# 2. Equilibrio Hardy-Weinberg y endocría

Enrique Lessa

2024-08-09

## Introducción

Vamos a usar R, un lenguaje de programación de uso libre para estadística y un gran conjunto de aplicaciones relacionadas (análisis de datos, gráficas, presentación de resultados). Como interfase usamos RStudio, y para producir los archivos de salida usamos Rmarkdown. RStudio puede ser instalado en una computadora personal o utilizarse abriendo una cuenta en rstudio.cloud. En Rmarkdown, intercalamos texto común (incluyendo ecuaciones en Latex) con bloques ("chunks") de código (las líneas de código en R propiamente dichas). Si producimos una salida (en nuestro caso en html), esta intercalará texto, código y resultados del código.

El primer bloque de código activa los paquetes necesarios para las salidas en html.

```
library("knitr")
knitr::opts_chunk$set(echo = TRUE)
library("markdown")
library("lattice")
#no se si era necesario pero activé readbitmap y saqué las comillas
a RColorbrewer
library("readbitmap")
# algunos detalles (algo oscuros) para usar en lattice
library("RColorBrewer")
 MisColores = brewer.pal(6, "Accent")
  my.settings = list(col = MisColores[],
superpose.polygon=list(col=MisColores[1:3]),
strip.background=list(col=MisColores[6]))
```

## Distribución binomial y frecuencias alélicas

Ya realizamos una exploración de la distribución binomial, de la que solamente repetimos la introducción.

Si conocemos la frecuencia real del alelo $A_1$ *p=f(*$A_1$*)* en la población, podemos aplicar la binomial para calcular la probabilidad de observar *i* alelos de tipo A en una muestra de tamaño *n*. Como es lógico, dicha probabilidad depende de la frecuencia del alelo y del tamaño de la muestra. En concreto, toma la siguiente forma:

$$P(i) = \binom{n}{i} p^i (1-p)^{n-i}$$

,

$$i = 0, 1, \ldots, n$$

Notamos que:

$p^i$ es la probabilidad de muestrear *n* veces el alelo $A_1$.

$(1-p)^{n-i}$ es la probabilidad de muestrear *n-i* veces el alelo $A_1$.

$\binom{n}{i} = n!/[i!(n-i)!]$ es el número de formas de obtener *i* alelos de tipo A en una muestra de tamaño *n*.

Este es un ejercicio deductivo: dada una frecuencia alélica conocida (*p*) y un tamaño de muestra, deducimos con qué probabilidad podemos obtener todos los resultados posibles, desde *i=0* hasta *i=n*.

Por ejemplo, tomamos una muestra de *n=50* alelos, y queremos saber cual es la probabilidad de que 20 de ellos sean de tipo $A_1$ (*i=20*).

**Aplicación al modelo de Hardy-Weinberg**

Recordemos que, en el modelo de Hardy-Weinberg, consideramos una población de reproducción sexuada de tamaño infinito, sin inmigración, sin mutación, con frecuencias alélicas idénticas en los dos sexos, en la que los genotipos de una generación son combinaciones al azar de los alelos de la generación precedente. La oferta de alelos está dada por los gametos, en los que las frecuencias alélicas son idénticas en los dos tipos de gametos.

