LE ANOMALIE DI FREQUENZA NEL SECONDO ESPERIMENTO DI MENDEL

R. Prisco, G. Caramia

1. INTRODUZIONE

I risultati ottenuti da Mendel hanno incuriosito numerosi studiosi, con successive rivisitazioni dei dati. L'alternarsi delle interpretazioni ha suscitato anche il nostro interesse[1]. Per questo, non certo per giudicarne le conclusioni, proponiamo la tecnica delle anomalie con la quale si può cercare di evidenziare aspetti non immediatamente percepibili delle tabelle di frequenza.

2. DEFINIZIONI

Un insieme di n oggetti, sottoposti ad una misurazione congiunta relativa a due caratteri, R e C, dà luogo alla tabella a doppia entrata RxC delle frequenze congiunte. Nella casella ij, ottenuta come intersezione delle modalità marginali R_i e C_j dei caratteri marginali *Riga* e *Colonna* divisi rispettivamente in n_r e n_c modalità, viene scritta la sua frequenza n_{ij}. Inoltre n_i ed n_j sono le frequenze delle modalità marginali. L'estensione al caso delle tabelle a tre dimensioni può essere fatta per analogia.

Le tabelle di frequenza sono usate per verificare se una certa ipotesi probabilistica[2], fatta sulle distribuzioni marginali e sulla congiunta, è statisticamente compatibile con i dati ottenuti dalla misurazione di fatti reali.

La discrepanza di una distribuzione di frequenze osservate, da una distribuzione di frequenze costruita sulla base dell'ipotesi, viene calcolata con diversi indici; alcuni di questi si possono ottenere dalla formula di Cressie e Read (1988) e si distribuiscono approssimativamente come una distribuzione di χ^2. In particolare in questo lavoro ricorriamo all'indice G^2.

---

[1] L'articolo è il risultato del lavoro comune dei due autori, che quindi ne condividono interamente la responsabilità; Roberto Prisco ha comunque scritto i paragrafi 3, 5 e 7; Giovanna Caramia i paragrafi 1, 2, 4, 6 e 8.

[2] Mendel nel suo lavoro avanzò l'ipotesi che i caratteri fossero indipendenti e che le distribuzioni marginali avessero entrambe la struttura probabilistica 1/4; 1/2; 1/4.

Chiamiamo G^2_T la formula per il calcolo della discrepanza relativa ad una tabella a doppia entrata con caratteri indipendenti

$$G^2_T = 2n\sum_{ij} n_{ij} \ln \frac{n_{ij}}{n\rho_i\gamma_j}$$

dove ρ_i e γ_j sono le probabilità marginali ipotizzate.

Questo indice si può ottenere ponendo $\lambda = 1$ nella formula di Cressie e Read e si distribuisce come un χ^2 con $n_r\, n_c$ - 1 gradi di libertà.

Una proprietà interessante è la scomponibilità per addizione (Agresti, 1984 p. 14). Si ottiene infatti:

$$G^2_T = 2n\sum_i n_i \ln \frac{n_i}{n\rho_i} + 2n\sum_j n_j \ln \frac{n_j}{n\gamma_j} + 2n\sum_{ij} n_{ij} \ln \frac{n_{ij}n}{n_i n_j}$$

Cioè

$$G^2_T = G^2_R + G^2_C + G^2_I \qquad (1)$$

Tutto questo permette di costruire quattro test di bontà di accostamento che verificano se le frequenze osservate possono provenire da una popolazione la cui distribuzione di probabilità:

– è come quella complessivamente posta per ipotesi (H_T)

– per il carattere riga è come quella posta per ipotesi alle righe della tabella (H_R)

– per il carattere colonna è come quella posta per ipotesi alle colonne della tabella (H_C)

– nella quale i caratteri Riga e Colonna sono indipendenti (H_I)

L'ipotesi H_T è vera quando sono vere tutte e tre le ipotesi componenti; in altri termini

$$H_T = H_R \cap H_C \cap H_I \qquad (2)$$

Per quanto riguarda i gradi di libertà, si nota facilmente che $\nu_T = n_r\, n_c$ - 1 e poi $\nu_R = n_r$ - 1, $\nu_C = n_c$ - 1, $\nu_I = (n_r$ - 1)(n_c - 1) e quindi

$$\nu_T = \nu_R + \nu_C + \nu_I$$

Una forma di sostegno dell'ipotesi H_k (dove k vale T, C, R oppure I) può essere data dal *valore-p* (Azzalini, 1992)

$$\Pr(\chi^2_{\nu_{H_k}} > G^2_{H_k}) = S(H_k \mid X)$$

che vale 1 se il campione rispecchia esattamente la distribuzione posta in H_k ed è tanto più piccolo quanto più il campione è difforme secondo l'indice adottato.

