GENETICS AND MOLECULAR BIOLOGY FOR ENVIRONMENTAL ANALYSIS

MOLECULAR ECOLOGY

LESSON 10: POPULATION GENETICS

Prof. Alberto Pallavicini
pallavic@units.it

BREVE RIASSUNTO DELLA STATISTICA

- I genetisti che si occupano di ecologia utilizzano la statistica per identificare schemi ordinati tra la stupefacente variabilità del mondo naturale.
- Gli obbiettivi dell'analisi statistica sono quindi simili a quelli della modellistica, ma mentre quest'ultima si basa su astrazioni teoriche, la ***statistica rappresenta un modo per comprendere dati sperimentali ottenuti dal mondo reale.***

BREVE RIASSUNTO DELLA STATISTICA

- Questi dati generalmente sono raccolti da campioni casuali di una popolazione, e i "tests" statistici vengono utilizzati per stabilire se determinati schemi identificati nel campione siano generati da eventi casuali o da fenomeni biologici reali. Più grande è la dimensione del campione (spesso indicato con N o n), più sicuri possiamo essere del risultato dell'analisi.

BREVE RIASSUNTO DELLA STATISTICA

- Descriveremo tre tipi basilari di analisi statistica, i quali sono destinati a rispondere a tre diversi tipi di quesito.
- In genere, tali domande dipendono dalla natura dei dati raccolti e questi, a loro volta, possono essere **discreti** (o anche qualitativi) oppure **continui** (quantitativi).

BREVE RIASSUNTO DELLA STATISTICA

- I dati **discreti** possono essere facilmente suddivisi in gruppi distinti o classi.
- Tutti i marcatori genetici descritti in precedenza, sia visibili che molecolari, identificano genotipi distinti che rappresentano buoni esempi di dati discreti.

BREVE RIASSUNTO DELLA STATISTICA

- Molti caratteri fenotipici, invece, sono distribuiti in maniera **continua**; cioè, non vi sono limitate tipologie definite.
- Gli esempi comprendono virtualmente tutte le misure dimensionali di un carattere di un qualsiasi organismo, come il peso, l'altezza, la lunghezza e la larghezza. Questi caratteri sono chiamati quantitativi perché devono essere misurati

BREVE RIASSUNTO DELLA STATISTICA

I tre tipi di quesito e le corrispondenti analisi statistiche sono:

- 1. c'è una differenza per alcuni caratteri continui tra due o più gruppi distinti?

- Per esempio, ci sono differenze fra coleotteri maschi per la quantità di progenie che essi generano?
- Esistono differenze della forma del fiore tra popolazioni di piante che vivono nello stesso ambiente?

BREVE RIASSUNTO DELLA STATISTICA

- I test statistici utilizzati per questo tipo di domande comprendono i t-test e l'analisi della varianza (ANOVA);

BREVE RIASSUNTO DELLA STATISTICA

2. esiste una relazione tra due variabili continue?

- Esempi di questo genere di quesito sono: La lunghezza del tarso di un uccello è correlata a quanta progenie esso genera?
- La lunghezza delle zampe dei genitori è correlata a quella dei figli nelle rane?

- Questa tipologia di analisi è spesso studiata con gli strumenti statistici della **correlazione e della regressione**

BREVE RIASSUNTO DELLA STATISTICA

- 3. una popolazione può deviare da ciò che un determinato modello prevede?

- Questo è spesso denominato test della "bontà della corrispondenza" ("goodness-of-fit" test). Questo test si applica, per esempio, per determinare se le frequenze genotipiche di una popolazione reale si discostino da quelle attese dall'equilibrio di Hardy-Weinberg.

  - Il modo consueto per far ciò è attraverso l'uso del **test del chi-quadro**

BREVE RIASSUNTO DELLA STATISTICA

- Tutti i metodi statistici generano un test corrispondente.
- Generalmente, più ampio è il test statistico, maggiore è l'accuratezza che otteniamo nel considerare le differenze o le relazioni che stiamo analizzando non generate dal verificarsi di una semplice coincidenza casuale, ma piuttosto come il riflesso di un fenomeno reale nella popolazione.

BREVE RIASSUNTO DELLA STATISTICA

- Questo livello di accuratezza è quantificato dal valore di P (**P-value**), che *rappresenta la probabilità che tale differenza o relazione possa essere stata prodotta dal caso*.
- Più piccolo sarà il valore di P, più sicuri siamo di essere di fronte a differenze o relazioni reali.

BREVE RIASSUNTO DELLA STATISTICA

- Per quale valore di P possiamo concludere che uno schema identificato sia vero?

- Per convenzione, i tests con valori di P minori di 0.05 sono giudicati statisticamente significativi, in pratica, siamo sicuri che i fenomeni siano reali e non dovuti al caso.
- In ogni caso, un valore di P uguale a 0.05 indica che abbiamo il 5% di probabilità di commettere un errore nel concludere che esista un vero fenomeno.

BREVE RIASSUNTO DELLA STATISTICA

- Più il valore di P è **piccolo**, più siamo sicuri che i risultati riflettano differenze o relazioni reali.
- Più grandi sono le dimensioni del campione, **più basso sarà il valore di P rispetto alla media di quelli ottenuti per lo stesso modello**, infatti, siamo più sicuri quando analizziamo un campione più numeroso di popolazione.

EQUILIBRIO DI UNA POPOLAZIONE

Una popolazione è detta in equilibrio genetico se nel tempo non cambiano né le frequenze alleliche né quelle genotipiche.

Se siamo in una situazione in cui possiamo ritenere che la legge di
Hardy-Weinberg possa essere rispettata sarà stabile qualsiasi tipo di frequenza allelica e, come abbiamo visto, le frequenze genotipiche rispettano una distribuzione binomiale.

