3. Deriva genética y heterocigosidad

Enrique Lessa

2024-08-09

Introducción

La deriva genética es el cambio de frecuencias alélicas en la población a lo largo del tiempo debido al azar. En genética de poblaciones, existen varios modelos que especifican cómo se produce dicho cambio. Uno de los más usados es el modelo Wright-Fisher, desarrollado en forma más o menos paralela por Sewall Wright y Ronald Fisher, dos de los fundadores del campo. En dicho modelo, los alelos de una generación se obtienen por un muestreo al azar con reposición de los alelos de la generación precedente. Para un sistema diploide, la población tiene 2N alelos. Cada uno de los 2N alelos de la generación $t-1$ se tomaron al azar (imaginemos, para visualizarlo, uno a uno) de los 2N disponibles en la generación t_0. Cada uno de los 2N alelos de la generación parental en t_0 tiene igual probabilidad de ser muestreado al escoger uno de la generación $t-1$. Por lo tanto, la probabilidad de que un alelo sea de tipo A en t es igual a la frecuencia de A en t-1. El proceso se repite generación tras generación.

$$p_0 ==> p_1 ==> p_2 \ldots$$

El modelo básico de Wright-Fisher asume que no hay mutación, que no hay selección sobre el gen de interés, y que la población es homogénea (sin subdivisiones) y cerrada (no hay migración). Además, tiene un tamaño constante de N individuos. Para un sistema diploide, el número de alelos es 2N, pero el modelo puede aplicarse para cualquier sistema: es un modelo "haploide" en el sentido de que observamos los cambios en las frecuencias alélicas a lo largo del tiempo sin importarnos si se combinan para formar individuos (y en caso afirmativo de qué manera). Aquí usaremos 2N alelos porque asumimos en que estamos siguiendo un locus diploide autosómico.

Ejercicio 1. Completar la siguiente tabla, indicando el número de alelos que deben tenerse en cuenta para usar el modelo Wright-Fisher en loci con distintos modos de herencia en una población de N individuos de una especie de mamífero (diploide). Asumimos una proporción de sexos de 1:1.

Número de individuos	Locus autosómico	Locus mitocondrial	Cromosoma X	Cromosoma Y
N				

Distribución binomial

Antes de utilizar la distribución binomial para entender la deriva genética, puede ser recomendable estudiar la introducción disponible en "1. Distribución binomial en RStudio Cloud." Recordemos que hay dos formas de utilizar la binomial:

a) Usando la función *dbinom* ("d" de distribución) para calcular la probabilidad de observar de observar 0, 1, … n copias de A en una muestra de tamaño n, dada la frecuencia de A (p).

b) Usando la función *rbinom* ("r" de random) para simular el muestreo al azar de n copias de los alelos de una generación y registrar cuántos de ellos son de tipo A, dada la frecuencia de A en la población muestreada.

Para estudiar la deriva genética, usamos *dbinom* para calcular probabilidades de obtener un cierto número de copias de A en una generación a partir de la frecuencia de A en la generación precedente. A cada resultado posible (0, 1, ... n copias) asociamos de este modo una probabilidad.
Por otra parte, usamos *rbinom* para simular una o más realizaciones del proceso de muestreo de una generación a la siguiente. Si queremos simular el proceso a lo largo del tiempo, usaremos *rbinom* en cada paso (de la generación 0 a la 1, de 1 a la 2, de la 2 a la 3, etc.), actualizando los valores de p a lo largo del tiempo ($p_0, p_1, \ldots p_t$).

Ejemplo 1: distribución de probabilidad de frecuencias en una generación

Usamos *dbinom* para calcular las probabilidades como se explicó más arriba. Usamos $2N$ porque estamos estudiando un locus autosómico, con alelos A y a. Como estamos evaluando todos los resultados posibles, verificamos también que la suma de los *P(i)* es 1.