In particolare il test prevede che l'ipotesi venga accettata se

$$\Pr(\chi^2_{\nu_{H_k}} > G^2_{H_k}) > \alpha$$

3. ANOMALIE DEBOLI E FORTI

Le definizioni precedenti portano ad ordinare le ipotesi dalle meno alle più sostenute sulla base dei riscontri campionari (Agresti, 1984 p. 52). Il fatto che l'ipotesi H_T sia costituita dall'intersezione delle altre tre farebbe ritenere ad un primo esame che debba essere delle quattro ipotesi quella che, dato un certo campione di osservazioni, gode di un livello più basso di sostegno. Ove poi si faccia riferimento all'additività di G^2 ci si trova condotti a ritenere che sia raro il verificarsi di situazioni diverse da questa. Parrebbe di poter definire quindi come anomala la situazione in cui si trova un campione per il quale

$$S(H_T \mid X) > S(H_k \mid X)$$

con k si intende una delle ipotesi (R, C, I) componenti di T. Questa viene definita come anomalia debole.

Definiamo poi anomalia forte a destra la situazione in cui si trova un campione per il quale

$$\left[S(H_T \mid X) > \alpha\right] \bigcap \left[S(H_k \mid X) \leq \alpha\right].$$

Un campione cioè per il quale risulta significativa una delle ipotesi componenti (H_k) e non risulta significativa l'ipotesi composta (H_T).

È poi possibile definire come anomalia forte a sinistra la situazione per cui

$$\left[S(H_T \mid X) > 1-\alpha\right] \bigcap \left[S(H_k \mid X) \leq 1-\alpha\right].$$

In questo caso l'ipotesi totale si adatta "troppo bene" ai dati, mentre le ipotesi componenti hanno un adattamento più "regolare".

Le anomalie, avendo riguardo alle tre ipotesi componenti, sono suddivise in tipi diversi secondo la struttura descritta nella tabella 1.

TABELLA 1

Tipi diversi di anomalie

	H_I anomala		H_I non anomala	
	H_R anom.	H_R non anom.	H_R anom.	H_R non anom.
H_C anomala	Tipo A	Tipo B	Tipo C	Tipo D
H_C non anom.	Tipo E	Tipo F	Tipo G	Tipo H

Si potrebbe definire anomalia fortissima quella per cui

$$\left[S(H_T \mid X) > 1-\alpha\right] \bigcap \left[S(H_k \mid X) \leq \alpha\right]$$

si deve comunque notare che in tutto il lavoro di cui si presentano qui i risultati non si sono mai verificati casi di anomalia fortissima.

4. IL "PROBLEMA MENDEL"

"Ogni generazione, forse, ha trovato nel *paper* di Mendel solo quello che si aspettava di trovare... Ogni generazione, comunque, ha ignorato quello che non confermava le sue proprie aspettative." (Fisher, 1936). Molteplici e contraddittorie furono infatti le interpretazioni degli esperimenti di Mendel. Evoluzionista, non evoluzionista; "non-Darwiniano", buon "Darwiniano"; per alcuni i dati furono almeno in parte falsificati, per altri invece nessun dato fu falsificato, per altri ancora tutti i dati erano fittizi. Queste ed altre ancora le reazioni (Sapp, 1990). Al di là di tutte le possibili posizioni emerse nel corso di questa singolare disputa, il punto cruciale non è tanto il rapporto probabilistico 1/4: 1/2; 1/4: sulla cui validità c'è consenso, quanto la possibilità di ricostruire il percorso sperimentale che ha portato a quei risultati.

Al termine di alcuni anni di esperimenti, Mendel giunse alla formulazione della legge dell'indipendenza: *i caratteri si trasmettono, da una generazione alla successiva, indipendentemente gli uni dagli altri con frequenze proporzionali a 1:2:1 nel senso di dominante: ibridi: recessivo*. Sorprendentemente, le frequenze ottenute dalla sperimentazione di Mendel si adattano troppo bene ai valori attesi e ipotizzati da tale proporzione.

A partire dall'ormai classico lavoro di Fisher (1936), i dati comunicati da Mendel nella memoria presentata all'Accademia di Brno sono stati ripresi più volte ed esaminati con cura da statistici e genetisti, che con alterni risultati hanno rivisitato quei dati sottoponendoli all'esame più attento, ricorrendo a tecniche ed ipotesi differenti ed ottenendo i risultati più difformi[3].