Para el caso del modelo original (con dos tipos de alelos), podemos aplicar la distribución binomial para formar pares de alelos, tomados al azar en base a la frecuencia de las clases alélicas en la población. En cada par, puede haber 0, 1 o 2 copias del alelo de referencia $A_1$, lo que corresponde a los genotipos $A_2A_2$, $A_1A_2$, y $A_1A_1$, respectivamente.

```
# Definimos aquí los parámetros para las secciones siguientes:
x1 = 50 # número de genotipos por muestra
n1 = 2  # tamaño de la muestra en cada réplica (en este caso, los
dos alelos que forman un genotipo diploide)
p1 = 0.4 # frecuencia del alelo A1 en la población [Nota: el código
actual, provisorio, funciona mejor lejos del 0 y del 1; ver #C más
abajo]
y = 15 # número de muestras (número de veces que repetiremos el
muestreo al azar en la sección #F y siguientes)

#A. Frecuencias esperadas relativas y absolutas:
  Esperadas_relativas = c(p1**2, 2*p1*(1-p1), (1-p1)**2) #
frecuencias relativas esperadas de A1A1, A1A2, y A2A2
  Esperadas = Esperadas_relativas * x1 # frecuencias absolutas de
los tres genotipos
 # print(Esperadas)

#B. Una muestra al azar de genotipos obtenida con la binomial
(función *rbinom*)
  Muestra_1 = rbinom(x1,n1,p1)
```

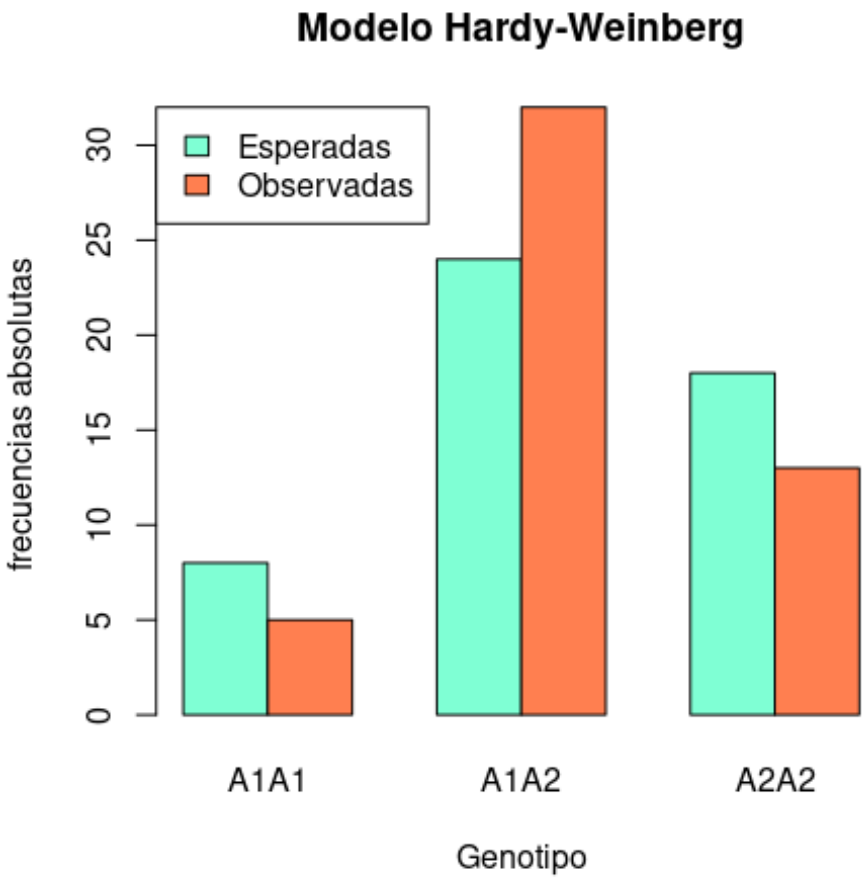
```r
  print(Muestra_1[1:30]) # examinamos los primeros 30 genotipos de
la muestra
```

```
##  [1] 0 1 2 1 1 1 1 1 1 0 0 0 1 1 1 1 2 1 1 1 0 1 0 0 1 0 1 1 0 1
```

```r
  # Vector de frecuencias en la muestra, en el orden A1A1, A1A2,
A2A2
  Observadas_1 = c(length(which(Muestra_1 == 2)),
                   length(which(Muestra_1 == 1)),
                   length(which(Muestra_1 == 0))
                    )
  # Nota: no usamos "table(Muestra_1)" porque solamente cuenta los
casos representados;
  # si falta uno de los genotipos, recuperará solamente la
frecuencia de los otros dos y la tabla estará incompleta.

#C. Combinamos las frecuencias observadas y esperadas en una misma
tabla y graficamos los resultados
  Resumen_1 = rbind(Esperadas, Observadas_1)
  colnames(Resumen_1) = c("A1A1", "A1A2", "A2A2")
  rownames(Resumen_1) = c("Esperadas HW", "Observadas")

  print(Resumen_1)
```

```
##              A1A1 A1A2 A2A2
## Esperadas HW    8   24   18
## Observadas      5   32   13
```

```r
  barplot(Resumen_1,
          main = "Modelo Hardy-Weinberg",
          xlab = "Genotipo",
          ylab = "frecuencias absolutas",
          col = c("aquamarine","coral"),
          beside = TRUE
    )

legend("topleft",
c("Esperadas", "Observadas"),
fill = c("aquamarine","coral")
)
```




