EJERCICIOS P1

Mathias Mangino

28 September, 2025

Genetica de poblaciones

Table of contents

1 Total de ejercicios de genetica de poblaciones

1. Completar la siguiente tabla, en la que y son las frecuencias de un alelo de referencia en las subpoblaciones 1 y 2, respectivamente:

p_1	H_1	p_2	H_2	$\bar{H}_s$	$\frac{(p_1-p_2)^2}{2}$	$\bar{H}_t$
0.4		0.6				
0.1		0.3				
0		1				

¿Cómo se comparan los dos primeros casos?

2. Ejemplo 2: Asignando genotipos a subpoblaciones con frecuencias alélicas conocidas Nos gustaría asignar un individuo a una de dos posibles subpoblaciones fuente S_1 y S_2 según su genotipo $G_1 = a_1a_1b_1b_1b_2c_1c_1 $, obtenido para tres loci bialélicos no ligados. Se conocen las frecuencias alélicas en estas dos subpoblaciones:

Tabla 1. Frecuencias alélicas en dos subpoblaciones

	S_1	S_2
$f(a_1)$	0.3	0.2
$f(b_1)$	0.2	0.25
$f(c_1)$	0.3	0.2

¿Cuáles son las probabilidades de observar este genotipo en particular en cada una de las dos subpoblaciones?

4. Para este ejercicio, seleccionamos un subconjunto de 20 individuos (10 de Pine Forest y 10 de la subpoblación Warner Mts.) tomados aleatoriamente de las secuencias generadas en el trabajo citado. De los 701 pares de bases reportados en el artículo, se seleccionaron 620 sitios tras excluir los sitios que no se secuenciaron en todos los individuos. De estos, 50 fueron polimórficos, y solo esos sitios se presentan a continuación. Observamos que el número total de sitios polimórficos (= variables o, usando un término mendeliano, segregantes) en las secuencias es de aproximadamente el 8%.

En la tabla resumen, cada tipo de secuencia (haplotipo) representa una clase alélica. El segundo y tercer bloque proporcionan los datos por separado para cada subpoblación.

Con base en los datos mitocondriales, complete la siguiente tabla con las siguientes estadísticas descriptivas para cada subpoblación:

A: número de clases de alelos (= haplotipos distintos) S: número de sitios segregantes H: heterocigosidad π: diversidad nucleotídica

Calcular es un poco engorroso, por lo que le recomendamos completar la siguiente tabla como punto de partida:

Número de diferencias entre pares de alelos (haplotipos):

	Haplo_1	Haplo_2	Haplo_3	Haplo_4	Haplo_5
Haplo_2		-			
Haplo_3			-		

	Haplo_1	Haplo_2	Haplo_3	Haplo_4	Haplo_5
Haplo_4	40	39	39	-	
Haplo_5	47	46	44	7	-
Haplo_6	48	47	45	12	5

La tabla incluye datos de heterocigosidad del artículo, basados en los 11 loci microsatélites. Complete la tabla con los valores obtenidos anteriormente.

	Pine Forest	Warner Mts
ADN mitocondrial		
A		
S		
H		0.64
π		22.74
Microsatélites		
$\bar{H}$	0.37	0.78

Ten en cuenta que usamos H para el ADNmt, ya que se trata de un solo locus, mientras que $\bar{H}$ para los microsatélites, ya que 0,37 es la heterocigosidad promedio de 11 loci de microsatélites.

¿Cómo se pueden interpretar estos resultados y qué nos indican sobre la variación entre subpoblaciones?

1. La frecuencia de un alelo autosómico A en una población es $p = 0.1$. ¿Cuál es el número esperado de alelos A en una muestra de 20 alelos? ¿Cuál es la probabilidad de no observar el alelo A en la muestra?
2. Utilizando la función rbinom, simule un gran número de muestras de tamaño $n = 20$ y caracterice la distribución de los resultados utilizando descriptores gráficos y numéricos apropiados.
3. Considere ahora el caso de $f(A) = p = 0.01$.¿Cuál es la probabilidad de observar el alelo A en una muestra de $n = 20$?
4. La frecuencia observada de un alelo autosómico diploide A es 3/10 (tamaño de la muestra $n = 10$). Utilice dbinom para calcular la probabilidad de observar esta frecuencia en la muestra para un rango de frecuencias subyacentes en la población entre $0,01$ y $0,99$.
5. Genotipos como muestras aleatorias de tamaño $n = 2$. Para un locus autosómico diploide con dos alelos, describa las proporciones de HW como esperanzas de la distribución binomial.
6. Para una muestra de $n = 10$ individuos, obtenga la frecuencia esperada de heterocigotos para $f(A) = 0.3$ y utilice rbinom para simular un gran número de muestras y caracterizar la variación aleatoria de la frecuencia observada de heterocigotos.
7. Para $n = 4$, obtener en R una simulación de todos los componentes del proceso coalescente. Asumir que $N = 5 \times 10^5$ y $\mu = 10 - 5$. Para la realización obtenida, calcular π, S y θ_W.
8. Structure: El programa structure se corrió con $K = 2$ y 90 genotipos, y la cadena de MCMC paró en la generación $m - 1$. Nos enfocamos en un sólo locus con alelos. El estado de la cadena en $m-1$ es el siguiente: Frecuencias del alelo de referencia en cada subpoblación: $P(m-1)1 = 0.5$, $P(m-1)2 = 0.8$. Ubicación de los genotipos a subpoblaciones:

	Subpoblación 1	Subpoblación 2
AA	13	25
Aa	24	13
aa	13	2
Total	50	40

a) Usando una distribución apropiada, obtener valores propuestos de para la P generación $m[P(m)]$ para cada subpoblación. Explicar, en cada caso, si deberían ser aceptadas como los nuevos valores $P(m)_1$ y $P(m)_2$ y con qué regla de probabilidad.
b) Ahora considerar un genotipo que puede moverse de la subpoblación 1 a la 2.⊠⊠ Explicar cómo se toma la decisión de moverlo o no. En las respuestas, incluir tanto las explicaciones como el código de R o los cálculos equivalentes.

2 Ejercicio 1: Frecuencias en subpoblaciones, practico Efecto Wahalund

1. Completar la siguiente tabla, en la que y son las frecuencias de un alelo de referencia en las subpoblaciones 1 y 2, respectivamente:

p_1	H_1	p_2	H_2	$\bar{H}_s$	$\frac{(p_1-p_2)^2}{2}$	$\bar{H}_t$
0.4		0.6				
0.1		0.3				
0		1				

¿Cómo se comparan los dos primeros casos?

