Equilibrio Hardy-Weinberg y endocría

Curso de Evolución 2019

16/04/2019

#Equilibrio Hardy-Weinberg

##Introducción

Vamos a usar R, un paquete de programas de uso libre para estadística y un gran conjunto de aplicaciones relacionadas (análisis de datos, gráficas, presentación de resultados). Como interfase usamos RStudio, ampliamente utilizado, y para producir los archivos de salida usamos Rmarkdown. 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 "knitr" y "markdown", 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.b
ackground=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, ..., 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. $\binom{n}{i}$ es el número de formas de obtener los resultados anteriores.

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*.

Aplicación al modelo de Hardy-Weinberg

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*).

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 ge
notipo diploide)
p1 = 0.6 # frecuencia del alelo A1 en la población [Nota: el código actual, provisorio, funci
ona 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 secció
n #F y siguientes)
```

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

```
## [1]  8 24 18
```

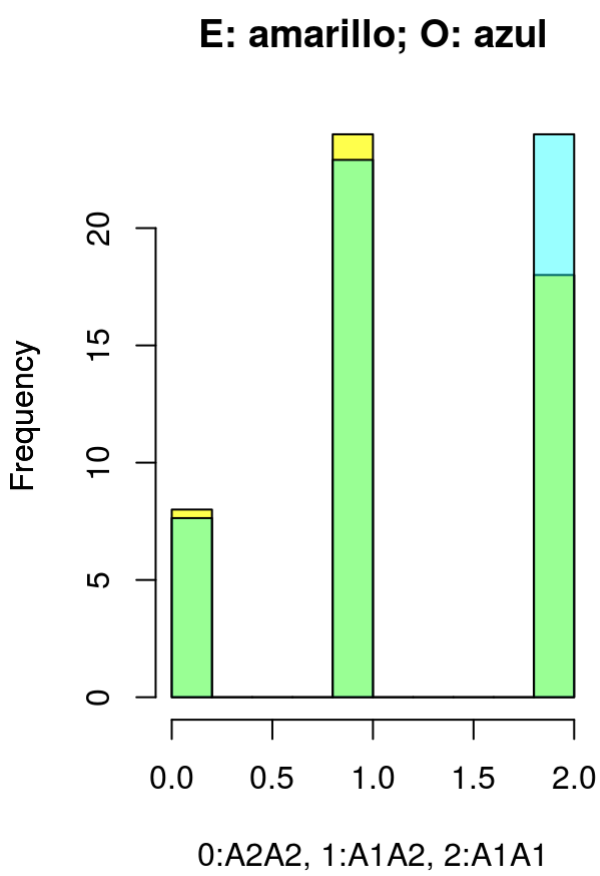
```
#B. Una muestra al azar de genotipos obtenida con la binomial (función *rbinom*)
  Muestra_1 = rbinom(x1,n1,p1)
  print(Muestra_1)
```

```
##  [1] 0 2 1 1 1 2 1 2 1 1 1 1 2 0 2 2 0 2 1 2 2 1 2 2 2 2 1 1 0 1 2 2 1 1 2
## [36] 2 2 1 0 2 2 0 1 1 1 2 0 2 1 1
```

```
  Observadas_1 = table(Muestra_1) # tabla de frecuencias en la muestra
  print(Observadas_1)
```

```
## Muestra_1
##  0  1  2
##  7 21 22
```

```
#C. Combinamos las frecuencias observadas y esperadas en una misma tabla
  # [Nota: falla si, por azar, el muestreo no incluye los tres genotipos]
  Esp = c(rep(0, Esperadas[1]), rep(1, Esperadas[2]), rep(2, Esperadas[3]))
  par(mfrow=c(1,2))
  hist(Esp, breaks = 10, col=rgb(1,1,0,0.7), main = "E: amarillo; O: azul", xlab = "0:A2A2,
 1:A1A2, 2:A1A1")
  par(new = TRUE)
  hist(Muestra_1, breaks = 10, col=rgb(0,1,1,0.4), main ="", xlab = "", axes=FALSE)

# Frec. esperadas en amarillo y observadas en azul, con la superposición en verde por transpa
rencia)
```



**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; - en el bloque anterior, seguimos el orden natural de los genotipos (0, 1, y 2 copias de A); ahora usamos el orden más usual en genética, A_1A_1, A_1A_2, A_2A_2, usando esos 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].