Essendo ben d'accordo con Cox (1997 p. 272) non intendiamo giudicare le conclusioni ottenute da Mendel, confermate peraltro da ripetute sperimentazioni (Olby, 1965 p. 183) ma fornire un esempio di come può essere usata la tecnica qui proposta. Rielaborare dati rilevati da altri studiosi, infatti, (usando le sue stesse parole) "può gettare luce sui metodi piuttosto che sul problema empirico che aveva generato i dati".

5. LA SIMULAZIONE

Nell'ambito del primo esperimento, Mendel ha dimostrato l'attendibilità della sua ipotesi per quanto riguarda due caratteri; nel secondo ha preso in esame la sua attendibilità per tre caratteri considerati congiuntamente[4]. La tabella 2, composta di 27 caselle, mostra i risultati dell'esperimento (Mendel, 1865; Fisher, 1936). L'altra tabella (la n. 3) riporta invece le probabilità desunte dall'ipotesi fatta da Mendel.

---

[3] Non ha contato tutte le piante ma si è fermato quando ha raggiunto il risultato che intendeva perseguire. Non c'è stata quindi nessuna alterazione, i dati sono semplicemente incompleti (Olby, 1965). La classificazione è stata difficile, le generazioni venivano ottenute da linee pure o da ibridi? (Piegorsch, 1983) Un giardiniere assistente di Mendel ha falsificato i dati (Fisher, 1936)....

[4] Il carattere A è la forma, B il colore e C la membrana.

TABELLA 2

Risultati del secondo esperimento

	CC			Cc			cc		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
BB	8	14	8	22	38	25	14	18	10
Bb	15	49	19	45	78	36	18	48	24
bb	9	20	10	17	40	20	11	16	7

TABELLA 3

Probabilità delle celle

	CC			Cc			cc		
	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
BB	0.0156	0.0313	0.0156	0.0313	0.0625	0.0313	0.0156	0.0313	0.0156
Bb	0.0313	0.0625	0.0313	0.0625	0.1250	0.0625	0.0313	0.0625	0.0313
bb	0.0156	0.0313	0.0156	0.0313	0.0625	0.0313	0.0156	0.0313	0.0156

Trattare direttamente la tabella a tre caratteri porta allo studio di un numero notevole di anomalie dato che le ipotesi possibili sono molte. I modelli di dipendenza sono (Agresti, 1984), qualora ci si limiti ai soli modelli gerarchici, ben nove (al posto dei due possibili per le tabelle doppie: dipendenza ed indipendenza); infatti bisogna considerare il modello di indipendenza, i tre con una interazione a due, i tre con due interazioni a due, quello con tre interazioni a due e quello con l'interazione a tre. Questi vanno combinati con i modelli sulle distribuzioni marginali. Ci si può immaginare quale complessità presenti poi la definizione delle anomalie in una struttura di ipotesi così elaborata. La nostra scelta è stata allora di trattare separatamente le tabelle condizionate alle tre modalità del carattere A, le tre condizionate alle modalità del B e le tre condizionate alle modalità del C. Queste tabelle condizionate ai nove genotipi sono state sottoposte all'elaborazione dell'indice G^2.

Le ipotesi sottoposte a test sono elencate nella tabella 4 che contiene nella penultima colonna l'indicazione delle anomalie riscontrate. Queste anomalie sono riassunte nella tabella 5; tenendo poi conto che la stessa variabile talvolta compare come riga e talvolta come colonna le anomalie possono essere riunite nella tabella 8.

Da diverse simulazioni pilota di 60000 campioni è emerso che le distribuzioni marginali delle anomalie sono dipendenti e quindi la distribuzione congiunta non può essere rappresentata da una multinomiale, dalla quale poi sia possibile passare alle statistiche di difformità (Cressie and Read, 1988). Una possibilità di soluzione del problema consiste nell'ottenere da una congrua simulazione la distribuzione empirica dei vettori di anomalie. Questa verrà utilizzata per valutare i vettori di anomalie (tabella 8) risultati dalle tabelle di frequenza ottenute dalla misurazione (tabella 2).

La procedura di simulazione deve essere almeno altrettanto precisa dello strumento probabilistico con cui la si intende confrontare; in particolare deve essere in grado di distinguere (con probabilità pari ad almeno 0.95) tra i diversi percentili di G^2 ottenuti nelle tabelle della sperimentazione di Mendel.