H_1 y $H_2 \longrightarrow$ heterocigosidad esperada dentro de cada subpoblación.

$\bar{H}_s \longrightarrow$ heterocigosidad promedio dentro de subpoblaciones.

$\bar{H}_t \longrightarrow$ heterocigosidad total si las subpoblaciones se trataran como una sola (población combinada).

$\frac{(p_1 - p_2)^2}{2} \longrightarrow$ exceso de varianza en frecuencias alélicas entre subpoblaciones, que contribuye a la pérdida de heterocigotos al combinar poblaciones.

Para calcular H usamos Hardy-Weinberg para heterosigotas $H = 2 \cdot p \cdot (1 - p)$ Para H1 $p = 0.4$ $H = 2 \cdot 0.4 \cdot (1 - 0.4)$, para $p = 0.1$ $H = 2 \cdot 0.1 \cdot (1 - 0.1)$, para $p = 0$ $H = 2 \cdot 0 \cdot (1 - 0)$
Para H2 $p = 0.6$ $H = 2 \cdot 0.6 \cdot (1 - 0.6)$, para $p = 0.3$ $H = 2 \cdot 0.3 \cdot (1 - 0.3)$, para $p = 1$ $H = 2 \cdot 1 \cdot (1 - 1)$

Para calcular $\bar{H}_s = p_1 + p_2 - p_1^2 - p_2^2$ Para $\bar{H}_s$ $\bar{H}_s = 0.4 + 0.6 - 0.4^2 - 0.6^2$, para Para $\bar{H}_s$ $\bar{H}_s = 0.1 + 0.3 - 0.1^2 - 0.3^2$ y finalmente Para $\bar{H}_s$ $\bar{H}_s = 0 + 1 - 0^2 - 1^2$

Para calcular el termino $\frac{(p_1 - p_2)^2}{2}$ simplemente sustituimos

y finalmente usamos los terminos $\bar{H}_s$ y $\frac{(p_1 - p_2)^2}{2}$ para calcular $\bar{H}_t$

p_1	H_1	p_2	H_2	$\bar{H}_s$	$\frac{(p_1-p_2)^2}{2}$	$\bar{H}_t$
0.4	0.48	0.6	0.48	0.48	0.02	0.50
0.1	0.18	0.3	0.42	0.30	0.02	0.32
0 1	0	1	0	0	0.50	0.50

En el primer caso ($p_1 = 0.4$, $p_2 = 0.6$), la heterocigosidad dentro de cada subpoblación es relativamente alta ($H_1 = H_2 = 0.48$). Al promediarlas, $\bar{H}_s = 0.48$. Pero cuando juntamos ambas subpoblaciones, la heterocigosidad total es un poco mayor ($\bar{H}_t = 0.50$). La diferencia es pequeña, porque las frecuencias alélicas no difieren demasiado.

En el segundo caso ($p_1 = 0.1$, $p_2 = 0.3$), las subpoblaciones están más diferenciadas: $\bar{H}_s = 0.30$, pero la heterocigosidad total sube a 0.32. La diferencia es otra vez pequeña, pero se ve más clara la contribución de la varianza de frecuencias (0.02).

En el tercer caso ($p_1 = 0$, $p_2 = 1$), cada subpoblación es completamente homocigótica ($H_1 = H_2 = 0$), pero al combinar ambas el total tiene la máxima heterocigosidad posible ($\bar{H}_t = 0.50$). Este es el ejemplo extremo del **efecto Wahlund**: dentro de cada subpoblación no hay heterocigotos, pero si juntamos los datos parecería que la población combinada debería tener un 50% de heterocigotos.

¿Cómo se comparan los dos primeros casos?

Caso 1 ⟶ más heterocigotos porque p está cerca de 0.5.

Caso 2 ⟶ menos heterocigotos porque p está más lejos de 0.5.

Aunque las heterocigosidades absolutas difieren entre los dos casos, el **déficit de heterocigotos** al combinar subpoblaciones es el mismo. Esto se debe a que el **efecto Wahlund depende únicamente de la varianza en frecuencias alélicas entre subpoblaciones**. En ambos casos, esa varianza produce el mismo término.

$\frac{(p_1-p_2)^2}{2}$

3 Ejercicio 2: Asignando genotipos a subpoblaciones con frecuencias alélicas conocidas

2. Ejemplo 2: Asignando genotipos a subpoblaciones con frecuencias alélicas conocidas Nos gustaría asignar un individuo a una de dos posibles subpoblaciones fuente S_1 y S_2 según su genotipo $G_1 = a_1a_1b_1b_2c_1c_1 $, obtenido para tres loci bialélicos no ligados. Se conocen las frecuencias alélicas en estas dos subpoblaciones:

Tabla 1. Frecuencias alélicas en dos subpoblaciones

	S_1	S_2
$f(a_1)$	0.3	0.2
$f(b_1)$	0.2	0.25
$f(c_1)$	0.3	0.2

¿Cuáles son las probabilidades de observar este genotipo en particular en cada una de las dos subpoblaciones?

Modelo de probabilidades de genotipos

Bajo equilibrio de Hardy–Weinberg, la probabilidad de un genotipo depende de la frecuencia del alelo p:

Homocigota (AA): p^2

Heterocigota (Aa): $2p(1-p)$

Homocigota alternativo (aa): $(1-p)^2$

Dado que los loci son independientes, la probabilidad total del genotipo en una subpoblación es el producto de las probabilidades en cada locus.

$L_1 = P(AA|S_1) \cdot P(Aa|S_1) \cdot P(AA|S_1) \simeq P^2 \cdot 2 \cdot P \cdot (1-P) \cdot P^2$

Probabilidad en S_1

Subpoblacion 1

$Locus1 = P(a_1a_1|S_1) = P^2 = 0.3^2$

$Locus2 = P(b_1b_2|S_1) = 2 \cdot P \cdot (1-P) = 2 \cdot 0.2 \cdot (1-0.2)$

$Locus3 = P(c_1c_1|S_1) = P^2 = 0.3^2$

$L_1 = 0.3^2 \cdot 2 \cdot 0.2 \cdot (1-0.2) \cdot 0.3^2 = 0.002592$

Probabilidad en S_2

Subpoblacion 2

$Locus1 = P(a_1a_1|S_2) = P^2 = 0.2^2$

$Locus2 = P(b_1b_2|S_2) + 2 \cdot P \cdot (1-P) = 2 \cdot 0.25 \cdot (1-0.25)$

$Locus3 = P(c_1c_1|S_1) = P^2 = 0.2^2$

$L_2 = 0.2^2 \cdot 2 \cdot 0.25 \cdot (1-0.25) \cdot 0.2^2 = 0.0006$

Razon de verosimilitud

Comparamos qué subpoblación explica mejor el genotipo:

$\frac{L1}{L2} = \frac{0.002592}{0.0006} = 4.32$ es 4.32 veces mas probable la subpoblacion 1 que la 2

El genotipo observado es 4.32 veces más probable en S_1 que en S_2.