```
#D. 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
  Esperadas_relativas_2 = c(p1**2, 2*p1*(1-p1), (1-p1)**2)
  Muestra_2 = sample(Genotipos,  x1, replace = TRUE, Esperadas_relativas_2) # obtenemos una m
uestra de tamaño x1, muestreando al azar con reposición los genotipos, cada uno con una proba
bilidad definida más arriba (Esperadas_relativas_2)
  Observadas_2 = table(Muestra_2) # tabla de frecuencias en la muestra
  print(Observadas_2)
```

```
## Muestra_2
## A1A1 A1A2 A2A2
##   18   22   10
```

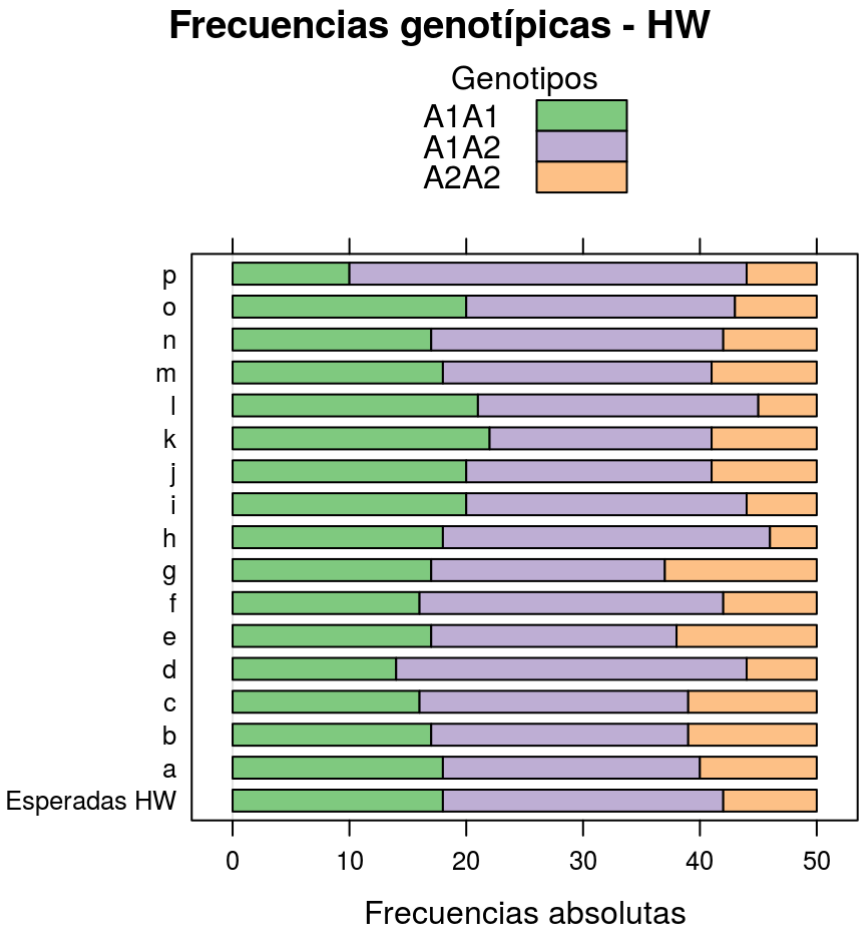
```
  Esperadas_2 = Esperadas_relativas_2 * x1

#E. Combinamos las frecuencias observadas y esperadas en una misma tabla
  Resumen_2 = rbind(Esperadas_2, Observadas_2)
  print(Resumen_2)
```

```
##              A1A1 A1A2 A2A2
## Esperadas_2    18   24    8
## Observadas_2   18   22   10
```

```
 # probemos crear un resumen sin depender del formato generado en table
  Resumen_3 = matrix(data = NA, nrow = y+2 , ncol= 3)
  colnames(Resumen_3) = c("A1A1", "A1A2", "A2A2")
  rownames(Resumen_3) = c("Esperadas 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_3[1,] = Esperadas_2
  Resumen_3[2,] = Observadas_2