TABELLA 4

Ipotesi provate sulle tabelle 3 x 3

					G^2 Anomalie	
Genotipo	Ipotesi	Valore	v	valore-p	Tipo	valore-p
CC	A	1.653940	2	0.437373		
CC	B	2.468568	2	0.291043		
CC	indip.	1.755268	4	0.780655		
CC	totale	5.877775	8	0.660921	C debole	0.61518
Cc	A	0.306918	2	0.857736		
Cc	B	0.423260	2	0.809264		
Cc	indip.	2.258530	4	0.688329		
Cc	totale	2.988708	8	0.935064	A debole	0.37249
cc	A	0.071721	2	0.964775		
cc	B	2.025794	2	0.363165		
cc	indip.	3.732597	4	0.443403		
cc	totale	5.830112	8	0.666255	B debole	0.39158
non cond.	A	0.014085	2	0.992982		
non cond.	B	1.137963	2	0.566102		
non cond.	indip.	2.851208	4	0.583026		
non cond.	totale	4.003255	8	0.856830		
AA	C	2.128874	2	0.344922		
AA	B	0.662301	2	0.718097		
AA	indip.	2.098859	4	0.717582		
AA	totale	4.890034	8	0.769262	A debole	0.36115
Aa	C	0.258430	2	0.878785		
Aa	B	2.870158	2	0.238097		
Aa	indip.	3.491023	4	0.479245		
Aa	totale	6.619611	8	0.578173	B debole	0.62598
aa	C	0.261825	2	0.877294		
aa	B	0.456712	2	0.795841		
aa	indip.	3.084963	4	0.543709		
aa	totale	3.803501	8	0.874403	B debole	0.26869
non cond.	C	0.630636	2	0.729557		
non cond.	B	1.137963	2	0.566102		
non cond.	indip.	3.779658	4	0.436646		
non cond.	totale	5.548258	8	0.697688		
BB	A	1.855873	2	0.395369		
BB	C	3.087030	2	0.213629		
BB	indip.	0.953779	4	0.916722		
BB	totale	5.896682	8	0.658804	C debole	0.61518
Bb	A	0.982752	2	0.611784		
Bb	C	0.873851	2	0.646020		
Bb	indip.	4.533778	4	0.338561		
Bb	totale	6.390381	8	0.603591	F debole	0.11443
bb	A	0.026667	2	0.986755		
bb	C	0.449413	2	0.798750		
bb	indip.	1.426124	4	0.839642		
bb	totale	1.902205	8	0.983865	B forte a sin.	0.07200
non cond.	A	0.014085	2	0.992982		
non cond.	C	0.630636	2	0.729557		
non cond.	indip.	2.018494	4	0.732357		
non cond.	totale	2.663214	8	0.953685		

TABELLA 5

Anomalie deboli riscontrate nelle 9 tabelle condizionate

	H_I anomala		H_I non anomala	
	H_R anom.	H_R non anom.	H_R anom.	H_R non anom.
H_C anomala	2	4	2	0
H_C non anom.	0	1	0	0

TABELLA 6

Anomalie forti a sinistra riscontrate nelle 9 tabelle condizionate

	H_I anomala		H_I non anomala	
	H_R anom.	H_R non anom.	H_R anom.	H_R non anom.
H_C anomala	0	1	0	0
H_C non anom.	0	0	0	8

TABELLA 7

Anomalie forti a destra riscontrate nelle 9 tabelle condizionate

	H_I anomala		H_I non anomala	
	H_R anom.	H_R non anom.	H_R anom.	H_R non anom.
H_C anomala	0	0	0	0
H_C non anom.	0	0	0	9

TABELLA 8

Frequenza delle anomalie e valore-p

Descrizione	Tipi	deboli	forti Sx	forti Dx
Tutte	A	2	0	0
1 Marg + Dip	B∪E	4	1	0
Dipendenza	F	1	0	0
2 Marginali	C	2	0	0
1 Marginale	D∪G	0	0	0
Nessuna	H	0	8	9
valore-p stimato		0.52033	0.10367	0.54431

Nella tabella 9 la coppia di *valori-p* che è a maggior rischio di valutazione erronea è la coppia 0.603591 e 0.611784 della tabella condizionata al genotipo Bb. La numerosità della simulazione dovrà essere quindi di almeno tanti elementi che permettano all'estremo inferiore dell'intervallo di stima del quantile 0.611784 di essere maggiore dell'estremo superiore dell'analogo intervallo di 0.603591. Per una maggiore cautela si è arrotondato rispettivamente a 0.604 e 0.611. La numerosità che permette di soddisfare questa condizione (Rohatgi, 1984 p. 616 e 617) è di 74772 unità.

Con un programma Matlab (versione 5.3) abbiamo steso una procedura di calcolo che ha simulato 100000 tabelle 3x3x3, prodotte con le probabilità teoriche ipotizzate da Mendel. Delle tabelle abbiamo contato quante sono state le anomalie deboli e forti di ciascun tipo ottenendone la distribuzione per ciascuna numerosità delle tabelle osservate (vedi tabelle 9, 10 e 11).