**Usando la función "sample"**

En el siguiente bloque repetimos el muestreo del anterior con un código ligeramente distinto:
- definimos los objetos (genotipos) a ser muestreados; - definimos las probabilidades correspondientes usando las frecuencias "Esperadas_relativas" definidas; - usamos los genotipos como objetos de muestra en la función *sample*; - (notar la opción "replace = TRUE", que corresponde al muestreo con reposición [indica que las probabilidades se mantienen constantes: no cambian a medida que vamos muestreando los distintos genotipos].

```
#D1. Una muestra al azar, usando la función *sample* de manera
equivalente a *rbinom*.
  Genotipos = c("A1A1", "A1A2", "A2A2") # creamos la lista de
genotipos de interés
  Muestra_2 = sample(Genotipos,  x1, replace = TRUE,
Esperadas_relativas) # obtenemos una muestra de tamaño x1,
muestreando al azar con reposición los genotipos, cada uno con una
probabilidad definida en #A (Esperadas_relativas)
 # Observadas_2 = table(Muestra_2) # tabla de frecuencias en la
muestra
#  print(Observadas_2)

#D2. Obtenemos un vector con los conteos de los tres genotipos
```

```r
  Observadas_2 = c(length(which(Muestra_2 == "A1A1")),
                   length(which(Muestra_2 == "A1A2")),
                   length(which(Muestra_2 == "A2A2"))
                    )

#E. Combinamos las frecuencias observadas y esperadas en una misma
tabla
  Resumen_2 = rbind(Esperadas, Observadas_2)
  colnames(Resumen_2) = c("A1A1", "A1A2", "A2A2")

  print(Resumen_2)
```

