Apunts complementaris de genetètica mendeliana

02/10/2023

Probabilitat

El concepte clàssic de probabilitat consisteix en una fracció, el numerador de la qual seria el nombre d'oportunitats de què es produïsca un cert event, i el denominador, el nombre total d'oportunitats de què l'event es produïsca o no. Aquesta fracció seria la probabilitat de l'event, sempre que les *oportunitats* que estem comptant siguen *equiprobables*. Aquesta noció de **probabilitat** és adequada en moltes situacions, però no és l'única.

Existeix també la **interpretació subjectiva** de probabilitat: el grau de certesa o confiança amb què creiem alguna cosa. Aquesta és la interpretació pròpia de l'estadística Bayesiana. Sota aquesta interpretació té sentit assignar probabilitats a les hipòtesis, per exemple.

Tant una interpretació com l'altra (n'hi ha d'altres) poden ser formalitzades mitjançant una mesura numèrica entre 0 i 1, on la probabilitat 0 indica impossibilitat i la probabilitat 1 indica certesa absoluta. Matemàticament, la probabilitat és una *funció* $P()$ definida sobre qualsevol subconjunt dels events considerats possibles, la totalitat dels quals es representa amb la llegra grega Ω. Aquesta funció té les propietats següents:

$$\begin{aligned} P(A) \geq & 0 \qquad \text{per a tot subconjunt } A \text{ d'}\Omega \\ P(\Omega) = & 1 \\ P(A \cup B) = & P(A) + P(B) \qquad \text{per a tot } A \text{ i } B \text{ mútuament excloients.} \end{aligned}$$

La *probabilitat condicional* es calcula així:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{Sempre que } P(B) > 0.$$

Per exemple, la probabilitat de què siga heterozigot (**Aa**) un individu de fenotip dominant (pésol groc, per exemple), els progenitors del qual eren ambdós heterozigots per a l'únic gen que determina aquest fenotip es pot calcular així:

$$\begin{aligned} P(\text{heterozigot}|\text{dominant}) = & \frac{P(\text{heterozigot} \cap \text{dominant})}{P(\text{dominant})} \\ P(\text{heterozigot}|\text{dominant}) = & \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3} \end{aligned}$$

(Com tots els heterozigots expressen el fenotip dominant, la probabilitat de ser alhora heterozigot i de fenotip dominant és igual a la probabilitat de ser heterozigot).

Reorganitzant la fòrmula de la probabilitat condicional, obtenim una expressió d'una altra propietat important de les probabilitats:

$$P(A|B) \cdot P(B) = P(A \cap B)$$

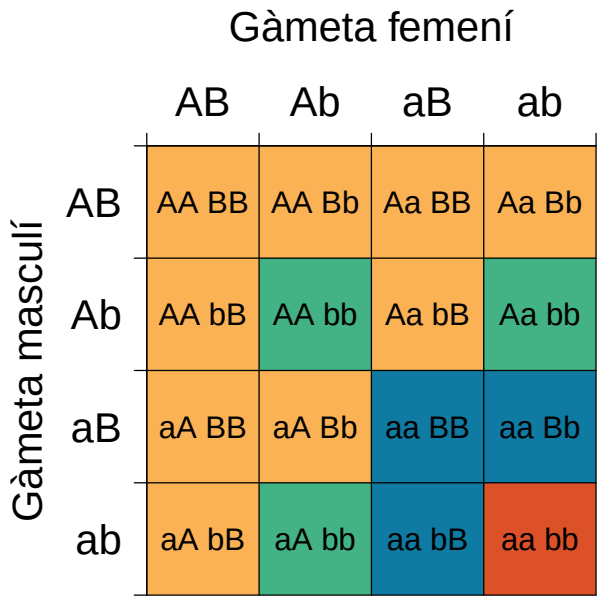
La probabilitat de què es produïsquen dos event A i B seria igual al producte de les seues probabilitats ($P(A \cap B) = P(A) \cdot P(B)$) si i només si els dos events són independents. És a dir, si $P(A|B) = P(A)$.