Groenlandia: pM = 0.913, H_{MN} = 0.156

Islanda : pM = 0.569, H_{MN} = 0.515

Le frequenze attese degli eterozigoti saranno pertanto rispettivamente:

Groenlandia: $2pq=2\cdot0.913\cdot(1-0.913)= 0.159$

Islanda: $2pq = 2\cdot0.569\cdot(1 - 0.569) = 0.490$

Nel campione groenlandese troviamo una mancanza dello 0.3% di eterozigoti, in quello islandese un eccesso del 2.5%.

IL TEST X 2

Tali discrepanze possono essere sottoposte ad un'analisi statistica per vedere se sono significative.

Un test statistico atto a vedere la corrispondenza dei dati sperimentali con quelli teorici è il test $\chi 2$.

Per la popolazione **islandese** questi sono i dati da considerare:

	M	MN	N	totale
val. osserv. (f_o)	233	385	129	747
val. attesi (f_e)	242	366	139	747
differenze (f_o-f_e)	9	19	10	

Il valore di χ^2 si calcola secondo $\sum \frac{(f_o - f_e)^2}{f_e}$

Nel nostro caso $\chi^2 = \frac{9^2}{242} + \frac{19^2}{366} + \frac{10^2}{139} = 2.046$

Come in tutti i tests statistici bisogna considerare le condizioni poste al nostro sistema; in questo caso le condizioni sono 2:

1 - il totale dei dati attesi deve essere 747

2 - i rapporti tra i tre fenotipi devono soddisfare un determinato valore della frequenza allelica; in particolare

$$\frac{M_o + \frac{1}{2} \cdot MN_o}{tot} = 0.569$$

Gli elementi considerati sono 3 (le tre frequenze fenotipiche), per cui con 2 limiti 1 solo può essere posto indipendentemente.

Si esprime tale concetto dicendo che abbiamo 3-2=1 grado di libertà.

Ciò significa che, con i limiti posti (un dato valore di p ed un determinato totale), possiamo stabilire arbitrariamente uno solo dei tre valori, essendo gli altri due determinati da p e dal totale. Infatti abbiamo stimato p come (233 + 385/2) / 747 = 0.5696. I valori che possiamo dare a M_0, MN_0 e N_0 devono essere tali che siano rispettate le seguenti equazioni:

A

$$\frac{\mathbf{M}_0 + \frac{1}{2} \cdot \mathbf{MN}_0}{747} = 0.5696 = p$$

B

$$\mathbf{M}_0 + \mathbf{MN}_0 + \mathbf{N}_0 = 747$$

MN_o		M_o	N_o		tot	p
0	$\rightarrow$	425.5	321.5	$\rightarrow$	747	0.5696
643	$\rightarrow$	104	0	$\rightarrow$	747	0.5696
385	$\rightarrow$	233	129	$\rightarrow$	747	0.5696
366	$\rightarrow$	242	139	$\rightarrow$	747	0.5696

Per ogni livello di gradi di libertà (g.d.l.) il valore di $\chi 2$ è funzione della probabilità: un valore di $\chi 2$ pari a 0 ha probabilità pari a 1 (la certezza) di poter essere superato in un altro campionamento casuale. Solitamente si assume che soltanto quando la probabilità scende sotto il livello del 5% possiamo ritenere che i valori sperimentali si discostino da quelli attesi in una misura non casuale.

Esistono tabelle che danno il valore di $\chi 2$ in funzione della probabilità e dei g.d.l.; esistono programmi di calcolo che forniscono il valore di probabilità associato al valore dei g.d.l. e di $\chi 2$.

Con 1 g.d.l. il valore 2.046 corrisponde ad un livello di probabilità pari a 0.153.

Ciò significa che abbiamo una probabilità del 15.3% di ottenere valori di $\chi 2$ superiori a quello riscontrato, cioè di avere valori che si discostano ancora maggiormente dai dati attesi.