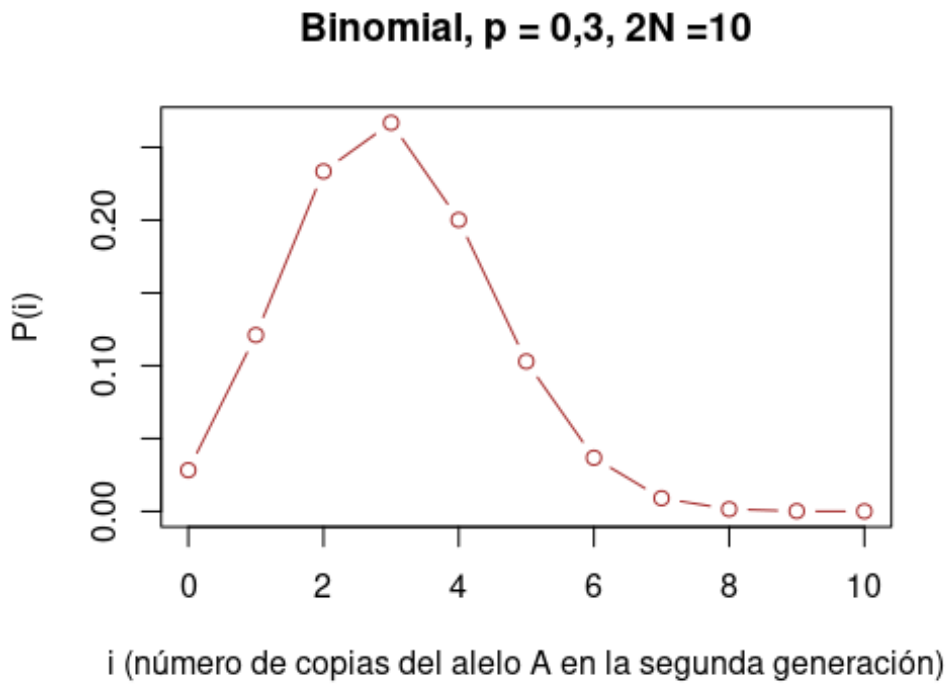
```
N  = 5  # número de individuos (diploides) en la población (asumimos
que N es constante).
pr = 0.3 # valor inicial de p (frecuencia del alelo A) en la
población

#A. Probabilidad de observar 0, 1, ... 2N copias del alelo A en una
muestra de n=10, dado que la frecuencia del
#   alelo en la población es pr
  serie= c(0:(2*N))
  dist = dbinom(serie, 2*N, pr) # c(0:2*N) es el rango de resultados
cuyas probabilidades queremos calcular
  print(dist)
```

```
##  [1] 0.0282475249 0.1210608210 0.2334744405 0.2668279320
0.2001209490
##  [6] 0.1029193452 0.0367569090 0.0090016920 0.0014467005
0.0001377810
## [11] 0.0000059049
```

```
  sum(dist) # como calculamos para todos los resultados posibles,
verificamos que la suma da 1
```

```
## [1] 1
```

```
#B. Graficamos los valores obtenidos
  plot(c(0:(2*N)), dist, type = "b",
       main = "Binomial, p = 0,3, 2N =10",
       xlab = "i (número de copias del alelo A en la segunda
generación)",
       ylab = "P(i)",
       col = "brown"
       )
```



Observamos que, si $0 < p_0 < 1, p_1$ puede tomar cualquier valor entre 0 1 y (el número de copias de A puede ser 0, 1, … 2N). Notemos también que el proceso de muestreo binomial se puede repetir de la generación 1 a la 2, y así sucesivamente. Las frecuencias seguirán fluctuando según las reglas de la distribución binomial, excepto si A se pierde (p=0) o se fija (p=1). En estos casos, como el modelo básico asume que no hay mutación, la variación se agota y no habrá más cambios en frecuencias. Estos estados se llaman estados absorbentes.

Ejercicio 2

a) Los valores iniciales usados para *dbinom* en el ejemplo de arriba son $N = 5$ y $p = 0.3$. ¿Cuál es la probabilidad de alcanzar un estado abosrbente en una generación en este caso?

b) usar el código provisto, modificando los valores iniciales, para averiguar cómo pueden modificarse los parámetros anteriores para aumentar o reducir la probabilidad de alcanzar un estado absorbente en una generación.