Probabilidad posteriori (Formula de Bayes)

Si asumimos probabilidades a priori iguales (es decir, que antes de ver el genotipo ambos orígenes son igualmente probables), la probabilidad de que el individuo pertenezca a S_1 dado su genotipo es:

$P(S_1|G_1) = \frac{L_1}{L_1+L_2} = \frac{0.002592}{0.002592+0.0006} = 0.812$

Con la información genética disponible, el individuo tiene un 81.2% de probabilidad de pertenecer a S_1.

4 Ejercicio 4: Ejemplo 2 practico 1 HW y F

4. Para este ejercicio, seleccionamos un subconjunto de 20 individuos (10 de Pine Forest y 10 de la subpoblación Warner Mts.) tomados aleatoriamente de las secuencias generadas en el trabajo citado. De los 701 pares de bases reportados en el artículo, se seleccionaron 620 sitios tras excluir los sitios que no se secuenciaron en todos los individuos. De estos, 50 fueron polimórficos, y solo esos sitios se presentan a continuación. Observamos que el número total de sitios polimórficos (= variables o, usando un término mendeliano, segregantes) en las secuencias es de aproximadamente el 8%.

En la tabla resumen, cada tipo de secuencia (haplotipo) representa una clase alélica. El segundo y tercer bloque proporcionan los datos por separado para cada subpoblación.

Con base en los datos mitocondriales, complete la siguiente tabla con las siguientes estadísticas descriptivas para cada subpoblación:

A: número de clases de alelos (= haplotipos distintos) S: número de sitios segregantes H: heterocigosidad π: diversidad nucleotídica

Calcular es un poco engorroso, por lo que le recomendamos completar la siguiente tabla como punto de partida:

Número de diferencias entre pares de alelos (haplotipos):

	Haplo_1	Haplo_2	Haplo_3	Haplo_4	Haplo_5
Haplo_2	1	-			
Haplo_3	3	2	-		
Haplo_4	40	39	39	-	
Haplo_5	47	46	44	7	-
Haplo_6	48	47	45	12	5

La tabla incluye datos de heterocigosidad del artículo, basados en los 11 loci microsatélites. Complete la tabla con los valores obtenidos anteriormente.

	Pine Forest	Warner Mts
ADN mitocondrial		
A		
S		
H		0.64
π		22.74

Microsatélites	Pine Forest	Warner Mts
$\bar{H}$	0.37	0.78

Ten en cuenta que usamos H para el ADNmt, ya que se trata de un solo locus, mientras que $\bar{H}$ para los microsatélites, ya que 0,37 es la heterocigosidad promedio de 11 loci de microsatélites.

¿Cómo se pueden interpretar estos resultados y qué nos indican sobre la variación entre subpoblaciones?

Frecuencias haplotípicas

$P = \frac{\text{Numero de diferencias}}{\text{Total de diferencias}}$

Pine forest

$P_1 = 1/10\ P_2 = 7/10\ P_3 = 2/10$

Warner Mts.

$P_1 = 5/10\ P_4 = 3/10\ P_5 = 1/10\ P_6 = 1/10$

Heterocigosidad (H)

Pine forest

$H = 1 - \sum_i P_i^2\ H = 1 - (0.1)^2 + (0.7)^2 + (0.2)^2$

Warner Mts.

$H = 1 - \sum_i P_i^2\ H = 1 - (0.5)^2 + (0.3)^2 + (0.1)^2 + (0.1)^2$

Diversidad nucleotídica (π)

Pine forest

$\pi = \sum_{i<j} 2P_iP_j\pi_{ij}$

haplotipo 1 vs 2 $\pi = 2(0.1)(0.7)1$

haplotipo 1 vs 3 $\pi = 2(0.1)(0.2)3$

haplotipo 2 vs 3 $\pi = 2(0.2)(0.7)2$

Warner Mts.

$\pi = \sum_{i<j} 2P_iP_j\pi_{ij}$

haplotipo 1 vs 4 $\pi = 2(0.5)(0.3)40$

haplotipo 1 vs 5 $\pi = 2(0.5)(0.1)47$

haplotipo 1 vs 6 $\pi = 2(0.5)(0.1)48$

haplotipo 4 vs 5 $\pi = 2(0.3)(0.1)7$

haplotipo 4 vs 6 $\pi = 2(0.3)(0.1)12$

haplotipo 5 vs 6 $\pi = 2(0.1)(0.1)5$

Resumen de resultados

	Pine Forest	Warner Mts
ADN mitocondrial		
A	3	4
S	3	50

	Pine Forest	Warner Mts
H	0.46	0.64
π	0.82	22.74

Microsatélites	Pine Forest	Warner Mts
$\bar{H}$	0.37	0.78

Mayor diversidad en Warner Mts.: Tanto el número de haplotipos (A), como el número de sitios segregantes (S), la heterocigosidad (H) y la diversidad nucleotídica (π) son mayores en Warner Mts. que en Pine Forest.

ADN mitocondrial vs microsatélites: El patrón es consistente: ambas fuentes de datos muestran más variabilidad genética en Warner Mts.

Implicancia biológica: La subpoblación Warner Mts. conserva mayor variabilidad genética, lo que sugiere un historial poblacional distinto (p. ej. mayor tamaño efectivo, menor efecto fundador o menos deriva genética). En contraste, Pine Forest muestra menor diversidad, posiblemente por un cuello de botella o mayor aislamiento.

5 Ejercicio 1: Frecuencia de alelo p=0.1

1. La frecuencia de un alelo autosómico A en una población es $p = 0.1$. ¿Cuál es el número esperado de alelos A en una muestra de 20 alelos? ¿Cuál es la probabilidad de no observar el alelo A en la muestra?

Para responder a estas preguntas, tratamos la extracción de cada alelo como un ensayo de Bernoulli, donde "éxito" es obtener el alelo A.

