CAPITOLO III
CONFRONTI TRA DISTRIBUZIONI

3.1 CONFRONTI TRA DISTRIBUZIONI OSSERVATE E DISTRIBUZIONI TEORICHE OD ATTESE.

Nella teoria statistica e nella pratica sperimentale, è frequente la necessità di verificare se esiste accordo tra una distribuzione osservata e la corrispondente distribuzione attesa o teorica. Il test viene definito **test per la bontà dell'adattamento** (in inglese, goodness of fit test). Sia per dati qualitativi che possono essere classificati in categorie nominali, sia per dati quantitativi distribuiti in classi di frequenza, nella ricerca ambientale è spesso necessario **saggiare la concordanza tra fatto ed ipotesi. E' lo scopo per il quale storicamente è stato costruito il test** $\chi^2_{(\text{g. d. l.})}$ (chi-quadro o chi-quadrato).

E' un metodo di **inferenza statistica** che **non richiede ipotesi "a priori" sul tipo e sulle caratteristiche della distribuzione**, come invece avviene per la statistica parametrica che fa riferimento alla distribuzione normale. Sono **metodi non parametrici** (detti anche "distribution free), con i quali è **possibile stabilire se una serie di dati, raccolti in natura od in laboratorio, è in accordo con una specifica ipotesi sulla loro distribuzione o sulla loro frequenza relativa**.

Il test $\chi^2_{(\text{g. d. l.})}$ serve anche per in **confronto tra 2 o più distribuzioni osservate**; in queste condizioni, il suo uso più frequente è per la **verifica dell'associazione tra le varie modalità di due o più caratteri qualitativi**. Risulta particolarmente utile nella fase iniziale dell'analisi, quando si ricercano le variabili più significative e le relazioni di associazione tra esse.

Per l'applicazione di questo tipo di inferenza, le distribuzioni di frequenze osservate delle classi fenotipiche e quelle attese secondo le leggi di Mendel forniscono un esempio classico.

Tabella 1 A. Distribuzioni fenotipiche (osservate ed attese) di *pisum sativum* in alcuni esperimenti di Mendel per un carattere.

A Segregazione di un ibrido:			
Carattere	dominante	recessivo	Totale
a) colore del fiore	*rossi 705*	*bianchi 224*	*929*
distribuzione attesa (3:1)	696,75	232,25	929
b) lunghezza del fusto	*alte 787*	*basse 277*	*1064*
distribuzione attesa (3:1)	798	266	1064
c) colore del seme	*gialli 6022*	*verdi 2001*	*8023*
distribuzione attesa (3:1)	6017,25	2005,75	8023
d) forma del seme	*lisci 5474*	*rugosi 1850*	*7324*
distribuzione attesa (3:1)	5493	1831	7324

Nella loro analisi, si pone il problema di verificare se la distribuzione della progenie degli ibridi rispetta la distribuzione teorica attesa di 3 a 1 per un solo carattere, oppure quella di 9:3:3:1 quando si seguono due caratteri.

E' evidente che tra distribuzioni osservate e distribuzioni attese non si ha mai una perfetta coincidenza, anche quando si possono constatare valori molto simili. In tutti i casi in cui si fanno prove ripetute per verificare una legge di distribuzione, è quasi impossibile ottenere esattamente i medesimi risultati sperimentali. Tra l'altro, mentre ogni classe di una distribuzione osservata è un conteggio ed è sempre formata da numeri interi, una distribuzione attesa segue una legge teorica di ridistribuzione dell'ammontare totale ed è spesso formata da classi con numeri frazionali. E' ovvio che **piccole differenze, accidentali, non sono tali da negare un sostanziale accordo tra osservato ed atteso, mentre grandi differenze lasciano supporre la presenza di fattori differenti da quelli ipotizzati**.

Il problema statistico è di poter **dedurre scientificamente ed in modo universalmente accettato se le differenze sono trascurabili** e quindi probabilmente dovute solo al caso (**ipotesi nulla**, indicata con $\mathbf{H_0}$); oppure se sono di dimensioni tali da fare più **ragionevolmente supporre una distribuzione realmente diversa da quella attesa** (**ipotesi alternativa**, indicata con $\mathbf{H_1}$), anche se le cause sono ignote.

L'interesse consiste nel **trarre conclusioni generali dal singolo esperimento**; in altri termini, nel **conoscere la probabilità con cui le differenze tra una distribuzione osservata e quella attesa possono riprodursi per caso, in una serie di esperimenti analoghi**.

Tabella 1 B. Distribuzioni fenotipiche (osservate ed attese) di *pisum sativum* in un esperimento di Mendel per due caratteri.

