



# FUNDAMENTOS DE GENÉTICA DE POPULAÇÕES USANDO R

Carlos Guerra Schrago  
Dep. Genética, Universidade Federal do Rio de Janeiro  
carlos.schrago@gmail.com

Versão 23/11/2023

Licença de utilização: Distribuição e utilização livres, sem modificação do conteúdo e com referência da fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)



## Conteúdo

|   |     |
|---|-----|
| Pensando estatisticamente.....  | 2   |
| Panmixia em populações infinitas (equilíbrio Hardy-Weinberg).....                               | 6   |
| A distribuição de chi-quadrado .....  | 10  |
| Teste do equilíbrio de Hardy-Weinberg usando dado de SNP humanos.....                           | 13  |
| Populações finitas: modelo Wright-Fisher e deriva genética.....                                 | 19  |
| Estudando fenômenos estocásticos por cadeias de Markov I.....                                   | 23  |
| Estudando fenômenos estocásticos por cadeias de Markov II.....                                  | 32  |
| Quantificação do decaimento da heterozigosidade por deriva.....                                 | 44  |
| Decaimento da heterozigosidade por deriva genética & equilíbrio mutação-deriva.....             | 49  |
| Efeito da variação do número de fêmeas e machos na diversidade genética e tamanho efetivo.....  | 60  |
| Efeito da variação do tamanho da população na diversidade genética e tamanho efetivo.....       | 64  |
| Redução da heterozigosidade devido à estruturação populacional.....                             | 77  |
| Consequências do fluxo gênico nas frequências alélicas e diversidade genética.....              | 84  |
| Calculando as probabilidades IBD em 1 geração de duas cópias amostradas da população.....       | 92  |
| A ideia de seleção natural.....   | 97  |
| Quantificação da ação da seleção natural nas populações.....                                    | 103 |
| Tipos de seleção natural.....   | 109 |
| Formalização da seleção e variação das frequências alélicas.....                                | 125 |
| Seleção natural em populações finitas.....  | 129 |
| Custo da seleção natural.....   | 135 |
| Carga genética, manutenção do polimorfismo genético e o debate balanceado-clássico.....         | 141 |
| Destino das mutações I: Tempos médios para fixação e perda de novos alelos mutantes.....        | 148 |
| Destino das mutações II: Origem do polimorfismo entre espécies e probabilidades de fixação..... | 151 |
| Evolução multi-loci: O relógio da recombinação.....   | 158 |
| Evolução multi-loci: Efeito da seleção no polimorfismo genômico com ligação genética.....       | 162 |
| Evolução multi-loci: Interação de múltiplos loci e a hipótese <i>shifting balance</i> .....     | 165 |
| Tempo de espera até a coalescência de um par de alelos.....                                     | 170 |
| Tempos de espera até as diversas coalescências de vários alelos: a coalescência Kingman.....    | 174 |
| Unificando genética de populações e filogenética: o modelo da coalescência multi-espécies.....  | 181 |
| Interpretando filogenias num contexto populacional.....   | 192 |

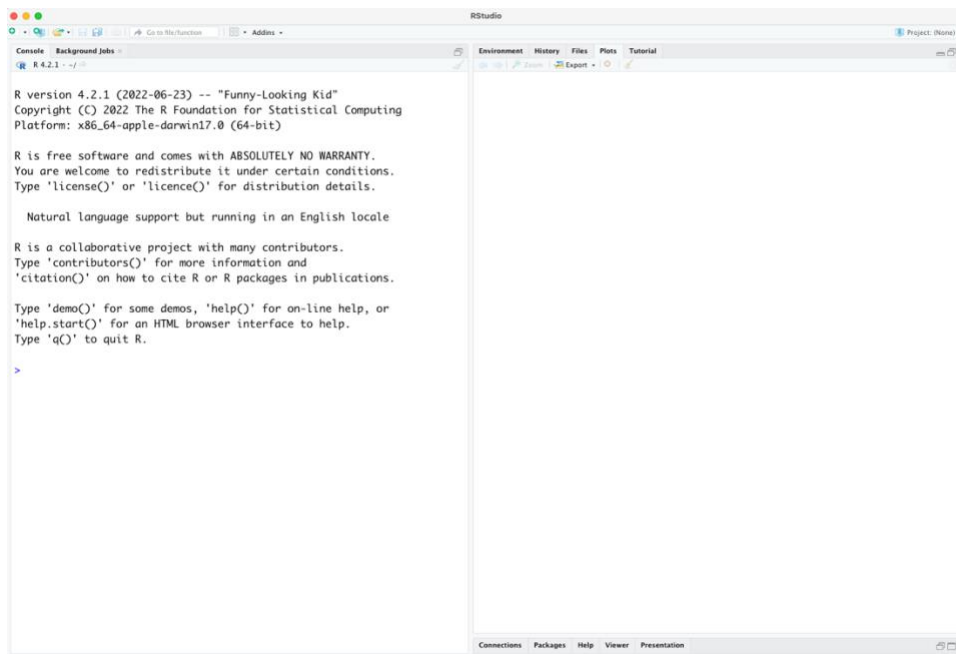
## Pensando estatisticamente

Para realização dessa prática, precisamos instalar a linguagem de programação R. Acessaremos às funções desta linguagem através da interface RStudio Desktop:

- Instalação do R: <https://www.r-project.org/>
- Instalação do RStudio Desktop: <https://www.rstudio.com/products/rstudio/>

Existem diversos textos online e livros publicados introduzindo a linguagem R. O manual introdutório oficial está disponível em <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>. Você também encontrará diversos cursos no YouTube introduzindo os conceitos básicos do R.

Após instalação do R e RStudio, abra o RStudio. Você será apresentado a uma interface semelhante a esta:



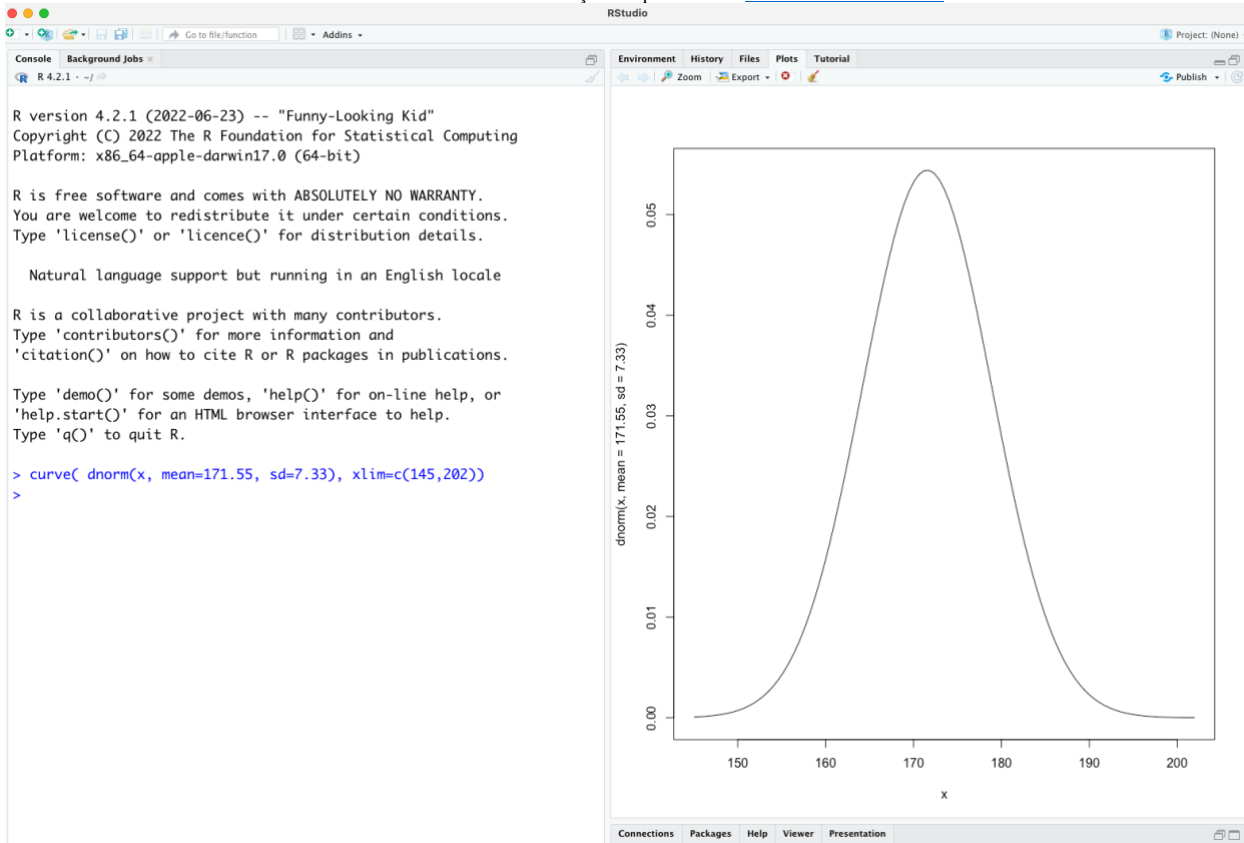
As áreas que nos interessam são a aba “Console” à esquerda e “Plots” à direita.

Na área do Console, digite o seguinte comando e dê enter/return:

```
curve(dnorm(x, mean=171.55, sd=7.33), xlim=c(145,202))
```

Rapidamente aparecerá o gráfico da curva normal com média em 171.55 cm e desvio padrão de 7.33 que representa a distribuição probabilística da altura dos humanos obtida de <https://ourworldindata.org/human-height>

Licença de utilização: Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)



O comando acima plota uma curva (função **curve**) de uma lista valores retornados pela função **dnorm**. Esta função fornece os valores das densidades (eixo y) dos valores das alturas (eixo x) de uma distribuição normal com média = 171.55 e desvio = 7.33 cm. A curva será plotada entre os valores de 145 e 202 cm (*xlim*).

Vamos agora sortear alguns valores desta distribuição normal. Suponhamos que queiramos sortear 10 valores de alturas desta normal. A função **rnorm** fará isso:

```
rnorm(n=10, mean=171.55, sd=7.33)
```

Esse comando retornará 10 valores de alturas sorteados respeitando as probabilidades estabelecidas pela função. Ou seja, valores próximos da média (171.55 cm) terão maior chance de serem sorteados.

Vamos agora calcular a média de 10 valores de altura sorteados. Isso será feito aplicando a função **mean** aos valores retornados pelo comando anterior:

```
mean(rnorm(n=10, mean=171.55, sd=7.33))
```

Observe que como a função **rnorm** retorna valores sorteados (ao invés de fixos), a cada vez que o comando acima for acionado, o valor da média será diferente.

Vamos agora calcular a diferença entre dois sorteios independentes de 10 valores de alturas da mesma distribuição probabilística:



```
mean(rnorm(n=10, mean=171.55, sd=7.33)) - mean(rnorm(n=10, mean=171.55,
sd=7.33))
```

O que fizemos no comando acima foi simplesmente calcular a diferença (-) entre as médias (**mean**) duas chamadas independentes de **rnorm** (com 10 valores sorteados). Perceba que essa diferença pode ser negativa ou positiva. Note também que, embora tenhamos feito sorteios da mesma distribuição probabilística, a diferença dificilmente será exatamente igual a zero (o que significaria que os mesmos 10 números foram sorteados ao acaso nas duas chamadas independentes).

Suponhamos agora que queiramos repetir o comando acima 1000 vezes. Podemos fazer isso construindo o seguinte *loop for*:

```
for(i in c(1:1000)){
  mean(rnorm(n=10, mean=171.55, sd=7.33)) - mean(rnorm(n=10, mean=171.55, sd=7.33))
}
```

Esse comando fará 1000 chamadas independentes do comando da diferença executado anteriormente. Entretanto, ao entrar com esse comando, você verá que nada aparecerá como resultado.

É necessário solicitar que o R imprima os valores das diferenças obtidas nas 1000 chamadas independentes do *loop*. Isso será feito com a função **print**.

```
for(i in c(1:1000)){
  print (mean(rnorm(n=10, mean=171.55, sd=7.33)) - mean(rnorm(n=10, mean=171.55,
sd=7.33)))
}
```

Agora, você verá que uma lista de 1000 valores de diferenças aparecerá na tela.

Por fim, iremos guardar estes 1000 valores numa variável. Chamarei essa variável de *D* (de diferença), mas você pode criar o nome que desejar.

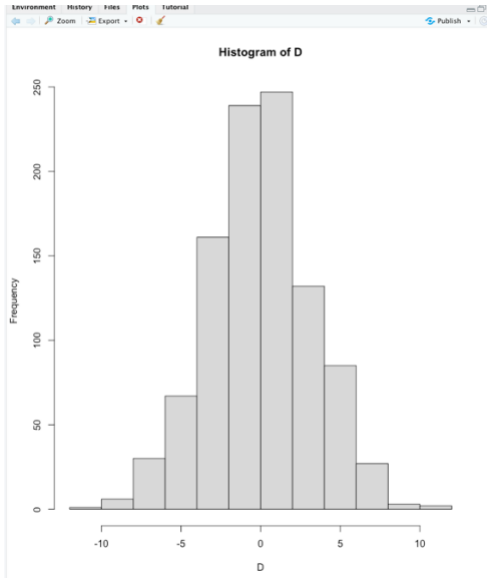
```
D = vector()
for(i in c(1:1000)){
  D = c(D, mean(rnorm(n=10, mean=171.55, sd=7.33)) - mean(rnorm(n=10, mean=171.55,
sd=7.33)) )
}
```

Ao invés de imprimir os 1000 valores de diferença na tela, nós os atribuímos à variável *D*, que é um vetor (**vector**). Ao digitar *D* e apertar enter/return, você verá os 1000 valores diferença sorteados.

Por fim, vamos plotar um histograma desses 1000 valores usando a função **hist**:

```
hist(D)
```

Um gráfico semelhante a este abaixo aparecerá na área direita do RStudio:



O que significa este gráfico? Ele mostra que, ao realizarmos 1000 experimentos (chamado tecnicamente de réplicas ou replicatas) em que amostramos 10 valores de altura duas vezes da mesma distribuição e calculamos a diferença entre as médias, eventualmente encontramos diferenças tão grandes quanto 5 cm ou mesmo até 10 cm (para mais ou para menos)!

Embora, em média, o valor da diferença fique próximo a zero, mesmo quando duas amostras (de 10 pessoas) vieram da mesma população, valores de diferença tão grandes quanto 10 cm podem ser obtidos **ao acaso**. É claro que eles são mais raros, mas podem acontecer.

Esse histograma das diferenças que obtivemos é chamado de **distribuição nula**. Ele servirá para testar uma **hipótese nula**. No nosso caso, a hipótese nula é que as duas médias são iguais ( $H_0: \mu_1 = \mu_2$  ou  $H_0: D = 0$ )

Com este histograma, podemos calcular qual é probabilidade de obtenção de valores de diferença ao acaso com um nível de significância  $\alpha = 5\%$  (0.05), quando duas amostras vieram da mesma distribuição (população). Usaremos a função **quantile**:

```
quantile(D, c(0.025, 0.975))
```

Os valores que este comando retornará estabelecerão os valores mínimos e máximos de diferença para o *critério de rejeição* da hipótese nula – valores fora deste intervalo, indicarão rejeição num nível de confiança de 5%. No geral, temos

```
quantile(D, c( alpha/2 , 1 - alpha/2))
```

Afinal, o valor de diferença observado em sala de aula é significativo para rejeição da hipótese nula de que as médias são iguais ou não?



## Panmixia em populações infinitas (equilíbrio Hardy-Weinberg)

Nesta prática iremos simular as frequências dos alelos e dos genótipos ao longo das gerações. Em todos os casos, estudaremos um locus com apenas dois alelos – A e a.

Nosso primeiro exemplo é a situação em que a população inicial (geração 1, ou seja,  $G_1$ ) contém  $N_m$  machos e  $N_f$  fêmeas, com número de indivíduos para cada genótipo conforme a seguir:

Conformação da população em  $G_1$

| Genótipos | Fêmeas      | Machos      |
|-----------|-------------|-------------|
| <b>AA</b> | 46          | 18          |
| <b>Aa</b> | 58          | 27          |
| <b>Aa</b> | 21          | 74          |
|           | $N_f = 125$ | $N_m = 119$ |

No R, podemos calcular as frequências genotípicas e alélicas em cada sexo nesta primeira geração (você pode copiar e colar o código abaixo ou entrar o comando linha por linha. As linhas que começam com o símbolo # não serão lidas pelo R. São chamadas de comentários):

```
Nf = 125
Nm = 119

#genotípicas
fAA_f = 46/Nf
fAA_m = 18/Nm
fAa_f = 58/Nf
fAa_m = 27/Nm
faa_f = 21/Nf
faa_m = 74/Nm

#alélicas
fA_f = fAA_f + (fAa_f/2)
fA_m = fAA_m + (fAa_m/2)
fa_f = faa_f + (fAa_f/2)
fa_m = faa_m + (fAa_m/2)
```

Desta forma, no *pool* gamético, encontraremos óvulos contendo o alelo A na frequência  $fA_f$  e o alelo a na frequência  $fa_f$ . Entraremos também espermatozoides com o alelo A na frequência  $fA_m$  e a na frequência  $fa_m$ .

Para calcular as frequências dos genótipos da próxima geração **assumindo que não ocorrerá erro amostral no pool gamético** (ou seja, ele será amostrado infinitas vezes), podemos aplicar uma simples tabela de multiplicação:

|                 |      | óvulos |      |
|-----------------|------|--------|------|
|                 |      | fA_f   | fa_f |
| espermatozoides | fA_m | fAA    | fAa  |
|                 | fa_m | fAa    | faa  |

As frequências dos genótipos e alelos em  $G_2$  será, portanto:

```
#genótipos
fAA = fA_f * fA_m
fAa = (fa_f * fA_m) + (fa_m * fA_f)
faa = fa_f * fa_m
```



```
#alelos  
fA = fAA + (fAa/2)  
fa = faa + (fAa/2)
```

Assumindo que a probabilidade de formação de um zigoto que será uma fêmea ou um macho é igual (pense na razão desse pressuposto ser razoável), a partir desta geração, não ocorrerá mais variação das frequências dos alelos, pois o próximo pool gamético conterá:

```
#gametas no pool gamético dos parentais de G1  
fA_f = fA  
fA_m = fA  
fa_f = fa  
fa_m = fa
```

Note que, acima, para calcular a frequência dos óvulos com o alelo  $A$  ( $fA_f$ ), não precisamos dividir a frequência do alelo  $A$  na população parental  $G_1$  por 2 (metade machos e metade fêmeas)!

Em  $G_1$ , sabemos que existe  $fA$  alelos  $A$ . Como a frequência é uma grandeza relativa, no *pool* gamético também teremos  $fA$  dos óvulos carregando esse alelo.