TABELLA 9

Anomalie deboli riscontrate su 100000 tabelle

	H_I anomala		H_I non anomala	
	H_R anom.	H_R non anom.	H_R anom.	H_R non anom.
H_C anomala	133682	106794	132607	111649
H_C non anom.	106039	168522	111033	29674

TABELLA 10

Anomalie forti a destra riscontrate su 100000 tabelle

	H_I anomala		H_I non anomala	
	H_R anom.	H_R non anom.	H_R anom.	H_R non anom.
H_C anomala	0	0	211	31206
H_C non anom.	0	25898	30899	811786

TABELLA 11

Anomalie forti a sinistra riscontrate su 100000 tabelle

	H_I anomala		H_I non anomala	
	H_R anom.	H_R non anom.	H_R anom.	H_R non anom.
H_C anomala	20609	4226	11136	1314
H_C non anom.	4196	633	1282	856604

La bontà della simulazione è stata testata in diversi modi. Una prima verifica è stata posta sul generatore di numeri casuali che ha superato il test di lunghezza mostrando che dopo 100 milioni di numeri non aveva ancora ripetuto il numero iniziale. La procedura è stata testata contando il numero di tabelle per le quali è risultato significativo il test di difformità per ciascuna delle quattro ipotesi (vedi tabella 12).

TABELLA 12

Tabelle risultate significative al test G^2

Num.	H_R	H_C	H_I	H_T
Destra	45184	45589	48866	48079
Sinistra	42495	42226	44132	43509

Un'altra verifica è stata fatta prendendo alcuni valori del test G^2 e calcolandone il percentile empirico, il suo confronto con la cumulata di probabilità calcolata con la distribuzione χ^2 è stato positivo (vedi tabella 13).

TABELLA 13

Probabilità teoriche ed empiriche di G^2

	H_R	H_C	H_I	H_T
G2	0.0147	1.1431	2.8639	4.0205
ν	2	2	4	8
Cum. Teo.	0.00733	0.43536	0.41914	0.14473
Cum. Emp.	0.00734	0.43318	0.41013	0.14042

Un'altra possibilità di errore può annidarsi nel fatto che per numerosità basse la tabella simulata ospiti caselle con frequenza uguale a zero; in questo caso G^2 si trova a dover calcolare il logaritmo di zero con conseguenze non sempre controllabili. La scelta a questo riguardo è stata di porre la frequenza di quella casella uguale a 0.01. La sostituzione ha interessato soltanto 35 delle 100000 tabelle.

Si può, quindi, concludere che la simulazione sia stata condotta rispettando quelle condizioni di casualità che sono richieste per la sua validità.

6. APPLICAZIONE AI DATI DI MENDEL

Il quesito a cui abbiamo cercato risposta è il seguente "La distribuzione di anomalie riscontrata nelle nove tabelle di Mendel è compatibile con la distribuzione riscontrata nelle 100000 tabelle generate dalla simulazione?"[5]

La distribuzione di frequenza delle anomalie viene usata per stimare quella delle probabilità allo scopo di sottoporre a test il profilo di anomalie deboli, forti a destra e forti a sinistra. Applichiamo il test interpretandolo come una FRPS (*Falsifying*[6] *Rule for Probability Statements*) (Gillies, 1973 p. 171). Il *valore-p* viene calcolato dopo aver ordinato gli elementi dello spazio campionario in funzione della loro probabilità, dal più al meno probabile. Quelli con probabilità più elevata appartengono alla zona di accettazione, mentre quelli con probabilità più piccola costituiscono la zona di rigetto. Una volta individuato il campione in questo elenco ordinato, la somma delle probabilità dei campioni, che di questo sono meno probabili, costituisce appunto il *valore-p*. Nel nostro caso la distribuzione di probabilità è di difficile elaborazione, le anomalie, infatti, sono dipendenti tra loro e dal valore di G^2_T, è stato quindi necessario stimarla per mezzo della simulazione.

Il *valore-p* per le anomalie deboli è (tabella 8) 0.52033 e fa accettare l'ipotesi che il campione possa essere stato generato casualmente. La stessa procedura è stata applicata alle anomalie forti sia a destra sia a sinistra, in questo caso la probabilità vale rispettivamente 0.10367 e 0.54431 e porta ancora ad accettare l'ipotesi nulla che la presenza di anomalie forti riscontrate sui dati di Mendel possa essere dovuta al caso.

7. ANOMALIE CONDIZIONATE

Il valore di G^2_T influenza fortemente la presenza di anomalie, che sono infatti molto più frequenti per valori bassi di questo indice. La figura 1 rappresenta le combinazioni di G^2_R e G^2_I che determinano la presenza di anomalie[7].

---

[5] Da una serie di simulazioni fatte per campioni di numerosità comprese tra 150 e 300 si è notato che la frequenza di anomalie non varia in funzione della numerosità.