Muestras al azar y cambios a lo largo del tiempo

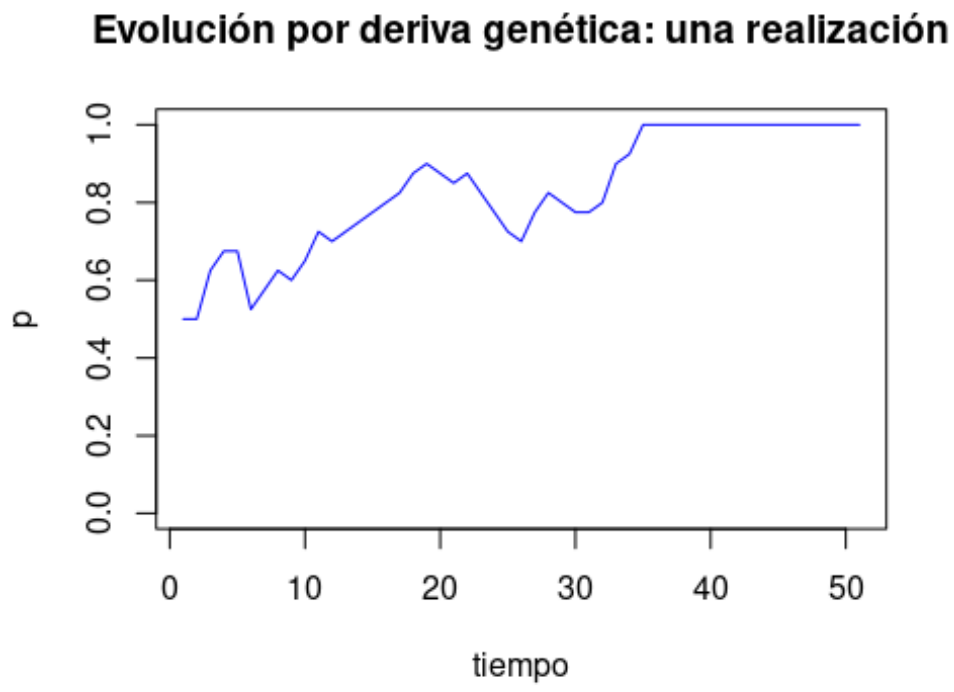
En esta aplicación de la binomial, en cada generación muestreamos todos los alelos de la población (*i=2N*) usando la frecuencia *p* de la generación inmediatamente precedente. Para obtener una realización aleatoria del proceso, usamos la función *rbinom*. para obtener trayectorias a lo largo del tiempo, definimos además el número de generaciones en las que queremos seguir el proceso.

```
# Condiciones de la simulación
  N  = 20 # número de individuos; como simulamos loci diploides, hay
2N alelos en la población
  p0 = 0.5 # frecuencia inicial de un alelo (A), cuya frecuencia en
```

```
la población seguimos a lo largo del tiempo.
  t   = 50 # número de generaciones
  pvec = as.vector(p0) # inicializando un vector de frecuencias

# Simulación
  p=p0 #frecuencia inicial
  for(i in seq(1:t)){
    p = rbinom(1, 2*N, p)/(2*N)
    # print(p)
    pvec = append(pvec, p)
  }
# Frecuencias inicial y final
  print(c(p0, pvec[length(pvec)]))
```

```
## [1] 0.5 1.0
```

```
# Gráfica de la trayectoria de evolución por deriva genética de la
simulación precedente
  plot(pvec, type = "l" ,main = "Evolución por deriva genética: una
realización", xlab = "tiempo", ylab = "p", ylim = range(0,1), col =
"blue")
```



Podemos usar la simulación de más arriba para ganar cierta intuición sobre las características de la deriva genética. Por ejemplo, podemos probar distintos tamaños poblacionales (modificando el valor de N) para ver cómo cambia el proceso. Del mismo modo, podemos experimentar con diferentes frecuencias iniciales o cambiar el número de generaciones.