- ¿Cuál es el número esperado de alelos A en una muestra de 20 alelos? La esperanza de una distribución binomial se calcula como $E(X) = n \cdot p$. En este caso, el tamaño de la muestra (n) es 20 y la frecuencia alélica (p) es 0.1. $E(X) = 20 \cdot 0.1 = 2$ Se espera que, en promedio, haya 2 alelos A en una muestra de 20.
- ¿Cuál es la probabilidad de no observar el alelo A en la muestra? Aquí buscamos la probabilidad de obtener 0 éxitos (k=0) en 20 ensayos. Usamos la fórmula de la función de la probabilidad binomial: $P(X = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$. $P(X = 0) = \binom{20}{0}(0.1)^0 \cdot (1-0.1)^2 0 = 1 \cdot 1 \cdot (0.9)^2 0 \simeq 0.1216$. Hay un 12.16% de probabilidad de no encontrar ningún alelo A en la muestra.

6 Ejercicio 2: Simulación con rbinom

2. Utilizando la función rbinom, simule un gran número de muestras de tamaño $n = 20$ y caracterice la distribución de los resultados utilizando descriptores gráficos y numéricos apropiados.

El objetivo de este ejercicio es simular el proceso de muestreo miles de veces para ver cómo los resultados se distribuyen alrededor del valor esperado. rbinom en R genera números aleatorios basados en una distribución binomial.

- rbinom simula el número de alelos A en muestras de tamaño fijo.
- Esperamos ver que la media $\simeq 2$ y la desviación estándar $\simeq 1.34$.
- El histograma muestra que la mayoría de las muestras tienen entre 0 y 5 alelos A, con el 2 como valor más probable.

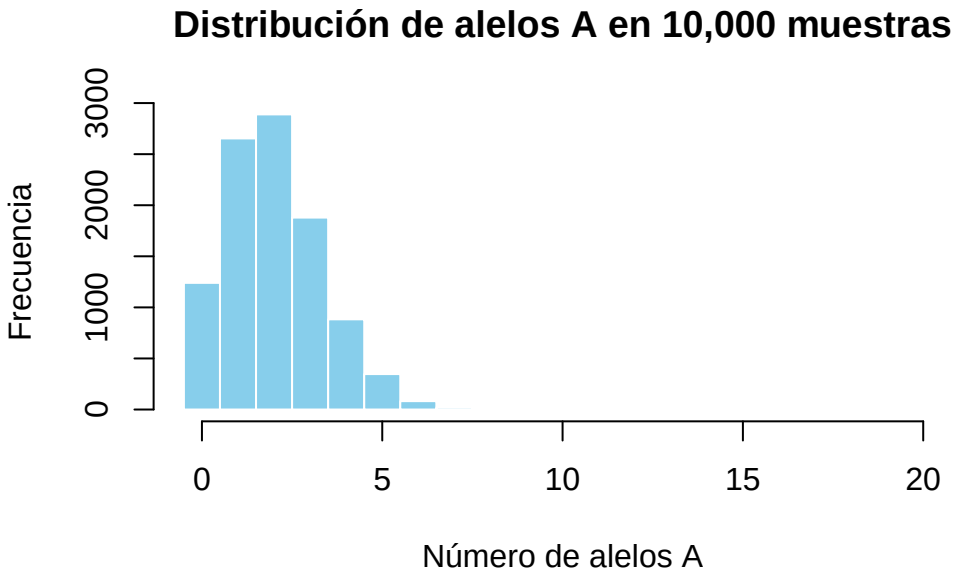
```
# Definir los parámetros
n <- 20  # Tamaño de la muestra
p <- 0.1  # Frecuencia del alelo A

# Simular 10,000 muestras
num_simulaciones <- 10000
muestras_simuladas <- rbinom(n = num_simulaciones, size = n, prob = p)

# Descriptores numéricos
media_simulada <- mean(muestras_simuladas)
print(paste("La media de las simulaciones es:", round(media_simulada, 2)))
```

```
[1] "La media de las simulaciones es: 2"
```

```
print(paste("La desviación estándar de las simulaciones es:", round(sd(muestras_simuladas), 2
```

```
[1] "La desviación estándar de las simulaciones es: 1.35"
```

```
# Gráfico de la distribución
hist(muestras_simuladas,
     breaks = seq(-0.5, n + 0.5, by = 1),
     col = "skyblue",
     border = "white",
     main = "Distribución de alelos A en 10,000 muestras",
     xlab = "Número de alelos A",
     ylab = "Frecuencia")
```



7 Ejercicio 3: Frecuencia de alelo p=0.01

3. Considere ahora el caso de $f(A) = p = 0.01$.¿Cuál es la probabilidad de observar el alelo A en una muestra de $n = 20$?

Para encontrar la probabilidad de observar el alelo A, es más fácil calcular la probabilidad de no observarlo (0 alelos) y restarla de 1. ¿Cuál es la probabilidad de que, en esas 20 copias, aparezca al menos una vez el alelo A?.

$P(observarA) = 1 - P(noobservarA)$ $P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{20}{0}(0.01)^0 \cdot (0.99)^20 = 1 - (0.99)^20 \simeq 1 - 0.8179 = 0.1821$. La probabilidad de observar al menos un alelo A es de aproximadamente 18.21%.

8 Ejercicio 4: Frecuencia observada de 3/10

4. La frecuencia observada de un alelo autosómico diploide A es 3/10 (tamaño de la muestra $n = 10$). Utilice dbinom para calcular la probabilidad de observar esta frecuencia en la muestra para un rango de frecuencias subyacentes en la población entre $0,01$ y $0,99$.

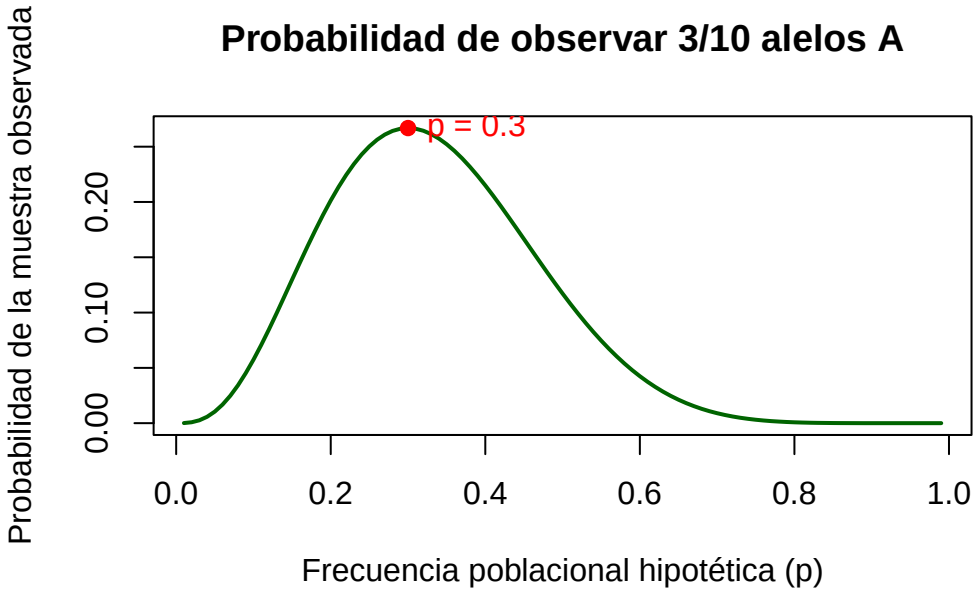
Aquí invertimos el problema: tenemos la muestra y queremos saber qué tan probable es esta observación para diferentes frecuencias poblacionales. Usamos dbinom, que calcula la probabilidad de un número específico de éxitos.