Exemplo: Se 20% dos alelos totais da população são  $A$  e ele está dividido igualmente entre machos e fêmeas (pense, novamente, na razão disso ser razoável), sabemos que 20% dos óvulos carregarão o alelo  $A$ . Talvez você queira testar isso no papel para se convencer.

Desta forma, aplicando o mesmo quadro de multiplicação de óvulos e espermatozoides, temos na geração  $G_2$ :

```
#genótipos  
fAA = fA_f * fA_m  
fAa = (fa_f * fA_m) + (fa_m * fA_f)  
faa = fa_f * fa_m  
  
#alelos  
fA = fAA + (fAa/2)  
fa = faa + (fAa/2)
```

Você perceberá que as frequências dos alelos em  $G_2$  serão as mesmas de  $G_1$ . A partir de  $G_2$ , as frequências dos genótipos também irão estabilizar, pois o *pool* gamético que irá compor  $G_3$  será idêntico em frequências àquele com originou  $G_2$ .

Vamos agora fazer um *loop* para acompanhar a mudança das frequências alélicas e genotípicas ao longo de 20 gerações com populações infinitas a partir de  $G_1$ :

```
# população em G1 (estado inicial):  
Nf = 125  
Nm = 119  
  
#genotípicas em cada sexo em G1  
fAA_f = 46/Nf  
fAA_m = 18/Nm  
fAa_f = 58/Nf  
fAa_m = 27/Nm  
faa_f = 21/Nf  
faa_m = 74/Nm  
  
#alélicas em cada sexo em G1  
fA_f = fAA_f + (fAa_f/2)
```



```
fA_m = fAA_m + (fAa_m/2)
fa_f = faa_f + (fAa_f/2)
fa_m = faa_m + (fAa_m/2)
```

```
#genotípicas e alélicas totais em G1
fAA = (46 + 18) / (Nf + Nm)
fAa = (58 + 27) / (Nf + Nm)
faa = (21 + 74) / (Nf + Nm)
```

```
fA = fAA + (fAa/2)
fa = faa + (fAa/2)
```

#Abaixo, começa o loop de 20 gerações:

```
for(i in c(1:20)){
  fAA = c(fAA, fA[length(fA)]*fA[length(fA)])
  fAa = c(fAa, 2*(fA[length(fA)]*fa[length(fa)]) )
  faa = c(faa, fa[length(fa)]*fa[length(fa)])

  fA = c(fA, fAA[length(fAA)] + (fAa[length(fAa)]/2) )
  fa = c(fa, faa[length(faa)] + (fAa[length(fAa)]/2) )
}
```

Provavelmente, você achará o comando interno do *loop* muito complicado. Aqui vai a razão: como queremos guardar os valores das frequências alélicas e genotípicas ao longo de todas as gerações, as variáveis *fAA*, *fAa*, *faa*, *fA* e *fa* são agora vetores.

Conforme vimos no estudo passado, no R, a forma de acrescentar um novo elemento *x2* a um vetor *x* é *x = c(x, x2)*.

Desta forma, se o vetor possui os seguintes valores:

```
> x
[1] 1 2 3 4 5 6
```

Ao usar o comando *x = c(x, 1526251)*, obteremos o seguinte:

```
> x
[1] 1 2 3 4 5 6 1526251
```

Mostrando que o número 1526251 foi adicionado no último elemento do vetor.

Finalmente, temos que entender a razão de acrescentarmos um valor às frequências usando a forma *fAA[length(fAA)]*.

Essa é a maneira de acessarmos o último elemento de um vetor. Por exemplo, *x[length(x)]* retornará o último valor contido em *x*:

```
> x[length(x)]
[1] 1526251
```

Assim, garantimos que, ao calcular as frequências de, por exemplo, *G<sub>14</sub>*, usaremos os valores de *G<sub>13</sub>*.

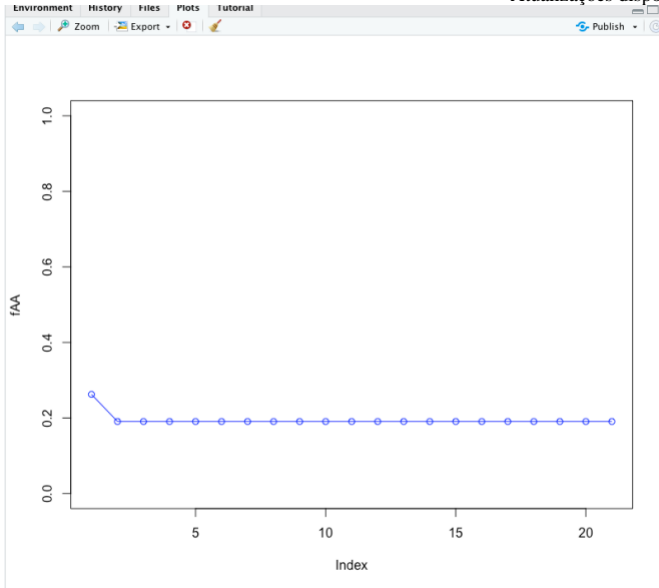
Por fim, podemos plotar gráficos com a evolução das frequências alélicas e genotípicas assumindo panmixia em populações infinitas a partir de *G<sub>1</sub>*. Essa é a propriedade do Equilíbrio Hardy-Weinberg:

```
plot(fAA, ylim=c(0,1), col="blue")
lines(fAA, col="blue")
```

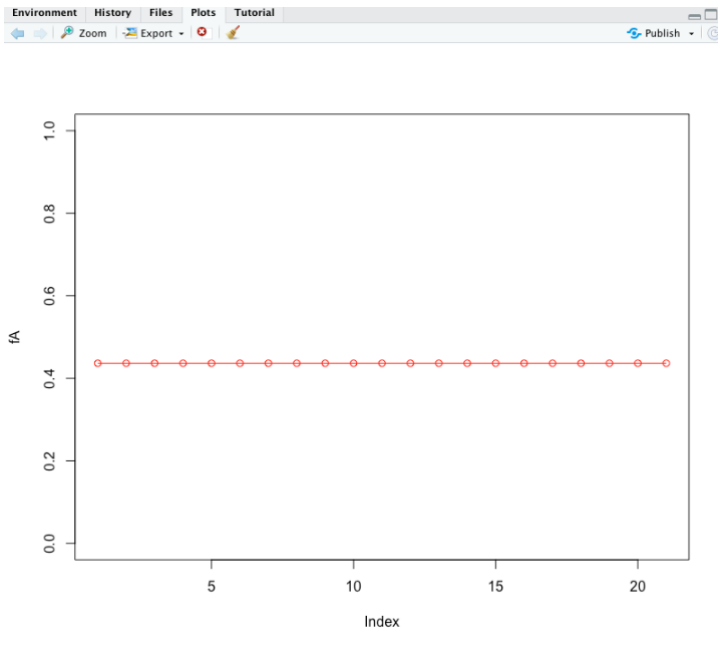




*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)



```
plot(fA, ylim=c(0,1), col="red")  
lines(fA, col="red")
```





## A distribuição de chi-quadrado

A estatística de  $\chi^2$  mede o desvio entre o número de ocorrências observadas e a esperada, normalizado pelo esperada. Esta normalização faz que a métrica ofereça a magnitude do desvio comparado ao valor esperado. Assim, um 0.5 indica que o desvio quadrático é de 50% do valor esperado.

Para cada classe comparada, é contabilizado o número de ocorrências em cada classe e, posteriormente, é calculado o quadrado do desvio do número de ocorrências observada. Esse valor é então normalizado. Ao final, para as  $n$  classes comparadas, soma-se todos os valores de desvios normalizados:

$$\chi^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

Suponhamos que queiramos obter a distribuição nula entre o número observado e esperado da ocorrência dos três genótipos possíveis em um locus dialélico ( $A$  e  $a$ ) –  $AA$ ,  $Aa$  e  $aa$ .

Para tal, temos que saber de antemão que as frequências reais desses genótipos (lembre-se: nas distribuições nulas temos a probabilidade dos desvios ocorrerem apenas por erro amostral). Portanto, precisamos obter exatamente isso – valores de chi-quadrado que ocorrem em consequência de erro amostral apenas.

Por exemplo, vamos trabalhar com uma população em que as frequências genotípicas são

$f_{AA} = 0.25$   
 $f_{Aa} = 0.5$   
 $f_{aa} = 0.25$

Se quisermos amostrar  $n=10$  indivíduos dessa população, podemos fazer:

```
n=100  
sample(c("AA", "Aa", "aa"), n, replace = T, prob = c(fAA, fAa, faa))
```

Neste comando **sample**, informamos que os elementos do vetor `c("AA", "Aa", "aa")` devem ser amostrados  $n = 100$  vezes com reposição (`replace = T`). Ou seja, ao amostrar o elemento “AA” por exemplo, ele será devolvido ao vetor original. Por fim, cada um dos elementos será amostrado com probabilidades iguais às suas respectivas frequências - `prob = c(fAA, fAa, faa)`.

Ao fazer isso, sabemos que as **frequências esperadas** de cada um dos genótipos serão iguais àquelas fornecidas em `prob = c(fAA, fAa, faa)`. Desta forma, qualquer desvio entre as frequências encontradas na amostra de  $n = 100$ , ou seja, as **observadas**, serão oriundas de erro amostral. Se, ao invés de 100 vezes, amostrássemos, digamos, 800 de vezes, reduziríamos a diferença entre as frequências observadas e as esperadas, corroborando que o desvio é oriundo de erro amostral.

Como poderíamos obter a distribuição dos valores de diferença quadrática ( $\chi^2$ ) entre observado e esperado quando sabemos, de antemão, que eles são iguais (hipótese nula)?

Para isso, basta fazer várias amostragens independentes de  $n = 100$  (replicatas) e, para cada replicata, calcular a fórmula de  $\chi^2$ . Suponhamos que façamos 100.000 replicatas de sorteios de  $n = 100$ :

```
chi = vector()
n = 100
replicatas = 100000

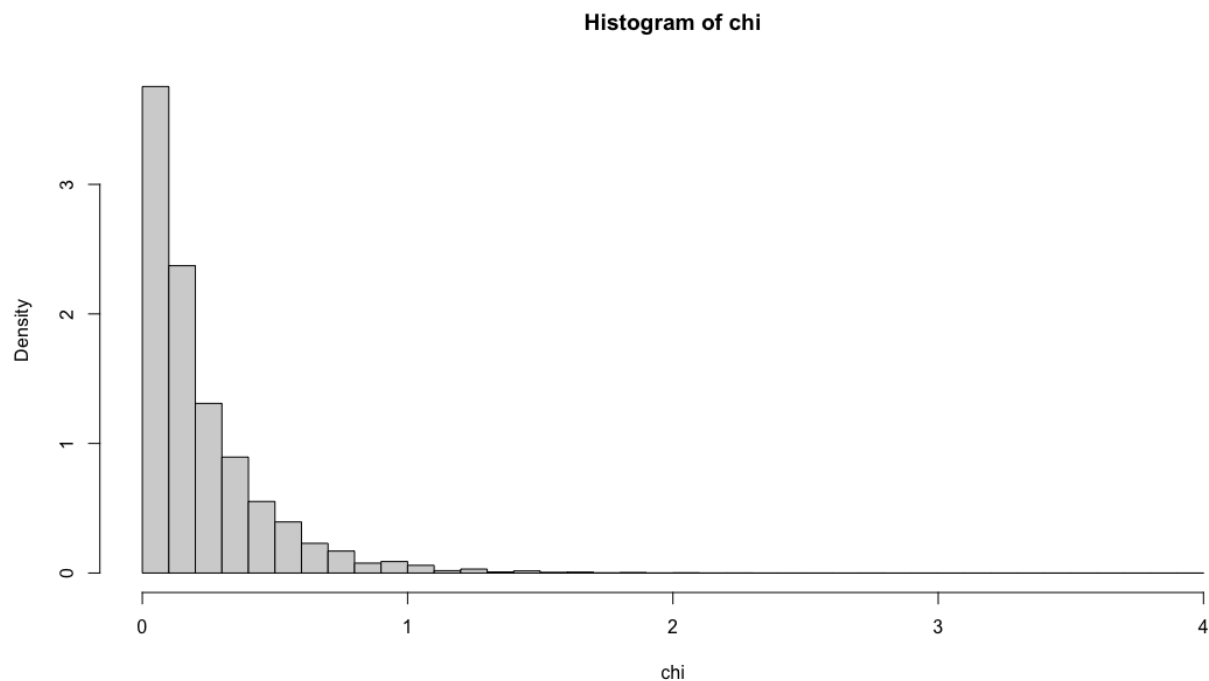
for(i in c(1:replicatas)){

  x = sample(c("AA","Aa","aa"), n, replace = T, prob = c(0.25,0.5,0.25))

  chi[i] = ( ( length(which(x == "AA")) - n*0.25)^2/n*0.25 ) + ( ( length(which(x == "Aa")) - n*0.5)^2/n*0.5 ) + ( ( length(which(x == "aa")) - n*0.25)^2/n*0.25 )

}
```

Ao plotar os valores da variável **chi**, que guarda os 100.000 valores de  $\chi^2$ , obtemos:



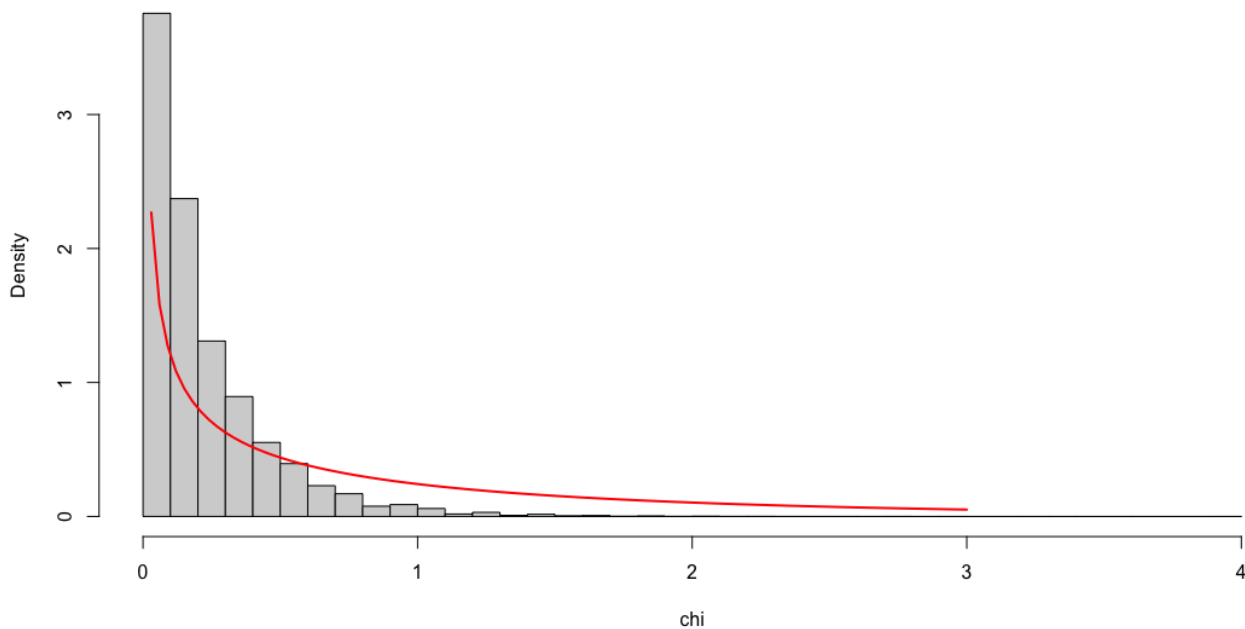
Essa é a distribuição nula da métrica de  $\chi^2$ .

Como o  $\chi^2$  é uma distribuição muito popular, ela está disponível no R e pode ser plotada através do comando **dchisq**.

Vamos então comparar o nosso histograma aproximado com a distribuição formal de  $\chi^2$ :

```
curve(dchisq(x, df = 1), xlim=c(0,3), col="red", add=T, lwd=2)
```

### Histogram of chi



A curva vermelha é a distribuição formal de  $\chi^2$  enquanto o histograma é a nossa aproximação.

Embora nossa aproximação pareça razoável, ela tem uma quantidade maior de valores de  $\chi^2$  próximos de zero.

Isso indica que, por exemplo, se  $\chi^2 = 1$ , a hipótese nula de que o observado é igual ao esperado seria possivelmente rejeitada. Por outro lado, na curva vermelha, o mesmo  $\chi^2 = 1$  não seria suficiente para rejeitar  $H_0$ .

Essa é uma das razões do  $\chi^2$  ser um **teste conservador**. Ele tende a não rejeitar  $H_0$ .

## Teste do equilíbrio de Hardy-Weinberg usando dado de SNP humanos

Demonstraremos aqui o teste de HW no SNP **rs699**. Trabalharemos com os dados da população total. Esse SNP possui dois alelos segregando nas populações humanas A/G. O alelo G é o ancestral do alelo A:

**rs699** SNP

Most severe consequence: **missense variant** | [See all predicted consequences](#)

Alleles: **A/G** | Ancestral: G | MAF: 0.29 (A) | Highest population MAF: 0.50

Change tolerance: CADD: 0.347 | GERP: -2.88

Location: **Chromosome 1:230710048** (forward strand) | VCF: 1 230710048 rs699 A G

Co-located variants: HGMD-PUBLIC [CM920010](#) ; COSMIC [COSV64184214](#) ; dbSNP [rs1553314015](#) (A/-)

Evidence status:

Clinical significance:

HGVS names: This variant has 26 HGVS names - [Show](#)

Synonyms: This variant has 10 synonyms - [Show](#)

Genotyping chips: This variant has assays on 11 chips - [Show](#)

Original source: Variants (including SNPs and indels) imported from dbSNP (release 154) | [View in dbSNP](#)

About this variant: This variant overlaps [12 transcripts](#), has [3009 sample genotypes](#), is associated with [9 phenotypes](#) and is mentioned in [296 citations](#).

### Explore this variant

Ao entrar no link “Population genetics”, veremos que o projeto 1000 Genomes, que trabalha com uma amostra de 2504 indivíduos, lista as seguintes ocorrências genotípicas para esse SNP:

### 1000 Genomes Project Phase 3 (32)

| Population | Allele: frequency (count)          | Genotype: frequency (count)                               |
|------------|------------------------------------|---|
| ALL        | A: 0.295 (1477)    G: 0.705 (3531) | A/A: 0.126 (315)    A/G: 0.338 (847)    G/G: 0.536 (1342) |

Portanto, temos:

AA = 315 indivíduos ( $f_{AA} = 0.126$ )

AG = 847 indivíduos ( $f_{AG} = 0.338$ )

GG = 1342 indivíduos ( $f_{GG} = 0.536$ )

As frequências para cada um dos alelos também está listada:

$f_A = 0.295$

$f_G = 0.705$

Lembremos que o cálculo da frequência alélica é feito assim:

$$f_A = \frac{2N_{AA} + N_{Aa}}{2N_{total}} = \frac{2(315) + 847}{2(2504)} = 0.295$$

Ou

$$f_A = f_{AA} + \frac{f_{AG}}{2} = 0.126 + \frac{0.338}{2} = 0.295$$

Queremos agora testar se as frequências genotípicas estão em HW.