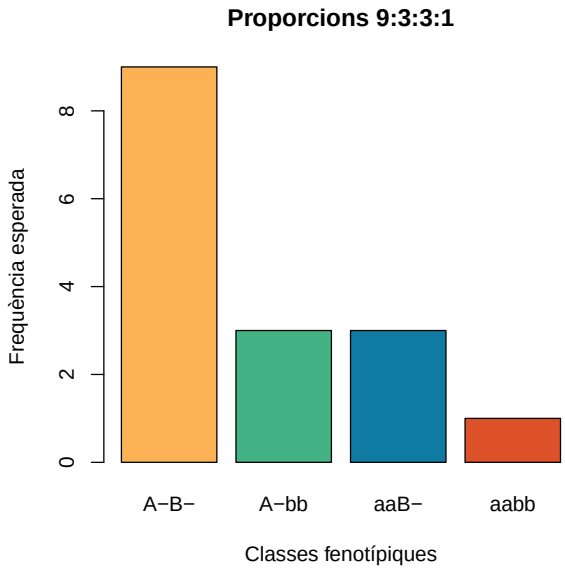
És fàcil demostrar que $P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$, la qual cosa ens porta a una altra expressió de la probabilitat condicional, coneguda com el *teorema de Bayes*:

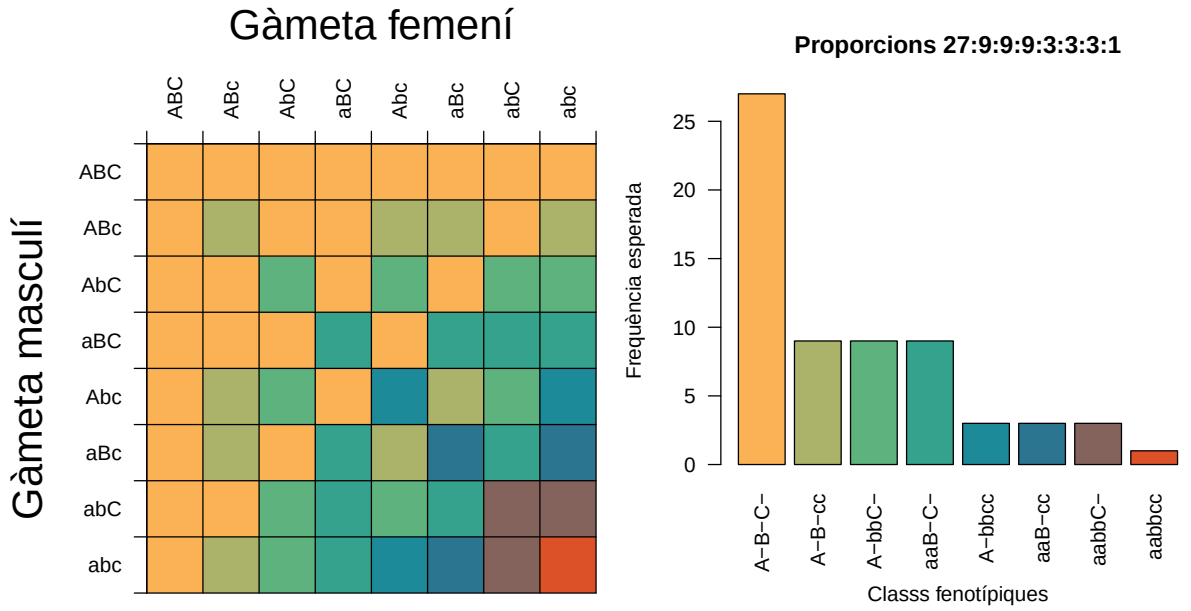
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Quadre de Punnet

Una taula de doble entrada, coneguda com el *quadre de Punnet* ens permet mostrar i calcular les probabilitats de totes les combinacions dels resultats possibles de dos processos *independents*. Per exemple: quins al · lels transmet el pare i quins, la mare. O dit d'una altra manera, quin dels tipus gamètics possibles de cada progenitor es combinen per formar el zigot. En un creuament dihíbrid, **Aa Bb** × **Aa Bb**, els descendents poden resultar de les 16 combinacions següents:



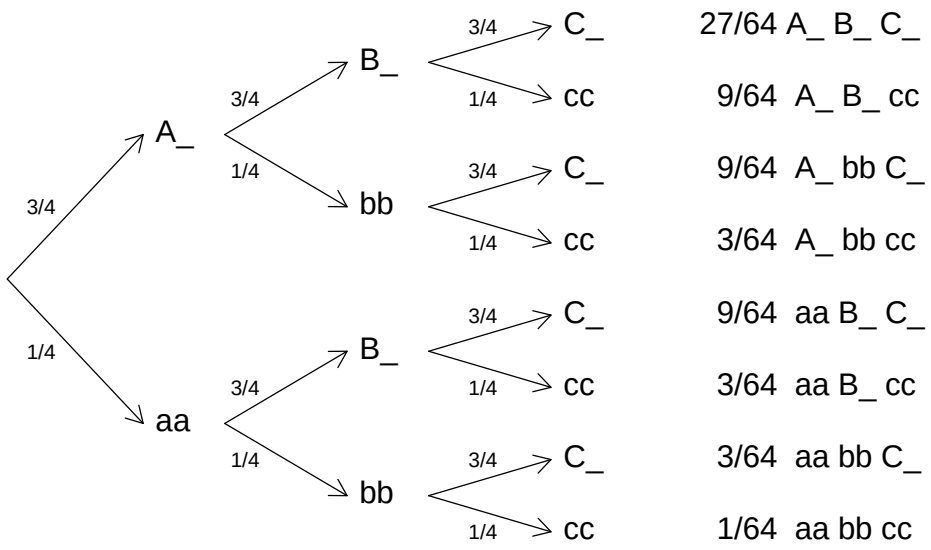




Les $2^3 \times 2^3 = 64$ combinacions possibles d'un creuament trihíbrid (**Aa Bb Cc** $\times$ **Aa Bb Cc**) poden representar-se en un quadre de 8×8 cel · les, com un tauler d'escacs. Però els quadres de PUnnet són molt redundants, perquè dediquen una casella a cada genotip possible, distingint fins i tot els heterozigots en funció del progenitor del qual procedeix cada al · lel (**Aa BB** i **aA BB** tenen caselles diferents, encara que normalment es consideren el mateix genotip). Existeix una altra manera més compacta de representar els possibles resultats d'un creuament di- o trihíbrid i de calcular-ne les probabilitats: els diagrames ramificats.

Diagrama ramificat

Els diagrames ramificats representen els diferents processos independents com una seqüència de decisions. En el diagrama de la figura inferior, cada un dels tres gens *no lligats* (amb segregació independent) pot determinar un fenotip dominant o un de recessiu. Com es tracta de la descendència d'un creuament dihíbrid, en cada un dels gens la probabilitat de què l'embrió siga de fenotip dominant és $\frac{3}{4}$, i la probabilitat de què siga de fenotip recessiu és d'$\frac{1}{4}$. La probabilitat (o freqüència esperada) d'un fenotip total, com per exemple **A_ _ B_ _ cc** (és a dir, dominant per a A i B, i recessiu només per a C) es calcula multiplicant les probabilitat de les branques que ens condueixen a eixe fenotip: $\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{3 \times 3 \times 1}{4 \times 4 \times 4} = \frac{9}{64}$.



Graus de llibertat

Abans d'introduir el test de la χ^2, és convenient aclarir algunes confusions comunes respecte dels **graus de llibertat**. El concepte procedeix de l'àlgebra i s'usa també en física per designar un nombre de dimensions: en quantes dimensions es pot moure un objecte, quantes dimensions té el subespai definit per un nombre de vectors, etc.

En estadística, els graus de llibertat representen la quantitat d'informació amb què es calcula un **estadístic**. Un estadístic és qualsevol resultat d'aplicar una fòrmula a unes dades. Per exemple, si les meues dades són el nombre de pésols verds i el nombre de pésols grocs que ha produït una planta, tinc dos números, x_1 i x_2. Qualsevol operació que faça amb ells la realitzaré amb dos graus de llibertat, en el sentit de què qualsevol dels dos números (com a *dades* o *variables aleatòries* que són) podria haber sigut diferent.

Aquesta obvietat no sembla tenir cap importància en el càlcul d'un estadístic qualsevol. Per exemple, la proporció de pésols verds, $p_1 = \frac{x_1}{x_1+x_2}$ la calcule amb dos graus de llibertat. Tanmateix, quan l'objectiu de calcular un estadístic és estimar un **paràmetre** poblacional, aleshores pot ser important revisar i conéixer exactament quants graus de llibertat hem utilitzat. Seguint amb l'exemple, la proporció de pésols verds és un estimador del paràmetre p de la distribució binomial: $\hat{p} = \frac{x}{n}$, on n és l'*altre paràmetre* de la distribució binomial, a saber, el nombre d'*intents* o número total d'experiments de Bernoulli. En el nostre cas, $n = x_1 + x_2$ és el nombre total de pésols que hem comptat i classificat. Resulta que si pretenem utilitzar $p_1 = \frac{x_1}{x_1+x_2}$ com un estimador del paràmetre p de la binomial $B(p, n)$, aleshores $\hat{p} = p_1$ està calculat amb només un grau de llibertat, perquè en fixar la mida de la mostra, n, ja no és cert que tant x_1 com x_2 podrien haver pres qualsevol altre valor.

En general, quan un estadístic pretén estimar un paràmetre de la població d'on s'ha tret la mostra, aleshores el nombre de graus de llibertat amb què es calcula l'estadístic és igual al nombre de dades menys el nombre de paràmetres addicionals que hem hagut d'estimar amb les mateixes dades.