A continuación se presenta código para comparar dos procesos de deriva con el mismo punto de partida e idénticas condiciones generales.

```r
# Condiciones de la simulación
  # Nota: tomamos los valores de N, p0 y t del bloque anterior
  # En cambio, creamos dos vectores para registrar las frecuencias a lo largo del tiempo, comenzando por asignarle a cada uno el valor de p0
    pvec1 = pvec2 = as.vector(p0) # inicializando dos vectores de frecuencias

# Simulación
  p1 = p2 = p0 # frecuencias iniciales

  for(i in seq(1:t)){
    p1 = rbinom(1, 2*N, p1)/(2*N) # muestreo de la binomial (en número de copias del alelo), divido por 2N para obener la frecuencia relativa
    pvec1 = append(pvec1, p1) # el valor obtenido de p1 se agrega al final del vector de frecuencias

    p2 = rbinom(1, 2*N, p2)/(2*N) # lo mismo de arriba pero para la segunda simulación
    pvec2 = append(pvec2, p2)
  }

# Frecuencias inicial y final
  print(c("p(0) =", p0, "p(t) =", pvec1[length(pvec1)]))
```

```
## [1] "p(0) =" "0.5"    "p(t) =" "0"
```

```r
  print(c("p(0) =", p0, "p(t) =", pvec2[length(pvec2)]))
```

```
## [1] "p(0) =" "0.5"    "p(t) =" "0"
```

```r
# Gráfica de la trayectoria de evolución por deriva genética de la simulación precedente
  plot(pvec1, type = "l" ,main = "Dos realizaciones delproceso de deriva genética", xlab = "tiempo", ylab = "p", ylim = range(0,1), col = "orange")
  lines(pvec2, col = "blue")
```

Comentarios

Las ideas principales que queremos reforzar simulando dos realizaciones (y repitiendo este experimento varias veces) del mismo proceso son:

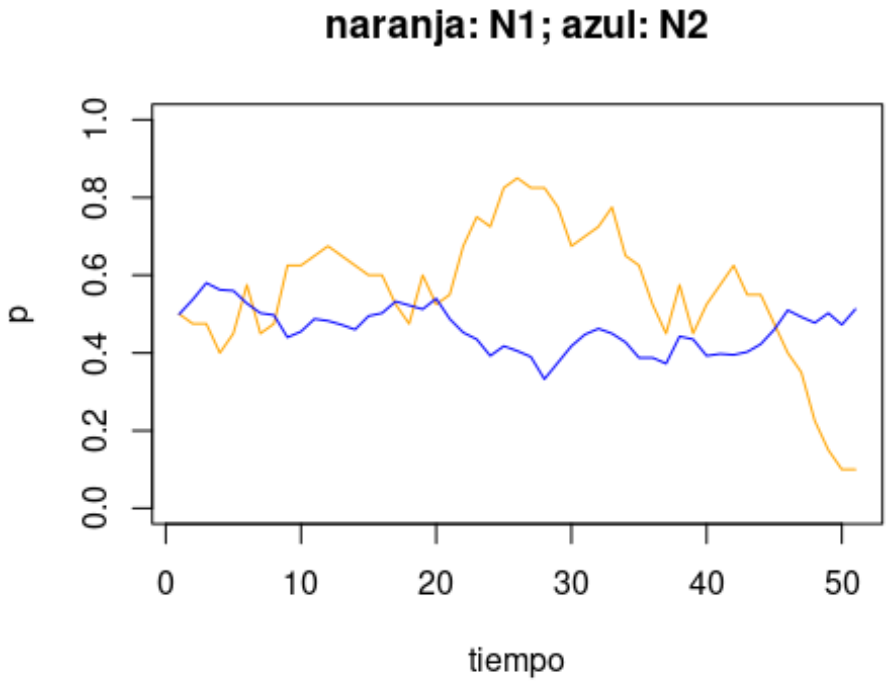
- Concepto de proceso aleatorio: el proceso tiene reglas probabilísticas, no deterministas, por lo que dos procesos con las mismas reglas y el mismo punto de partida tienen diferentes trayectorias. Se dice que son dos realizaciones de un mismo proceso.
- Proceso markoviano, sin memoria: en cada paso, el proceso depende únicamente de su estado (en nuestro caso, la frecuencia del alelo de referencia) y de las reglas para pasar de dicho estado al siguiente.
- Estados absorbentes: una vez que se llega a la fijación (p=1) o eliminación (p=0) del alelo, no hay más cambios; esos dos estados son absorbentes (se puede llegar a ellos, pero no se puede salir de ellos).