#F. COMBINANDO MÚLTIPLES MUESTRAS ADICIONALES ()
  Resumen_y = (Resumen_3) # comenzamos incorporando las observadas y esperadas generadas más
 arriba
  for (i in seq(1:y)){
    Muestra_y = sample(Genotipos,  x1, replace = TRUE, Esperadas_relativas_2) # obtenemos una
muestra de tamaño x1, muestreando al azar con reposición los genotipos, cada uno con una prob
abilidad definida más arriba (Esperadas_relativas)
    Observadas_y = table(Muestra_y)
     # if((length(Observadas_y) =3)
        Resumen_y[i+2,] = Observadas_y
    }
    print(Resumen_y)
```

```
##              A1A1 A1A2 A2A2
## Esperadas HW   18   24    8
## a              18   22   10
## b              17   22   11
## c              16   23   11
## d              14   30    6
## e              17   21   12
## f              16   26    8
## g              17   20   13
## h              18   28    4
## i              20   24    6
## j              20   21    9
## k              22   19    9
## l              21   24    5
## m              18   23    9
## n              17   25    8
## o              20   23    7
## p              10   34    6
```

```
# Pruebas con gráficas básicas de "lattice"
   Fig_1 = barchart(Resumen_y[1:(y+2),],
          xlab = "Frecuencias absolutas",
          main = "Frecuencias genotípicas - 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:

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

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

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

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:

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

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

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

Sumando las frecuencias, verificamos el resultado:

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

Notamos, 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_3 = 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_3) # obtenemos una m
uestra de tamaño x1, muestreando al azar con reposición los genotipos, cada uno con una proba
bilidad definida más arriba (Esperadas_relativas_2)
  Observadas_3 = table(Muestra_3) # tabla de frecuencias en la muestra
  print(Observadas_3)
```

```
## Muestra_3
## A1A1 A1A2 A2A2
##   24   13   13
```

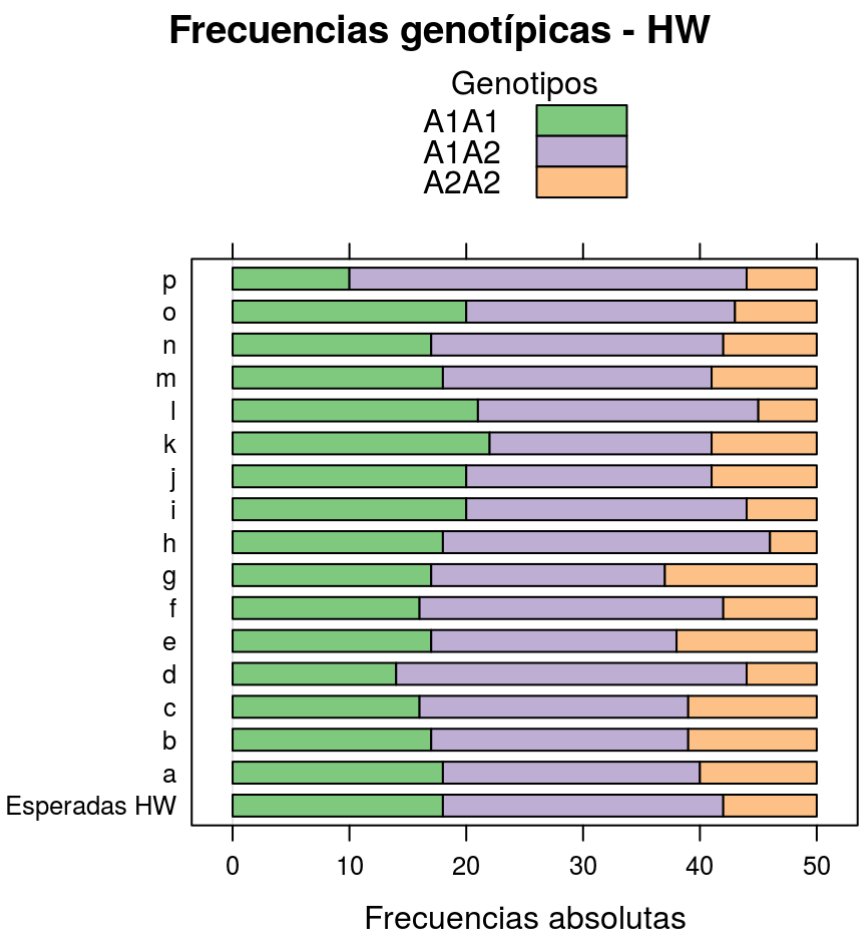
```
  Esperadas_2 = Esperadas_relativas_2 * x1 # mantenemos como referencia las frecuencias esper
adas según HW del bloque                                                  # anterior