Tenemos una muestra de 10 copias de un gen, y en ellas observamos que 3 son del alelo A. La pregunta es: ¿qué tan probable es esta observación (3 de 10) si asumimos distintas frecuencias del alelo en la población?

La distribución binomial nos dice cuál es la probabilidad de observar un número exacto de “éxitos” (en este caso, alelos A) en un número fijo de ensayos (10 copias).

La función dbinom(k, n, p)

```
# Definir los parámetros de la muestra
k_observado <- 3
n_muestra <- 10

# Rango de frecuencias poblacionales para evaluar
frecuencias_poblacionales <- seq(0.01, 0.99, by = 0.01) # Esto permite que recorra el vector

# Calcular la probabilidad para cada frecuencia poblacional
probabilidades <- dbinom(x = k_observado, size = n_muestra, prob = frecuencias_poblacionales)

# Gráfico de la verosimilitud
plot(frecuencias_poblacionales,
     probabilidades,
     type = "l",
     lwd = 2,
     col = "darkgreen",
     main = "Probabilidad de observar 3/10 alelos A",
     xlab = "Frecuencia poblacional hipotética (p)",
     ylab = "Probabilidad de la muestra observada")

# Marcar el punto de máxima probabilidad
max_p <- frecuencias_poblacionales[which.max(probabilidades)]
points(max_p, max(probabilidades), col = "red", pch = 19)
text(max_p, max(probabilidades), labels = paste("p =", round(max_p, 2)), pos = 4, col = "red"
```



La curva muestra cómo cambia esa probabilidad cuando vamos variando p. La curva tiene un pico en $p = 0.3$, porque justamente $3/10 = 0.3$.

Lo que estás dibujando es la función de verosimilitud de una binomial:

$L(p) = P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$ con $k = 3$ y $n = 10$.

El valor de p que maximiza esta función se llama estimador de máxima verosimilitud (MLE).

Si derivamos respecto a p: $\frac{d}{dp}[p^k(1-p)n-k] = 0$ el resultado es $\widehat{p} = k/n$ En este caso seria $\widehat{p} = 3/10$ Esto significa que, si tu muestra observada son 3 éxitos en 10 intentos, el valor de p que hace esa observación más probable es exactamente la frecuencia muestral (la proporción observada).

9 Ejercicio 5: Proporciones de Hardy-Weinberg (HW)

5. Genotipos como muestras aleatorias de tamaño $n = 2$. Para un locus autosómico diploide con dos alelos, describa las proporciones de HW como esperanzas de la distribución binomial.

Las proporciones de Hardy-Weinberg describen la distribución de genotipos en una población en equilibrio. Podemos ver la formación de un genotipo como el resultado de tomar una muestra de 2 alelos, uno de cada padre.

Genotipo homocigoto dominante (AA): Para obtener este genotipo, ambos alelos deben ser A. La probabilidad de que esto ocurra es la de obtener 2 éxitos $(k = 2)$ en 2 ensayos, lo que es igual a p^2. $P(X = 2) = \binom{2}{2} \cdot p^2 \cdot (1-p)0 = p^2$. siendo que $\binom{2}{2} = \frac{2!}{2!(2-2)!} = 1$

Genotipo heterocigoto (Aa): Para obtener este genotipo, necesitamos 1 alelo A y 1 alelo a. La probabilidad de obtener 1 éxito (k=1) en 2 ensayos es $2p(1-p)$. $P(X = 1) = \binom{2}{1} \cdot p^1 \cdot (1-p)^1 = 2p(1-p)$. siendo que $\binom{2}{1} = \frac{2!}{1!(2-1)!} = \frac{2}{1} = 2$

Las proporciones de HW $(p2, 2pq, q2)$ son, de hecho, las esperanzas de una distribución binomial con n=2 ensayos, modelando la selección aleatoria de alelos para formar un genotipo.

10 Ejercicio 6: Frecuencia esperada de heterocigotos

Primero calculamos la frecuencia esperada de heterocigotos en la población usando las proporciones de HW, y luego usamos rbinom para simular su variación en muestras.

```
Frecuencia esperada de heterocigotos:
Si la frecuencia de un alelo es p=0.3, la frecuencia de heterocigotos en la población es 2p(1

Simulación de la variación:
El tamaño de la muestra es de 10 individuos, y la probabilidad de que un individuo sea hetero
```