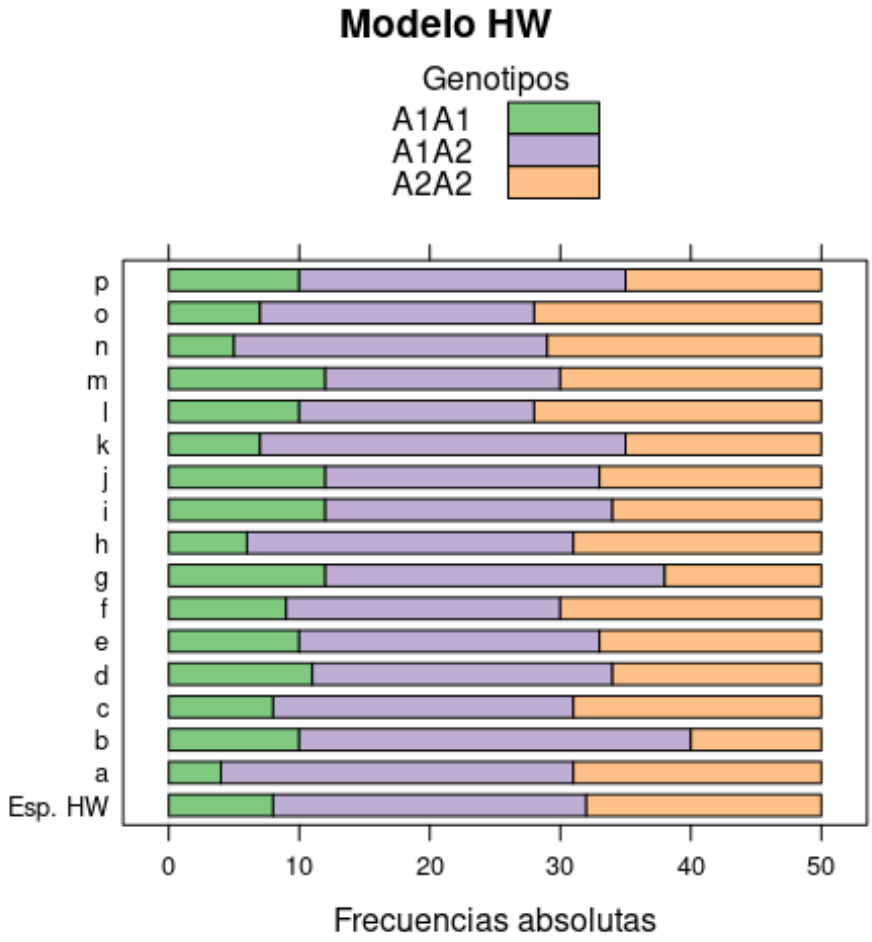
```
##              A1A1 A1A2 A2A2
## Esperadas       8   24   18
## Observadas_2    4   27   19
```

```r
#F. COMBINANDO MÚLTIPLES MUESTRAS ADICIONALES

 # Creamos una matriz para incluir los datos de "Resumen_2" y
agregarle luego *y* muestras adicionales
  Resumen_y = matrix(data = NA, nrow = y+2 , ncol= 3)
  colnames(Resumen_y) = c("A1A1", "A1A2", "A2A2")
  rownames(Resumen_y) = c("Esp. HW", letters[1:(y+1)]) # la primera
fila toma el nombre "Esp", y las siguientes se etiquetan
                                                      # con letras
hasta el valor definido en "y", usado más abajo para
                                                      # replicar los
muestreos
 # Comenzamos incorporando las observadas y esperadas generadas más
arriba
  Resumen_y[1,] = Esperadas
  Resumen_y[2,] = Observadas_2

 # Luego usamos un *for loop* para agregar *y* muestras adicionales
  for (i in seq(1:y)){
    Muestra_y = sample(Genotipos,  x1, replace = TRUE,
Esperadas_relativas) # obtenemos una muestra de tamaño x1,
muestreando al azar con reposición los genotipos, cada uno con una
probabilidad definida más arriba (Esperadas_relativas)
    Observadas_y = c(length(which(Muestra_y == "A1A1")),
                   length(which(Muestra_y == "A1A2")),
                   length(which(Muestra_y == "A2A2"))
                    )
     Resumen_y[i+2,] = Observadas_y
    }

    print(Resumen_y)
```

```
##          A1A1 A1A2 A2A2
## Esp. HW     8   24   18
## a           4   27   19
## b          10   30   10
```

```
## c           8   23   19
## d          11   23   16
## e          10   23   17
## f           9   21   20
## g          12   26   12
## h           6   25   19
## i          12   22   16
## j          12   21   17
## k           7   28   15
## l          10   18   22
## m          12   18   20
## n           5   24   21
## o           7   21   22
## p          10   25   15
```

```
# Pruebas con gráficas básicas de "lattice"
   Fig_1 = barchart(Resumen_y[1:(y+2),],
          xlab = "Frecuencias absolutas",
          main = "Modelo HW",
          auto.key= list(space = "top", columns = 1, points = FALSE,
          rectangles = TRUE, title = "Genotipos", cex.title = 1),
          panel = lattice.getOption("panel.barchart"),
          default.prepanel =
lattice.getOption("prepanel.default.barchart"),
          box.ratio = 2,
          par.settings = my.settings    )
  plot(Fig_1)
```