#H. Combinamos las frecuencias observadas y esperadas en una misma tabla
  Resumen_3 = rbind(Esperadas_2, Observadas_3)
  print(Resumen_2)
```

```
##              A1A1 A1A2 A2A2
## Esperadas_2    18   24    8
## Observadas_2   18   22   10
```

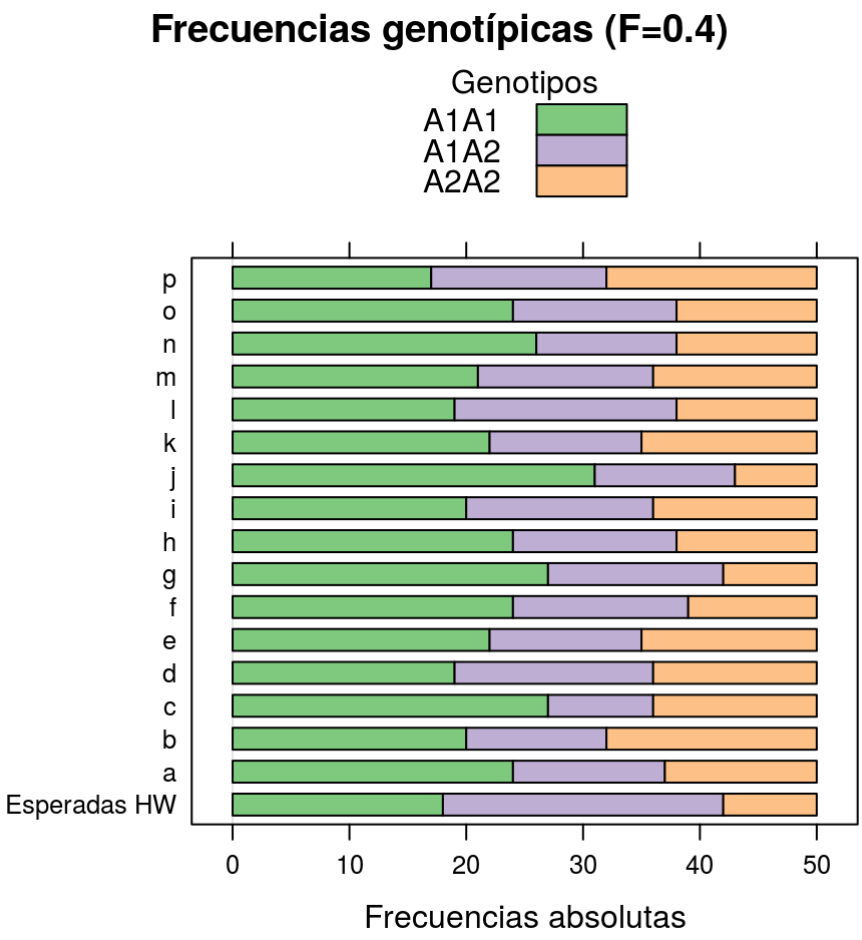
```
 # probemos crear un resumen sin depender del formato generado en table
  Resumen_4 = matrix(data = NA, nrow = y+2 , ncol= 3)
  colnames(Resumen_4) = c("A1A1", "A1A2", "A2A2")
  rownames(Resumen_4) = c("Esperadas HW", letters[1:(y+1)]) # la primera fila toma el nombre
 "Esp", y las siguientes se etiquetan
                                                          # con letras hasta el valor definido en
 "y", usado más abajo para
                                                          # replicar los muestreos
  Resumen_4[1,] = Esperadas_2
  Resumen_4[2,] = Observadas_3