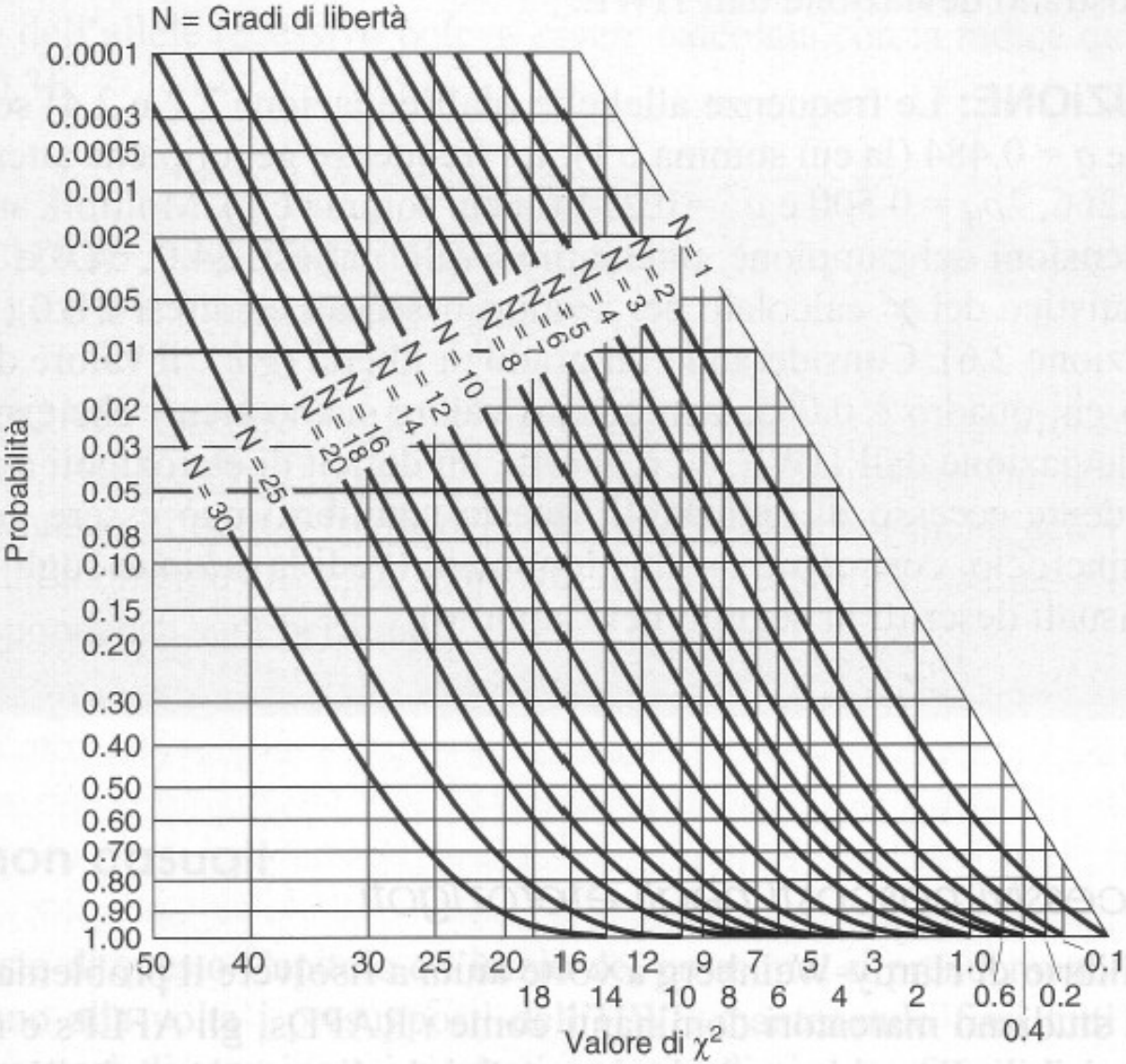


Figura 2.11 Grafico del χ^2. Per utilizzare il grafico, trova il valore del χ^2 lungo l'asse (orizzontale) delle x, quindi spostati verticalmente fino a che intersechi la linea diagonale corrispondente all'appropriato valore dei gradi di libertà. Il valore di P si legge sull'asse y (verticale) al punto d'intersezione. (Per concessione di James F. Crow.)

STIMA DELLE FREQUENZE ALLELICHE

Possiamo inoltre osservare come, nell'ipotesi che la legge di Hardy-Weinberg sia rispettata, possiamo valutare la frequenza degli eterozigoti anche nel caso di geni con dominanza.

Prendiamo il caso della sensibilità al feniltiocarbammato (PTC) negli uomini; per alcuni questa sostanza è assolutamente insapore, altri la trovano notevolmente amara.

L’esame di alberi geneaologici dimostra che questa sensibilità è dovuta ad un gene con due alleli; T - t. Gli individui TT e quelli Tt percepiscono il sapore del PTC, mentre quelli con genotipo tt non lo percepiscono (T è dominante su t).

Un campionamento ha fornito questi dati:

percepiscono il sapore 218 individui, mentre 102 non lo percepivano.

Questi ultimi dovevano essere di genotipo tt e, chiamando con q la frequenza dell'allele t, se si pensa che le condizioni di Hardy Weinberg siano rispettate, ci si aspetta una quantità q^2 di insensibili. Possiamo allora stimare

$$q = \sqrt{\frac{102}{102+218}} = 0.5646$$

Ovviamente la stima di p sarà 1-0.5646 = 0.4354.

Consideriamo ora il problema dell'attendibilità statistica di queste stime.

Ogni dato affetto da errore statistico di campionamento presenta una possibilità di presentare ad ogni campionamento un valore che varia attorno ad un valore medio; la misura di questa variazione prende il nome di **varianza** e viene indicata con σ^2; la radice quadrata della varianza viene chiamata **errore standard** e viene indicata con σ.
E' evidente come questa varianza sia tanto maggiore quanto minore è la dimensione del campione.

Nel caso di un rapporto p la varianza è uguale a

$$\frac{p \cdot (1-p)}{\mathrm{N}} \quad \text{dove } \mathbf{N} \text{ è il totale delle osservazioni}$$

Nel nostro caso il rapporto che abbiamo stimato è il valore

$$q^2 = \frac{102}{320} = 0.31875$$

pertanto la varianza di questa stima sarà

$$\sigma^2_{q^2} = \frac{0.31875 \cdot (1-0.31875)}{320} = 0.00067859$$

Osserviamo che in questo caso la varianza della stima delle persone sensibili alla PTC è identica.

$$\left(\frac{218}{320} = 0.68125\right)$$

Correttamente dovremo quindi indicare che la frequenza stimata
delle persone insensibili al PTC è 0.31875 ± 0.02605. Ciò significa che la frequenza di queste persone è compresa in un intorno del valore 0.31875, con dei limiti che saranno tanto più ampi quanto maggiore vuole essere la precisione della nostra affermazione.

Entra in gioco un ulteriore parametro, che ci esprime la probabilità della veridicità della nostra affermazione;

questo parametro è chiamato **t**; anche questo è funzione del numero delle osservazioni, ma per valori sufficientemente grandi vale 2 se il livello di probabilità richiesto è del 95%, 2.6 se tale livello è del 99%.

Allora dalle nostre osservazioni potremo affermare che nella popolazione campionata la frequenza delle persone insensibili alla PTC sarà compresa, con una probabilità del 95% tra 0.31875-2x0.02605 e 0.31875+2x0.02605, cioè tra 0.26665 e 0.37085.

Scegliendo il livello di probabilità al 99% otterremo invece i seguenti limiti: 0.31875-2.6x0.02605 e 0.31875+2.6x0.02605, cioè possiamo affermare con una sicurezza del 99 % che la frequenza è compresa tra 0.25102 e 0.38648.