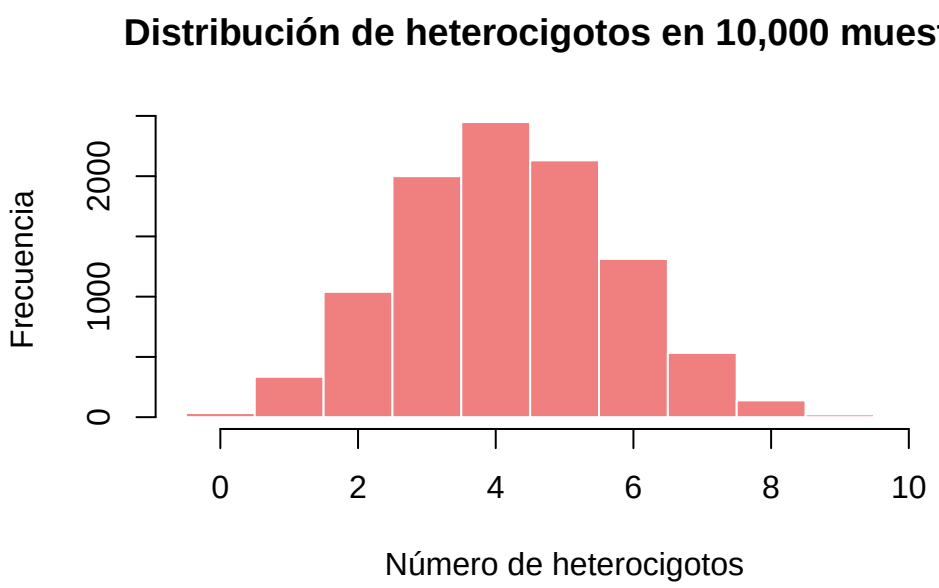
```r
# Definir los parámetros
p_alelo_A <- 0.3
prob_heterocigoto <- 2 * p_alelo_A * (1 - p_alelo_A)
n_individuos <- 10

# Simular la variación de la frecuencia de heterocigotos
num_simulaciones <- 10000
muestras_heterocigotos <- rbinom(n = num_simulaciones, size = n_individuos, prob = prob_heter

# Caracterizar la distribución
media_simulada_h <- mean(muestras_heterocigotos)
print(paste("La media de heterocigotos simulados es:", round(media_simulada_h, 2)))
```

```
[1] "La media de heterocigotos simulados es: 4.18"
```

```r
print(paste("La desviación estándar de las simulaciones es:", round(sd(muestras_heterocigotos
```

```
[1] "La desviación estándar de las simulaciones es: 1.56"
```

```r
# Gráfico de la distribución
hist(muestras_heterocigotos,
     breaks = seq(-0.5, n_individuos + 0.5, by = 1),
     col = "lightcoral",
     border = "white",
     main = "Distribución de heterocigotos en 10,000 muestras",
     xlab = "Número de heterocigotos",
     ylab = "Frecuencia")
```



11 Ejercicio 8: Simulación del proceso coalescente con n=4

8. Para $n = 4$, obtener en R una simulación de todos los componentes del proceso coalescente. Asumir que $N = 5 \times 10^5$ y $\mu = 10 - 5$. Para la realización obtenida, calcular π, S y θ_W.

Queremos simular el coalescente para n=4 alelos, con:

Tamaño efectivo de la población $N = 5 \times 10^5$

Tasa de mutación $\mu = 10^-5$

Y a partir de esa genealogía calcular:

π (diversidad nucleotídica promedio por par)

S (número de sitios segregantes)

θ_w (estimador de Watterson).

1. Para $n = 4$ existen dos topologias posibles que tienen sus probabilidades asociadas. Estas topologias se forman con la siguiente distribucion de alelos (2,2) con probabilidad $1/3$ y la distribucion (3,1) con la probabilidad de $2/3$. Esta ultima distribucion puede tener la bifurcacion mas reciente en uno de los ancestros hermanos por eso su probabilidad es de $2/3$. Para simular este proceso usaremos la funcion de R runif (random uniforme), esta funcion nos proporciona un muestreo de un numero al azar entre 0 y 1.

```
runif(1)
```

```
[1] 0.3121692
```

Si $runif > 0.33333$ nos quedamos con la topologia mas probable, si $runif \leq 0.33333$ nos quedamos con la manos probable de las topologias.

2. Los tiempos de coalescencia

Imaginemos que empezamos con $n = 4$ linajes. El proceso coalescente consiste en ver cuánto tardan en irse uniendo hasta llegar a un ancestro común.

La probabilidad de coalescencia por generación depende de:

El número de pares posibles: $\binom{n}{2} = \frac{n!}{2!(n-2)!} = n(n-1)/2$ Y de $1/(2N)$, que es la probabilidad de que un par coalesca en una generación. El tiempo de espera hasta el evento coalescente es el inverso de esa probabilidad.

- De 4 a 3 linajes: Pares posibles: $C(4, 2) = 6$ usando el coficiente binomial $\binom{4}{2} = \frac{4!}{2!(4-2)!}$ Probabilidad por generación: $\binom{4}{2}.\frac{1}{2N} = 6/(2N)$ Tiempo esperado: $E[T_4] = 2N/6 = 10^6/6 \approx 1.67 \times 10^5$
- De 3 a 2 linajes: Pares posibles: $C(3, 2) = 3$ usando el coficiente binomial $\binom{3}{2} = \frac{3!}{2!(3-2)!}$ Probabilidad por generación: $\binom{3}{2}.\frac{1}{2N} = 3/(2N)$ Tiempo esperado: $E[T_3] = 2N/3 \approx 3.33 \times 10^5$

- De 2 a 1 linajes (MRCA): Pares posibles: $C(2,2) = 1$ usando el coficiente binomial $\binom{2}{2} = \frac{2!}{2!(2-2)!}$ Probabilidad por generación: $\binom{2}{2}.\frac{1}{2N} = 1/(2N)$ Tiempo esperado: $E[T_2] = 2N = 10^6$

Entonces los tiempos de espera esperados son:

$T_4 \approx 1.67 \times 10^5$ $T_3 \approx 3.33 \times 10^5$ $T_2 = 10^6$

3. Tiempo total de ramas del árbol

Cada intervalo de tiempo tiene un cierto número de linajes que lo "recorren": $Ttotal = (n_1 \cdot T_4) + (n_2 \cdot T_3) + (n_3 \cdot T_2)$

- Con 4 linajes durante T_4 : $4 \cdot 1.67 \times 10^5 \approx 6.68 \times 10^5$
- Con 3 linajes durante T_3 : $3 \cdot 3.33 \times 10^5 \approx 9.99 \times 10^5$
- Con 2 linajes durante T_2 : $2 \cdot 10^6 = 2 \times 10^6$

Sumando: $Ttotal \approx 3.67 \times 10^6$

4. Número de sitios segregantes (S)

El número de sitios segregantes es simplemente: $S = Ttotal \cdot \mu$

$S \approx (3.67 \times 10^6) \cdot (10^-5) \approx 36.7$ mutaciones (sitios segregantes)

5. Cálculo de π (diversidad nucleotídica promedio por par)

π mide las diferencias promedio entre pares de secuencias.

En el coalescente, la esperanza de π es: $\pi = 4N\mu$

Con nuestros valores: $\pi = 4 \cdot (5 \times 10^5) \cdot 10^-5 = 20$

En la práctica, π dependerá de cómo se reparten las 37 mutaciones (sitios segregantes) en el árbol, pero el valor esperado es $\pi \approx 20$.

Si muchas mutaciones cayeron en ramas profundas (compartidas por muchos individuos), eso hará que muchos pares de secuencias difieran en esas posiciones π será grande. Si la mayoría de las mutaciones cayeron en ramas terminales (propias de un solo individuo), entonces solo unos pocos pares mostrarán esas diferencias π será más chico.

