Apuntes complementarios de genética mendeliana

02/10/2023

Probabilidad

El concepto clásico de probabilidad consiste en una fracción, el numerador de la cual sería el número de oportunidades de que suceda un evento, y el denominador, el número total de oportunidades de que el evento suceda o no suceda. Esa fracción sería la probabilidad del evento, siempre que las *oportunidades* que estamos contando sean *equiprobables*. Esta noción de **probabilidad** es adecuada en muchas situaciones, pero no es la única.

Existe también la **interpretación subjetiva** de probabilidad: el grado de certeza o confianza con la que creemos algo. Esta es la interpretación propia de la estadística Bayesiana. Bajo esta interpretación, tiene sentido asignar probabilidades a las hipótesis, por ejemplo.

Tanto una interpretación como otra (existen más) pueden ser formalizadas mediante una medida numérica entre 0 y 1, donde la probabilidad 0 indica imposibilidad, y la probabilidad de 1 indica certeza absoluta. Matemáticamente, la probabilidad es una *función* $P()$ definida en un conjunto de eventos posibles, la totalidad de los cuales se representa con la letra Ω. Esta función tiene las propiedades siguientes:

$$\begin{aligned} P(A) \geq & 0 \qquad \text{para todo subconjunto } A \text{ de } \Omega \\ P(\Omega) = & 1 \\ P(A \cup B) = & P(A) + P(B) \qquad \text{para todo } A \text{ y } B \text{ mutuamente excluyentes.} \end{aligned}$$

La *probabilidad condicional* se calcula así:

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

Por ejemplo, la probabilidad de que sea heterocigoto (**Aa**) un individuo de fenotipo dominante (guisante amarillo), cuyos dos progenitores eran heterocigotos para el único gen que determina ese fenotipo se puede calcular así:

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

(Como todos los heterocigotos expresan el fenotipo dominante, la probabilidad de ser al mismo tiempo heterocigoto y de fenotipo dominante es igual a la probabilidad de ser heterocigoto).

Reorganizando la fórmula de la probabilidad condicional, obtenemos una expresión de otra propiedad importante de las probabilidades:

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

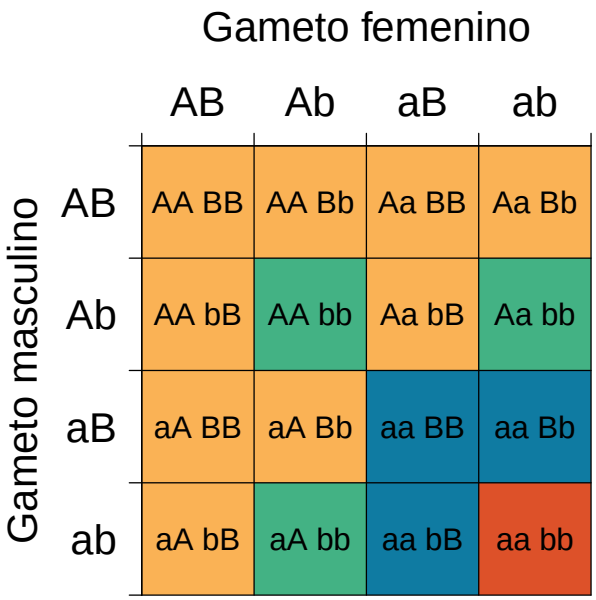
La probabilidad de que se produzcan dos eventos A y B seria igual al producto de sus probabilidades ($P(A \cap B) = P(A) \cdot P(B)$) si y solamente si los dos eventos son independientes. Es decir si $P(A|B) = P(A)$.

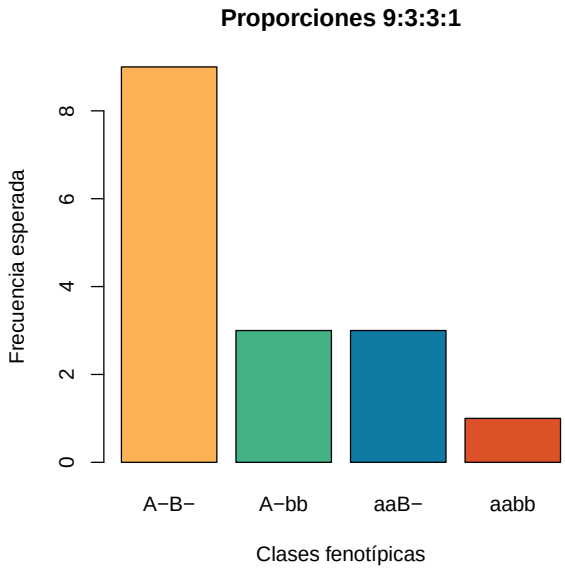
Es fácil demostrar que $P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$, lo cual nos lleva a otra expresión de la probabilidad condicional, conocida como el *teorema de Bayes*:

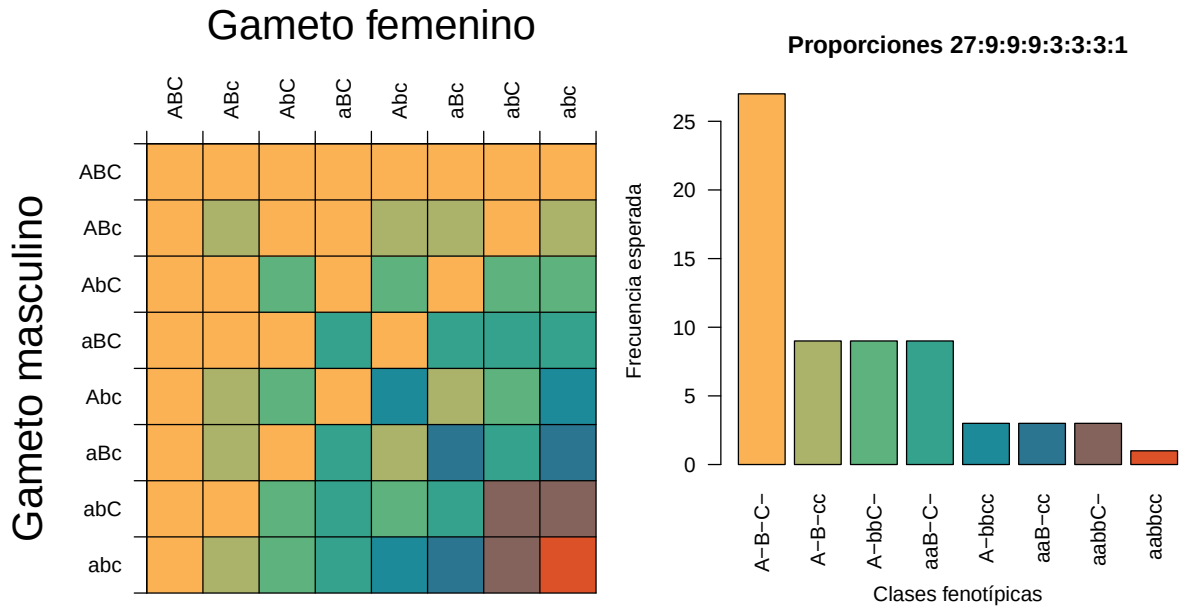
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Cuadro de Punnet

Una tabla de doble entrada, conocida como el *cuadro de Punnet* nos permite mostrar y calcular las probabilidades de todas las combinaciones de los resultados posibles de dos procesos *independientes*. Por ejemplo: qué alelos transmite el padre y qué alelos transmite la madre a la descendencia. O dicho de otro modo, cuál de los tipos gaméticos posibles de cada progenitor se combinan para formar el cigoto. Por ejemplo, en un cruce dihíbrido **Aa Bb** $\times$ **Aa Bb**, los descendientes pueden resultar de las 16 combinaciones siguientes:



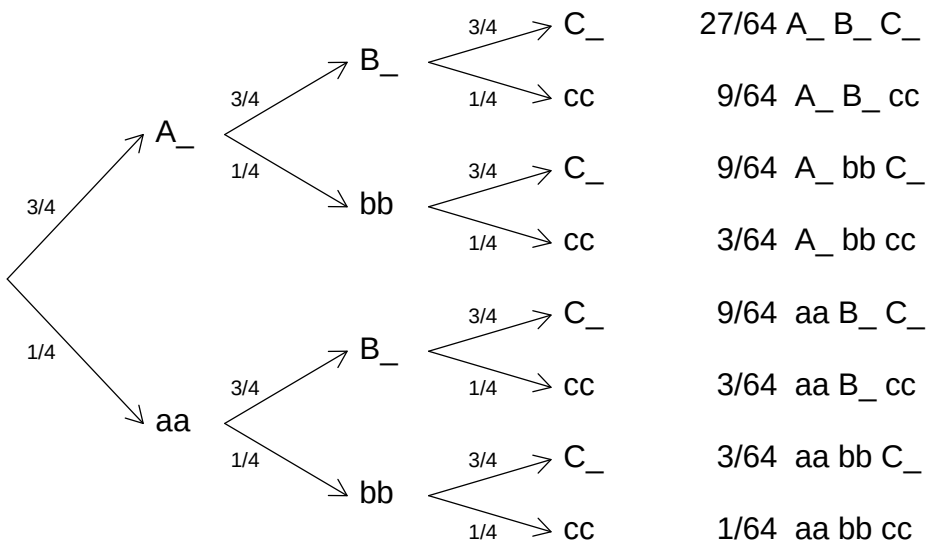




La $2^3 \times 2^3 = 64$ combinaciones posibles de un cruce trihíbrido (**Aa Bb Cc** $\times$ **Aa Bb Cc**) podrían representarse en un cuadro de 8×8 celdas, como un tablero de ajedrez. Pero los cuadros de Punnet son muy redundantes, porque dedican una casilla a cada genotipo posible, distinguiendo incluso los heterocigotos en función del progenitor del que procede cada alelo (**Aa BB** y **aA BB** tiene casillas diferentes, aunque a todos los efectos son el mismo genotipo). Existe otra manera más compacta de representar los posibles resultados de un cruce di- o trihíbrido y de calcular sus probabilidades: los diagramas ramificados.

Diagrama ramificado

Los diagramas ramificados representan los diferentes procesos independientes como una secuencia de decisiones. En el diagrama de abajo cada uno de los tres genes *no ligados* (con segregación independiente) puede determinar un fenotipo dominante o uno recesivo. Como se trata de la descendencia de un cruce dihíbrido, en cada uno de los genes la probabilidad de que el embrión sea de fenotipo dominante es $\frac{3}{4}$, y la probabilidad de que sea de fenotipo recesivo es de $\frac{1}{4}$. La probabilidad (o frecuencia) de un fenotipo total, como por ejemplo **A_ B_ cc** (es decir, dominante para A y B, y recesivo para C), se calcula multiplicando las probabilidades de las ramas que nos conducen a ese fenotipo: $\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{3 \times 3 \times 1}{4 \times 4 \times 4} = \frac{9}{64}$.



Grados de libertad

Antes de introducir el test de la χ^2, es conveniente aclarar algunas confusiones comunes respecto de los **grados de libertad**. El concepto procede del álgebra y se usa también en física para designar un número de dimensiones: en cuantas dimensiones se puede mover un objeto, cuántas dimensiones tiene el subespacio definido por un número de vectores, etc. En estadística, los grados de libertad representan la cantidad de información con la que se calcula un **estadístico**. Un estadístico es cualquier resultado de aplicar una fórmula a unos datos. Por ejemplo, si mis datos son el número de guisantes verdes y el número de guisantes amarillos que ha producido una planta, tengo dos números x_1 y x_2. Cualquier operación que haga con ellos la realizaré con dos grados de libertad, en el sentido de que cualquiera de los dos números (como *datos* o *variables aleatorias* que son) podrían haber sido diferentes. Esta obviedad no parece tener ninguna importancia en el cálculo de un estadístico cualquiera. Por ejemplo, la proporción de guisantes verdes $p_1 = \frac{x_1}{x_1+x_2}$ la calculo, por tanto, con dos grados de libertad. Sin embargo, cuando el objetivo de calcular un estadístico es estimar un **parámetro** poblacional, puede ser importante conocer y revisar exactamente cuantos grados de libertad hemos empleado. Siguiendo con el ejemplo, la proporción de guisantes verdes es un estimador del parámetro p de la distribución binomial: $\hat{p} = \frac{x}{n}$, donde n es el *otro parámetro* de la distribución binomial, a saber, el número de *intentos* o número total de experimentos de Bernoulli. En nuestro caso, $n = x_1 + x_2$ es el número total de guisantes que hemos contado y clasificado. Resulta que si pretendemos usar $p_1 = \frac{x_1}{x_1+x_2}$ como estimador del parámetro p de una binomial $B(p, n)$, entonces $\hat{p} = p_1$ está calculado con un solo grado de libertad, porque al fijar el tamaño de la muestra, n, ya no es cierto que tanto x_1 como x_2 podrían haber sido diferentes. En general, cuando un estadístico pretende estimar un parámetro de la población de donde se ha sacado la muestra, entonces el número de grados de libertad con los que se calcula el estadístico es igual al número de datos menos el número de parámetros adicionales que hemos tenido que estimar con los mismos datos. Y esto, ¿para qué sirve? La verdad es que muchos casos nos da igual saber o ignorar con cuántos grados de libertad estamos estimando una media, una varianza, etc. Pero hay algunas situaciones en las que sí importa. Por ejemplo, al realizar un test de la χ^2.