Para tal, precisamos calcular qual seriam as **frequências genóticas esperadas segundo o modelo** reprodutivo de Hardy-Weinberg.

O modelo HW prediz que, num locus com dois alelos em uma população de indivíduos diplóides (como é o caso aqui), os genótipos devem ser proporcionais aos encontros casuísticos dos gametas no pool gamético. Isso faz que as frequências possam ser obtidas através de um simples quadro de multiplicação gerando os valores:

Chamando  $f_A = p$  e  $f_G = q$ . Temos:

| Valores <b>esperados</b> pelo modelo HW      |
|--|
| $f_{AA} = p^2$ $f_{AG} = 2pq$ $f_{GG} = q^2$ |

Sabemos que  $f_A = p = 0.295$  e  $f_G = q = 0.705$ . Portanto:

| Valores <b>esperados</b> pelo modelo HW  |
|--|
| $f_{AA} = p^2 = (0.295)^2 = 0.087025$ $f_{AG} = 2pq = 2 \cdot 0.295 \cdot 0.705 = 0.41595$ $f_{GG} = q^2 = (0.705)^2 = 0.497025$ |

Como observamos  $N = 2504$  indivíduos, esperamos encontrar as seguintes contagens para cada genótipo:

| Números de indivíduos <b>esperados</b> pelo modelo HW  |
|--|
| $N_{AA} = 0.087025 \cdot 2504 = 217.9106$ $N_{AG} = 0.41595 \cdot 2504 = 1041.539$ $N_{GG} = 0.497025 \cdot 2504 = 1244.551$ |

Ao comparar com os valores **observados** nas populações humanas, percebemos que existe alguma diferença:

| Números de indivíduos <b>esperados</b> pelo modelo HW | Números de indivíduos <b>observados</b> | Diferença entre modelo ( <b>esperado</b> ) e realidade ( <b>observado</b> ) |
|---|---|---|
| $N_{AA} = 217.9106$                                   | $N_{AA} = 315$                          | 97.0894   |



|  |                                   |                       |
|--|-----------------------------------|-----------------------|
| $N_{AG} = 1041.539$<br>$N_{GG} = 1244.551$ | $N_{AG} = 847$<br>$N_{GG} = 1342$ | $194.539$<br>$97.449$ |
|--|-----------------------------------|-----------------------|

Temos que saber se essa diferença pode ser **oriunda de erro que naturalmente acontece, pois amostramos apenas 2504 indivíduos.**

Desta forma, usaremos um teste de ajuste do modelo aos dados coletados. Esse teste de *goodness of fit* é o  $\chi^2$ :

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

$$\chi^2 = \frac{(315 - 217.9106)^2}{217.9106} + \frac{(847 - 1041.539)^2}{1041.539} + \frac{(1342 - 1244.551)^2}{1244.551} = 87.22425$$

Para um lócus com dois alelos, o teste possui apenas 1 grau de liberdade, pois existe apenas um parâmetro efetivamente livre. Evidentemente, **quanto menor for o valor de  $\chi^2$ , melhor será o ajuste do modelo aos dados observados.** Se o  $\chi^2 = 0$ , a expectativa do modelo foi idêntica aos dados observados. Desta forma, não rejeitaríamos o modelo!

Fica evidente então, que a hipótese nula que está sendo testada, a  $H_0$ , é aquela em que os valores esperados e observados são iguais e, portanto, a diferença  $D$  entre ambos é nula:

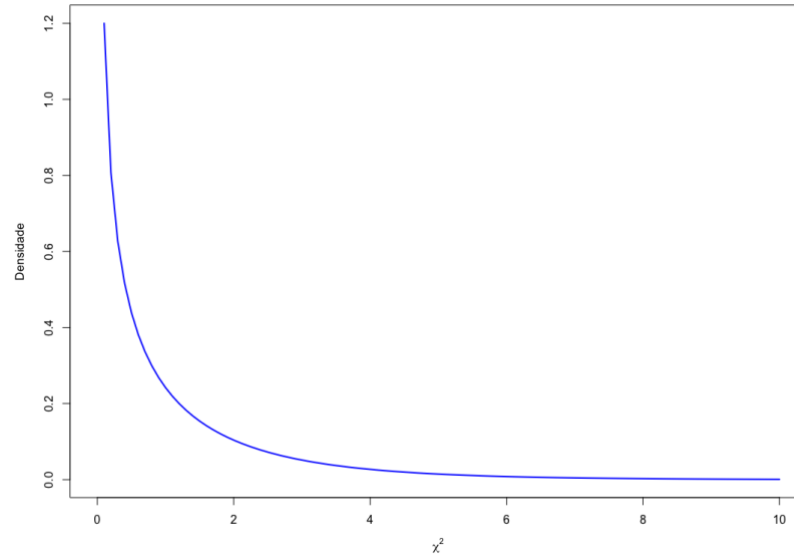
$$H_0 : D_{(esp - obs)} = 0$$

$$H_1 : D_{(esp - obs)} \neq 0$$

A distribuição nula do  $\chi^2$  com 1 grau de liberdade pode ser plotada no R com o seguinte comando:

```
curve(dchisq(x,df=1), xlim=c(0,10), xlab=expression(chi^2), ylab="Densidade", col="blue", lwd=2)
```

Note que especificamos a cor (**col**), espessura da linha (**lwd**), os nomes dos eixos x (**xlab**), y (**ylab**). Para plotar a letra grega  $\chi$ , usamos o comando **expression**.



Essa curva é uma **função de densidade de probabilidade**. Isso significa que a área total sob a curva é igual a 1.

No eixo x, temos os diversos valores de  $\chi^2$ , enquanto no y temos as densidades de cada valor de  $\chi^2$ .

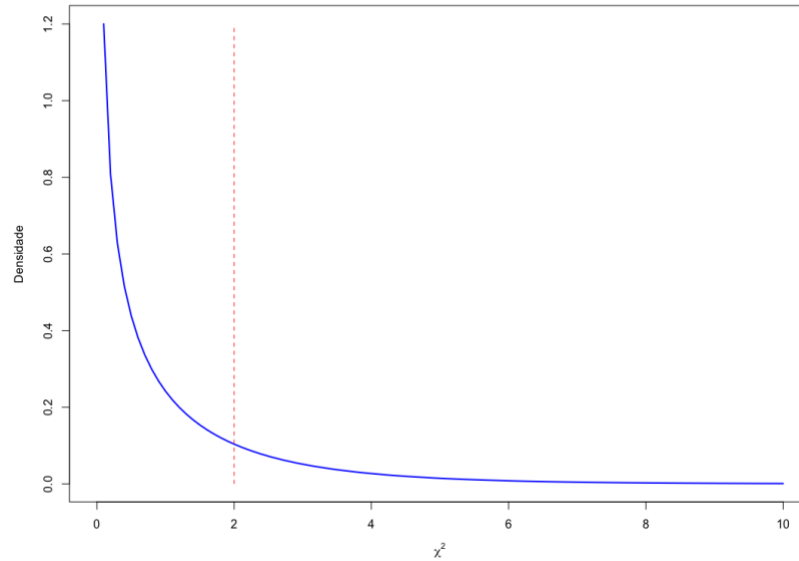
Para obter a probabilidade de um dado valor de  $\chi^2$  (eixo x) – que é a diferença entre observado e esperado – **ocorrer ao acaso**, precisamos calcular a área sob a curva azul deste valor até o infinito.

Por exemplo, suponhamos que  $\chi^2 = 2$ , **qual seria probabilidade de uma diferença entre modelo e realidade com no mínimo esta magnitude ocorrer simplesmente ao acaso (erro amostral)?**

Vamos encontrar o valor de  $\chi^2 = 2$  no gráfico acima:

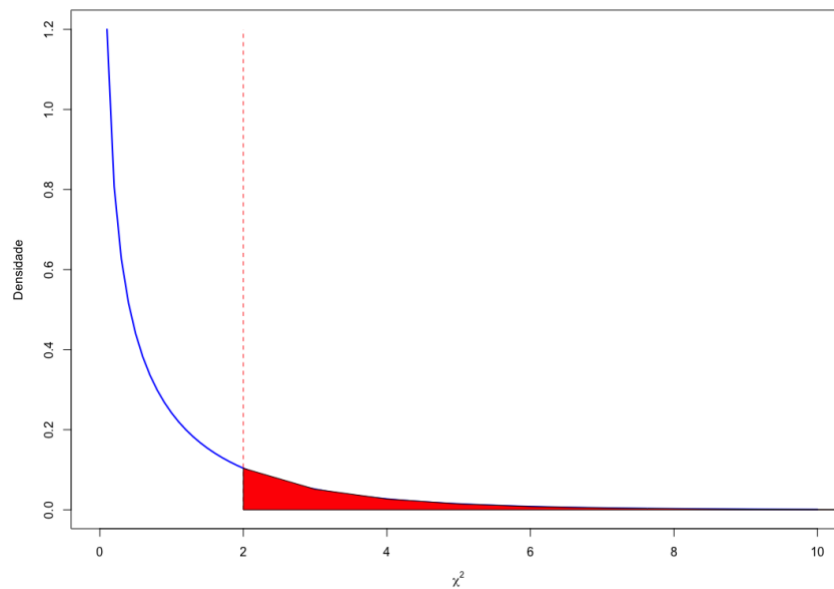
```
segments(2,0,2,1.2, col="red", lty=2)
```





Estamos interessados em calcular a área vermelha sob a curva azul que vai de  $\chi^2 = 2$  até  $\chi^2 = \infty$ :

```
x = seq(0,100)
y = dchisq(x,df=1)
polygon(c(x[x>=2], max(x), 2), c(y[x>=2], 0, 0), col="red")
```



O valor dessa área vermelha é justamente o **valor-p**. Ou seja, **o valor-p indica a chance de um valor de diferença entre esperado e observado ter no mínimo o valor de  $\chi^2$  encontrado** (no nosso caso, 2).

Podemos calcular o valor da área vermelha do gráfico acima (o valor-p) no R através do seguinte comando:

```
1 - pchisq(2 ,df=1)
```



Encontraremos o valor de 0.1572992. Isso significa que o valor- $p = 0.1572992$ .

Portanto, nossa interpretação será essa:

**“A chance de um valor de diferença entre esperado e observado ter no mínimo o valor de  $\chi^2 = 2$  é de 15,72992%”**

Perceba que o comando `1 - pchisq( $\chi^2$ ,df=1)` retornará os valores que encontramos plotados nas tabelas de  $\chi^2$  com 1 grau de liberdade ( $df = 1$ ). Simplesmente, agora, você sabe exatamente como esses valores tabelados foram calculados.

Será que 15,72992% é uma probabilidade razoável para manter  $H_0 : D_{(esp - obs)} = 0$ ?

Infelizmente, não existe resposta simples para essa questão.

Usualmente adota-se um valor de corte arbitrário para rejeição  $H_0$ . Esse é nível de significância ( $\alpha$ ). Em diversos estudos o valor de 5% ( $\alpha = 0,05$ ) é adotado. O nível de significância equivale a área sob a curva igual a 0,05.

Podemos calcular o valor de  $\chi^2$  correspondente à área de 0.05 com o seguinte comando:

```
qchisq(1 - 0.05,df=1)
```

Este comando tem a forma geral `qchisq(1 -  $\alpha$ ,df=1)`. O valor retornado foi 3.841459.

Portanto, temos 5% de chance de encontrar, ao acaso, um valor no mínimo tão grande quanto  $\chi^2 = 3.841459$ .

Voltemos aos SNP rs699. O valor que  $\chi^2$  que encontramos foi  $\chi^2 = 87.22425$ . Qual seria a probabilidade de um valor no mínimo desta magnitude ocorrer ao acaso?

```
1 - pchisq(87.22425,df=1)
```

Veremos que esse valor é, fundamentalmente, = 0.

Desta forma a chance de uma diferença desta magnitude entre modelo e realidade ocorrer por erro amostral apenas é nula.

Portanto, é seguro afirmar que, para o SNP rs699, o modelo Hardy-Weinberg deve ser rejeitado.



## Populações finitas: modelo Wright-Fisher e deriva genética

Nesta prática, iremos implementar um código em R que simule a dinâmica populacional de Wright-Fisher.

Ao contrário da Prática 2, a cada geração o pool gamético será efetivamente amostrado para gerar os  $N$  indivíduos da geração seguinte.

Novamente, temos dois alelos no locus ( $A$  e  $a$ ) e os indivíduos da população são diplóides. Portanto, se existem  $N$  indivíduos, existem  $(2 * N)$  cópias alélicas no total. Desta forma, a cada geração, o *pool* gamético será amostrado  $(2 * N)$  vezes.

No modelo Wright-Fisher, a frequência de um alelo no *pool* gamético é a mesma da população parental que o criou. O tamanho populacional  $N$  ficará constante ao longo das gerações.

Em nossa implementação em R, começaremos com uma população com 20 indivíduos diplóides. Na geração 1 ( $G_1$ ), a frequência do alelo  $A$  será igual a 50% ( $f_A = 0.5$ ) e o processo ocorrerá por 50 gerações (esses valores podem, obviamente, ser ajustados futuramente):

```
# Código implementando o modelo Wright-Fisher
# Variáveis iniciais:

N = 20
g = 50

# frequência de A em G1:
fA = vector()
fA[1] = 0.5

# número de alelos A em G1:
nA = vector()
nA[1] = fA[1] * (2*N)

possibilidades = seq(0, (2*N), 1)
```

Nas linhas acima, estabelecemos os valores iniciais do tamanho da população ( $N$ ), do número de gerações que a simulação correrá ( $g$ ). A frequência do alelo  $A$ ,  $f_A$ , é um vetor cuja primeira posição ( $G_1$ ) será ocupado pela frequência inicial ( $=0.5$ ). Também temos um outro vetor com o número de alelos  $A$  em cada geração. Esse valor é obtido multiplicando a frequência do alelo pelo número de cópias alélicas totais.

A variável `possibilidades` é um pouco abstrata, mas será essencial no nosso código. Após entrar com o código acima, vamos ver qual é o seu conteúdo:

```
> possibilidades
 [1] 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
[18] 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
[35] 34 35 36 37 38 39 40
```

Ou seja, `possibilidades` é um vetor com 41 posições que guardam os números de 0 a 40. O que isso significa?



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

Lembremos que existem  $N = 20$  indivíduos em nossa simulação. Conseqüentemente, existem  $2^*N = 40$  cópias alélicas no total. A variável `possibilidades` indica todas as possibilidades de composição do número de alelos  $A$  numa população de 20 diplóides.

Como assim? Se `possibilidades = 0`, indica que o número de alelos  $A$  é 0. Se `possibilidades = 35` indica que o número de alelos  $A$  é 35 e assim por diante. O número máximo é 40, pois existem  $(2^*N) = 40$  cópias alélicas no total.

Iniciaremos agora o *loop* para calcular dinâmica da frequência do alelo  $A$  ao longo das  $g = 50$  gerações:

```
for(i in c(2:g)){  
  probabilidades = dbinom(possibilidades, (2*N), fA[i-1] )  
  nA[i] = sample(possibilidades, 1, prob=probabilidades)  
  fA[i] = nA[i] / (2*N)  
}
```

Perceba que o *loop* começa em 2 a vai até  $g$  (ou seja, 50). Começa em 2, pois os valores de  $fA$  e  $nA$  da primeira geração já foram calculados antes do loop. Então, queremos os valores de 2 até 50.

Dentro do loop, o primeiro comando é:

```
probabilidades = dbinom(possibilidades, (2*N), fA[i-1] )
```

O que esse comando faz? A função **dbinom** retorna as probabilidades binomiais. Mas probabilidade de quê? Neste caso, retorna a probabilidade de cada uma das composições de número de alelos  $A$  do vetor `possibilidades`.

Como a frequência do alelo  $A$  em  $G_1$  é 0.5, a primeira volta do *loop* (ou seja,  $G_2$ ) calculará:

```
probabilidades = dbinom(possibilidades, 40, 0.5 )
```

O vetor `probabilidades` terá, então, os seguintes valores:

```
> probabilidades  
[1] 9.094947e-13 3.637979e-11 7.094059e-10 8.985808e-09  
[5] 8.311872e-08 5.984548e-07 3.490986e-06 1.695622e-05  
[9] 6.994440e-05 2.486912e-04 7.709428e-04 2.102571e-03  
[13] 5.081214e-03 1.094415e-02 2.110658e-02 3.658474e-02  
[17] 5.716365e-02 8.070163e-02 1.031187e-01 1.194007e-01  
[21] 1.253707e-01 1.194007e-01 1.031187e-01 8.070163e-02  
[25] 5.716365e-02 3.658474e-02 2.110658e-02 1.094415e-02  
[29] 5.081214e-03 2.102571e-03 7.709428e-04 2.486912e-04  
[33] 6.994440e-05 1.695622e-05 3.490986e-06 5.984548e-07  
[37] 8.311872e-08 8.985808e-09 7.094059e-10 3.637979e-11  
[41] 9.094947e-13
```

Mas o que isso significa? Significa as probabilidades de ocorrência de cada uma das composições de número de alelos (de 0 a 40) quando amostramos 40 vezes o alelo  $A$  de um pool gamético com  $fA = 0.5$ .

Em síntese, é o resultado da função de distribuição binomial:



$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Neste caso,  $k$  será cada um dos números de alelos contidos em possibilidades. Por exemplo, qual é probabilidade de obtenção de  $k = 0$  alelos  $A$  quando realizamos  $n = 40$  sorteios de um pool gamético em que  $fA = p = 0.5$ ? A distribuição binomial nos diz isso:

$$\Pr(k = 0) = \binom{40}{0} 0.5^0 (1 - 0.5)^{40-0} = 9.094947 \times 10^{-13}$$

Por isso, o primeira posição do vetor `probabilidades[1]`, é ocupada por  $9.094947e-13$ . Ela equivale à probabilidade de `possibilidades[1]`, que é justamente o estado de 0 alelos  $A$ .

Similarmente, podemos calcular a probabilidade de sortear  $k = 20$  alelos  $A$  quando realizamos 40 sorteios de um pool gamético com  $fA = p = 0.5$ :

$$\Pr(k = 20) = \binom{40}{20} 0.5^{20} (1 - 0.5)^{40-20} = 0.1253707$$

A 21ª posição do vetor `probabilidades[21]` é ocupada por  $1.253707e-01$ . Ela equivale à probabilidade de `possibilidades[21]`, que é o estado de 20 alelos  $A$ .