[6] Il termine falsificazione va inteso nel senso di Popper.

[7] La rappresentazione completa, comprendente anche G^2_C , richiede il ricorso alla terza dimensione e non è eseguibile graficamente.

TABELLA 14

Anomalie forti a sinistra condizionate a G^2_T

G^2_T	A	B∪E	F	C	D∪G	H	Totale
0.0	0	0	0	12	56	44	112
0.5	7	55	47	603	644	69	1425
1.0	895	1027	156	2181	731	0	4990
1.5	4057	2236	175	3154	558	0	10180
2.0	9400	3307	172	3469	449	0	16797
2.5	6250	1797	83	1717	158	13655	23660
3.0	0	0	0	0	0	30601	30601
≥ 3.5	0	0	0	0	0	812235	812235
Totale	20609	8422	633	11136	2596	856604	900000

Nella Figura 1, G^2_{*2} rappresenta il valore di G^2_R corrispondente nella distribuzione di $\chi^2_{\nu=2}$ al valore della funzione di ripartizione in G^2_T con $\nu = 8$, analogamente per G^2_{*4}. Se ad esempio abbiamo $G^2_T = 5.24$ con $\Pr(\chi^2 \leq 5.24) = 0.7316$ i valori corrispondenti degli altri due indici sono $G^2_{*2} = 0.62$ e $G^2_{*4} = 2.02$. Il trapezio in basso individua l'insieme delle coppie di valori dei due indici parziali per le quali si verifica una anomalia relativa alla distribuzione delle righe (anomalia di tipo G). Per una anomalia doppia (tipo E oppure A) l'insieme è dato dal triangolo; il quadrato individua infine le coppie di valori che non presentano anomalie (tipo D oppure H). Le probabilità delle anomalie potrebbero essere calcolate conoscendo la distribuzione congiunta di G^2_R, G^2_I e G^2_C condizionata a G^2_T.

Non disponendo di questa distribuzione condizionata è necessario seguire la procedura approssimata della simulazione.

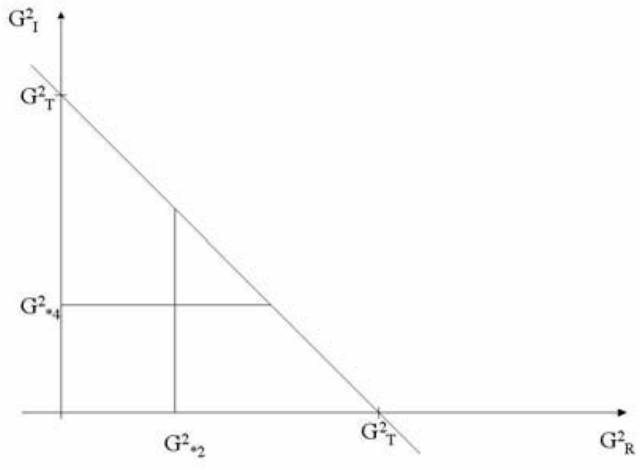

Figura 1 – Domini di G^2_k.

La simulazione opera attraverso la rilevazione delle anomalie classificate in funzione del valore di G^2_T. I suoi valori sono stati divisi in classi, per rendere possibile la rilevazione delle distribuzioni di frequenza. Dato l'elevato numero di campioni la scelta è potuta cadere su una divisione abbastanza fine, che è stata operata in

TABELLA 15

Anomalie deboli condizionate a G^2_T

G^2_T	A	B ∪ E	F	C	D ∪ G	H	Totale
0.0	110	2	0	0	0	0	112
0.5	1349	23	2	50	1	0	1425
1.0	4520	183	2	282	3	0	4990
1.5	8358	805	18	969	30	0	10180
2.0	12383	2179	63	2034	138	0	16797
2.5	14847	4464	190	3793	366	0	23660
3.0	16380	7463	538	5432	788	0	30601
3.5	16316	10722	965	7400	1569	0	36972
4.0	14871	14044	1717	8822	2534	0	41988
4.5	12946	16600	2676	10245	3782	0	46249
5.0	10386	18214	3703	10810	5269	0	48382
5.5	7950	19073	4898	11083	6560	0	49564
6.0	5698	19006	6007	11011	8071	0	49793
6.5	3569	18265	7186	10302	9512	0	48834
7.0	2258	16884	8071	9530	10799	0	47542
7.5	1107	14842	8871	8493	11855	0	45168
8.0	490	12844	9469	7277	12828	0	42908
8.5	141	10343	9895	5923	13424	0	39726
9.0	3	7832	10192	4756	13514	6	36303
9.5	0	5835	10041	3693	13517	78	33164
10.0	0	4408	9959	2826	13019	260	30472
10.5	0	3179	9122	2199	12564	477	27541
11.0	0	2153	8539	1577	11362	731	24362
11.5	0	1407	7685	1111	10282	988	21473
12.0	0	917	7012	865	9363	1198	19355
12.5	0	540	6168	654	8085	1367	16814
13.0	0	328	5454	462	7190	1551	14985
13.5	0	170	4758	357	6125	1626	13036
14.0	0	82	4199	201	5141	1672	11295
14.5	0	23	3439	151	4342	1646	9601
15.0	0	3	2987	92	3765	1682	8529
15.5	0	0	2524	82	3074	1714	7394
16.0	0	0	2100	49	2528	1597	6274
16.5	0	0	1731	38	2110	1498	5377
17.0	0	0	1459	12	1690	1409	4570
17.5	0	0	1232	7	1392	1245	3876
18.0	0	0	996	9	1149	1135	3289
18.5	0	0	835	5	901	989	2730
19.0	0	0	692	2	779	955	2428
19.5	0	0	567	2	632	854	2055
20.0	0	0	517	1	549	680	1747
≥ 20.5	0	0	2043	0	2080	4316	8439
Totale	133682	212833	168522	132607	222682	29674	900000