Notemos, de paso, que las trayectorias "azul" y "roja" son independientes, aunque les marcamos un mismo punto de partida y siguen las mismas reglas (muestreo binomial con reposición, idéntico tamaño poblacional). Estas dos trayectorias pueden pensarse como:

- Ejemplos de dos posibles trayectorias de la frecuencia de un mismo alelo, partiendo de p0. Si la línea azul representa la trayectoria de un alelo en una población real, podemos pensar en la línea roja como otra trayectoria igualmente probable… y podríamos seguir agregando más y más trayectorias.
- La historia de dos genes no ligados (esto es, con trayectorias independientes) en una misma población, con idénticas frecuencias iniciales.

Efecto del tamaño poblacional

```
# Condiciones de la simulación
  # Nota: tomamos los valores de p0 y t de los bloques previos

# Pero usaremos dos tamaños poblacionales diferentes:
N1 = 20
N2 = 200


  # En cambio, creamos dos vectores para registrar las frecuencias a lo largo del tiempo, comenzando por asignarle a cada uno el valor de p0
    pvec1 = pvec2 = as.vector(p0) # inicializando dos vectores de frecuencias

# Simulación
  p1 = p2 = p0 # frecuencias iniciales

  for(i in seq(1:t)){
    p1 = rbinom(1, 2*N1, p1)/(2*N1) # muestreo de la binomial (en número de copias del alelo), divido por 2N para obener la frecuencia relativa
    pvec1 = append(pvec1, p1) # el valor obtenido de p1 se agrega al final del vector de frecuencias

    p2 = rbinom(1, 2*N2, p2)/(2*N2) # lo mismo de arriba pero para la segunda simulación
    pvec2 = append(pvec2, p2)
  }

# Frecuencias inicial y final
  print(c("p(0) =", p0, "p(t) =", pvec1[length(pvec1)]))

## [1] "p(0) =" "0.5"    "p(t) =" "0.1"

  print(c("p(0) =", p0, "p(t) =", pvec2[length(pvec2)]))

## [1] "p(0) =" "0.5"    "p(t) =" "0.5125"

# Gráfica de la trayectoria de evolución por deriva genética de la simulación precedente
  plot(pvec1, type = "l" ,
       main = "naranja: N1; azul: N2",
       xlab = "tiempo", ylab = "p",
       ylim = range(0,1), col = "orange")
  lines(pvec2, col = "blue")
```



Heterocigosidad y panmixia

El término heterocigosidad se usa de un modo ligeramente distinto en diferentes contextos. Así, hablamos de heterocigosidad esperada H_e para referirnos a la frecuencia esperada de heterocigotas bajo el modelo Hardy-Weinberg. Dadas las frecuencias alélicas en la población $p_1, p_2, \ldots, p_k$,

$$H_e = \sum_{i<j}^{k} 2p_i p_j$$

Notamos, de paso, que la heterocigosidad esperada resulta del supuesto del modelo de que la población es panmíctica, es decir que los alelos se aparean al azar. La heterocigosidad observada H_o es la frecuencia observada de individuos heterocigotas. Comparar el ajuste entre H_e y H_o es, por tanto, una forma de evaluar la hipótesis de panmixia. Cualquiera haya sido el proceso de la población en el pasado, dado que las frecuencias alélicas son, en la generación actual, p_1, p_2, etc., bajo panmixia (apareamientos al azar) esperamos las frecuencias genotípicas predichas por el modelo Hardy-Weinberg y, si combinamos todos los genotipos heterocigotas, obtenemos H_e.