**Incorporando la endocría**

Muchas poblaciones reales se apartan de la panmixia, de modo que los apareamientos ocurren con una mayor probabilidad entre individuos emparentados. Aunque con frecuencia la reproducción entre individuos fuertemente emparentados se evitan, efectos de vecindario, organización social, y otros tienen a hacer que los apareamientos entre individuos con un parentesco mayor que el promedio de la población sean más frecuentes de lo esperado por azar.

Para incorporar de manera sencilla este fenómeno general (sin obligarnos por ello a estudiar pedigrís de manera directa), introducimos *F*, el coeficiente de endocría (o endogamia) de la población. Se trata de un único parámetro adicional que procura capturar el efecto neto de la endocría sobre las frecuencias genotípicas esperadas.

Si los alelos se aparean con sus símiles ($A_1$ con $A_1$, y $A_2$ con $A_2$) con probabilidad *F*, entonces aumentarán las frecuencias de homocigotas $A_1A_1$ y $A_2A_2$ a expensas de las frecuencias de heterocigotas ($A_1A_2$).

Planteamos ahora lo razonado más arriba de manera explícita:

(1) Una fracción *1-F* de las combinaciones gaméticas se realizan al azar, produciendo los siguientes resultados parciales:

$$frec.\left(A_1A_1\right) = p^2(1-F)$$

$$frec.\left(A_1A_2\right) = 2pq(1-F)$$

$$frec.\left(A_2A_2\right) = q^2(1-F)$$

(2) Por otra parte, la restante fracción *F* de las combinaciones gaméticas corresponden a la endocría, es decir que no son al azar, sino que combinan gaméticos idénticos. Por tanto, estas combinaciones estarán en proporción a las frecuencias de los alelos, y serán *pF* y *qF* para los genotipos AA y aa, respectivamente. En combinación con el resultado anterior, obtenemos:

$$frec.\left(A_1A_1\right) = p^2(1-F) + pF$$

$$frec.\left(A_1A_2\right) = 2pq(1-F)$$

$$frec.\left(A_2A_2\right) = q^2(1-F) + qF$$

Sumando las frecuencias, verificamos el resultado:

$$p^2(1-F) + 2pq(1-F) + q^2(1-F) + pF + qF = (1-F)\left(p^2 + 2pq + q^2\right) + F\left(p+q\right) = 1$$

Notamos también, de paso, que la expresión de frecuencias esperadas obtenida en (2) es también válida cuando F = 0, en cuyo caso se simplifica a la expresión en (1).

Podemos hacer una simulación de muestreo de genotipos como la de más arriba, solamente que introduciendo a *F* para ponderar las probabilidades asociadas a cada genotipo.