**Ou seja, ao realizar 40 sorteios independentes de um *pool* gamético contendo alelos  $A$  em frequência de 0.5, é muito mais provável obter 20 alelos  $A$  do que 0!**

Voltemos agora ao *loop*.

O segundo comando `loop` é:

```
nA[i] = sample(possibilidades, 1, prob=probabilidades)
```

O que esse comando faz? A função **sample** vai retornar para a variável `nA` o número de alelos  $A$  que foi sorteado do *pool*. Apenas 1 sorteio de composição de alelos contida no vetor `possibilidades` será realizado. Por isso o número 1 no meio.

Entretanto, a chance de sorteio de cada uma das composições do vetor `possibilidades` será modulada pelo vetor `probabilidades`. Por exemplo, ao chamar o comando desta linha, obtemos o seguinte:

```
> sample(possibilidades, 1, prob=probabilidades)
[1] 26
```

O valor que foi retornado, 26, significa que sorteamos do *pool* gamético 26 alelos  $A$  para compor a população da geração  $i$ .

Por fim, o comando da próxima linha calcula a frequência do alelo  $A$  na geração atual do *loop*:

```
fA[i] = nA[i] / (2*N)
```



Observe que, na próxima chamada do *loop*, para compor a próxima geração, trabalharemos com a composição da geração imediatamente anterior. Note que a próxima chamada de *probabilidades* no *loop* será baseada em  $fA[i - 1]$ , a frequência do alelo *A* na geração imediatamente anterior:

```
probabilidades = dbinom(possibilidades, (2*N), fA[i-1] )
```

Após o *loop*, o vetor *fA* conterá as frequências do alelo *A* em cada uma das 50 gerações. O alelo poderá ser perdido (terminará com frequência = 0), fixado (terminará com frequência = 1) ou continuará entre 0 e 1:

#### Perdido:

```
> fA
 [1] 0.500 0.500 0.450 0.400 0.525 0.525 0.500 0.400
 [9] 0.400 0.525 0.350 0.400 0.550 0.525 0.600 0.400
[17] 0.425 0.400 0.450 0.500 0.425 0.475 0.550 0.450
[25] 0.350 0.275 0.350 0.425 0.400 0.325 0.225 0.150
[33] 0.025 0.025 0.000 0.000 0.000 0.000 0.000 0.000
[41] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[49] 0.000 0.000
```

#### Fixado:

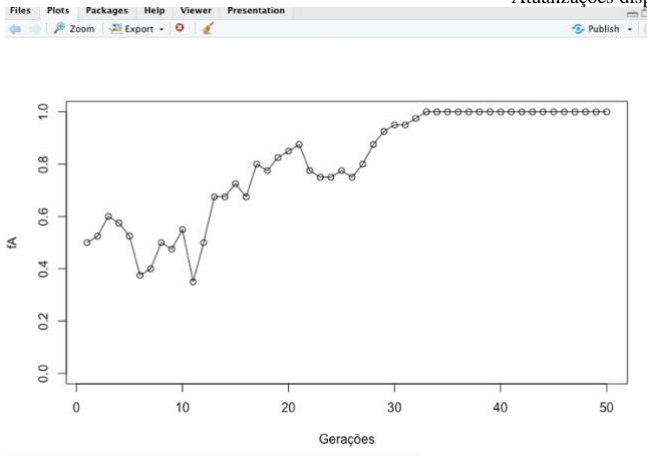
```
> fA
 [1] 0.500 0.525 0.600 0.575 0.525 0.375 0.400 0.500
 [9] 0.475 0.550 0.350 0.500 0.675 0.675 0.725 0.675
[17] 0.800 0.775 0.825 0.850 0.875 0.775 0.750 0.750
[25] 0.775 0.750 0.800 0.875 0.925 0.950 0.950 0.975
[33] 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
[41] 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000
[49] 1.000 1.000
```

#### Ou entre 0 e 1:

```
> fA
 [1] 0.500 0.400 0.325 0.300 0.300 0.425 0.500 0.575
 [9] 0.450 0.600 0.650 0.725 0.675 0.650 0.550 0.525
[17] 0.675 0.550 0.650 0.675 0.700 0.700 0.600 0.525
[25] 0.575 0.675 0.700 0.675 0.650 0.700 0.750 0.700
[33] 0.775 0.825 0.725 0.850 0.850 0.825 0.800 0.700
[41] 0.750 0.750 0.650 0.500 0.450 0.500 0.450 0.425
[49] 0.425 0.600
```

Podemos enfim plotar a evolução da frequência do alelo *A* ao longo das 50 gerações:

```
plot(fA, ylim=c(0,1), xlab="Gerações")
lines(fA)
```



## Estudando fenômenos estocásticos por cadeias de Markov I

### Parte 1: Estudo qualitativo

Nesta prática continuaremos a análise da dinâmica das frequências alélicas dentro do modelo de reprodução de Wright-Fisher. Entretanto, agora, ao invés de aplicarmos a dinâmica apenas 1x, iremos estudar o que ocorre quando **diversas populações idênticas com o mesmo estado inicial** (replicatas) são simuladas **independentemente**.

Verificaremos que esta replicação do **mesmo processo** estocástico várias vezes resulta no **surgimento de padrões** que, em última análise, permite uma predição probabilística dos estados futuros possíveis mesmo na presença do acaso. É uma estratégia de análise chamada de cadeia de Markov.

Esta prática está dividida em duas partes. Na **Parte 1** (este documento), iremos apenas observar a dinâmica das diversas replicatas (ou seja, evoluções independentes de populações com o mesmo estado inicial). Na **Parte 2**, um pouco mais avançada, veremos como quantificar essa dinâmica usando matrizes de probabilidade de transição.

Nesta primeira parte, apenas vamos repetir a mesma dinâmica de Wright-Fisher várias vezes. Essas várias vezes são exatamente as replicatas. Lembremos do código do modelo reprodutivo Wright-Fisher:

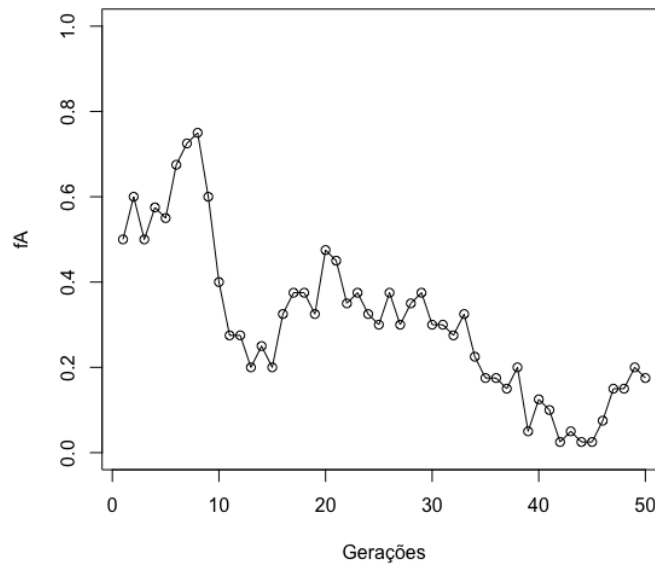
```
# Código implementando o modelo Wright-Fisher
# Variáveis iniciais:
N = 20
g = 50
# frequência de A em G1:
fA = vector()
fA[1] = 0.5
# número de alelos A em G1:
nA = vector()
nA[1] = fA[1] * (2*N)
possibilidades = seq(0, (2*N), 1)

for(i in c(2:g)){
  probabilidades = dbinom(possibilidades, (2*N), fA[i-1] )
```

```
nA[i] = sample(possibilidades, 1, prob=probabilidades)
fA[i] = nA[i] / (2*N)
}
```

Exatamente como fizemos antes, o vetor de frequências ao longo das gerações  $fA$  pode ser plotado usando os comandos:

```
plot(fA, ylim=c(0,1), xlab="Gerações")
lines(fA)
```



Observamos que, nesta dinâmica específica, após  $g = 50$  gerações, o alelo  $A$  não está fixado ou perdido.

Vamos agora repetir esse mesmo processo 10x. Ou seja, o número de replicatas será = 10. Para cada vez que o processo for repetido, a população começará exatamente no mesmo estado em  $G_1$ , com  $fA = 0.5$  ( $fA[1] = 0.5$ ).

Toda vez que fizermos uma replicata, iremos plotar os resultados no gráfico:

```
# Código implementando o modelo Wright-Fisher - cadeia de Markov Parte 1
# Variáveis iniciais:

N = 20
g = 50
replicatas = 10
possibilidades = seq(0, (2*N), 1)
fA_G1 = 0.5      # frequência de A em G1:
nA_G1 = fA_G1 * (2*N) # número de alelos A em G1

fA = matrix(nrow=replicatas, ncol=g)

for(r in (1: replicatas)){

  fA[r,1] = fA_G1
  nA = NULL
```





```
nA[1] = nA_G1

for(i in c(2:g)){
  probabilidades = dbinom(possibilidades, (2*N), fA[r,i-1] )
  nA[i] = sample(possibilidades, 1, prob=probabilidades)
  fA[r,i] = nA[i] / (2*N)
}
}
```

Vamos entender o código acima. Basicamente, **fizemos o modelo Wright-Fisher 10x independentemente**. Ao contrário do código simples de Wright-Fisher, agora as frequências do alelo A ao longo das  $g = 50$  gerações estão guardadas numa **matriz**. As matrizes podem ser entendidas como tabelas do Excel, por exemplo.

Esta matriz possui as seguintes dimensões: o número de colunas é igual ao número de gerações (50) e o número de linhas é igual ao número de replicatas (10). Ela foi criada com o comando:

```
fA = matrix(nrow=replicatas, ncol=g)
```

**Após rodar** o Wright-Fisher para as 10 replicatas, podemos visualizar as primeiras 25 colunas (gerações) das 10 replicatas:

|    | V1  | V2    | V3    | V4    | V5    | V6    | V7    | V8    | V9    | V10   | V11   | V12   | V13   | V14   | V15   | V16   | V17   | V18   | V19   | V20   | V21   | V22   | V23   | V24   | V25   |       |
|----|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1  | 0.5 | 0.650 | 0.800 | 0.725 | 0.725 | 0.700 | 0.775 | 0.825 | 0.900 | 0.925 | 0.975 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 2  | 0.5 | 0.450 | 0.525 | 0.425 | 0.500 | 0.650 | 0.600 | 0.550 | 0.500 | 0.525 | 0.475 | 0.350 | 0.275 | 0.275 | 0.300 | 0.325 | 0.225 | 0.275 | 0.375 | 0.350 | 0.275 | 0.275 | 0.225 | 0.150 | 0.275 |       |
| 3  | 0.5 | 0.450 | 0.525 | 0.525 | 0.525 | 0.450 | 0.375 | 0.350 | 0.250 | 0.300 | 0.350 | 0.250 | 0.325 | 0.350 | 0.300 | 0.250 | 0.300 | 0.350 | 0.225 | 0.150 | 0.150 | 0.025 | 0.000 | 0.000 | 0.000 |       |
| 4  | 0.5 | 0.500 | 0.450 | 0.500 | 0.525 | 0.350 | 0.525 | 0.550 | 0.675 | 0.575 | 0.625 | 0.675 | 0.650 | 0.625 | 0.675 | 0.675 | 0.625 | 0.700 | 0.700 | 0.800 | 0.775 | 0.800 | 0.825 | 0.825 | 0.800 |       |
| 5  | 0.5 | 0.475 | 0.550 | 0.625 | 0.625 | 0.525 | 0.575 | 0.475 | 0.475 | 0.550 | 0.750 | 0.750 | 0.700 | 0.700 | 0.725 | 0.600 | 0.500 | 0.450 | 0.450 | 0.475 | 0.500 | 0.575 | 0.600 | 0.575 | 0.550 |       |
| 6  | 0.5 | 0.450 | 0.450 | 0.425 | 0.300 | 0.375 | 0.275 | 0.300 | 0.425 | 0.500 | 0.450 | 0.575 | 0.575 | 0.700 | 0.600 | 0.600 | 0.650 | 0.625 | 0.675 | 0.775 | 0.850 | 0.850 | 0.825 | 0.875 | 0.925 |       |
| 7  | 0.5 | 0.625 | 0.600 | 0.750 | 0.800 | 0.725 | 0.675 | 0.625 | 0.750 | 0.575 | 0.425 | 0.325 | 0.225 | 0.275 | 0.275 | 0.225 | 0.250 | 0.300 | 0.300 | 0.325 | 0.450 | 0.475 | 0.525 | 0.700 | 0.725 |       |
| 8  | 0.5 | 0.550 | 0.825 | 0.650 | 0.650 | 0.650 | 0.600 | 0.750 | 0.625 | 0.800 | 0.900 | 0.900 | 0.950 | 0.975 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |       |
| 9  | 0.5 | 0.475 | 0.275 | 0.325 | 0.250 | 0.175 | 0.225 | 0.375 | 0.400 | 0.525 | 0.500 | 0.450 | 0.425 | 0.450 | 0.450 | 0.500 | 0.475 | 0.500 | 0.525 | 0.525 | 0.500 | 0.500 | 0.500 | 0.475 | 0.350 |       |
| 10 | 0.5 | 0.400 | 0.275 | 0.300 | 0.350 | 0.350 | 0.250 | 0.250 | 0.175 | 0.150 | 0.275 | 0.250 | 0.300 | 0.175 | 0.125 | 0.075 | 0.100 | 0.100 | 0.075 | 0.100 | 0.025 | 0.000 | 0.000 | 0.000 | 0.000 |       |

Conforme mostrado, na geração 1, todas as 10 replicatas estão com frequência = 0.5, pois este foi o valor de frequência inicial. Na segunda geração (coluna V2), cada população tomou um caminho evolutivo diferente, afinal, trata-se de um processo estocástico.

Em síntese, nesta matriz, as linhas indicam toda a evolução das frequências alélicas ao longo das 50 gerações. Cada processo Wright-Fisher ocupa uma linha diferente.

Ao final dos 10 processos independentes (linhas), podemos verificar as frequências alélicas em cada uma das 10 replicatas:



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

| V43   | V44   | V45   | V46   | V47   | V48   | V49   | V50   |
|-------|-------|-------|-------|-------|-------|-------|-------|
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.075 | 0.025 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.425 | 0.250 | 0.325 | 0.300 | 0.250 | 0.275 | 0.325 | 0.375 |
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.425 | 0.400 | 0.475 | 0.525 | 0.525 | 0.550 | 0.550 | 0.475 |
| 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.400 | 0.475 | 0.400 | 0.250 | 0.225 | 0.200 | 0.250 | 0.275 |
| 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Após 50 gerações, encontramos 4 replicatas em que o alelo *A* foi **fixado** ( $f_A = 1.0$ ) e 3 replicatas em que ele foi **perdido** ( $f_A = 0.0$ ). Em outras 3 replicatas, a frequência está entre 0-1 e, portanto, *A* e *a* ainda estão segregando nas populações.

Resumindo, ao invés de rodar o Wright-Fisher simples 10x, construímos um *loop* que vai preenchendo a matriz *fA* com essas informações.

Para isso, tivemos que fazer um *loop* dentro de outro *loop*. O primeiro *loop* é o das replicatas, que vai de 1 até 10, com valores guardados na variável *r*. E o segundo *loop* é o da geração, que vai de 2 até 50, com valores guardados na variável *i*.

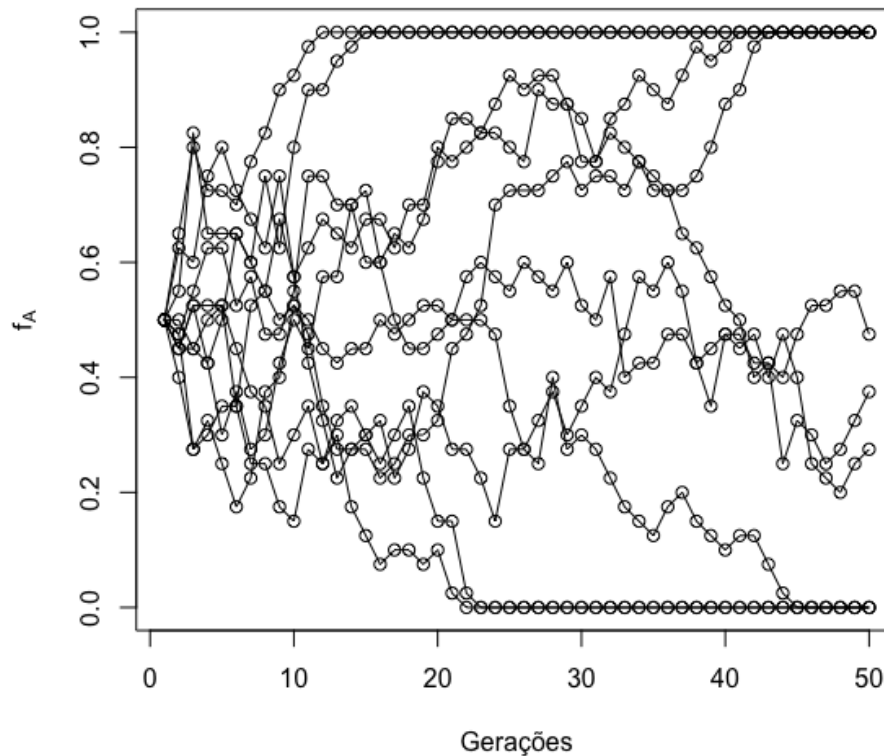
Então, o código inicialmente preenche as 50 colunas da linha 1 ( $r = 1$ ) com as frequências. Depois reinicia no mesmo estado inicial e preenche as 50 colunas da linha 2 ( $r = 2$ ) com as frequências, assim por diante.

Por isso, quando a matriz recebe o valor de frequência, indicamos exatamente a posição (célula da tabela) em que o valor entrará:  $f_{A[r,i]}$ . Ou seja, linha *r* e coluna *i*.