#I. COMBINANDO MÚLTIPLES MUESTRAS ADICIONALES ()
  Resumen_z = (Resumen_4) # comenzamos incorporando las observadas y esperadas generadas más
 arriba
  for (i in seq(1:y)){
    Muestra_z = sample(Genotipos,  x1, replace = TRUE, Esperadas_relativas_3) # obtenemos una
muestra de tamaño x1, muestreando al azar con reposición los genotipos, cada uno con una prob
abilidad definida más arriba (Esperadas_relativas)
    Observadas_z = table(Muestra_z)
     # if((length(Observadas_y) =3)
        Resumen_z[i+2,] = Observadas_z
    }
    print(Resumen_z)
```

```
##              A1A1 A1A2 A2A2
## Esperadas HW   18   24    8
## a              24   13   13
## b              20   12   18
## c              27    9   14
## d              19   17   14
## e              22   13   15
## f              24   15   11
## g              27   15    8
## h              24   14   12
## i              20   16   14
## j              31   12    7
## k              22   13   15
## l              19   19   12
## m              21   15   14
## n              26   12   12
## o              24   14   12
## p              17   15   18
```

```
# Pruebas con gráficas básicas de "lattice"
   Fig_2 = barchart(Resumen_z[1:(y+2),],
         xlab = "Frecuencias absolutas",
         main = "Frecuencias genotípicas (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)
```


Frecuencias genotípicas - HW
Genotipos
A1A1
A1A2
A2A2
p
o
n
m
l
k
j
i
h
g
f
e
d
c
b
a
Esperadas HW
0
10
20
30
40
50
Frecuencias absolutas

```
plot(Fig_2)
```


Frecuencias genotípicas (F=0.4)
Genotipos
A1A1
A1A2
A2A2
p
o
n
m
l
k
j
i
h
g
f
e
d
c
b
a
Esperadas HW
0
10
20
30
40
50
Frecuencias absolutas

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:

$$f(A_1A_2) = 2pq(1-\hat{F}) = H_o = H_e(1-\hat{F})$$

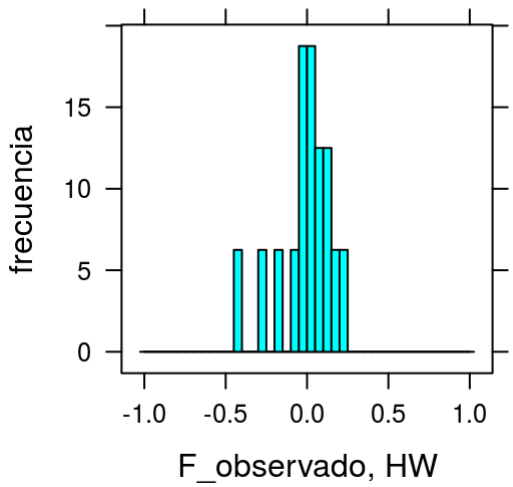
(Notar que usamos $\hat{F}$ para indicar que vamos a obtener una estimación de *F", cuyo verdadero valor desconocemos).

Despejando, tenemos que

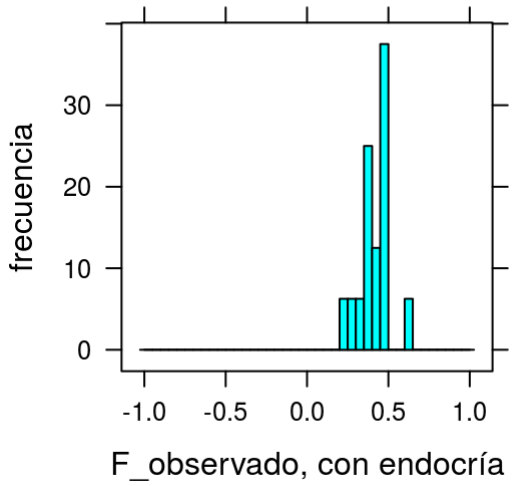
$$\hat{F} = (H_o - H_e)/H_e$$

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 coficente 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.