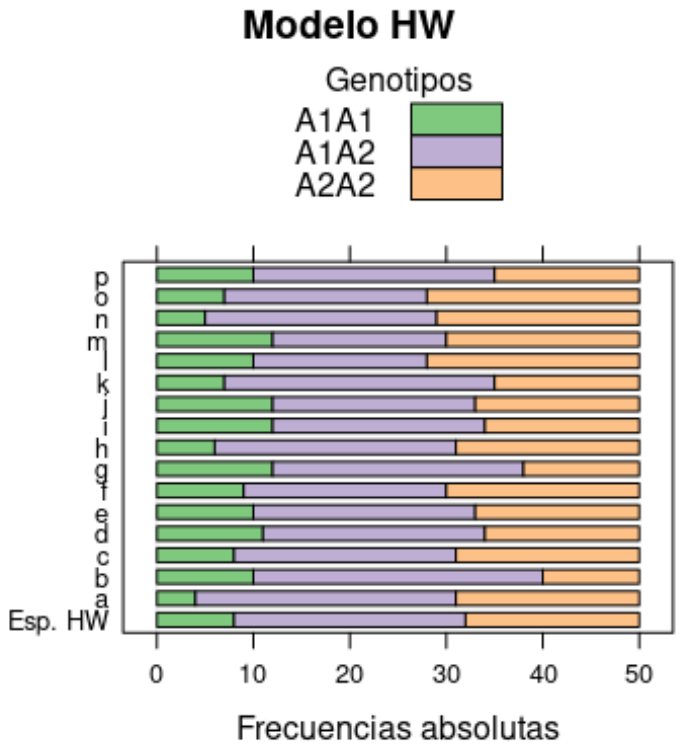
```
# Definimos el parámetro adicional F
 F = 0.40 # usamos un F alto para poner en evidencia el fenómeno

#G. Una muestra al azar, usando la función *sample*
  Genotipos = c("A1A1", "A1A2", "A2A2") # creamos la lista de
genotipos de interés
  Esperadas_relativas_F = c(p1**2*(1-F) + p1*F, 2*p1*(1-p1)*(1-F),
(1-p1)**2*(1-F) + (1-p1)*F)
  Muestra_3 = sample(Genotipos,  x1, replace = TRUE,
Esperadas_relativas_F) # obtenemos una muestra de tamaño x1,
muestreando al azar con reposición los genotipos, cada uno con una
probabilidad definida más arriba (Esperadas_relativas_F)
  Observadas_3 = c(length(which(Muestra_3 == "A1A1")),
                   length(which(Muestra_3 == "A1A2")),
                   length(which(Muestra_3 == "A2A2"))
                  )
#H. Combinamos las frecuencias observadas y esperadas en una misma
tabla
  Resumen_3 = rbind(Esperadas, Observadas_3)
  print(Resumen_3)

##              [,1] [,2] [,3]
## Esperadas       8   24   18
## Observadas_3    9   20   21
```

```r
#I. COMBINANDO MÚLTIPLES MUESTRAS ADICIONALES ()
  # Creamos una matriz para incluir los valores observados y
esperados recién generados y otros nuevos
  Resumen_z = matrix(data = NA, nrow = y+2 , ncol= 3)
  colnames(Resumen_z) = c("A1A1", "A1A2", "A2A2")
  rownames(Resumen_z) = c("Esp. HW", letters[1:(y+1)]) # la primera
fila toma el nombre "Esp", y las siguientes se etiquetan
                                                    # con letras
hasta el valor definido en "y", usado más abajo para
                                                    # replicar los
muestreos
  Resumen_z[1,] = Esperadas
  Resumen_z[2,] = Observadas_3


  # Usamos un *for loop* para agregar *y* muestras al azar
adicionales
  for (i in seq(1:y)){
    Muestra_z = sample(Genotipos,  x1, replace = TRUE,
Esperadas_relativas_F) # obtenemos una muestra de tamaño x1,
muestreando al azar con reposición los genotipos, cada uno con una
probabilidad definida más arriba (Esperadas_relativas)
    Observadas_z = c(length(which(Muestra_z == "A1A1")),
                     length(which(Muestra_z == "A1A2")),
                     length(which(Muestra_z == "A2A2"))
                      )
    Resumen_z[i+2,] = Observadas_z
   }
    print(Resumen_z)
```