- ESERCIZIO: Per il locus PCI, nella popolazione di *Daphnia* dei bacino di Ojibway, Spitz trovò due alleli, S e S^-, e determinò che il numero di individui di ciascun genotipo era:

  42 SS, 48 SS^- *e* 38 S^-S^-

- Verifica se i genotipi mostrano deviazione dall'HWE.

FREQUENZA DEGLI ETEROZIGOTI

Su caratteri recessivi possiamo fare altre considerazioni sulla frequenza dell'allele negli eterozigoti e negli omozigoti recessivi.

Nell'uomo l'albinismo è un carattere recessivo e sono albini circa un individuo su 20,000.

$$\mathbf{R} = q^2 = 1/20000 = 0.00005 \qquad q = \sqrt{q^2} = 0.00707 = 1/141$$

$$\mathbf{H} = 2 \cdot q \cdot (1\text{-}q) = 0.014 = 1/71$$

Il rapporto tra gli alleli dell'albinismo nascosti negli eterozigoti, che sono detti "portatori", e quelli manifesti nei recessivi sarà:

$$\frac{\mathbf{H}/2}{\mathbf{R}} = \frac{p \cdot q}{q^2} = \frac{p}{q}$$

Pertanto nel caso dell'albinismo per ogni allele manifestato nei recessivi ve ne saranno 140 nascosti nei portatori.

Cioè gli alleli recessivi rari saranno portati per lo più in una forma “nascosta”, nell’eterozigote, nel quale non si manifestano.

Le tre funzioni

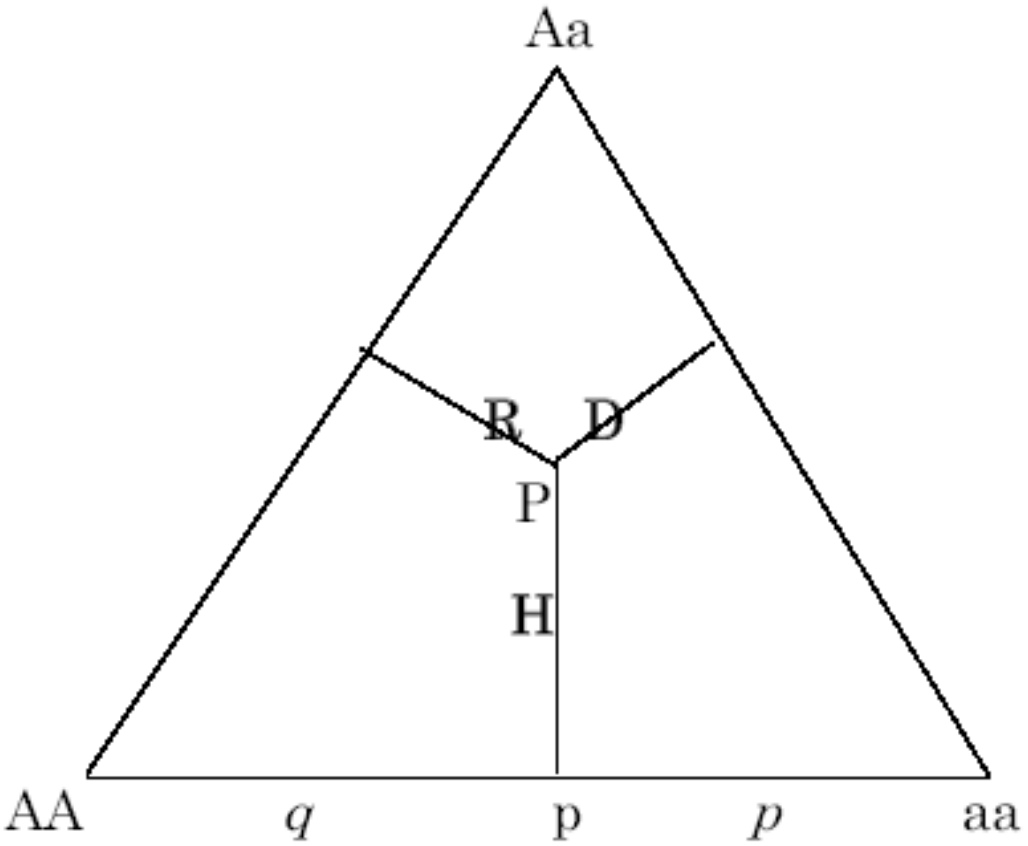
$$D = p^2 \qquad H = 2p(1-p) \qquad R = (1-p)^2$$

rappresentano 3 parabole, definite in ($0 \leq p \leq 1$).

H presenta un massimo in dH/dp = 0, cioè per p =0.5, cui corrisponde
il valore di H =0.5.

Cioè il numero massimo di eterozigoti che potremo trovare in una popolazione con due alleli, che rispetta le condizioni della legge di Hardy-Weinberg, è la metà degli individui, valore che si ottiene per p = q = 0.5.

Per un gene con due alleli a volte può essere utile rappresentare la situazione di una popolazione P in un grafico a coordinate triangolari (in un triangolo equilatero).