6. Cálculo de θ_W (Theta de Watterson)

Se define como: $\theta_W = \frac{S}{a_n}$ donde $a_n = \sum_{i=1}^{n-1} 1/i$ entonces $\theta_W = \frac{S}{\sum_{i=1}^{n-1} 1/i}$

Para $n = 4$: $a_4 = 1 + 1/2 + 1/3 \approx 1.833$

Entonces: $\theta_W = 37/1.833 \approx 20.18$

Resultado final

$S = 37$

$\pi \approx 20$

$\theta_W \approx 20.18$

```r
# ----------------------------
# Parámetros
# ----------------------------
N <- 5e5        # tamaño efectivo
mu <- 1e-5      # tasa de mutación
n <- 4          # número de secuencias

# ----------------------------
# 1. Simular tiempos de coalescencia
# ----------------------------
# Función para simular tiempo de espera (geométrico)
simular_T <- function(n_linajes, N) {
  pares <- choose(n_linajes, 2)
  p <- pares / (2 * N)     # probabilidad de coalescencia por generación
  rgeom(1, p)              # tiempo de espera ~ geométrica
}

# Guardamos los tiempos
T4 <- simular_T(4, N)
T3 <- simular_T(3, N)
T2 <- simular_T(2, N)

cat("T4 =", T4, "\n")
```

```
T4 = 223198
```

```r
cat("T3 =", T3, "\n")
```

```
T3 = 4178
```

```r
cat("T2 =", T2, "\n")
```

```
T2 = 25694
```

```r
# ----------------------------
# 2. Tiempo total de ramas
# ----------------------------
T_total <- (4 * T4) + (3 * T3) + (2 * T2)
cat("Tiempo total de ramas =", T_total, "\n")
```

```
Tiempo total de ramas = 956714
```

```r
# ----------------------------
# 3. Número de mutaciones (S)
# ----------------------------
S <- rpois(1, lambda = T_total * mu)  # Poisson con media T_total * mu
cat("Número de mutaciones S =", S, "\n")
```

```
Número de mutaciones S = 13
```

```r
# ----------------------------
# 4. Diversidad nucleotídica (□)
# ----------------------------
pi_est <- 4 * N * mu
cat("□ esperado =", pi_est, "\n")
```

```
□ esperado = 20
```

```r
# ----------------------------
# 5. Theta de Watterson (□W)
# ----------------------------
a_n <- sum(1 / (1:(n - 1)))
theta_W <- S / a_n
cat("□W estimado =", theta_W, "\n")
```

```
□W estimado = 7.090909
```

12 Ejercicio 9: Sobre el algoritmo MCMC (Markov chain Monte Carlo)

9. Structure: El programa structure se corrió con $K = 2$ y 90 genotipos, y la cadena de MCMC paró en la generación $m - 1$. Nos enfocamos en un sólo locus con alelos. El estado de la cadena en $m-1$ es el siguiente: Frecuencias del alelo de referencia en cada subpoblación: $P(m-1)1 = 0.5$, $P(m-1)2 = 0.8$. Ubicación de los genotipos a subpoblaciones:

	Subpoblación 1	Subpoblación 2
AA	13	25
Aa	24	13
aa	13	2
Total	50	40

a) Usando una distribución apropiada, obtener valores propuestos de para la P generación $m[P(m)]$ para cada subpoblación. Explicar, en cada caso, si deberían ser aceptadas como los nuevos valores $P(m)_1$ y $P(m)_2$ y con qué regla de probabilidad.

b) Ahora considerar un genotipo que puede moverse de la subpoblación 1 a la 2.⊠⊠ Explicar cómo se toma la decisión de moverlo o no. En las respuestas, incluir tanto las explicaciones como el código de R o los cálculos equivalentes.

Idea general de los pasos qué hay que hacer

Estás en una iteración del MCMC. Para cada subpoblación:

1. Proponés un nuevo valor de frecuencia alélica p' (por ejemplo: $p' \sim U(0,1)$)
2. Calculás la verosimilitud de los datos (los genotipos observados en esa subpoblación) con p' propuesto y con el p actual.
3. Aceptás p' como nuevo p con probabilidad $\alpha = \frac{p'}{p}$

a) Obtener y decidir sobre las nuevas frecuencias alélicas

El objetivo aquí es simular el Paso 1 del algoritmo MCMC, donde se actualizan las frecuencias alélicas de cada subpoblación.

Cálculo de Verosimilitud

La verosimilitud de los genotipos bajo una frecuencia alélica p puede escribirse de dos formas equivalentes:

1. Como producto de probabilidades de genotipos (modelo de Hardy–Weinberg).

$L(p) \propto [p^2]^{n_{AA}} \cdot [2p(1-p)]^{n_{Aa}} \cdot [(1-p)^2]^{n_{aa}}$

2. Como probabilidad binomial de observar k alelos A en $2n$ intentos, dado p.

Como la probabilidad binomial la subpoblación hay n individuos ($= 2n$ alelos), y observaste k alelos A, la probabilidad de ver k éxitos bajo p es

$Pr\ (K = k \mid p) = (\binom{2n}{k} p^k (1-p)^{2n-k}$

En **R** eso se calcula con **dbinom(k, size=2*n, prob=p)**.

Subpoblación 1

Datos de la **subpoblacion 1**: Frecuencia $P1\ = 0.5$, Genotipos: 13 AA, 24 Aa, 13 aa, $N = 50$, $2N = 100$

a mano la ecuacion seria $L(p) = Pr\ (datos \mid p) = \binom{(2 \cdot 50)}{50} 0.5^5 0 (1-0.5)^{(2 \cdot 50)-50}$ el propuesto $L(p') = Pr\ (datos \mid p') = \binom{(2 \cdot 50)}{50} p'^5 0 (1-p')^{(2 \cdot 50)-50}$

Numero de alelos $A : k = 2.13 + 24 = 50$

Recordemos:

Cada individuo tiene 2 alelos (porque estamos en organismos diploides). Un AA aporta 2 copias del alelo A. Entonces $A = 2 \times 13$

Un Aa aporta 1 copia de A y 1 de a. Entonces $A = 1 \times 24$

Un aa aporta 0 copias de A. Entonces $A = 0 \times 13$

En **Metropolis-Hastings**:

Si $\alpha \geq 1$, aceptás siempre. Si $\alpha < 1$, aceptás con probabilidad α.

Si $u < \alpha$, aceptás. Si $u \geq \alpha$, rechazás.