B Segregazione di un diibrido per colore e forma del seme:		
	Osservati	Attesi
gialli-lisci	***315***	9/16 = 312,75
gialli-rugosi	***101***	3/16 = 104,25
verdi-lisci	***108***	3/16 = 104,25
verdi-rugosi	***32***	1/16 = 34,75
Totale	*556*	556,00

Per affrontare questo problema di inferenza statistica, si ricorre al test $\chi^2_{(\text{g. d. l.})}$ (chi-quadrato), proposto da Pearson nel 1900, che utilizza non le frequenze relative o percentuali ma le **frequenze assolute**, con la formula

$$\chi^2_{(g.d.l.)} = \sum_{i=1}^{n} \frac{\left(f_i^{oss} - f_i^{att}\right)^2}{f_i^{att}}$$

dove:

f_i^{oss} = frequenza osservata i-esima

f_i^{att} = frequenza attesa i-esima

g.d.l. = numero di gruppi meno uno

e la sommatoria Σ è estesa a tutti i gruppi o a tutte le classi a confronto.

La distribuzione della densità di probabilità del $\chi^2_{(\text{g. d. l.})}$ dipende dai suoi gradi di libertà o g.d.l. (in inglese, degrees of freedom o d.f.). **Conteggiati nel calcolo delle frequenze attese, per definizione i gradi di libertà sono il numero di classi che restano indipendenti, conoscendo il numero totale dei dati**. Nell'esempio delle classi fenotipiche sono **n-1**, dove **n** è il numero di gruppi o classi che si analizzano.

Il numero di g.d.l. viene riportato tra parentesi, ai piedi del simbolo: corrisponde al numero di osservazioni indipendenti. Infatti **i valori attesi di ogni gruppo, che sono calcolati dal totale ed attribuiti ad ogni gruppo secondo la legge di distribuzione, sono liberi di assumere qualsiasi valore; ma fa eccezione il valore atteso dell'ultimo gruppo**, la cui frequenza è totalmente determinata dalla differenza tra la somma di tutti i gruppi precedenti, già definiti, ed il totale.

Negli esempi fino ad ora presentati, il numero di gradi di libertà corrisponde al numero di gruppi meno uno. Ma **quando tra n variabili casuali sussistono k vincoli lineari, cioè relazioni che riducono il numero di osservazioni indipendenti, i gradi di libertà del corrispondente χ^2 diminuiscono di un numero pari a k**.

Il numero dei gradi di libertà è determinato dai vincoli, di qualsiasi natura, che esistono fra le frequenze dei vari gruppi. Per esempio, in genetica delle popolazioni le frequenze attese fenotipiche dei gruppi sanguigni A, B, AB e O sono calcolate dalle frequenze relative p, q, ed r (il cui totale è sempre uguale a 1) dei geni I^A, I^B ed i, mediante lo sviluppo di

$$\left(p + q + r\right)^2 = 1;$$

pertanto, i 4 gruppi fenotipici attesi, calcolati da 3 frequenze geniche, hanno 2 gradi di libertà.

Per la stessa legge, anche i 6 gruppi genotipici (I^AI^A, I^Ai, I^BI^B, I^Bi, I^AI^B, ii) hanno 2 gdl.

Secondo uno schema valido per tutti i test statistici, il procedimento logico che deve essere seguito nell'applicazione del χ^2 comprende diverse fasi, che possono essere riassunte in 7 passaggi:

1 - stabilire l'ipotesi nulla (H_0) e l'eventuale ipotesi alternativa (H_1);

2 - scegliere il test più appropriato per saggiare l'ipotesi nulla H_0, secondo le finalità della ricerca e le caratteristiche statistiche dei dati (in questo caso, ovviamente, è il test chi quadrato);
3 - specificare il livello di significatività (i cui criteri saranno discussi nel capitolo 4), l'ampiezza del campione e i gradi di libertà;
4 - trovare la distribuzione di campionamento del test statistico nell'ipotesi nulla H_0, di norma fornita da tabelle;
5 - stabilire la zona di rifiuto (che negli esercizi di norma sarà prefissata al 5%);
6 - calcolare il valore del test statistico sulla base dei dati sperimentali, stimando il valore di probabilità ad esso associato;
7 - sulla base della probabilità, trarre le conclusioni: se la probabilità risulta superiore a quella prefissata, concludere che non è possibile rifiutare l'ipotesi nulla H_0; se la probabilità risulta inferiore a quella prefissata, rifiutare l'ipotesi nulla e quindi implicitamente accettare l'ipotesi alternativa H_1.