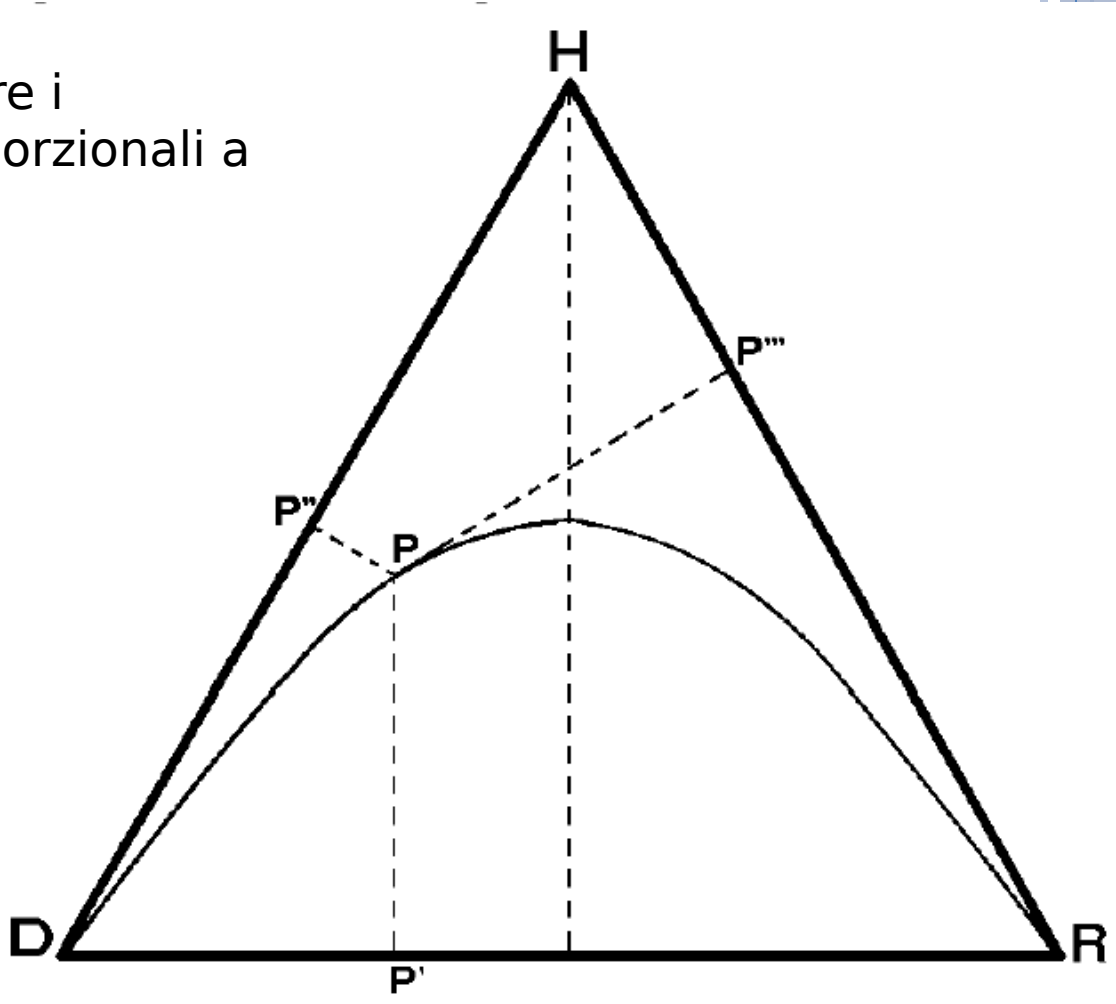
In condizioni di equilibro di Hardy-Weinberg P non potrà trovarsi in un punto qualsiasi, ma solo là dove sono soddisfatte le tre funzioni precedenti cioè là dove

$$\mathbf{H^2 = 4DR}.$$

Questa equazione identifica una parabola che passa per AA e per aa; presenta un massimo per H = 0.5, dove AA-p = p-aa.

Tale parabola è nota come parabola di De Finetti.

In queste coordinate i segmenti PP’, PP”, PP”’ sono rispettivamente proporzionali a H, R, D, mentre i segmenti DP’ e RP’ sono proporzionali a q e p.



- Il diagramma di De Finetti è stato portato ad un uso esteso alla Genetica delle popolazioni da A.W.F. Edwards nel suo libro *Fondamenti della Genetica Matematica*.
- Nella sua forma più semplice il diagramma può essere usato per mostrare la gamma di frequenze del genotipo perché l'equilibrio di Hardy-Weinberg sia soddisfatto. A.W.F. Edwards e C. Cannings estesero il suo uso per dimostrare i cambiamenti che capitano nelle frequenze alleliche sotto la selezione naturale.

Cannings C., Edwards A.W.F. (1968) "Natural selection and the de Finetti diagram" Ann Hum Gen **31**:421-428