cento classi di ampiezza uguale a 0.5. Soltanto le prime novanta hanno comunque ospitato valori dell'indice di aderenza. Questa rilevazione ha reso possibile la stima delle seguenti probabilità di anomalie forti a sinistra (nella tabella 15 sono riportati i valori rilevanti)

$$\Pr\left(B \cup E \mid 1.5 \le G^2{}_T < 2.0\right) = \frac{2236}{10180} = 0.21965$$

$$\Pr\left(1.5 \le G^2{}_T < 2.0\right) = \frac{10180}{900000} = 0.01131$$

quest'ultima concorda con quella ottenibile dal calcolo fatto con la distribuzione di χ^2 con 8 gradi di libertà

$$\Pr\left(1.5 \le G^2{}_T < 2.0\right) = 0.011696$$

Nell'ultima colonna della Tabella 9 sono mostrati i *valori-p* stimati delle anomalie condizionate relativi alle nove tabelle. I loro valori sono stati calcolati con i dati della Tabella 15. In tutte le tabelle le anomalie riscontrate sono risultate accettabili. Le distribuzioni condizionate sono rappresentate anche nella Figura 3.

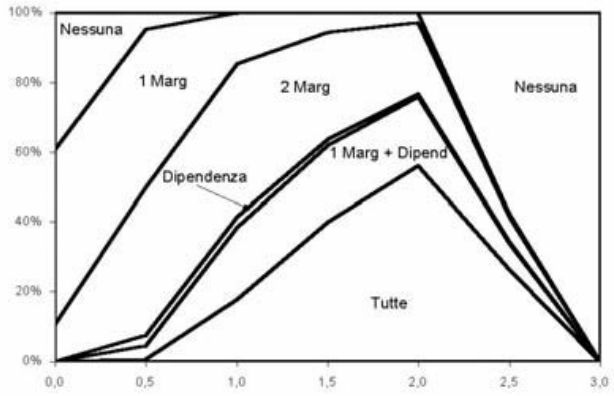

Figura 2 – Distribuzioni delle anomalie forti a sinistra condizionate ai diversi valori di $G^2{}_T$.

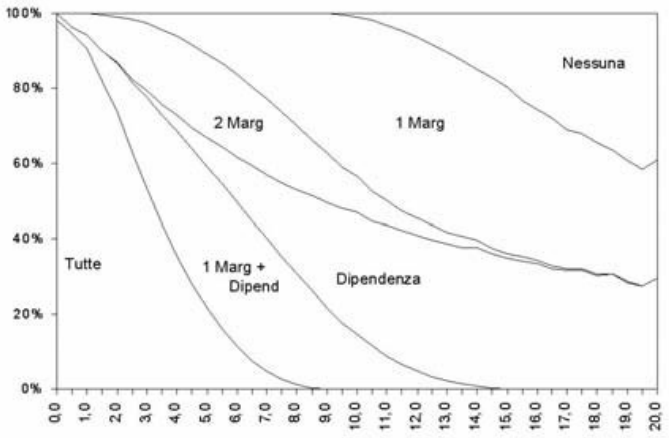

Figura 3 – Distribuzioni delle anomalie deboli condizionate ai diversi valori di $G^2{}_T$.