ESEMPIO 1. Utilizzando i dati sulla segregazione mendeliana della precedente tabella 1B, il calcolo del χ^2 è semplice:

$$\chi^2_{(3)} = \frac{(315-312{,}75)^2}{312{,}75} + \frac{(101-104{,}25)^2}{104{,}25} + \frac{(108-104{,}25)^2}{104{,}25} + \frac{(32-34{,}75)^2}{34{,}75}$$

$$\chi^2_{(3)} = \frac{(2{,}25)^2}{312{,}75} + \frac{(-3{,}25)^2}{104{,}25} + \frac{(3{,}75)^2}{104{,}25} + \frac{(-2{,}75)^2}{34{,}75} = 0{,}47$$

Con l'aiuto delle tavole, è possibile stimare con precisione la probabilità di trovare differenze uguali o superiori a quelle riscontrate tra distribuzione osservata e distribuzione attesa, nell'ipotesi nulla (H_0) che le differenze siano dovute esclusivamente a fattori casuali.
Nella tavola a 2 entrate della distribuzione dei valori critici del χ^2 per 3 gradi di libertà (indicato sulla riga) e per probabilità 0.05 (indicato sulla colonna), il valore del χ^2 approssimato alla seconda cifra decimale risulta uguale a 7,81. Il valore calcolato nell'esercizio è sensibilmente minore ($\chi^2_{(3)}$ = 0,47) di quello tabulato. La probabilità che le differenze siano imputabili solo al caso è alta, superiore al valore prefissato del 5%; di conseguenza, non si può rifiutare l'ipotesi nulla, secondo la quale le differenze riscontrate tra distribuzione osservata e distribuzione attesa sono dovute esclusivamente a fattori casuali. Si afferma che le differenze tra distribuzione osservata e distribuzione attesa non sono significative.

Per la comprensione dell'inferenza statistica con il test chi quadrato, è utile ricordare che quanto più le differenze tra osservato ed atteso sono grandi, tanto più il valore del χ^2 sarà elevato; di conseguenza, la probabilità che tali differenze siano dovute solo al caso sarà bassa e si rifiuterà l'ipotesi nulla, accettando implicitamente l'ipotesi alternativa H_1. Al contrario,

quando le differenze tra osservato ed atteso sono ridotte, ugualmente basso sarà il valore del χ^2; pertanto, sarà elevata la probabilità che esse siano imputabili esclusivamente al caso e si accetterà l'ipotesi nulla H_0.

ESEMPIO 2 . In una popolazione di *Mixodiaptomus Kupelwieseri* (Copepode, Calanoide) di pozza temporanea (Lagastro - Val d'Aveto) sono state osservate le seguenti frequenze di 4 alleli del locus MPI (Mannoso fosfato isomerasi)

tipo di allele	frequenza osservata
allele 1	**26**
allele 2	**38**
allele 3	**62**
allele 4	**118**
Totale	244

Le differenze riscontrate fra le frequenze dei vari alleli possono essere imputate al caso (H_0) oppure è possibile pensare ragionevolmente che esistano uno o più fattori che li rendono effettivamente differenti (H_1)?

Risposta.
Se fosse vera l'ipotesi nulla espressa (equidistribuzione delle frequenze), la frequenza attesa per ogni allele sarebbe 244/4 = 61. Il valore del chi quadrato con 3 gradi di libertà per saggiare tale ipotesi risulta uguale a 82,03:

$$\chi^2_{(3)} = \frac{(26-61)^2}{61} + \frac{(38-61)^2}{61} + \frac{(62-61)^2}{61} + \frac{(118-61)^2}{61} = \frac{1225}{61} + \frac{529}{61} + \frac{1}{61} + \frac{3249}{61} = 82{,}03$$

Consultando la tabella del chi-quadrato per 3 gradi di libertà, alla probabilità 0.05 corrisponde un valore di 7,82 mentre alla probabilità 0.01 corrisponde un valore critico di 11,34 e alla probabilità 0.001 un valore critico di 16,27. Il valore del chi quadrato calcolato sui dati sperimentali è molto più grande. La probabilità che le differenze (tra le frequenza riscontrate e quella attesa secondo l'ipotesi nulla) siano imputabili esclusivamente al caso è molto piccola, inferiore non solo al 5% ma addirittura al 0,1%; di conseguenza, si rifiuta l'ipotesi nulla e si accetta l'ipotesi alternativa.
Con probabilità inferiore a 0,1% di commettere un errore, si può sostenere che i 4 alleli hanno frequenze tra loro molto differenti.