Ya señalamos que el modelo de Wright-Fisher no se ocupa de las combinaciones alélicas y es, en este sentido, "haploide": sigue las frecuencias de los alelos a lo largo del tiempo. Los loci que estudiamos pueden ser diploides, haploides, o estar localizados en el ADN mitocondrial, el cromosoma Y, etc. Una población bacteriana haploide se puede seguir igual que los sistemas diploides usados aquí, simplemente usando N en lugar de $2N$ alelos. Dicha población también tiene heterocigosidad, que definimos como la *probabilidad de que 2 alelos tomados al azar sean distintos (pertenezcan a distintas clases alélicas)*. Esta definición se aplica a cualquier población y locus, independientemente de

la ploidía y modo de herencia. En las clases teóricas, usamos H_D para destacar ese uso del término heterocigosidad cuando estamos considerando su evolución solamente por deriva genética. En la mayor parte de los textos científicos se usa simplemente H y el sentido se infiere del contexto.

Para el caso de un gen diploiode con dos alelos A y a, con frecuencias p y $p-1$, respectivamente en una generación determinada,

$$H_D = \sum_{i<1}^{k} 2p(1-p)$$

Del mismo modo, para k alelos con frecuencias $p_1, p_2, \ldots p_k$, en una generación dada.

$$H_D = \sum_{i<j}^{k} 2p_i p_j$$

En otras palabras, H_e nos dice la frecuencia esperada de heterocigotas bajo el modelo Hardy-Weinberg, y obviamente se aplica a un sistema diploide autosómico, mientras que H_D nos dice la probabilidad de que dos alelos tomados al azar sean diferentes, independientemente de la ploidía y modo de herencia, en una generación determinada. Si la población es de tipo Hardy-Weinberg, $H_e = H_{DM}$ porque los pares de alelos de los genotipos son tomados al azar de los alelos de la población.

Pérdida esperada de heterocigosidad por deriva genética

Para modelar el comportamiento de una población sometida a deriva genética, utilizamos el modelo de Wright-Fisher. Se trata de un modelo haploide, en el sentido de que describe cómo se muestrean los alelos de una población desde una generación a la siguiente. Si el gen de interés resulta está ubicado en un autosoma de organismos diploides, entonces podemos estudiar, como acabamos de plantear, cómo se combinan los alelos para formar los genotipos de la segunda generación. Pero el modelo funciona cualquiera sea la ploidía, y más en general el modo de transmisión del gen de interés (puede ser, por ejemplo, mitocondrial, o estar localizado en un cromosoma sexual; puede también tratarse de organismos haploides).

La deriva genética ocasiona fluctuaciones aleatorias (que modelamos, siguiendo a Wright-Fisher como un proceso de muestreo al azar con reposición de los alelos de una generación para formar la siguiente) de las frecuencias alélicas. Vamos a definir la heterocigosidad H de una población como la probabilidad de que dos alelos tomados al azar sean distintos, es decir pertenezcan a distintas clases de alelos. Bajo esta definición, existe heterocigosidad en la población aunque no haya heterocigotas, como en el caso de un sistema haploide. Es una definición general de una medida de la variación genética de la población.

Existen, naturalmente, otras medidas de variación. Por ejemplo, el número de clases alélicas k en la población es también una medida de variación, y tiene su interés. Sin embargo, la medida favorita es la heterocigosidad, por varias razones, que incluyen el hecho de que es una medida continua (entre 0 y 1), y captura más información sobre la variación genética. Para un gen con dos alelos, $k = 2$, pero H puede estar muy cerca de 0, si uno de los alelos tiene una frecuencia muy baja, o cerca de su máximo para $k = 2$,

que es 0.5. En el primer caso (H cercano a 0), la situación es la de un alelo casi fijado, y una variante rara que contribuye poco a la heterocigosidad.

La deriva genética es un proceso aleatorio que, mientras exista variación, puede hacer subir o bajar la frecuencia de un alelo cualquiera. En consecuencia, mientras H_0 sea mayor que 0 y menor que 1, H_1, al pasar de la generación en t_0 a t_1 H_1 puede ser mayor, igual o menor que H_0.