Para plotar todas as 10 evoluções independentes da população a partir do mesmo estado inicial, vamos usar o seguinte comando:

```
plot(fA[1,], ylim=c(0,1), xlab="Gerações", ylab=expression('f' [A]))
lines(fA[1,])

for(i in c(2:replicatas)){
  points(fA[i,])
  lines(fA[i,])
}
```

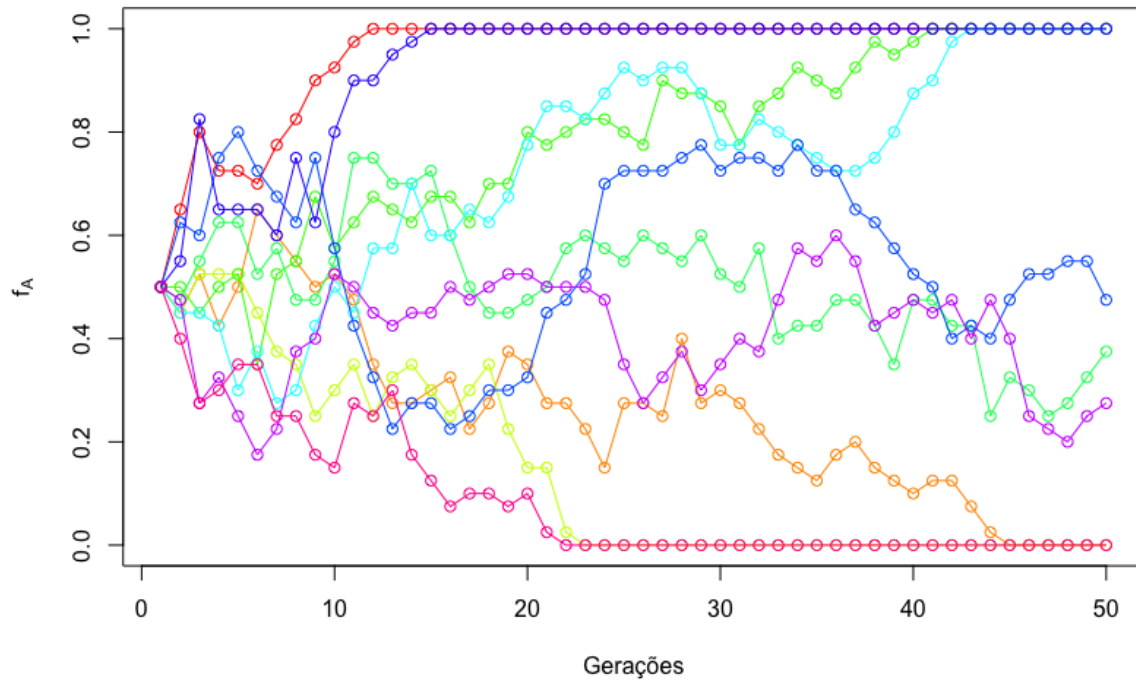


Podemos verificar que 4 replicatas atingiram a fixação, enquanto 3 foram perdidas. Exatamente conforme visualizamos na tabela.

Podemos usar cores diferentes para cada replicata para tornar a visualização de cada processo Wright-Fisher independente mais fácil:

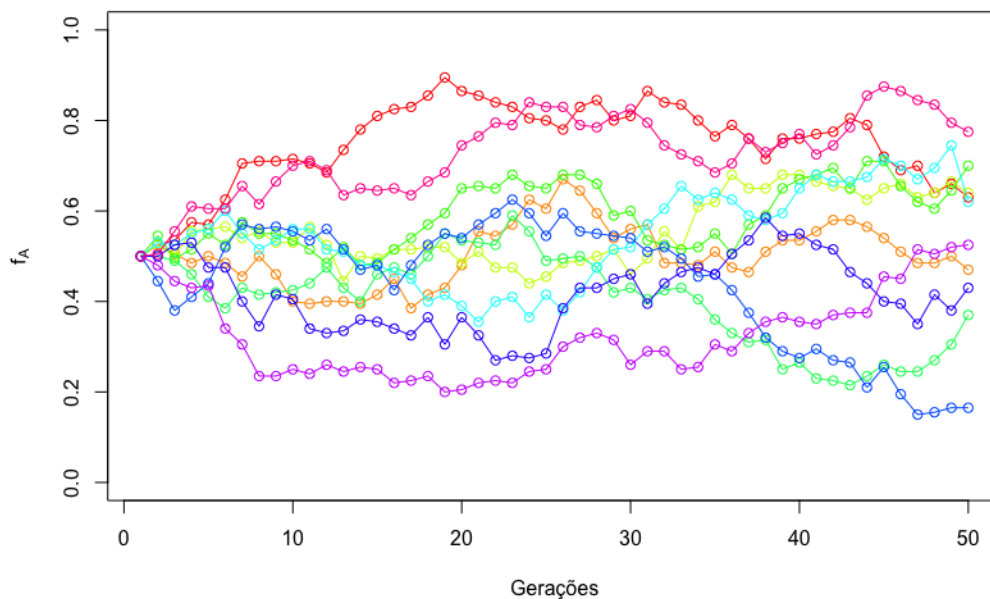
```
cores = rainbow(replicatas)
plot(fA[1,], ylim=c(0,1), xlab="Gerações", ylab=expression('f'[A]), col=cores[1])
lines(fA[1,],col=cores[1])

for(i in c(2:replicatas)){
  points(fA[i,], col=cores[i])
  lines(fA[i,], col=cores[i])
}
```



A função **rainbow** retornou replicatas = 10 cores diferentes retiradas da escala cromática do arco-íris.

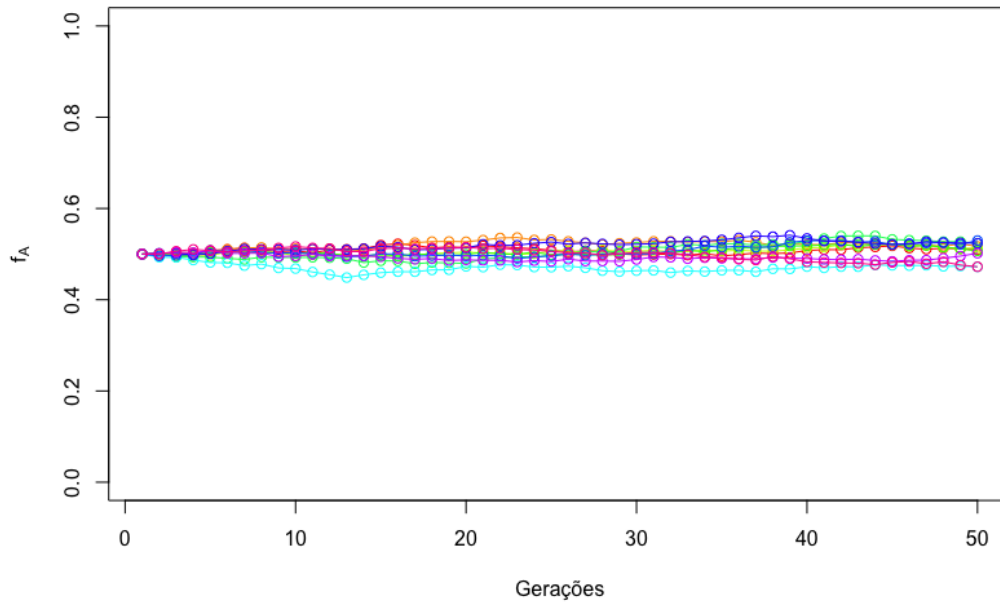
Tente agora aumentar o tamanho da população para **N = 100**. Aqui está o gráfico que eu obtive:



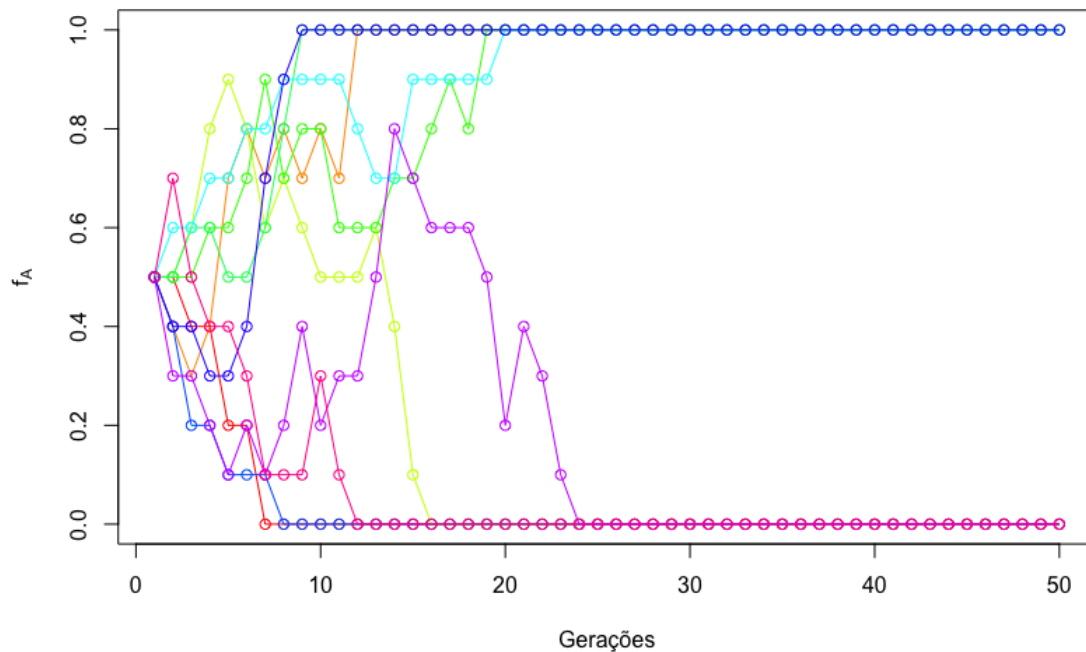


Quais são as diferenças mais notáveis? Em 50 gerações, alguma população (replicata) está com o alelo A fixado ou perdido?

Tentemos agora **N = 10000**. Aqui está o que eu obtive:



Por fim, que tal tentar uma população bem pequena. **N = 5**:



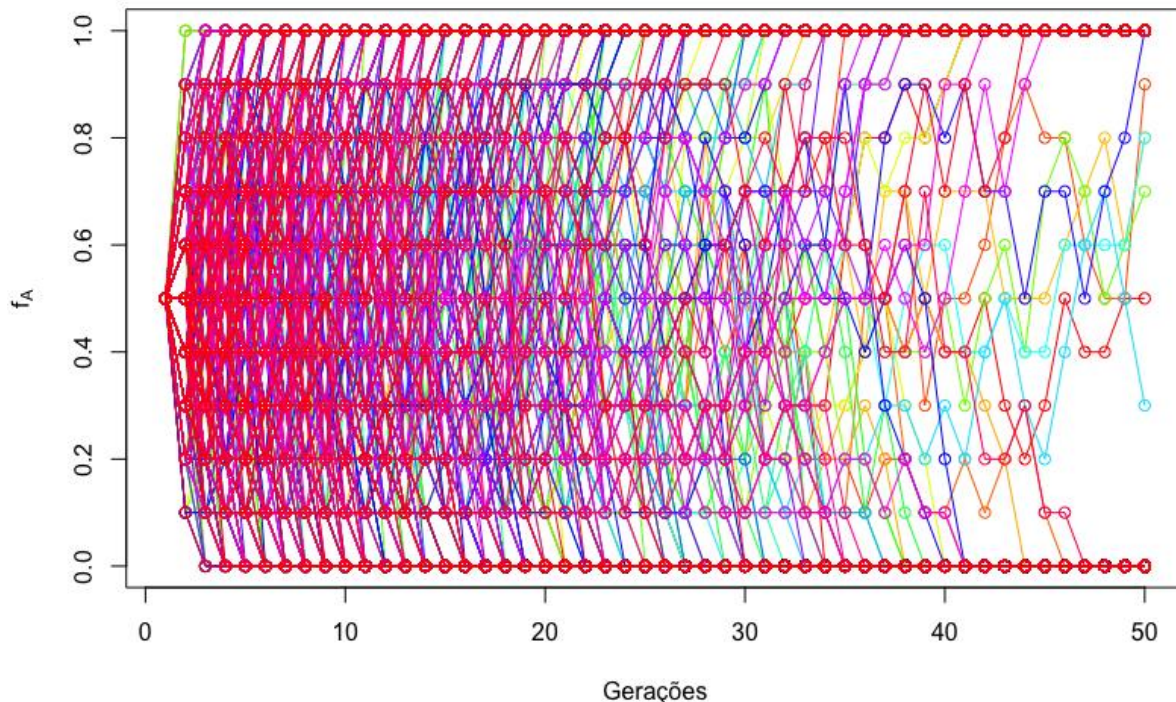
De forma geral, é notável que, quanto menor o tamanho populacional, maior será a variação da frequência do alelo A entre as gerações. Desta forma, a chance do alelo A ser fixado ou perdido em poucas gerações é alta.

Um resultado importante da abordagem que vimos acima é que **a frequência de replicatas em que o alelo A atinge fixação é aproximadamente igual à frequência inicial do alelo A**. Por isso, concluímos que

*“A probabilidade de fixação de um alelo neutro por deriva é igual a sua frequência inicial”*

Esta é uma previsão sobre um sistema estocástico que foi possível pela abordagem da cadeia de Markov.

Vejamus isso na prática. Ao invés de *replicatas* = 10, faremos agora *replicatas* = 1000. Desta forma, teremos 1000 evoluções independentes da população a partir do mesmo inicial. Continuemos com  $N = 5$ . Pela propriedade acima, como a frequência inicial do alelo A nas replicatas é 0.5, **esperamos que  $0.5 \times 1000 = 500$  replicatas tenham o alelo A fixado**. Vejamos:



Podemos obter as frequências na geração 50 de cada uma das 1000 replicatas chamando a última coluna (a 50ª coluna) da variável matriz  $fA$ :

```
fA[, 50]
```



```
> fA[,50]
 [1] 0.0 0.0 0.0 1.0 1.0 1.0 1.0 1.0 1.0 0.0 0.0 0.0 0.0 1.0 0.0 1.0 0.0
 [17] 0.5 1.0 1.0 0.0 0.0 1.0 0.0 1.0 0.0 1.0 1.0 0.0 0.0 1.0 1.0 0.0
 [33] 0.0 0.0 0.0 1.0 1.0 0.0 0.0 0.0 1.0 0.0 1.0 0.0 1.0 0.0 0.0 0.0
 [49] 0.0 0.0 0.0 1.0 0.0 0.0 1.0 1.0 0.0 1.0 1.0 1.0 1.0 0.0 1.0 0.9
 [65] 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 1.0 0.0 1.0 0.0 0.0
 [81] 0.0 0.0 1.0 0.0 1.0 1.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
 [97] 0.0 1.0 1.0 1.0 1.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 1.0 0.0 1.0
 [113] 0.0 1.0 1.0 1.0 1.0 0.0 0.0 1.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 1.0
 [129] 1.0 1.0 0.0 0.8 1.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 0.0
 [145] 0.0 0.0 0.0 0.0 0.0 1.0 1.0 1.0 1.0 1.0 0.0 0.0 0.0 0.0 1.0 1.0
 [161] 0.0 1.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 1.0 1.0 1.0 0.0 1.0
 [177] 1.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0 1.0 1.0 0.0 1.0 0.0 0.0 1.0 0.0
 [193] 1.0 0.0 1.0 0.0 0.0 0.0 1.0 0.0 1.0 1.0 0.0 0.0 0.0 1.0 1.0 1.0
 [209] 1.0 1.0 1.0 0.0 0.0 1.0 0.0 1.0 1.0 1.0 0.0 1.0 1.0 1.0 0.0 0.0
 [225] 0.0 0.0 1.0 0.0 0.0 1.0 1.0 1.0 1.0 1.0 0.0 1.0 0.0 0.0 0.0 1.0
 [241] 1.0 1.0 0.0 0.0 1.0 1.0 0.0 1.0 1.0 0.0 0.0 1.0 0.7 1.0 0.0 0.0
 [257] 0.0 1.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 1.0 1.0 0.0 0.0 1.0 0.0
```

Conforme observamos, quase todas as replicatas estão com o alelo *A* fixado (1.0) ou perdido (0.0). Existem algumas poucas replicatas em que *A* e *a* continuam segregando na  $g = 50$ .

Para contar quantas replicatas apresentam  $f_A = 1.0$  (fixado), basta usar o comando:

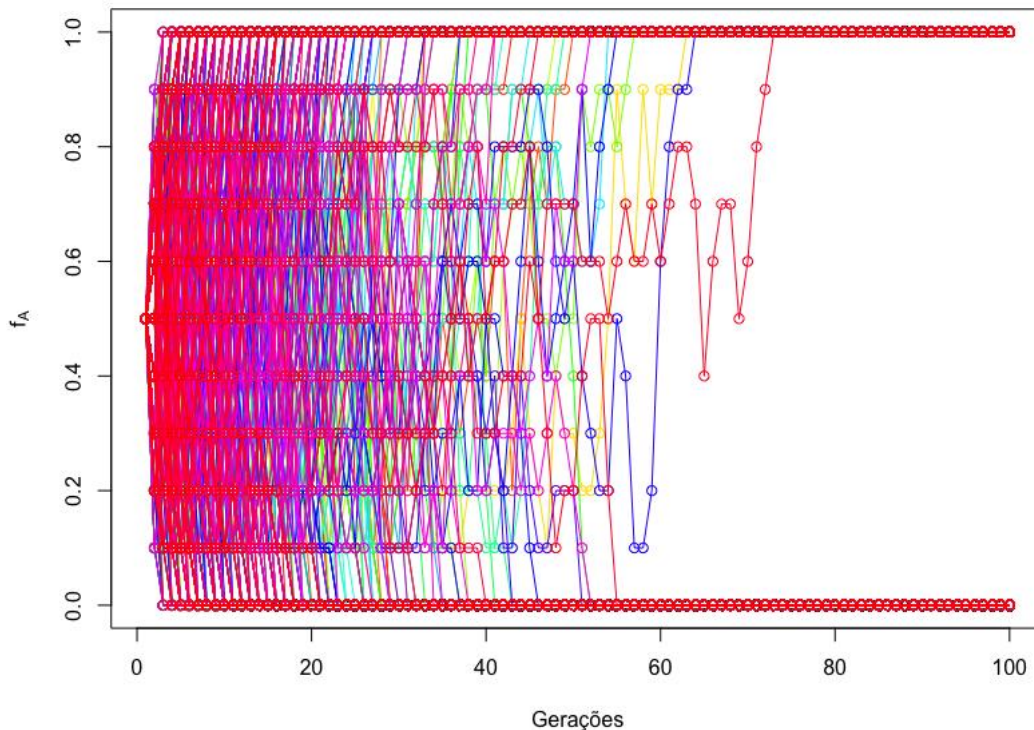
```
length( which(fA[,50] == 1) )
```

Neste comando, `which(fA[,50] == 1)` retornará quais linhas da coluna 50 apresentam o valor de 1.0. E `length` retornará o comprimento do vetor retornado por `which`.

Após o comando, no meu experimento, obtive 485 replicatas com *A* fixado. Um número bem próximo da previsão de 500.

Notemos que, na geração 50, ainda existem replicatas com ambos os alelos segregando. Para obter todas as replicatas nos estados fixado (1.0) ou perdido (0.0), devemos aumentar o número de gerações em que o Wright-Fisher é rodado.

Façamos  $g = 100$  (ainda com  $N = 5$  e  $replicatas = 1000$ ):



Agora, todas as replicatas estão com o alelo *A* fixado ou perdido. Aplicando o comando `length(which(fA[,100] == 1))` verificamos que 480 replicatas tiveram o alelo *A* fixado (o esperado era 500).

