *b* viene effettuata in modo che la somma dei quadrati degli scarti fra i valori *predetti* dalla equazione e quelli *sperimentali* sia minima *(metodo dei minimi quadrati).*

Il parametro *a* (intercetta) rappresenta il valore di y quando x=0, mentre *b* rappresenta la pendenza della retta. Essi sono dati da

$$a = \bar{y} - b \cdot \bar{x} \qquad b = \frac{s_{xy}}{s_x^2}$$

dove s_x^2 è la varianza di *x* mentre s_{xy} è la covarianza fra *x* e *y* che si ottiene con la formula

$$s_{xy} = \frac{\sum_{1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n - 1}$$

Dalla formula per il calcolo di *b* è evidente che i parametri cambiano se come variabile dipendente si considera *x* anzichè *y*.

L'efficienza della rappresentazione è misurata dal *coefficiente di determinazione* **R²** che è il rapporto fra la SSQ dovuta all'equazione e la SSQ complessiva del campione. Esso può variare da 0 a 1, dato che la varianza spiegata dall'equazione può essere al massimo uguale alla varianza totale.

Un altro indice della bontà dell'approssimazione dei dati con l'equazione di regressione è rappresentato dal **coefficiente di correlazione *r***, che si ottiene da

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

dove s_x e s_y sono le deviazioni standard di x e y mentre s_{xy} è la covarianza fra x e y .
Elevando al quadrato il coefficiente di correlazione r si ottiene il valore R^2.
Il coefficiente di correlazione misura quanto sia stretta la relazione fra la variabile indipendente x e la variabile dipendente y.

Nel caso che vi siano *più variabili* indipendenti il coefficiente R^2 si otterrà sempre come rapporto fra la SSQ dovuta al modello e la SSQ totale, mentre i coefficienti di correlazione fra la variabile dipendente y e ciascuna variable x saranno tanti quante sono le variabili indipendenti.

È possibile effettuare un test statistico sulla efficienza del modello lineare. L'ipotesi da dimostare è che l'equazione del

modello rappresenti in modo efficiente la relazione lineare fra i dati.
Il test si ottiene calcolando il rapporto

$$F = \frac{\text{var del modello}}{\text{var residua}}$$

Analysis of Variance

	DF	Sum of Square	Mean of Sq
Regression Residual	1 3	52.72232 .47768	52.72232 .15923
F=331.11589		Sig. F = .0004	

Multiple R .99550
R Square .99102

Variable	B	Sig
TEST1 Constant	.54241 7.375	.0004 .0054

ANALISI MULTIVARIATA

Quando un fenomeno è caratterizzato da più variabili dipendenti non è lecito effettuare più analisi univariate in quanto in tal modo non si tiene conto delle mutue relazioni fra le variabili.
L'intervallo di confidenza multivariato non può essere semplicemente dedotto da quelli multivariati.

Non sarebbe inoltre possibile ottenere una risposta unitaria sulla falsificabiltà dell'ipotesi nulla.

La formalizzazione matematica sia del modello relativo all'analisi della varianza che di quello relativo alla regressione permette tuttavia una facile estensione delle tecniche algebriche necessarie alla determinazione dei parametri.

Caratteristica principale delle tecniche multivariate è quella di saggiare contemporaneamente più variabili dipendenti. Vengono a volte erroneamente catalogate come multivariate la regressione multipla, l'analisi discriminante e l'analisi fattoriale. Quando queste metodiche di analisi non producono un test unico su più variabili dipendenti è opportuno classificarle come tecniche multiple.

ANALISI PER PROVE RIPETUTE

Quando una variabile dipendente viene misurata sugli stessi soggetti più volte in diverse condizioni sperimentali, ciò che interessa valutare sono le differenze, soggetto per soggetto, dei valori ottenuti. La significatività di queste differenze può essere valutata sul termine costante di una *analisi della varianza univariata* (con un opportuno metodo che tenga conto della ridotta variabilità all'interno del gruppo). Questo procedimento richiede tuttavia che siano verificate alcune condizioni riguardanti i dati sperimentali.

Un metodo con condizioni meno restrittive e con maggior potenzialità è quello dell'*analisi multivariata*. Come variabili dipendenti vengono utilizzate le differenze fra le diverse misure. Queste nuove variabili devono essere fra loro indipendenti per una corretta applicazione del metodo. Questo può essere ottenuto facendo la differenza fra il primo valore e il secondo, la differenza fra il terzo e la media dei primi due, etc.

È infine possibile, con questo approccio, effettuare anche l'analisi di più variabili dipendenti, ognuna ripetuta più volte.

TEST NON PARAMETRICI

Quasi tutti i test esaminati eseguono dei confronti basati sui *parametri* (media e varianza) della distribuzione gaussiana. Quando la variabile in studio non ha una distribuzione gaussiana in genere si operano alcune trasformazioni (logaritmiche, sinusoidali, etc.) tendenti a rendere i dati più vicini alla distribuzione richiesta. Inoltre alcune analisi statistiche (come per esempio l'analisi della varianza) sono poco sensibili alla deviazione dalla normalità. Quando però la deviazione dalla normalità è marcata e non facilmente rimediabile, non è possibile utilizzare comunque le procedure analizzate.

In questi casi si ricorre a **test non parametrici**, a test cioè che non fanno riferimento ai parametri della distribuzione. Riportiamo per esemplificazione il

TEST DEL SEGNO. Supponiamo di voler valutare se un certa condizione sperimentale migliora o peggiora una specifica caratteristica dell'individuo. Possiamo assegnare a ciascun soggetto il simbolo **+** **o** **-** a seconda che abbia migliorato o peggiorato. Se è vera l'ipotesi nulla che il miglioramento e il peggioramento abbiano uguale probabilità, mi debbo aspettare come risultato una distribuzione binomiale con p=.5.

La verifica viene allora effettuata valutando se il numero di miglioramenti ottenuti é esterno all'area della distribuzione binomiale contenente il 95% delle probabilità.

FINE