Sin embargo, la tendencia es hacia la pérdida de heterocigosidad. El caso extremo de pérdida o fijación de un alelo es obvio. Pero la tendencia vale en general, del modo que sigue:

Pensemos en la esperanza de H, que llamamos $E(H)$ en una generación a partir del valor de H en la generación precedente.

Si tomamos 2 alelos al azar, tenemos 2 situaciones resultantes:

1) Los dos alelos descienden de un mismo alelo en t_0; en ese caso son necesariamente idénticos (nuestro modelo no permite mutaciones), y esta situación ocurre con probabilidad $1/2N$. Si esta fuese la única situación, $H_1 = 0$ sin importar el valor de H_0.

2) Los dos alelos descienden de dos alelos distintos en t_0 con probabilidad que es el complemento de la anterior, o sea $1 - 1/2N$. Si esta fuese la única situación, $H_1 = H_0$.

Por definición, la esperanza es el producto de los valores que puede tomar la variable por sus probabilidades respectivas, de modo que:

$$\begin{aligned} E(H_1) &= 0 + (1 - 1/2N)H_0 \\ E(H_1) &= H_0(1 - 1/2N) \\ &= H_0 - H_0(1/2N) \end{aligned}$$

Es decir que, en promedio, H_1 es igual a H_0 menos una quita proporcional a H_0, en concreto el producto de H_0 y el inverso del número de alelos ($2N$) de la población.

Si tengo un gran número de loci, todos con el mismo valor H_0, cada uno puede tener un valor H_1 igual, mayor o menor que H_0 luego de una generación, pero el valor promedio (en el límite, si el número de loci tiende a infinito) tiende a la esperanza, que es menor que el valor inicial. De manera equivalente, si tomo un único locus y repito el proceso desde el mismo punto de partida un gran número de veces, el comportamiento promedio de H en esas distintas realizaciones del proceso tiende a la esperanza.

Naturalmente, si pasamos ahora de t1 a t2:

$$E(H_2) = H_1(1 - 1/2N)$$

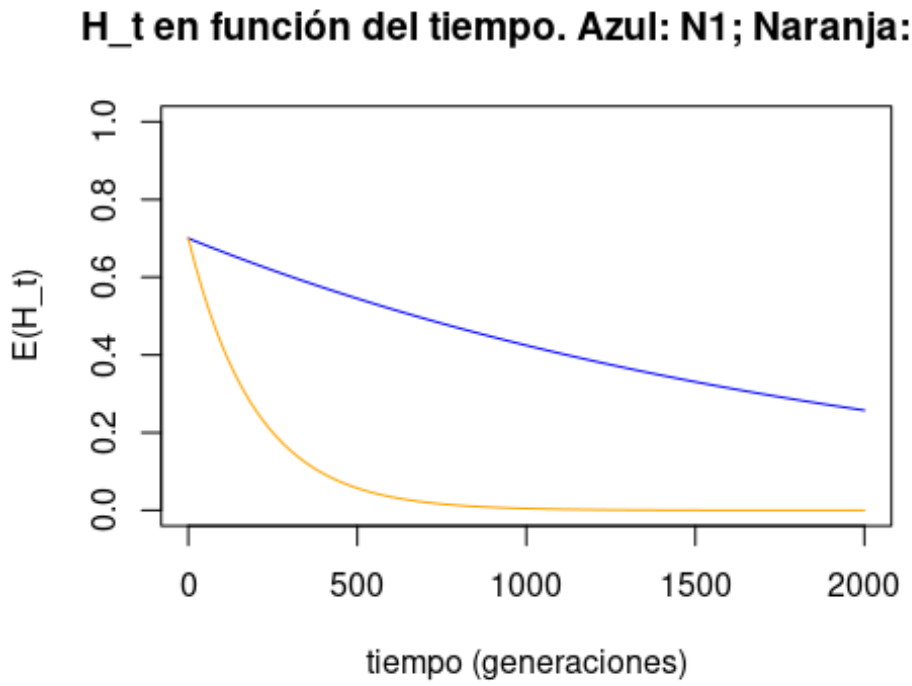
Reemplazando H_1 por $E(H_1)$, cuyo valor obtuvimos más arriba:

$$E(H_2) = H_0(1 - 1/2N)^2$$

Generalizando, luego de *t* generaciones

$$E(H_t) = H_0(1 - 1/2N)^t$$

```
    Ho = 0.7
    N1  = 1000
  curve( Ho*((1 - (1/(2*N1)))^x) , from= 0, to=(2*N1), n=300,
xlab="tiempo (generaciones)", ylab="E(H_t)",col="blue", main= "H_t
en función del tiempo. Azul: N1; Naranja: N2", N2, ylim = c(0, 1))

    N2 = 100
  curve( Ho*((1 - (1/(2*N2)))^x) , col="orange", add = TRUE  )
```



Notamos cómo la esperanza de la heterocigosidad decrece con el tiempo, y también cómo la curva de caída depende del tamaño poblacional. Observamos que, aún para una población de tamaño modesto, como $N1 = 1000$, la pérdida de heterocigosidad es muy lenta. Para ese tamaño poblacional, se requieren unas 2000 generaciones que la esperanza de H baje a la mitad.

Recursos adicionales

Existen muchos programas para simular procesos genético-poblacionales. Para una interfase amigable que se puede usar también para agregar selección natural, recomendamos la "shiny app" de $LearnPopGen$ (también disponible como paquete en R) que se encuentra en https://phytools.shinyapps.io/drift-selection/.

En los prácrticos usamos la página https://faculty.washington.edu/herronjc/a1/ (AlleleA1), de John Herron para realizar las simulaciones.

Ejercicios resueltos

Ejercicio 1. Completar la siguiente tabla, indicando el número de alelos que deben tenerse en cuenta para usar el modelo Wright-Fisher en loci con distintos modos de herencia en una población de N individuos de una especie de mamífero (diploide). Asumimos una proporción de sexos de 1:1.

Número de individuos	Locus autosómico	Locus mitocondrial	Cromosoma X	Cromosoma Y
N	2N	N/2	3N/2	N/2

Recordamos que aceptamos que la mitad de la población son machos y la mitad hembras. Notamos que el número de alelos (tamaño de la población para el modelo haploide de Wright-Fisher), tanto en loci mitocondriales como en el caso del cromosoma Y, es la cuarta parte del número de alelos de un locus autosómico (2N vs. N/2). Esos loci son transmitidos solamente por uno de los sexos (hembras para loci mitocondriales; machos para el cromosoma Y), que transmite una copia de dichos loci (esto merece cierta discusión en el caso del ADN mitocondrial).

En el caso del cromosoma X, hay 3 copias de X cada 4 de los loci autosómicos. Por tanto, para 2N autosomas el número de X es $2N(3/4) = 3N/2$.

Ejercicio 2

a) Los valores iniciales usados para *dbinom* en el ejemplo de arriba son $N = 5$ y $p = 0.3$. ¿Cuál es la probabilidad de alcanzar un estado absorbente en una generación en este caso?

El resultado se obtiene sumando las probabilidades de observar 0 o 2N (en el ejemplo 2N=10) alelos en una generación partiendo de $p = 0.3$. *Es decir 0.0282475249 + 0.0000059049. Notamos que el segundo valor, que es la probabilidad de pasar de 3 a 10 copias de A en una generación, es extremadamente bajo, o sea que la suma es casi igual al primer valor, que es la probabilidad de pasar de 3 a 0 copias de A en una generación.*

b) usar el código provisto, modificando los valores iniciales, para averiguar cómo pueden modificarse los parámetros anteriores para aumentar o reducir la probabilidad de alcanzar un estado absorbente en una generación.

El código se basa en dos valores, N y p (que llamamos "pr" en el código porque "p" tiene otros usos). Cambiando esos valores (de a uno o en combinación), es fácil constatar que los estados absorbentes se obtienen con mayor probabilidad reduciendo N y/o acercando p a uno de los valores extremos. De manera complementaria, se puede reducir la probabilidad de fijación o pérdida de A aumentando el tamaño de la población N y/o alejando p de 0 o 1, es decir acercándolo al valor medio de 0.5.