Para obtermos um valor mais próximo do esperado, devemos aumentar o número de replicatas. No meu experimento, quando `replicatas = 10000`, o número de replicatas com o alelo *A* fixado foi 4997 (quase idêntico ao esperado de 5000).

Finalmente, é importante observar que **quanto maior for o tamanho da população, mais gerações de Wright-Fisher deverão ser rodadas para obter todas as replicatas no estado de fixação ou perda.**

## Estudando fenômenos estocásticos por cadeias de Markov II

### Parte 2: Quantificação

Nesta prática continuaremos a análise da dinâmica das frequências alélicas dentro do modelo de reprodução de Wright-Fisher. Entretanto, agora, ao invés de aplicarmos a dinâmica apenas 1x, iremos estudar o que ocorre quando **diversas populações idênticas com o mesmo estado inicial** (replicatas) são simuladas **independentemente**.

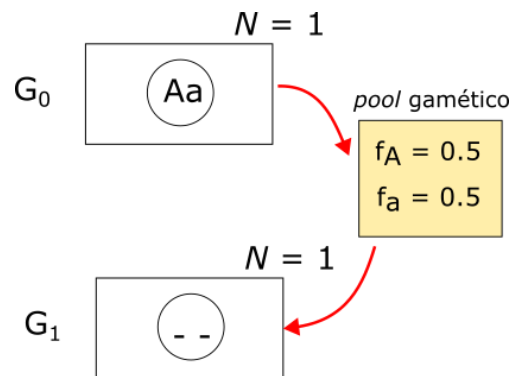


Verificaremos que esta replicação do **mesmo processo** estocástico várias vezes resulta no **surgimento de padrões** que, em última análise, permite uma predição probabilística dos estados futuros possíveis mesmo na presença do acaso. É uma estratégia de análise chamada de cadeia de Markov.

Esta prática está dividida em duas partes. Na **Parte 1**, iremos apenas observar a dinâmica das diversas replicatas (ou seja, evoluções independentes de populações com o mesmo estado inicial). Na **Parte 2** (este documento), um pouco mais avançada, veremos como quantificar essa dinâmica usando matrizes de probabilidade de transição.

Nesta segunda parte, iremos usar a distribuição binomial para calcular a probabilidade de transição de estados entre todas as possibilidades de composição alélica de uma população entre duas gerações.

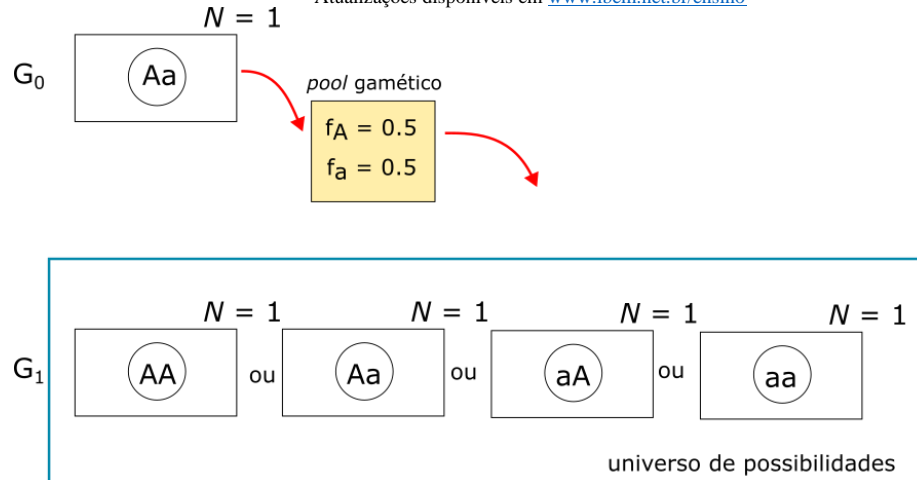
Estudemos um exemplo muito simples: uma população de apenas  $N = 1$  indivíduo heterozigoto  $Aa$ . Neste caso, o pool gamético composto por essa população terá  $f_A = 0.5$  e  $f_a = 0.5$ . O único indivíduo da próxima geração será composto amostrando esse *pool* gamético:



Essa estratégia nada mais é que o processo reprodutivo Wright-Fisher. Aprendemos que, neste processo, a probabilidade das diversas composições alélicas da próxima geração pode ser obtida através das probabilidades calculadas pela distribuição binomial. Isso foi o que fizemos na **Parte 1**.

Nesta **Parte 2**, iremos adotar uma estratégia diferente. Ao invés de amostrar efetivamente o pool gamético, **vamos considerar TODAS as possibilidades de composição genotípica de  $G_1$** .

Ora, no caso de  $N = 1$ , o universo de possibilidades de composição para a geração seguinte é relativamente pequeno. Vejamos no esquema:



Ou seja, com  $N = 1$ , a geração seguinte é composta de:

1. 1 possibilidade em que existem 2 alelos  $A$  ( $AA$ )
2. 2 possibilidades em que existe 1 alelo  $A$  ( $Aa$  ou  $aA$ )
3. 1 possibilidade em que existe 0 alelo  $A$  ( $aa$ ).

A probabilidade de obtenção de cada uma das possibilidades através da amostragem de um *pool* gamético com  $f_A = 0.5$  é obtida pela distribuição binomial:

$$\Pr(X = k) = \binom{n}{k} f_A^k (1 - f_A)^{(n-k)}$$

No nosso caso, amostramos o pool  $n = 2$  vezes e  $f_A = 0.5$ :

$$\Pr(X = k) = \binom{2}{k} 0.5^k (1 - 0.5)^{(2-k)}$$

O valor de  $k$  será igual a 2 para o caso (1) acima (2 alelos  $A$  amostrados);  $k = 1$  no caso (2) e  $k = 0$  no caso (3) acima.

Façamos então os cálculos para cada uma das possibilidades:

1. 01 possibilidade em que existem 2 alelos  $A$  ( $AA$ ):

$$\Pr(X = 2) = \binom{2}{2} 0.5^2 (1 - 0.5)^{(2-2)} = 0.25$$

2. 02 possibilidades em que existe 1 alelo  $A$  ( $Aa$  ou  $aA$ ):

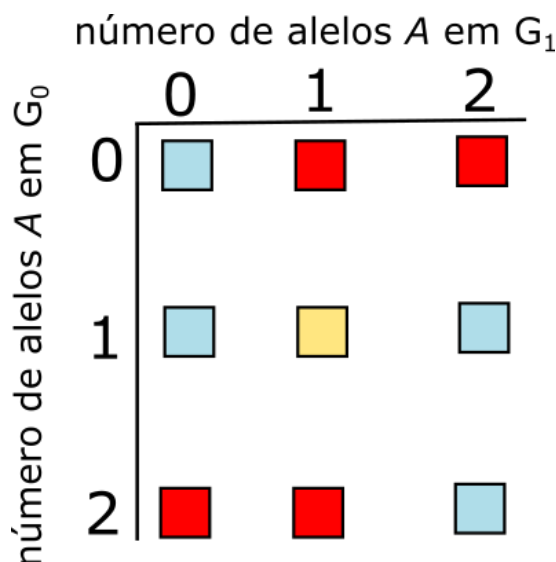
$$\Pr(X = 1) = \binom{2}{1} 0.5^1 (1 - 0.5)^{(2-1)} = 0.5$$

3. 01 possibilidade em que existe 0 alelo  $A$  ( $aa$ ).

$$\Pr(X = 0) = \binom{2}{0} 0.5^0 (1 - 0.5)^{(2-0)} = 0.25$$

Desta forma, a chance de amostrarmos dois alelos *A* na geração seguinte é 0.25. De amostrar apenas 1 alelo *A* é 0.5 e não amostrar alelo *A* algum é 0.25.

Esses valores que calculamos acima podem ser armazenados numa matriz especial, onde nas linhas temos os valores da composição alélica inicial (os estados iniciais) e, nas colunas, as composições alélicas na geração seguinte. Os elementos da matriz (as áreas quadriculadas) indicarão a **probabilidade de sair de uma dada composição em  $G_0$  e ir para uma composição seguinte em  $G_1$** :



O elemento da primeira linha com a primeira coluna (marcado com a cor azul) indica, portanto, a probabilidade de uma composição de 0 alelos *A* em  $G_0$  mudar para a conformação de 0 alelos *A* em  $G_1$ . Se em  $G_0$  não existem mais alelos *A*, significa que nesta geração este alelo está perdido (o alelo *a* foi fixado). Assim, a probabilidade de sair de uma composição de 0 alelos *A* em  $G_0$  mudar para a conformação de 1 ou 2 alelos *A* em  $G_1$  (marcados em vermelho na primeira linha) é evidentemente nula **na ausência de mutação ou migração**. Desta forma, a primeira linha desta matriz deverá conter (1,0,0).

Algo semelhante ocorre na última linha, aquela em que a população possui 100% de alelos *A* em  $G_0$  (*A* foi fixado). As mudanças de  $2 \rightarrow 0$  ou de  $2 \rightarrow 1$  não são possíveis (probabilidade nula). Apenas o último elemento da linha não é nulo (=1). Desta forma, a última linha contém (0,0,1).

A linha do meio é a mais interessante. Quando temos apenas 1 alelo *A* em  $G_0$  (que é justamente o exemplo que foi mostrado na figura acima), existe a possibilidade de fixação do *a* ( $1 \rightarrow 0$ ); manutenção dos dois alelos ( $1 \rightarrow 1$ ) ou fixação do *A* ( $1 \rightarrow 2$ ). Exatamente como no exemplo acima. Já calculamos a chance de cada uma dessas possibilidades: (0.25,0.5,0.25).

De forma geral, indicamos em azul os elementos da matriz que fixação, em vermelho aqueles que possui probabilidade nula e, em amarelo, o único estado em que ambos os alelos continuam segregando.

Portanto, de  $G_0$  para  $G_1$ , temos a seguinte **matriz de probabilidade de transição**:



$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0.25 & 0.5 & 0.25 \\ 0 & 0 & 1 \end{pmatrix}$$

Perceba que, nesta matriz, todos os elementos das linhas devem necessariamente somar 1. Além disso, é fundamental lembrar que cada um dos elementos da matriz foi obtido usando as probabilidades binomiais. De forma geral, no caso mais simples de  $N = 1$ , podemos representar a matriz assim:

$$Q = \begin{pmatrix} P_{0 \rightarrow 0} & P_{0 \rightarrow 1} & P_{0 \rightarrow 2} \\ P_{1 \rightarrow 0} & P_{1 \rightarrow 1} & P_{1 \rightarrow 2} \\ P_{2 \rightarrow 0} & P_{2 \rightarrow 1} & P_{2 \rightarrow 2} \end{pmatrix}$$

Onde

$$P_{i \rightarrow j} = \binom{n}{j} \left(\frac{i}{n}\right)^j \left(1 - \frac{i}{n}\right)^{(n-j)}$$

Note que a razão  $i/n$  indica a frequência do alelo  $A$  na geração  $G_0$ . Além disso,  $n$  é o número total de alelos da população ( $2N$  se for diplóide),  $i$  o número de alelos  $A$  em  $G_0$  e  $j$  o número de alelos  $A$  em  $G_1$ .

Vamos obter essa mesma matriz  $Q$  usando o R:

```
N = 1
estados <- seq(0, (2*N), 1)

P = NULL
for(i in estados){
  P = c(P, dbinom(estados, size=2*N, prob=i/(2*N) ) )
}

Q = matrix(P, ncol=(2*N)+1, byrow=T)
```

Neste código, o vetor  $P$  foi recebendo os valores das probabilidades binomiais de mudanças entre os estados (ou seja, o número de alelos  $A$ ) entre as gerações. Depois, através do comando `matrix(P, ncol=(2*N)+1, byrow=T)` transformamos esse vetor na matriz  $Q$ :

```
> P
[1] 1.00 0.00 0.00 0.25 0.50 0.25 0.00 0.00 1.00
```

Virou:

```
> Q
      [,1] [,2] [,3]
[1,] 1.00 0.0 0.00
[2,] 0.25 0.5 0.25
[3,] 0.00 0.0 1.00
```

Para  $N = 1$ ,  $Q$  tem o mesmo formato indicado acima neste texto.

No nosso exemplo, começamos com apenas  $N = 1$  heterozigoto  $Aa$  em  $G_0$ . O estado inicial desta população pode ser representado pelo vetor  $x = (0,1,0)$ . Indicando que **existe 100% de probabilidade de encontrar apenas 1 alelo  $A$  nesta população**. Afinal, estamos observando o estado inicial de  $N = 1$  heterozigoto  $Aa$ . Então sabemos que existe 1 alelo  $A$ . Se, por acaso o estado



inicial fosse  $N = 1$  homocigoto  $AA$  em  $G_0$ , o vetor seria  $\mathbf{x} = (0,0,1)$ , indicando 100% de chance de ter 2 alelos  $A$ . Por fim, com  $N = 1$  homocigoto  $aa$  em  $G_0$  teríamos  $\mathbf{x} = (1,0,0)$ , indicando 100% de chance de ter 0 alelos  $A$  em  $G_0$ .

Suponhamos que deste estado inicial  $\mathbf{x} = (0,1,0)$  em  $G_0$ , queremos saber como a probabilidade de cada uma das composições alélicas (0, 1 e 2 alelos  $A$ ) evolui ao longo de  $g = 20$  gerações.

Vamos então criar uma matriz  $\mathbf{x}$  no R em que as linhas indicam as gerações (de 1 até 20) e as colunas indiquem as composições alélicas (0, 1 e 2 alelos  $A$ ). Os elementos da matriz indicarão a probabilidade de cada uma dessas composições em cada geração:

```
g = 20
x = matrix(nrow=g, ncol=(2*N)+1)

> x
      [,1] [,2] [,3]
[1,] NA  NA  NA
[2,] NA  NA  NA
[3,] NA  NA  NA
[4,] NA  NA  NA
[5,] NA  NA  NA
[6,] NA  NA  NA
[7,] NA  NA  NA
[8,] NA  NA  NA
[9,] NA  NA  NA
[10,] NA  NA  NA
[11,] NA  NA  NA
[12,] NA  NA  NA
[13,] NA  NA  NA
[14,] NA  NA  NA
[15,] NA  NA  NA
[16,] NA  NA  NA
[17,] NA  NA  NA
[18,] NA  NA  NA
[19,] NA  NA  NA
[20,] NA  NA  NA
```

Ao ser criada, a matriz está vazia (elementos NA). Vamos então preencher a primeira linha (geração  $G_0$ ) com os valores da composição alélica original, ou seja,  $(0,1,0)$  – 1 alelo  $A$ :

```
x[1,] = c(0,1,0)
```

Agora é hora de calcularmos qual serão as probabilidades de cada uma das composições alélicas em  $G_1$  (a segunda linha), a geração seguinte.

Isso é obtido **multiplicando o estado inicial pela matriz Q**:

```
x[2,] = x[1,] %*% Q
```

O símbolo “%\*%” indica a multiplicação de matrizes no R. Após esta operação, a matriz  $x$  fica assim:



```
> x
      [,1] [,2] [,3]
[1,] 0.00 1.0 0.00
[2,] 0.25 0.5 0.25
[3,] NA NA NA
[4,] NA NA NA
[5,] NA NA NA
[6,] NA NA NA
[7,] NA NA NA
[8,] NA NA NA
[9,] NA NA NA
[10,] NA NA NA
[11,] NA NA NA
[12,] NA NA NA
[13,] NA NA NA
[14,] NA NA NA
[15,] NA NA NA
[16,] NA NA NA
[17,] NA NA NA
[18,] NA NA NA
[19,] NA NA NA
[20,] NA NA NA
```

Indicando que, em G1 (segunda linha), existe 25% de chance da composição 0 (ou seja, alelo A perdido), 50% de chance de ter 1 alelo A e 25% de chance de ter o alelo A fixado (2 alelos A).

Vamos agora construir um *loop* para calcular as probabilidades das gerações seguintes (linhas 3 até 20). A estratégia será sempre **multiplicar o valor da geração anterior pela matriz de transição Q**:

```
for(i in c(3:g)){
  x[i,] = x[i-1,] %*% Q
}
```

Ao final do loop, a matriz x ficará assim:

```
> x
      [,1]      [,2]      [,3]
[1,] 0.0000000 1.000000e+00 0.0000000
[2,] 0.2500000 5.000000e-01 0.2500000
[3,] 0.3750000 2.500000e-01 0.3750000
[4,] 0.4375000 1.250000e-01 0.4375000
[5,] 0.4687500 6.250000e-02 0.4687500
[6,] 0.4843750 3.125000e-02 0.4843750
[7,] 0.4921875 1.562500e-02 0.4921875
[8,] 0.4960938 7.812500e-03 0.4960938
[9,] 0.4980469 3.906250e-03 0.4980469
[10,] 0.4990234 1.953125e-03 0.4990234
[11,] 0.4995117 9.765625e-04 0.4995117
[12,] 0.4997559 4.882812e-04 0.4997559
[13,] 0.4998779 2.441406e-04 0.4998779
[14,] 0.4999390 1.220703e-04 0.4999390
[15,] 0.4999695 6.103516e-05 0.4999695
[16,] 0.4999847 3.051758e-05 0.4999847
[17,] 0.4999924 1.525879e-05 0.4999924
[18,] 0.4999962 7.629395e-06 0.4999962
[19,] 0.4999981 3.814697e-06 0.4999981
[20,] 0.4999990 1.907349e-06 0.4999990
```

Podemos observar que na vigésima linha (geração 20) os estados 0 alelos A (a primeira coluna) e 2 alelos A (a terceira coluna) possuem probabilidades próximas de 50%, enquanto o estado de 1 alelo A (segunda coluna) está com probabilidade próxima ao nulo.



O que isso significa? A interpretação é esta:

“A iniciar com um dado estado, usando as multiplicações sucessivas pela matriz de probabilidade de transição  $Q$ , chegamos à **distribuição estacionária dos estados.**”

No exemplo dado, observa-se que os estados de 0 alelos  $A$  (ou seja, 100% de homocigotos  $aa$ ) e 2 alelos  $A$  (100% de homocigotos  $AA$ ) são as condições estacionárias. E cada uma delas ocorre com chance de 50%. Mais importante: a chance de encontrar heterocigotos  $Aa$  decai com o tempo até alcançar o valor nulo.

A estratégia que fizemos acima chama-se de cadeia de Markov, pois foi estabelecida pelo matemático [Andrey Markov](#) no início do século 20.

Conforme observamos acima, o processo de Markov funciona basicamente em cima da matriz de probabilidade de transição. A cada etapa do tempo apenas o estado imediatamente anterior é relevante para o cálculo do estado imediatamente posterior. Assim, dizemos que o processo de Markov **não tem memória**.