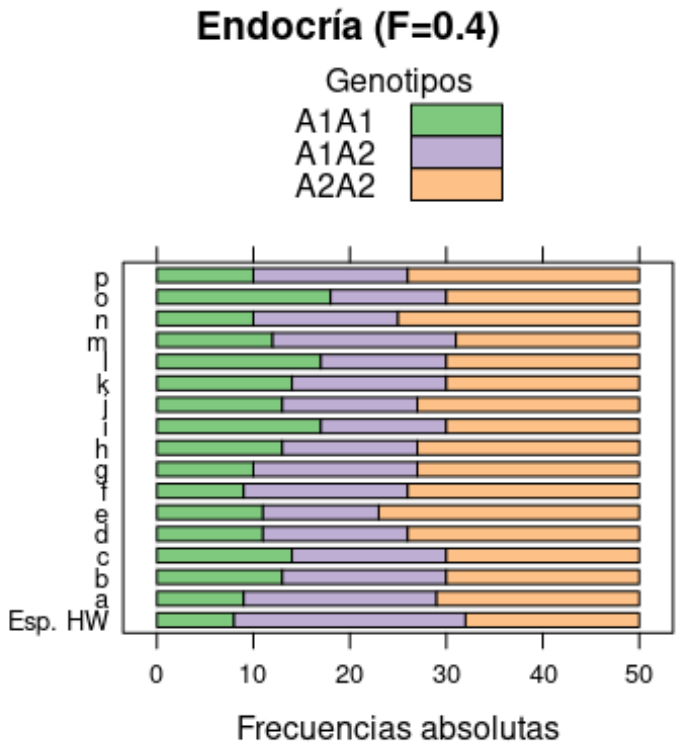
```
##         A1A1 A1A2 A2A2
## Esp. HW    8   24   18
## a          9   20   21
## b         13   17   20
## c         14   16   20
## d         11   15   24
## e         11   12   27
## f          9   17   24
## g         10   17   23
## h         13   14   23
## i         17   13   20
## j         13   14   23
## k         14   16   20
## l         17   13   20
## m         12   19   19
## n         10   15   25
## o         18   12   20
## p         10   16   24
```

```r
# Pruebas con gráficas básicas de "lattice"
   Fig_2 = barchart(Resumen_z[1:(y+2),],
```

```
         xlab = "Frecuencias absolutas",
         main = "Endocría (F=0.4)",
         auto.key= list(space = "top", columns = 1, points = FALSE,
         rectangles = TRUE, title = "Genotipos", cex.title = 1),
         panel = lattice.getOption("panel.barchart"),
         default.prepanel =
lattice.getOption("prepanel.default.barchart"),
         box.ratio = 2,
         par.settings = my.settings    )

  plot(Fig_1)
```




```
  plot(Fig_2)
```




Observamos, en general, un déficit de heterocigotas, aunque notamos que para que sea visible en una muestra relativamente pequeña definimos un valor de *F* alto.

**Coeficiente de endocría observado**

La frecuencia observada de heterocigotas $H_o$ nos permite estimar $F$, puesto que:

$$f(A_1A_2) = 2pq\left(1 - \hat{F}\right)$$

$$H_o = H_e\left(1 - \hat{F}\right)$$

(Notar que usamos $\hat{F}$ para indicar que vamos a obtener una estimación de $F$, cuyo verdadero valor desconocemos).

Despejando, tenemos que

$$\hat{F} = \left(H_e - H_o\right)/H_e$$

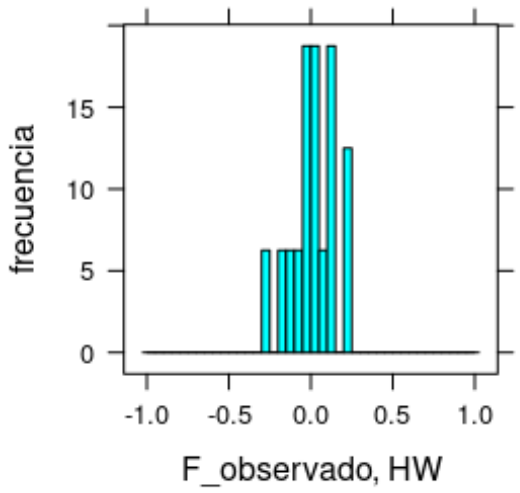
En palabras, el coeficiente de endocría es la diferencia entre la heterocigosidad esperada y la observada, normalizada al dividirla por la esperada. Cuando *F* es positivo, observamos menos heterocigosidad que la esperada por HW (endocría o endogamia). Cuando *F* es negativo, hay más heterocigosidad que la esperada por HW (exogamia). En condiciones de panmixia, $F = 0$.

