Ejercicio – Estimación básica de máxima verosimilitud

En una muestra de n = 10 individuos se parasitan k = 4 individuos. La probabilidad de obtener estos datos en un orden particular (por ejemplo, obtener primero 4 individuos parasitados, luego 6 individuos no parasitados) es $p^4(1-p)^6$, y la probabilidad de obtener estos datos en cualquier orden es $\frac{10!}{4!6!}p^4(1-p)^6$ (esta es la probabilidad binomial). La primera fracción, $\frac{10!}{4!6!}$, es solo el número de secuencias diferentes que puede obtener 4 individuos parasitados en una muestra de 10 y, a menudo, se reduce a $\binom{10}{4}$ and $4! - 1 \cdot 2 \cdot 3 \cdot 4$.

.

- Calcula la probabilidad de obtener estos datos dado que p = 0, 0.1, …, 1 (en R puedes usar la función 'dbinom(k,n,p)'). Luego haz un gráfico de esta probabilidad como función de p.
- Prueba diferentes valores de tamaño de muestra (n) y número de individuos parasitados (k) y observa cómo cambia el gráfico. ¿Cómo puede usar este gráfico para encontrar la "mejor suposición" de la prevalencia del parásito en toda la población con base en estos datos?
- Intenta aumentar tanto el tamaño de la muestra (n) como el número de individuos parasitados (k) manteniendo las mismas proporciones (ej., k/n = 20/8, 30/12, 40/16,…). ¿Cómo cambia la forma de la curva cuando aumenta el tamaño de la muestra? como interpretas esto?
- Calcula también la probabilidad de los datos dados los diferentes valores de p cuando los individuos parasitados aparecen en un orden particular (p. ej., primero obtienes 4 parasitados, luego 6 no parasitados) y grafica esto. ¿En qué se diferencian las dos curvas? ¿Esto tiene sentido? ¿Obtenemos alguna información sobre el parámetro p sabiendo el orden en el que tomamos muestras de individuos parasitados y no parasitados?

El último gráfico que hiciste se llama función de verosimilitud de p dados los datos (k y n) y se escribe $L(p;k,n) = p^k(1-p)^{n-k}$. Tenga en cuenta que esta es la misma expresión que la función de probabilidad binomial, excepto que vemos la expresión como una función de p en lugar de una función de k y nos hemos saltado el coeficiente binomial $\binom{n}{k}$ porque es solo una constante que solo afecta la elevación de la curva (no la forma o la ubicación del pico).

El valor de los parámetros que maximizan la función de verosimilitud se denomina estimación de máxima verosimilitud (a menudo abreviado como MLE). Esta es su "mejor suposición" para el valor del parámetro. En este caso podemos escribir $\hat{p}\ MLE = 0.4$ (el 'sombrero' sobre la p indica que se trata de una estimación del parámetro).

- Por lo general, el logaritmo de la función de verosimilitud, el log-verosimilitud, se usa para encontrar el MLE. La función de log-verosimilitud es en este caso $\ell(p; k, n) = k \ln(p) + (n - k)\ln(1 - p)$. La función logarítmica de verosimilitud y la función de verosimilitud siempre tienen el pico para el mismo valor de los parámetros. Puede confirmar esto trazando la función de probabilidad logarítmica.

C) En el ejemplo simple anterior, puede encontrar el MLE analíticamente encontrando el valor del parámetro donde la función de probabilidad es plana (es decir, en el pico), estableciendo la primera derivada en cero, $\partial\ell/\partial p = 0$, y resolviendo esto para p. Entonces obtendrías $\hat{p}\ MLE = k/n$. Sin embargo, en modelos más complejos, esto a menudo no se puede hacer de forma analítica, por lo que tiene que hacerse numéricamente mediante "prueba y error" (lo que a menudo es mucho más fácil de todos modos). Hay varias funciones de este tipo para dicha optimización numérica en R (por ejemplo, 'optim' y 'optimize').