ALLELISMO MULTIPLO

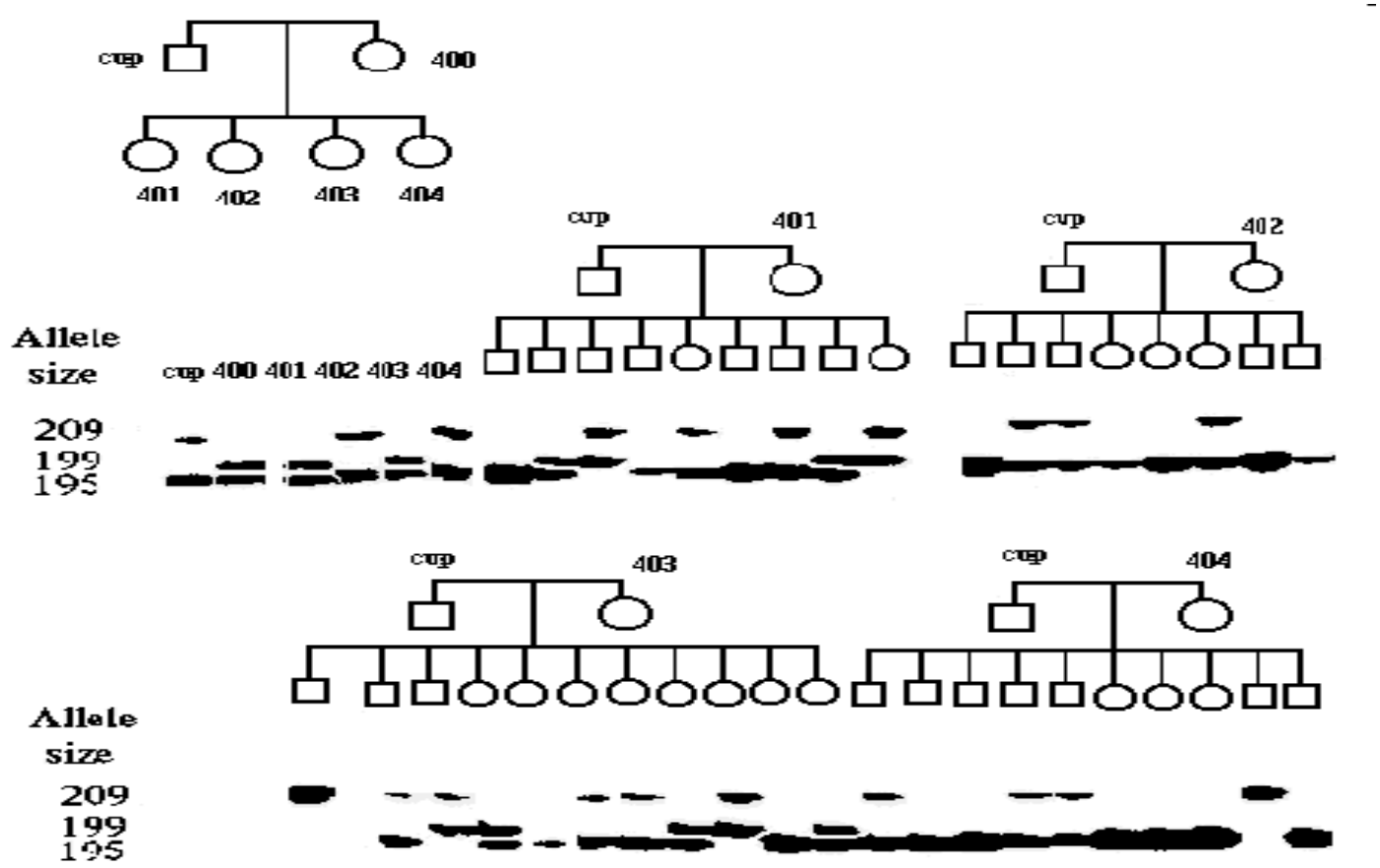
Se ad un locus possiamo riscontrare più di due alleli, avremo da calcolare più frequenze alleliche, genotipiche e fenotipiche.

Supponiamo di avere 3 alleli; li indicheremo, assieme alle loro frequenze, nel seguente modo:

a_1	a_2	a_3
p	q	r

Naturalmente sarà p + q + r = 1

Segregation of genetic variation (microsatellite) in a locus in a swine family detected by gel electrophoresis. The boar Cup carries the alleles 209 and 195 and sow 400 carries the alleles 199 and 195. Most of the animals are heterozygotes each with two bands. The homozygotes have only one band.



I genotipi omozigoti saranno 3:
a_1a_1, a_2a_2, a_3a_3;

quelli eterozigoti 3·2 /2 = 3:
a_1a_2, a_1a_3, a_2a_3.

Le loro frequenze rispetteranno lo sviluppo del quadrato del trinomio:

$$(p+q+r)^2$$

$$p^2 + q^2 + r^2 + 2pq + 2pr + 2qr$$

$$a_1a_1\ a_2a_2\ a_3a_3\ a_1a_2\ a_1a_3\ a_2a_3$$

Per cui il numero totale degli eterozigoti sarà:

$$H = 1 - (p^2 + q^2 + r^2)$$

e potrà anche superare il valore di 0.5.

Con tre alleli la frequenza massima possibile degli eterozigoti si ricava dal seguente sistema di equazioni differenziali:

$$\frac{\partial H}{\partial p} = 0 \quad ; \quad \frac{\partial H}{\partial q} = 0$$