Notamos, de paso, que F puede entenderse como un descriptor del apartamiento del equilibrio HW, independientemente de la causa. En ausencia de selección, F resulta de apareamientos no aleatorios (tal y como se dedujo más arriba), pero la selección natural

también puede, como veremos más adelante en el curso, causar apartamientos del equilibrio HW.

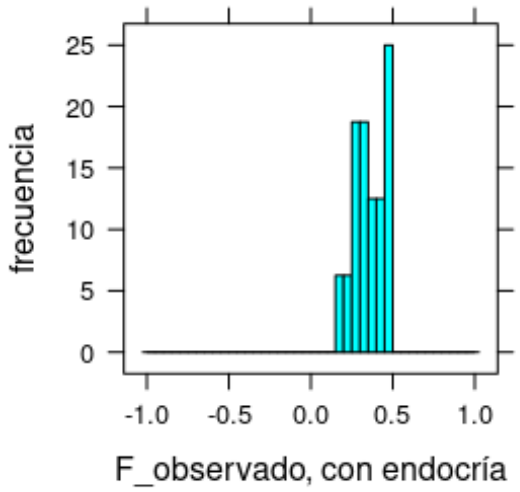
A continuación agregaremos las estimaciones de *F* para cada una de las muestras de genotipos obtenidas más arriba: una de ellas corresponde a muestreos realizados bajo el modelo de Hardy-Weinberg, y la otra a muestreos en los que las frecuencias esperadas están ajustadas asumiendo un valor de $F \neq 0$.

```
# Añadimos los valores observados de F a cada matriz, y, en cada
matriz,
# eliminamosla primera fila, que correspondía a los valores
esperados por HW

  Resumen_y = cbind (Resumen_y, (Resumen_y[1,2]-Resumen_y[,2]) /
Resumen_y[1,2])
  colnames(Resumen_y)[4] = "F_obs_HW"
  Resumen_y = Resumen_y[-1,]

  Resumen_z = cbind (Resumen_z, (Resumen_z[1,2]-Resumen_z[,2]) /
Resumen_z[1,2])
  colnames(Resumen_z)[4] = "F_obs_endocría"
  Resumen_z = Resumen_z[-1,]



# Graficamos los valores de F observados sin y con endocría
  histogram(Resumen_y[,4],
            xlab = "F_observado, HW",
            ylab = "frecuencia",
            main = "",
            breaks = seq(from = -1, to = 1, by = 0.05)
            )
```




```
  histogram(Resumen_z[,4],
            xlab = "F_observado, con endocría",
```

```
            ylab = "frecuencia",
            main = "",
            breaks = seq(from = -1, to = 1, by = 0.05)
            )
```




Observamos, naturalmente, variación al azar en los valores observados de *F* en cada realización, tanto en aquellas obtenidas por muestreo en base al modelo de Hardy-Weinberg (*F = 0*)como en las realizadas con un coeficente de endocría $F \neq 0$.

Sin embargo, las realizaciones en el primer caso varían en torno a cero, mientras que las segundas varían en torno al valor de *F* elegido para la simulación correspondiente. Notamos también, de paso, dos cuestiones más:

1) Focalizarnos en *F* permite visualizar los apartamientos de lo esperado de manera eficiente, sintetizando en un único valor una característica importante del régimen de apareamientos.

2) Al mismo tiempo, *F* es la diferencia entre frecuencias esperadas y observadas de heterocigotas, normalizada al dividir dicha diferencia por la frecuencia esperada. *F = 0.1* indica un 10% de déficit de heterocigotas, independientemente de si la frecuencia esperada es 5%, 50% o 70%. Los valores de *F* son comparables entre genes, aunque estos varíen en número de clases alélicas y frecuencias asociadas.