8. CONCLUSIONI

La presenza di anomalie, che ad un primo esame potrebbe sembrare un indizio di poca casualità, nei dati di Mendel ad un esame più attento risulta essere invece dovuta ad una dinamica interna dell'indice di aderenza. Combinando questi risultati con quelli di Fisher la conclusione da trarre è che si può ritenere come ipotesi plausibile quella di Olby, secondo il quale Mendel non avrebbe falsificato i risultati ma avrebbe troncato la registrazione dei dati avendo rilevato un buon numero di esiti favorevoli alla sua ipotesi.

Le anomalie forti risultano causate, per valori bassi di G^2_T, dalla dinamica interna dei valori componenti rispetto al valore totale; se questo assume un valore basso allora le anomalie sono quasi ineliminabili.

Dipartimento di Scienze economiche ROBERTO PRISCO
Università di Verona GIOVANNA CARAMIA

RIFERIMENTI BIBLIOGRAFICI

A. AGRESTI, (1984). *Analysis of ordinal categorical data*. John Wiley and Sons, New York.

A. AZZALINI, (1992). *Inferenza statistica*. Springer Verlag, Berlino.

D.R. COX, (1997). *The current position of Statistics: a personal view.* International Statistical Review, pp. 261-276.

N.A.C. CRESSIE, T.R.C. READ (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer, New York.

R.A. FISHER, (1936). *Has Mendel's work been rediscovered?* In Bennett, J. H., (a cura di), Collected papers of R. A. Fisher (1974), capitolo 144 (vol III), pp. 514-536. University of Adelaide.

D.A. GILLIES, (1973). *An objective theory of probability*. Methuen, London.

G. MENDEL, (1865). *Experiment in plant hybridization*. Naturforschenden Vereins – Brno (www.netspace.org/MendelWeb).

R.C. OLBY, (1965). *Origins of Mendelism*. Constable, London.

W.W. PIEGORSCH, (1983). *The questions of fit in the Gregor Mendel controversy.* Communications in Statistics, Part A – Theory and Methods, 12 pp. 2289-2304.

K.V. ROHATGI, (1984). *Statistical inference*. John Wiley and Sons, New York.

J. SAPP, (1990). *The nine lives of Gregor Mendel*. In Le Grand, H., (a cura di), Experimental Inquires, pp. 137-166. Kluwer Academic Press (www.netspace.org/MendelWeb).

RIASSUNTO

Le anomalie di frequenza nel secondo esperimento di Mendel

La situazione sperimentale da cui prende spunto questo articolo è rappresentata dal secondo esperimento eseguito da Mendel su tre caratteri congiuntamente (forma, colore, membrana). Con numerosi esperimenti, egli riuscì ad isolare piante che differivano per più di un carattere e osservò che ognuno di questi si trasmetteva indipendentemente dagli

altri seguendo il rapporto 1:2:1 (dominante: ibridi: recessivo). Tale rapporto e l'ipotesi di indipendenza, da verificare sulla base dei dati empirici, rappresentano il presupposto per il calcolo delle frequenze teoriche. Tra i numerosi indici che consentono di descrivere la differenza tra queste ultime e le frequenze osservate, abbiamo scelto G^2, che è stato calcolato per ognuna delle nove tabelle condizionate alle tre modalità di ciascun carattere. La scomponibilità per addizione di G^2 consente la formulazione di tre ipotesi sui caratteri e di una ipotesi totale, che, definita dall'intersezione delle sue componenti, risulterebbe essere la meno sostenuta.

Definiamo anomala la situazione di una tabella nella quale l'ipotesi totale risulta più "accettabile" di qualcuna delle componenti. Abbiamo applicato una procedura di simulazione a 100000 tabelle 3x3x3 costruite con i rapporti probabilistici teorizzati da Mendel. Il risultato più interessante è che le anomalie sono molto frequenti riguardando oltre il 96% delle tabelle esaminate.

SUMMARY

Frequency anomalies pertaining to Mendel's second experiment

This article is inspired by the second experiment executed by Mendel on three joint characters (shape, color, seed-coat). Through various experiments he managed to isolate plants which differed by more than one character and noted that they were transmitted independently from one another, following a relationship 1:2:1 (dominant: hybrid: recessive). Such relationship together with the hypothesis of independence, to be verified on the basis of empirical data, represent the basis for estimating the theoretical frequencies. Among the various indeces which may be used to describe the difference between theoretical and observed frequencies, we have chosen G^2, which was calculated for each of the nine tables conditioned to the three modalities of every character. The partition by addition of G^2 allows the definition of three hypotheses on the characters and a total hypothesis. The latter, defined by the intersection of its components, would seem to be the less supported. We defined as anomalous the situation in which, for a specific table the total hypothesis results more acceptable than that of one of the components. We have applied a simulation procedure to 100,000 tables 3x3x3 constructed with the probability relationships theorized by Mendel. The most interesting result is that anomalies are very frequent, as they regard more than 96% of the examined tables.