I açò, per a què serveix? La veritat és que en molts casos ens dóna igual saber o ignorar els graus de llibertat utilitzats en estimar una mitjana, una variança, etc. Però en algunes situacions sí que importa. Per exemple, a l'hora de realitzar un test de la χ^2.

Test de la χ^2

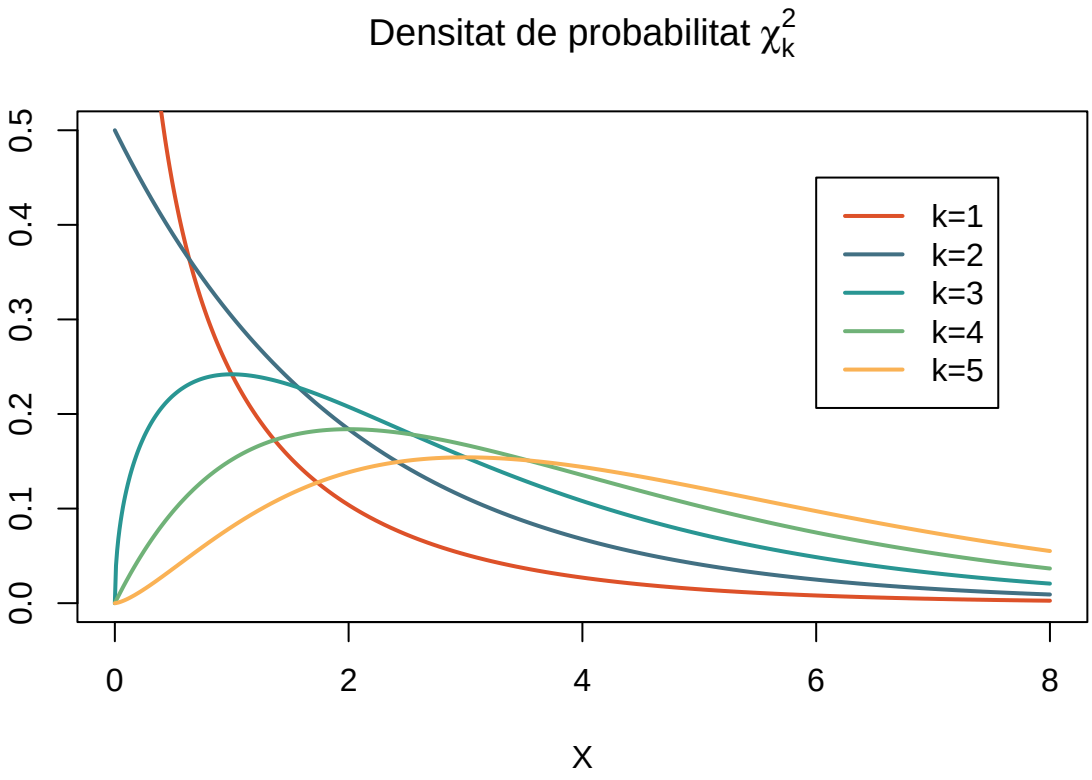
Utilitzem el test de la χ^2 per determinar si existeix una desviació estadística significativa entre les freqüències esperades i les freqüències observades de dues o més categories en què classifiquem els elements d'una mostra.

Per a tal fi, es calcula l'estadístic següent:

$$X = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

On O_i és el valor observat de freqüència **absoluta** de la categoria i (nombre de casos, no proporció!); i E_i és el nombre esperat de casos de la categoria i. L'estadístic X és més gran com majors són les diferències entre els valors observats i els esperats. Aquest estadístic ens permet comprovar si les desviacions són significativament majors que les esperades sota la hipòtesi de les freqüències esperades. Sempre què els nombres O_i són suficientment grans (almenys majors o iguals a 5) aleshores X hauria de tenir una distribució χ^2 sota la hipòtesi nul · la representada per les freqüències esperades.

Ara bé, la distribució χ^2 no és una única distribució, sinó una família d'elles, cada una caracteritzada per un valor (natural) del **paràmetre** k:



La distribució χ^2_k amb què cal comparar el nostre estadístic X és aquella el paràmetre k de la qual coincideix amb el nombre de graus de llibertat amb què hem calculat X. Per motius pràctics, al paràmetre k de la distribució χ^2_k se l'anomena *graus de llibertat*. Però és important adonar-se'n de què k és un paràmetre d'una distribució teòrica, un concepte molt diferent del que coneixem com els "graus de llibertat" d'un estadístic.