En este caso, α compara qué tan probable es observar los datos con la frecuencia propuesta p' respecto de la frecuencia actual p. Si α es grande (≥ 1), significa que p' explica igual o mejor los datos y se acepta siempre. Si $\alpha < 1$, se acepta con probabilidad proporcional a esa razón de verosimilitudes.

```
p1 <- 0.5 # La frecuencia actual
p2 <- runif(1)
p2 # La frecuencia propuesta
```

```
[1] 0.164172
```

```
# Verosimilitud bajo p1
conteo_p1 <- dbinom(50, 100, p1)
conteo_p1
```

```
[1] 0.07958924
```

```
# Verosimilitud bajo p2
conteo_p2 <- dbinom(50, 100, p2)
conteo_p2
```

```
[1] 7.493027e-15
```

```
# Razón de verosimilitudes (alpha)
alpha = conteo_p2/conteo_p1
alpha
```

```
[1] 9.414624e-14
```

```
# Decisión de aceptar o rechazar
u <- runif(1)
if (u < alpha) {
  decision <- "ACEPTAR"
} else {
  decision <- "RECHAZAR"
}
decision
```

```
[1] "RECHAZAR"
```

Como la probabilidad binomial la subpoblación hay n individuos ($= 2n$ alelos), y observaste k alelos A, la probabilidad de ver k éxitos bajo p es

$L(p) = Pr\ (K = k \mid p) = (\binom{2n}{k} p^k (1-p)^{2n-k}$

Subpoblación 2

Datos de la **subpoblacion 2**: Frecuencia $P1\ = 0.8$, Genotipos: 25 AA, 13 Aa, 2 aa, $N = 40$, $2N = 80$

En **R** eso se calcula con **dbinom(k, size=2*n, prob=p)**.

a mano la ecuacion seria $L(p) = Pr\ (datos \mid p) = \binom{(2 \cdot 63)}{63} 0.8^6 3 (1-0.8)^{(2 \cdot 63)-63}$ el propuesto $L(p') = Pr\ (datos \mid p') = \binom{(2 \cdot 63)}{63} p'^6 3 (1-p')^{(2 \cdot 63)-63}$

Numero de alelos $A : k = 2.25 + 13 = 63$

Recordemos:

Cada individuo tiene 2 alelos (porque estamos en organismos diploides). Un AA aporta 2 copias del alelo A. Entonces $A = 2 \times 25$

Un Aa aporta 1 copia de A y 1 de a. Entonces $A = 1 \times 13$

Un aa aporta 0 copias de A. Entonces $A = 0 \times 2$

En **Metropolis-Hastings**:

Si $\alpha \geq 1$, aceptás siempre. Si $\alpha < 1$, aceptás con probabilidad α.

Si $u < \alpha$, aceptás. Si $u \geq \alpha$, rechazás.

En este caso, α compara qué tan probable es observar los datos con la frecuencia propuesta p' respecto de la frecuencia actual p. Si α es grande (≥ 1), significa que p' explica igual o mejor los datos y se acepta siempre. Si $\alpha < 1$, se acepta con probabilidad proporcional a esa razón de verosimilitudes.

```r
p1 <- 0.8 # La frecuencia actual
p2 <- runif(1)
p2 # La frecuencia propuesta
```

```
[1] 0.4945759
```

```r
# Verosimilitud bajo p1
conteo_p1 <- dbinom(63, 100, p1)
conteo_p1
```

```
[1] 3.688153e-05
```

```r
# Verosimilitud bajo p2
conteo_p2 <- dbinom(63, 100, p2)
conteo_p2
```

```
[1] 0.002022911
```

```r
# Razón de verosimilitudes (alpha)
alpha = conteo_p2/conteo_p1
alpha
```

```
[1] 54.84889
```

```r
# Decisión de aceptar o rechazar
u <- runif(1)
if (u < alpha) {
  decision <- "ACEPTAR"
} else {
  decision <- "RECHAZAR"
}
decision
```

```
[1] "ACEPTAR"
```

b) Decidir si mover un genotipo

Este es el **Paso 2 del algoritmo MCMC**, donde se decide la reasignación de un individuo.

Genotipo considerado: Aa.

Subpoblaciones involucradas: De la **subpoblación 1 y 2**.

Frecuencias alélicas a usar: $P_{1m-1} = 0.5$ y $P_{2m-1} = 0.8$.

Cálculo de la Probabilidad de Pertenencia

El algoritmo compara la probabilidad del genotipo Aa en cada subpoblación, dada la frecuencia de alelos de cada una:

Probabilidad en **Subpoblación 1**: $Pr(Aa \mid Subpop1) = 2 \cdot P1 \cdot (1 - P1) = 2 \cdot 0.5 \cdot (1 - 0.5) = 0.5$

Probabilidad en **Subpoblación 2**: $Pr(Aa \mid Subpop2) = 2 \cdot P2 \cdot (1 - P2) = 2 \cdot 0.8 \cdot (1 - 0.8) = 2 \cdot 0.8 \cdot 0.2 = 0.32$.

El algoritmo no solo mueve el genotipo a la subpoblación más probable. Utiliza la misma regla de Metropolis-Hastings para decidir si aceptar o rechazar el movimiento.

Razon de verosimilitud

$\alpha = \frac{L(\text{genotipo en subpop 2})}{L(\text{Genotipo en subpop 1})}$ $\alpha = \frac{0.32}{0.50} = 0.64$.

En este ejemplo, $\alpha = 0.64$ significa que el genotipo Aa tiene un 64% de la probabilidad en la subpoblación 2 respecto de la subpoblación 1. Dicho de otra forma: la propuesta es menos consistente con los datos que el estado actual, pero no se descarta de inmediato: el algoritmo le da un 64% de chance de ser aceptada para permitir que la cadena explore también valores menos probables (evitando quedarse atrapada en un único modo).

Regla de Decisión **Metropolis-Hastings**:

Si $\alpha \geq 1$, aceptás siempre. Si $\alpha < 1$, aceptás con probabilidad α.

Si $u < \alpha$, aceptás. Si $u \geq \alpha$, rechazás.

```
# Probabilidad de observar un genotipo Aa en cada subpoblación
sub1H <- 2*0.5*(1-0.5)  # Subpoblación 1 (p = 0.5)
sub2H <- 2*0.8*(1-0.8)  # Subpoblación 2 (p = 0.8)

# Razón de verosimilitudes (alpha)
alpha <- sub2H / sub1H
alpha
```

```
[1] 0.64
```

```
# Decisión con Metropolis-Hastings
u <- runif(1)
if (u < alpha) {
  decision <- "ACEPTAR"
} else {
  decision <- "RECHAZAR"
}
decision
```

```
[1] "RECHAZAR"
```