da cui, essendo

$H = 1 - p^2 - q^2 - (1 - p - q)^2$

$= - 2\ p^2 - 2\ q^2 + 2\ p + 2\ q - 2\ pq$

Si ricava

$$\begin{cases} -4p + 2 - 2q = 0 \\ -4q + 2 - 2p = 0 \end{cases}$$

Che ha per soluzione

$$p = q = 1/3$$

e, ovviamente

pertanto si ricava

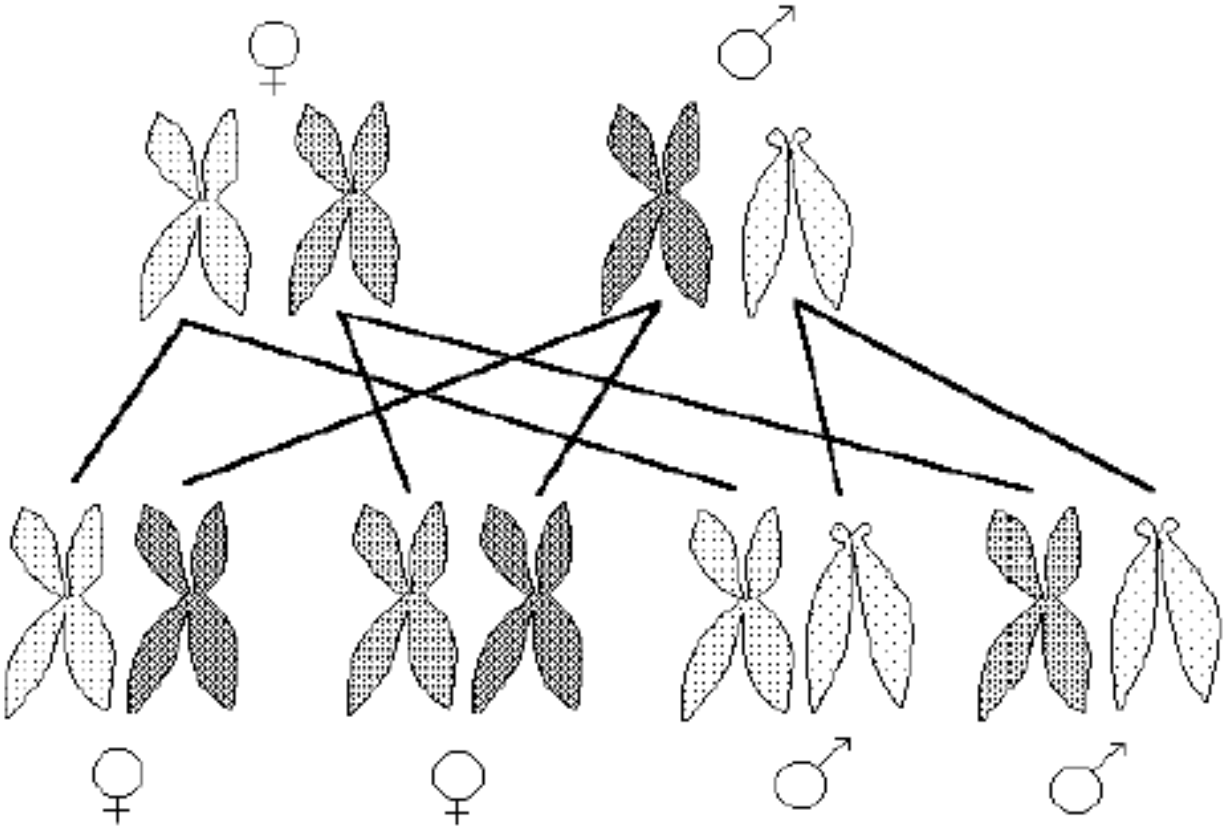
$$r = 1/3$$

$$H_{\max} = 2/3$$

GENI LEGATI AL SESSO

I geni localizzati sul cromosoma X hanno una trasmissione particolare. Poniamoci nella situazione in cui il sesso eterogametico sia quello maschile.

Allora il maschio può trasmettere il suo allele solo alle figlie femmine, mentre la femmina trasmetterà tale gene sia ai figli maschi che alle figlie.



Nel caso di due soli alleli le femmine presenteranno tre genotipi, le cui frequenze le indicheremo, al solito

D (AA), **H** (Aa), **R** (aa).

I maschi invece, dato che avranno una sola copia del gene sull'unico cromosoma X che possiedono, potranno avere due soli genotipi, le cui frequenze le indicheremo con **S** (A) e **T** (a).

Dovremo anche distinguere una frequenza dell'allele A

nelle femmine ($p_{♀}$ = **D** + ½**H**) e

nei maschi ($p_{♂}$ = **S**).

Potremo calcolare anche una *frequenza media* in tutta la popolazione dell'allele A;

nell'ipotesi che il numero di femmine sia eguale a quello dei maschi tale frequenza sarà:

$$p = \frac{2 \cdot p_{\female} + p_{\male}}{3}$$

dato che le *femmine possiedono due alleli e il maschio uno solo*

Dato che i maschi della generazione n ricevono l'allele A solo dalle madri sarà:

$$p_{\male,\mathrm{n}} = p_{\female,\mathrm{n-1}}$$

mentre nelle femmine, che possono ricevere tale allele sia dal padre che dalla madre, sarà:

$$p_{\female,\mathrm{n}} = \frac{(p_{\male,\mathrm{n-1}} + p_{\female,\mathrm{n-1}})}{2}$$

Quindi alla generazione **n** si può avere una differenza tra la frequenza dell'allele A nei maschi e nelle femmine che, riferita alla generazione precedente sarà:

$$p_{\male,n} - p_{\female,n} = p_{\female,n-1} - \frac{(p_{\male,n-1} + p_{\female,n-1})}{2}$$

$$= \frac{p_{\female,n-1}}{2} - \frac{p_{\male,n-1}}{2}$$

$$p_{\male,n} - p_{\female,n} = - \frac{(p_{\male,n-1} + p_{\female,n-1})}{2}$$

Ad ogni generazione la differenza della frequenza dell'allele tra maschi e femmine si dimezza e tale differenza inverte il segno.

Esprimendo la differenza alla generazione **0** come Δ_o e la differenza alla generazione **n** come Δ_n avremo:

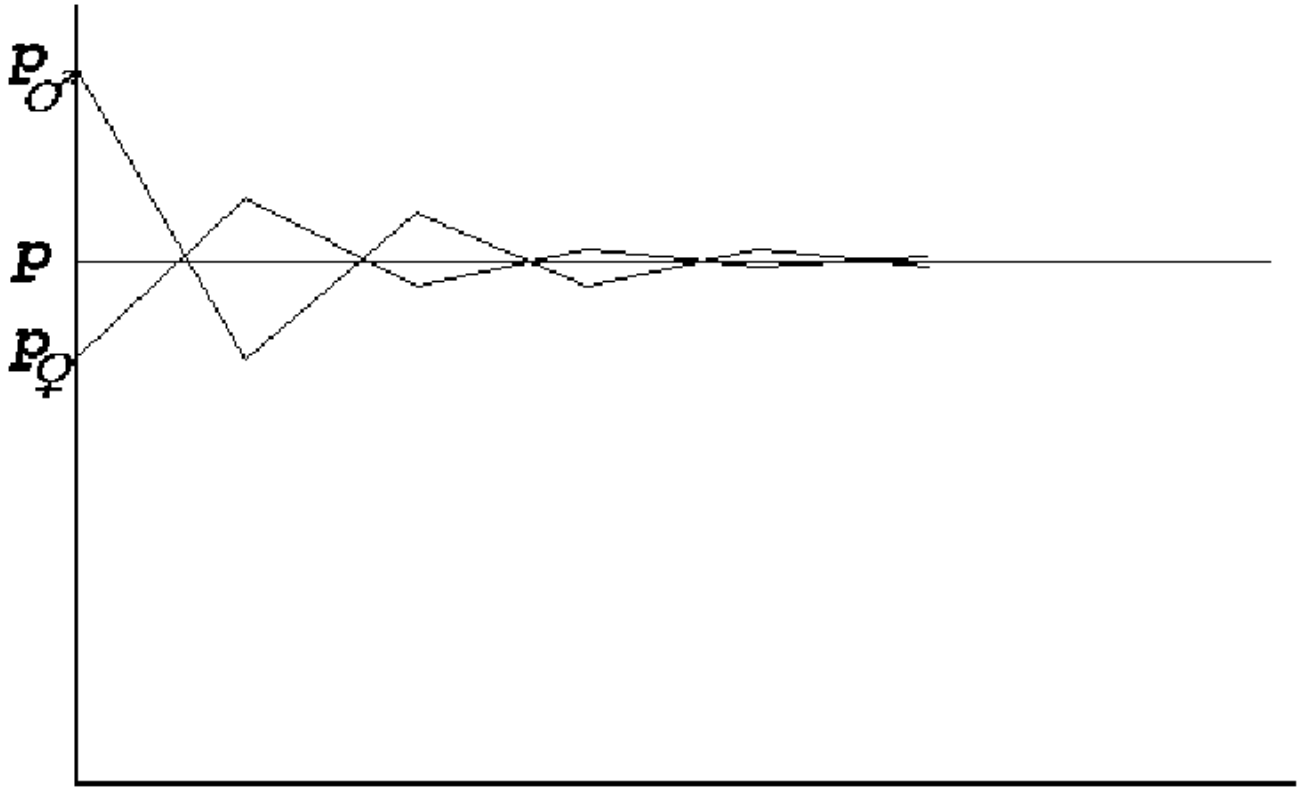
$$\Delta_n = (-1/2)^n \Delta_o$$

Per cui si tende $\lim_{n\to\infty} \Delta_n = 0$

Per cui:

$$p_{\female} = p_{\male} = p$$

La seguente figura mostra l'evoluzione delle frequenze di un gene legato al sesso.



Prendiamo ad esempio il gene y nei gatti, associato al sesso, che permette negli *eterozigoti* (solo femmine) l'espressione di altri colori nella pelliccia, oltre al colore giallo, mentre negli *omozigoti* inibisce l'espressione di altri colori.

- Nelle femmine sono possibili quindi 3 diversi fenotipi e 3 diversi genotipi:

genotipo	fenotipo
y+y+	non giallo, al massimo 2 colori
y+y	giallo e al massimo altri 3 colori
y y	giallo

Nei maschi sono possibili solo 2 diversi genotipi e fenotipi:

genotipo	fenotipo
y+	non giallo, al massimo 2 colori
y	giallo

Un'indagine su una popolazione di gatti dell'Inghilterra ha dato questi risultati:

femmine			tot.	maschi		tot.
y+y+	y+y	yy		y+	y	
277	54	7	338	311	42	353

Da questi dati risulta che la frequenza dell'allele y è:

$$p_{\female} = 0.101 \qquad p_{\male} = 0.119$$

senza una significativa differenza tra maschi e femmine, e con un valore medio p = 0.107.

Con questa frequenza ci si aspetterebbero, su 338 femmine e 353 maschi i seguenti valori:

Femmine				maschi		
y+y+	y+y	yy	tot.	y+	y	tot.
269	64	4	338	315	38	353

Anche in questo caso possiamo utilizzare il test χ^2 per verificare se la popolazione può essere considerata in equilibrio.

In questo caso però abbiamo una classe dei valori attesi che presenta un valore inferiore a 5 (gli individui yy). Il calcolo del valore di χ^2 è corretto solo per classi di dimensioni almeno pari a 5. Pertanto dovremo raggruppare tale classe ad un'altra.

Possiamo calcolare

$$\chi^2 = \frac{(277-269)^2}{269} + \frac{(54+7-64-4)^2}{(64+4)} + \frac{(315-311)^2}{315} + \frac{(42-38)^2}{38} = 1.430$$

Le classi considerate sono 4 (non 5 perché abbiamo raggruppato gli individui yy a quelli y+y); le condizioni poste sono 3 (numero di femmine, numero di maschi, valore di p); pertanto i gradi di libertà sono 4-3 = 1. Il
valore di probabilità associato è 0.23.

Anche in questo caso i valori teorici non sono significativamente diversi da quelli osservati.

Si deve notare che per caratteri legati al sesso pertanto, il sesso omogametico presenta un numero maggiore di fenotipi rispetto al sesso eterogametico e che, in presenza di dominanza, il sesso eterogametico presenta il fenotipo recessivo con una frequenza molto più elevata del sesso omogametico:

$$\mathrm{T} = q \gg \mathrm{R} = q^2$$

E' per questo motivo che tra gli uomini è molto più frequente incontrare maschi daltonici che femmine daltoniche.

Semplificando la situazione, che è più complessa, ammettiamo che il daltonismo sia dovuto ad un allele recessivo la cui frequenza sia q = 0.1; i maschi daltonici saranno presenti con frequenza q, cioè costituiranno il 10% della popolazione maschile, mentre le femmine daltoniche saranno presenti con frequenza q^2, cioè solo l'1% di tutte le femmine.

- **Haemophilia** in human is one of the best known examples of sex-linked recessive inheritance, the frequency in boys being 100 times larger than the one in girls.
- This occurs when the gene frequency is 0.01, corresponding to the frequency in boys. Whereas the gene in the girls has to come from both father and mother, each with a probability of 0.01, this corresponds to a frequency in girls of 1 in 10'000.