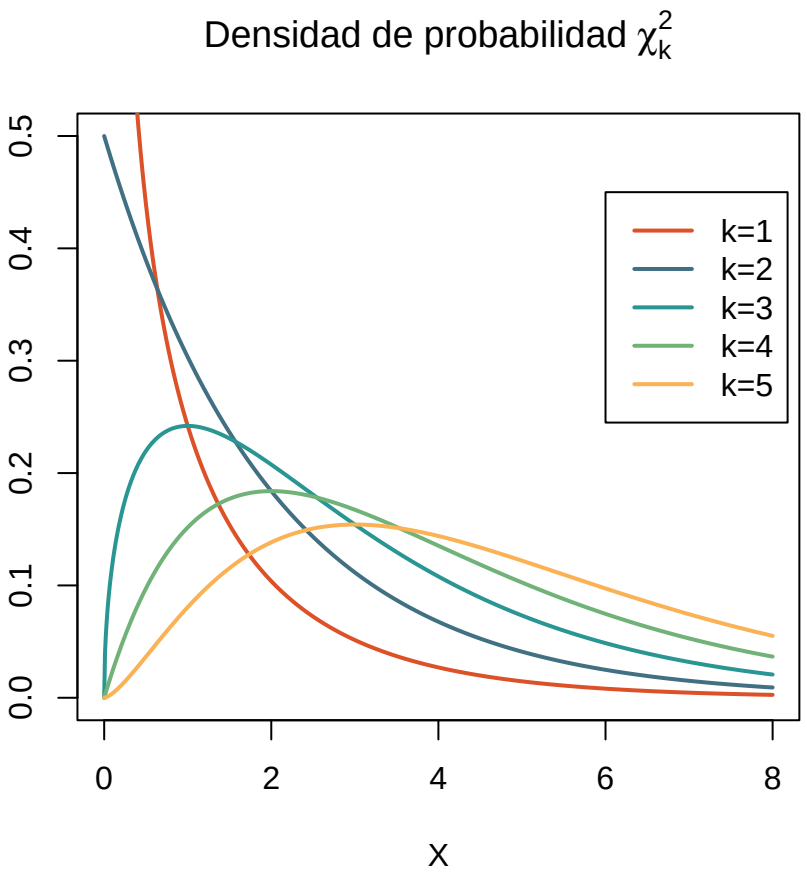
Test de la χ^2

Usamos el test de la χ^2 para determinar si existe una desviación estadística significativa entre las frecuencias esperadas y las frecuencias observadas de dos o más categorías en las que clasificamos una muestra. Para ello se calcula el estadístico siguiente:

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

Donde O_i es el valor observado de frecuencia **absoluta** de la categoría i (número de casos, no proporción); y E_i es el número esperado de casos de la categoría i. El estadístico X es más grande cuanto mayores son las diferencias entre los valores observados y esperados. Este estadístico nos permite comprobar si las desviaciones son significativamente mayores que las esperadas bajo la hipótesis de las frecuencias esperadas porque, si los número O_i son suficientemente grandes (al menos mayores o iguales a 5) entonces X debería tener una distribución χ^2 bajo la hipótesis nula representada por las frecuencias esperadas.

Ahora bien, la distribución χ^2 no es una única distribución, sino una familia de ellas, cada una caracterizada por un valor (natural) del **parámetro** k.



La distribución χ^2_k con la que debemos comparar nuestro estadístico X es aquella cuyo parámetro k coincide con el número de grados de libertad con el que hemos calculado X. Por motivos prácticos, al parámetro k de la distribución χ^2_k se le llama *grados de libertad*. Pero es importante reconocer que k es un parámetro de una distribución teórica: un concepto muy diferente de lo que conocemos como "grados de libertad" de un estadístico.