Vamos agora estudar um caso mais interessante. Uma população com  $N = 10$  indivíduos:

```
N = 10
estados <- seq(0, (2*N), 1)
```

O número de possibilidades de ocorrência do alelo  $A$  agora é muito maior. Das  $2N = 20$  cópias alélicas, temos as seguintes possibilidades:

```
> estados
[1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
```

Variando desde 0 (nenhum  $A$ ) até 20 (100%  $AA$ ). O valor de 10 indica obviamente que  $f_A = 0.5$ .

Novamente, vamos estabelecer nossa matriz de probabilidade de transição  $Q$  usando a distribuição binomial:

```
P = NULL
for(i in estados){
  P = c(P, dbinom(estados, size=2*N, prob=i/(2*N) ) )
}

Q = matrix(P, ncol=(2*N)+1, byrow=T)
```

A matriz  $Q$ , agora é bem maior! Ela tem dimensão de 21x21. Verifique isso. Lembre-se que as possibilidades de composição do alelo  $A$  vão de 0 até 20 (21 possibilidades no total):

```
> dim(Q)
[1] 21 21
```

Agora é hora de construir a matriz  $x$  que guardará as probabilidades dos estados ao longo das gerações. Vamos usar  $g = 100$  gerações:

```
g = 100
```



```
x = matrix(nrow=g, ncol=(2*N)+1)
```

A matriz x agora tem 100 linhas (gerações) e 21 colunas (possibilidades de composição do alelo A):

```
> dim(x)
[1] 100 21
```

Vamos agora **escolher** um estado inicial qualquer para  $G_0$ . Usaremos  $f_A = 0.5$ . Desta forma, a primeira linha da matriz x deverá ser

```
x[1,] = c(0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)
```

```
> x
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16]
[1,]    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0
[2,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[3,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[4,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[5,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[6,] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

Isso indica que na 11ª coluna, que corresponde ao estado de 10 alelos A num total de 20, adicionamos a probabilidade de 100%, afinal **estamos observando este valor**, ele é sabido.

Novamente, vamos multiplicar essa primeira linha pela matriz Q para obter as probabilidades de cada uma das composições alélicas de A na geração seguinte:

```
x[2,] = x[1,] %*% Q
```

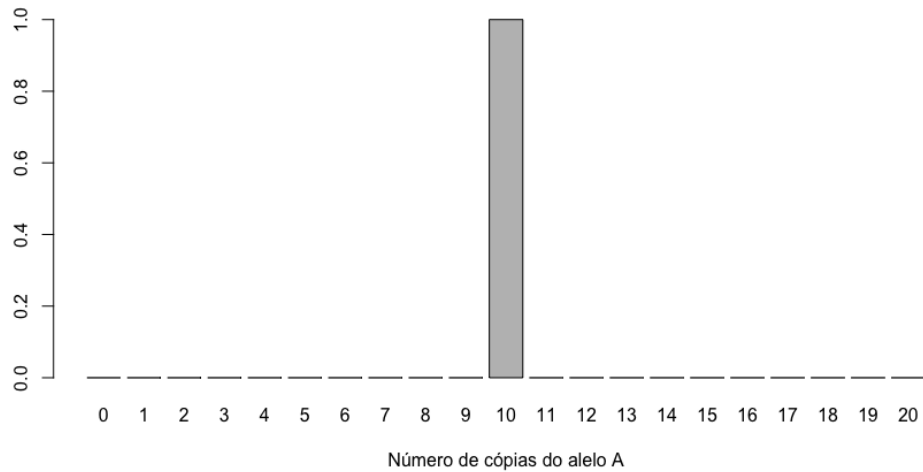
Finalmente, os outros valores serão preenchidos rodando o *loop*:

```
for(i in c(3:g)){
  x[i,] = x[i-1,] %*% Q
}
```

Como essa matriz é bem grande, é mais fácil observar as probabilidades dos estados em cada uma das gerações plotando gráficos. Por exemplo, em  $G_0$  (o início), temos:

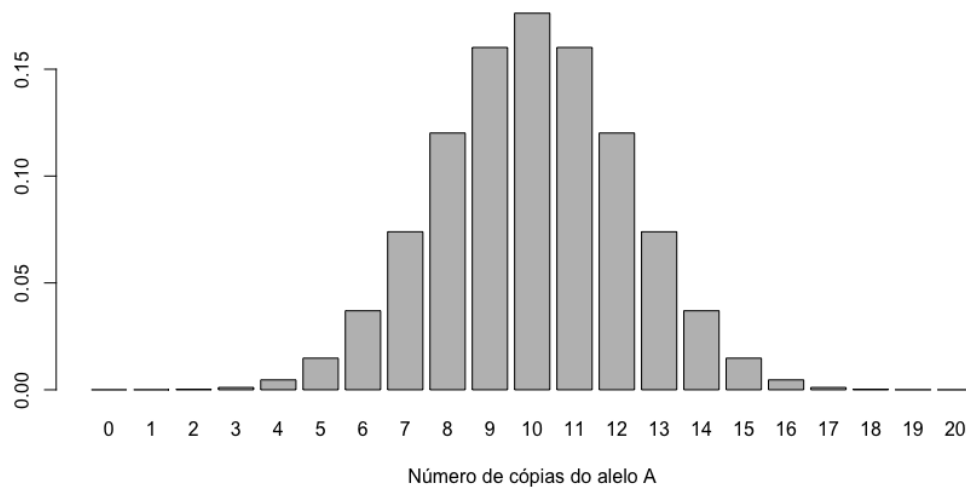
```
barplot(x[1,], names =c(0:20), xlab="Número de cópias do alelo A" )
```





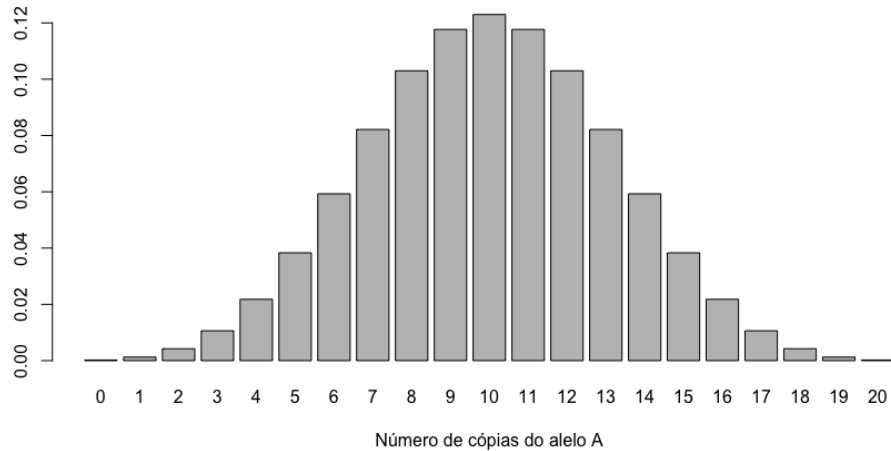
Isso indica que a probabilidade de 10 cópias alélicas é 100%.  
Em  $G_1$  temos:

```
barplot(x[2,], names =c(0:20), xlab="Número de cópias do alelo A" )
```



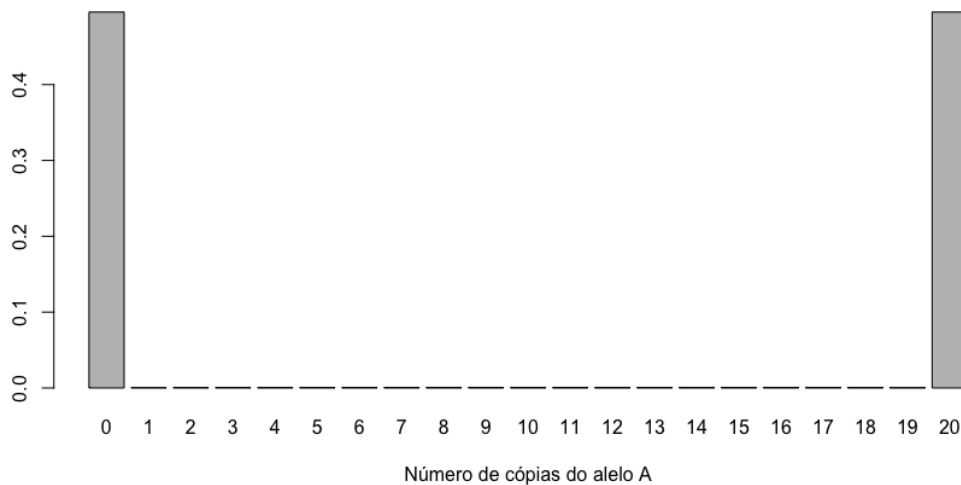
Na próxima geração (terceira linha), temos:

```
barplot(x[3,], names =c(0:20), xlab="Número de cópias do alelo A" )
```



Vamos agora para a última geração calculada (a 100ª linha):

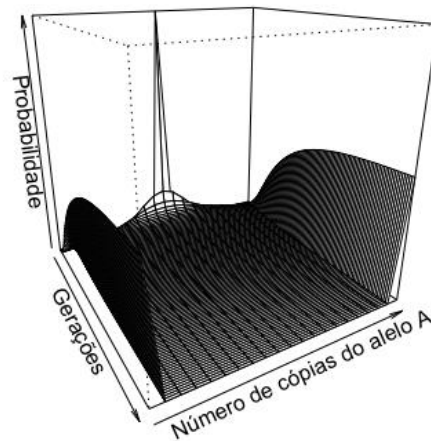
```
barplot(x[100,], names = c(0:20), xlab="Número de cópias do alelo A" )
```



Notamos que os estados 0 e 20 estão com probabilidade de 50% cada. Indicando que, após 100 gerações, começando com  $f_A = 0.5$ , temos 50% de ter fixação do alelo A (estado 20) e 50% de chance de fixar o a (estado 0).

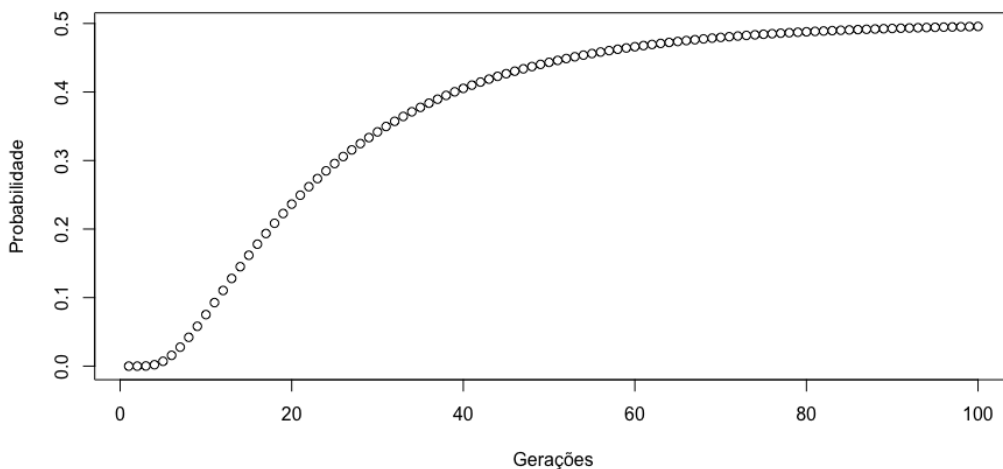
Enfim, podemos visualizar todo o processo de evolução da probabilidade dos 21 estados ao longo das 100 gerações num único gráfico:

```
persp(x, theta = 60, phi = 20 ,xlab = "Gerações", ylab = "Número de cópias do alelo A",zlab = "Probabilidade")
```



Podemos observar que os estados estacionários (0 e 20) rapidamente atingem a probabilidade de 50%. Vamos ver isso com mais detalhes:

```
plot(x[,21], xlab="Gerações", ylab="Probabilidade")
```



Por volta da 60ª geração, a fase estacionária já é atingida. O mesmo ocorre para o estado em que o número de A é 0. Verifique: `plot(x[,1], xlab="Gerações", ylab="Probabilidade")`.

Agora é hora de brincar com os valores de N e o estado inicial:

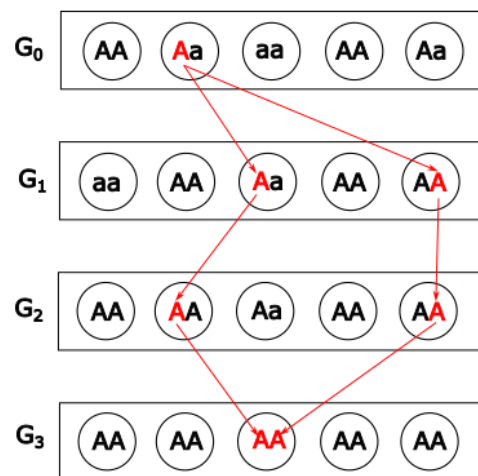
- O que acontece quando a população aumenta para N = 100? O tempo em que os estados estacionários atingem seu valor limite é o mesmo?
- O que acontece se utilizarmos como estado inicial apenas 1 alelo A? Quais são os estacionários e quais suas respectivas probabilidades?

## Quantificação do decaimento da heteroziguidade por deriva

Nesta prática iremos estudar como quantificar a queda da frequência de heterozigotos (heteroziguidade) pela ação da deriva genética.

Um conceito central da formulação utilizada para essa quantificação é a **identidade por descendência**. Numa população, os genótipos homocigotos possuem, por definição, duas cópias alélicas idênticas em relação às sequências de nucleotídeos num determinado locus de um cromossomo. Portanto, dizemos que homocigotos possuem alelos **idênticos por estado**. Aqui, definiremos um tipo mais específico de homociguidade: aquela em que os alelos são **idênticos por descendência** (além de serem idênticos por estado) ou simplesmente **IBD** (*identity by descent*). A frequência dos homocigotos IBD numa população é chamada de **F**.

O conceito de IBD pode ser compreendido com mais facilidade através de exemplos. Na figura abaixo, temos a evolução de uma população através do processo Wright-Fisher. Ou seja, a cada geração existe a produção e amostragem do *pool* gamético para formar os genótipos da geração seguinte.



No exemplo mostrado, na geração G<sub>3</sub>, todos os  $N = 5$  indivíduos são homocigotos AA. Entretanto, existe um indivíduo, marcado em vermelho, que além de ser um homocigoto AA, tem suas duas cópias alélicas oriundas da mesma molécula de DNA na geração G<sub>0</sub>.

Como isso é possível? Em G<sub>0</sub>, essa cópia específica do alelo A marcada de vermelho irá se **replicar por meiose** para a formação dos gametas deste indivíduo. Desta forma, **as cópias de A presentes nos gametas deste indivíduo também receberão a marcação vermelha**. Existem vários gametas com a cópia alélica A vermelha produzidos por este indivíduo no *pool* gamético. Em G<sub>1</sub>, esses gametas foram amostrados (aleatoriamente) duas vezes. E agora estão presentes em dois indivíduos.

Este processo continua ao longo das gerações até que, ao acaso (por simples amostragens sucessivas do *pool* gamético), na geração G<sub>3</sub>, um indivíduo (marcado pelo AA vermelho) recebeu **duas cópias alélicas cuja ancestralidade pode ser traçada três gerações atrás numa única molécula de DNA (cópia alélica) original comum**. Dizemos que este indivíduo possui alelos A idênticos por descendência (além, obviamente, de serem idênticos por estado).

Além disso, o evento de encontro da ancestralidade das cópias alélicas de  $G_3$  na cópia ancestral em  $G_0$  é denominado de **coalescência**.

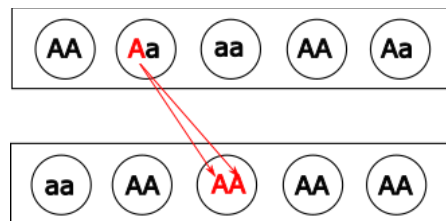
A quantificação do decaimento da heteroziguidade por deriva é realizado através da medida oposta – o aumento da homoziguidade.

Desta forma, para quantificar o aumento da homoziguidade, a primeira etapa é medir a frequência (= probabilidade) dos indivíduos que são homozigotos IBD. Ou seja, iguais ao indivíduo  $AA$  com as duas cópias vermelhas do esquema anterior.

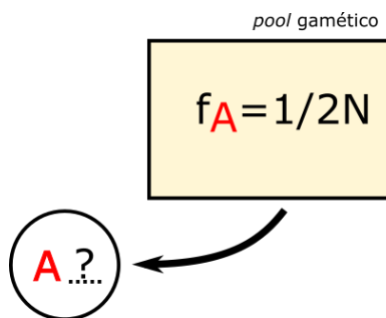
Assim, qual é a probabilidade de formar um indivíduo homozigoto IBD. Já mencionamos que esta probabilidade é chamada de  $F$ .

A **primeira parte** da quantificação diz respeito à probabilidade de duas cópias alélicas encontrarem sua cópia ancestral na **geração imediatamente anterior**. Essa probabilidade é dada por  $1/2N$ .

Como entender esse valor? Pensemos assim: encontrar seu ancestral na geração imediatamente anterior significa ocorrer algo assim:



Acabamos de afirmar que a chance de isto acontecer é  $1/2N$ . Isso acontece pois, quando amostramos o primeiro dos alelos  $A$  vermelhos no *pool* gamético, temos  $1/2N$  de amostrar um segundo alelo também vermelho do mesmo *pool*:



Isso ocorre pois a os alelos vermelhos  $A$  no *pool* estão com frequência igual a 1 sobre o total de alelos não-vermelhos (observe a composição da geração parental). Como existem um total de  $2N$  alelos, temos que a frequência dos gametas que carregam cópias do alelo  $A$  vermelho é exatamente  $1/2N$ .

Você possivelmente deve estar achando que este cálculo está incorreto. Afinal, se as cópias do alelo A vermelho foram amostradas duas vezes, **por que não consideramos os dois processos: a primeira amostragem E a segunda amostragem?** Resultando numa probabilidade de  $(1/2N)^2$ ?

De fato, essa observação está correta. Se efetivamente estivéssemos pensando na probabilidade de formar um homozigoto IBD **apenas para o alelo A vermelho**, deveríamos afirmar que  $F = (1/2N)^2$ .

Entretanto, estamos interessados na probabilidade (= frequência de ocorrência) de **qualquer homozigoto IBD**. É exatamente isso que mede  $F$ . Neste caso, **a primeira amostragem de alelos no pool gamético NÃO irá contar**. O fato de ser IBD será uma propriedade aplicada à segunda amostragem do *pool*. Na segunda amostragem, queremos pegar um gameta que carrega uma cópia alélica que veio exatamente da mesma molécula mãe na geração anterior.

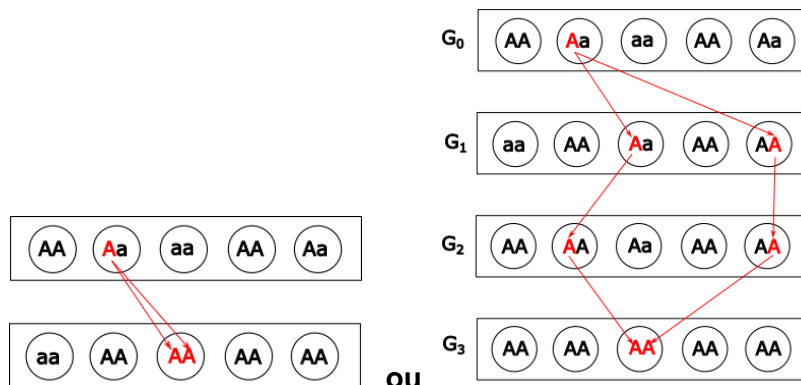
Assim, ao afirmar que  $F = 1/2N$ , estamos considerando **todas as possibilidades** de homozigotos IBD: sejam eles  $AA$  ou  $aa$  (vermelhos ou não). Afinal, *poderíamos marcar com a cor vermelha qualquer uma das cópias alélicas de  $G_0$* .

Assim, a primeira parte da nossa formulação é exatamente essa:  $F = 1/2N$ .

Note, entretanto, que um homozigoto IBD pode ser formado várias gerações no futuro. Esse é justamente o caso do esquema mostrado no início desta prática.

Devemos então contabilizar esse processo na nossa formulação de  $F$ . Essa é a **segunda parte** da formulação.

Ou seja, um homozigoto IBD pode ser formado de duas formas: (1) na geração imediatamente anterior ou (2) várias gerações atrás --



Vamos então para a segunda parte da formulação. **SE** não houve encontro da molécula ancestral na geração imediatamente anterior (a coalescência), isso ocorrerá com a seguinte probabilidade:  $(1 - 1/2N)$ . Afinal, se  $1/2N$  é a chance de encontrar o ancestral em 1 geração,  $(1 - 1/2N)$  é a chance de não encontrar.

Muito bem. Se não encontra o ancestral na geração imediatamente anterior, quando irá encontrar?

No nosso exemplo, esse encontro (coalescência) ocorreu três gerações atrás. Entretanto, poderia ocorrer **em qualquer momento do passado**. Como podemos incluir essa medida na nossa formulação de  $F$ , se não sabemos exatamente **em qual geração** o encontro com a cópia ancestral irá ocorrer?

A saída para esse problema é usar uma **formulação recursiva**. Essa recursividade é dada pela multiplicação do valor de  $F$  na geração imediatamente anterior. Vamos a formulação final de  $F$ , a probabilidade (= frequência de ocorrência) de homozigotos IBD:

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}$$

Vamos ler essa fórmula usando linguagem não matemática:

“A probabilidade de encontrar um homozigoto IBD numa população na geração  $t$  ( $F_t$ ) é composta de duas formas: (1) ele é IBD pois suas cópias alélicas encontram sua molécula ancestral na geração imediatamente anterior ( $\frac{1}{2N}$ ) **ou** (+) (2) elas não encontram a molécula ancestral na geração imediatamente anterior ( $1 - \frac{1}{2N}$ ), mas encontram em alguma geração do passado ( $F_{t-1}$ ).”

A recursividade ocorre justamente pelo uso de  $F_{t-1}$  para calcular  $F_t$ .

Portanto, já que calculamos  $F$ , a probabilidade dos homozigotos IBD, falta calcular a probabilidade dos homozigotos não-IBD. Afinal, queremos em última análise estimar a heterozigosidade (o oposto da homozigosidade).

Para calcular a probabilidade dos homozigotos não-IBD (e adicioná-la a  $F$ ), tomaremos uma decisão surpreendente:

### **Vamos assumir que todos os homozigotos são IBD!**

Possivelmente, você achará que fazer isso é uma saída não muito honesta. Afinal, existem homozigotos que possuem cópias alélicas idênticas por estado mas não por descendência.

É verdade. Portanto, para que façamos essa equivalência, devemos admitir que:

### **Toda mutação gera um novo alelo.**

Em outras palavras, um alelo ( $A$  ou  $a$  ou qualquer outro) **não pode aparecer mais de uma vez**. A mutação que originou  $A$  (ou qualquer outro alelo) só ocorreu, portanto, uma única vez. Desta forma, qualquer  $AA$  observado numa população **necessariamente possui cópias alélicas que encontrarão um ancestral comum em alguma geração do passado** – o primeiro  $A$  que surgiu!

Ou seja, todo homozigoto é, em última análise, um homozigoto IBD.

Essa estratégia de modelagem da probabilidade de encontrar um homozigoto é chamada de **modelo dos alelos infinitos**. Afinal, uma mutação nunca criará dois alelos iguais.



A partir desta decisão, o cálculo da **homozigidade numa dada geração**, ou seja, a frequência de ocorrência ou probabilidade de se encontrar um homozigoto, ficou muito mais fácil, pois:

$$\text{Homozigidade}_t = F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}$$

Por fim, como a heterozigidade é o oposto da homozigidade, temos:

$$\text{Heterozigidade}_t = 1 - \text{Homozigidade}_t$$

Ou seja,

$$H_t = 1 - F_t$$

Portanto,

$$H_t = 1 - \left[ \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1} \right]$$

Ou mesmo

$$H_t = 1 - \left[ \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) (1 - H_{t-1}) \right]$$

Ao simplificar a fórmula acima (use o [WolframAlpha.com](http://WolframAlpha.com)), veremos que:

$$H_t = H_{t-1} - \frac{1}{2N} H_{t-1}$$

Essa fórmula acima pode ser lida da seguinte forma:

“A frequência de heterozigotos numa geração  $t$  ( $H_t$ ) é igual à frequência de heterozigotos da geração imediatamente anterior ( $H_{t-1}$ ) subtraída (-) de um valor igual a  $\frac{1}{2N} H_{t-1}$ .”

Esta equação deixa evidente que, a cada geração, por ação da deriva genética, a frequência de heterozigotos é **reduzida** por um valor igual a  $\frac{1}{2N}$  da heterozigidade da geração anterior.

Podemos afirmar, portanto, que **a taxa de decaimento da heterozigidade por ação da deriva é igual a  $\frac{1}{2N}$  por geração.**



## Decaimento da heterozigiosidade por deriva genética & equilíbrio mutação-deriva

Nesta prática, iremos utilizar a quantificação derivada na prática anterior para verificar o impacto do tamanho da população ( $N$ ) na frequência de heterozigotos, assim como a influência da mutação neste processo.

Inicialmente voltemos às duas equações fundamentais que calculamos sobre a probabilidade (= frequência de ocorrência) de homozigotos ( $F$ ) e de heterozigotos ( $H$ ) numa dada geração:

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}$$

e

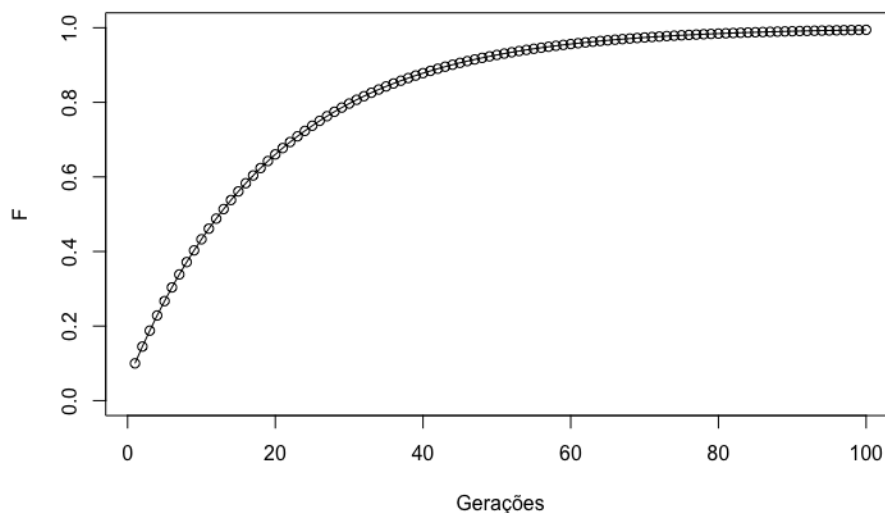
$$H_t = H_{t-1} - \frac{1}{2N} H_{t-1}$$

Podemos usar o R para plotar o gráfico para visualizar o decaimento de  $H$  e o aumento de  $F$ .

Vamos iniciar com uma frequência de homozigotos  $F = 10\%$  e verificar o seu amento em 100 gerações:

```
N = 10
F = 0.1
for(i in c(2:100)){
  F = c(F, (1/(2*N)) + (1-(1/(2*N)))*F[length(F)] )
}

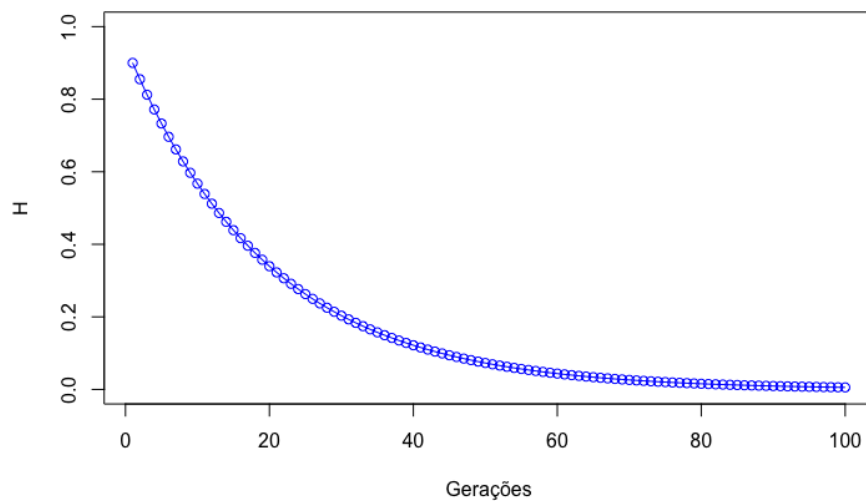
plot(F, xlab="Gerações", ylab="F", ylim=c(0,1))
lines(F)
```



É evidente que F aumentará até chegar na frequência de 100% de homozigotos. Afinal, **essa é consequência a longo prazo da deriva genética nas populações**. Vamos agora plotar o gráfico oposto: o decaimento da heterozigosidade:

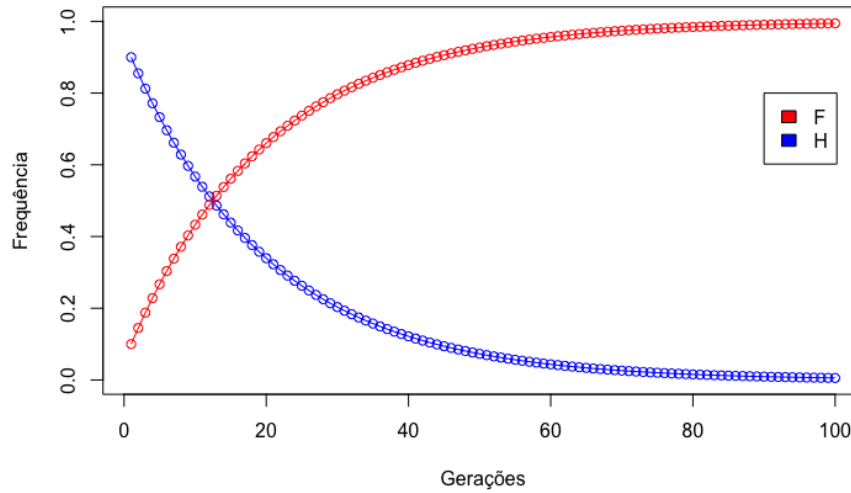
```
N = 10
F = 0.1
H = 1-F
for(i in c(2:100)){
  H = c(H, H[length(H)] - (1/(2*N))*H[length(H)] )
}

plot(H, col="blue")
lines(H, col="blue")
```



Podemos plotar ambas as linhas no mesmo gráfico:

```
plot(H, col="blue", xlab="Gerações", ylab=c("Frequência"), ylim=c(0,1))
lines(H, col="blue")
points(F, col="red")
lines(F, col="red")
legend(90,0.8,legend=c("F", "H"), fill=c("red", "blue"))
```



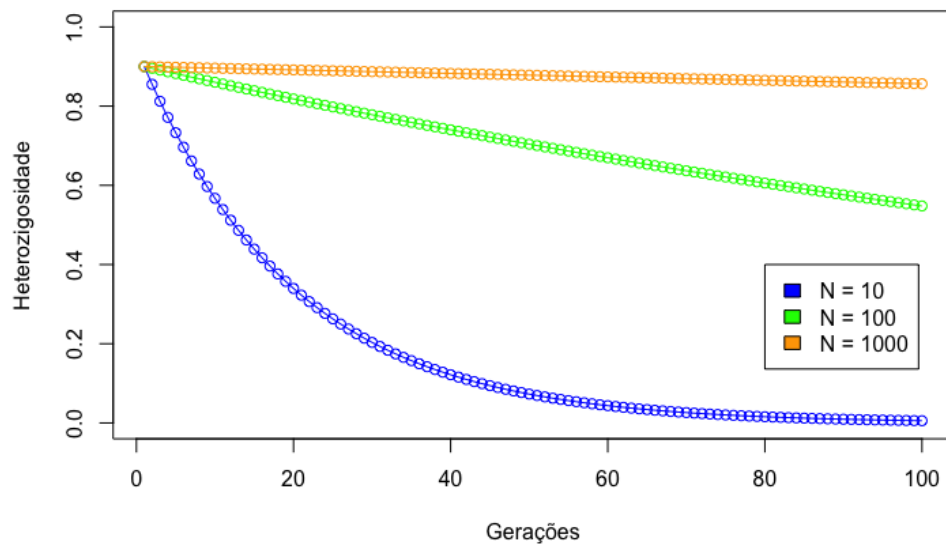
É notável pelo gráfico que  $F$  e  $H$  têm comportamentos opostos. Podemos também verificar o impacto do tamanho da população no decaimento da heterozigosidade. Vamos comparar populações de  $N = 10$ ,  $100$  e  $1000$  indivíduos. Vamos admitir que, na primeira geração  $G_0$ , temos  $H = 0.9$  em todas. Acompanharemos  $H$  por 100 gerações.

```
N = c(10, 100, 1000)
g = 100
H = matrix(nrow=3, ncol=g)

H[1,1] = 0.9
H[2,1] = 0.9
H[3,1] = 0.9

for(i in c(2:g)){
  H[1,i] = H[1,i-1] - (1/(2* N[1])) * H[1,i-1]
  H[2,i] = H[2,i-1] - (1/(2* N[2])) * H[2,i-1]
  H[3,i] = H[3,i-1] - (1/(2* N[3])) * H[3,i-1]
}

plot(H[1,], col="blue", xlab="Gerações", ylab=c("Heterozigosidade"), ylim=c(0,1))
lines(H[1,], col="blue")
points(H[2,], col="green")
lines(H[2,], col="green")
points(H[3,], col="orange")
lines(H[3,], col="orange")
legend(80,0.4,legend=c("N = 10", "N = 100", "N = 1000"), fill=c("blue", "green", "orange"))
```



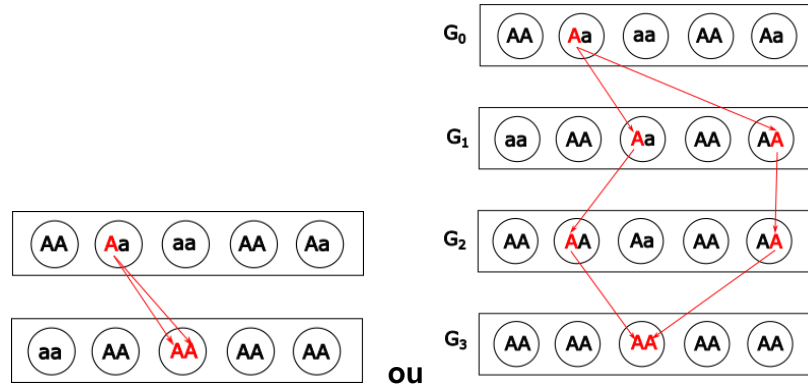
Pelo gráfico, observamos claramente o impacto do tamanho da população na queda da diversidade genética (heteroziguidade). Lembremos que isso já era esperado, pois a taxa de decaimento é  $1/2N$  – **quanto maior o  $N$ , menor será a taxa de decaimento.**

Quando a população chega ao ponto em que a heteroziguidade é nula ( $H = 0$ ), não há mais retorno. Entretanto, se novas mutações surgirem, evidentemente frequência de heterozigotos sai do valor nulo. Basta pensar assim: se temos 100% de indivíduos AA na população ( $H = 0$ ), a mutação ocorrerá num dos alelos A, aumento a chance de formação de genótipos heterozigotos.

Assim como fizemos anteriormente para derivar as equações de  $F$  e  $H$  ao longo das gerações, a inserção da **taxa de mutação** será estudada usando a formulação da homoziguidade ( $F$ ). A fórmula de  $F$  numa dada geração possui dois componentes:

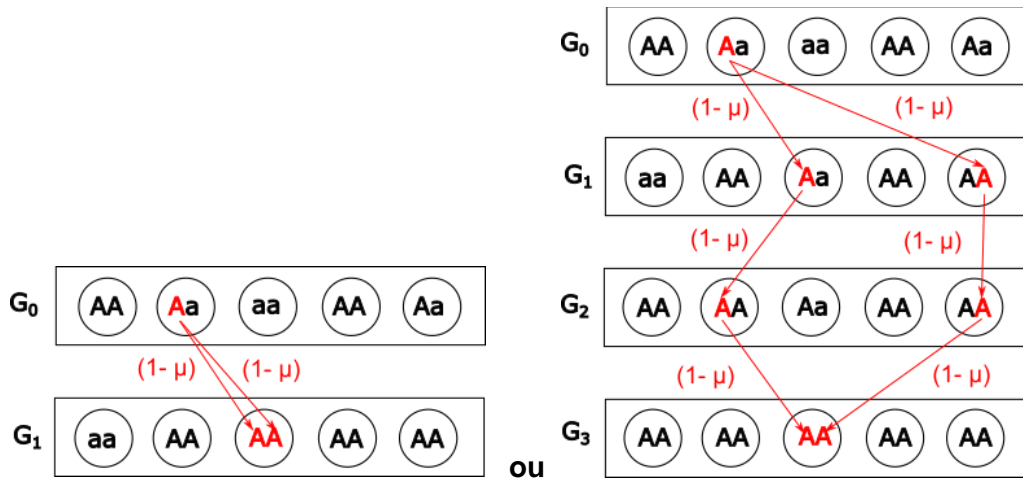
$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}$$

Correspondentes aos processos:



Para a manutenção da homoziguidade **não pode haver mutação em nenhuma das linhas de ancestralidade em vermelho dos esquemas acima**. Se considerarmos que, a cada geração, existe uma chance  $\mu$  da cópia alélica mutar, a **chance dela não sofrer mutação será  $(1 - \mu)$** .

Desta forma, para que a homoziguidade ocorra, não poderá ocorrer mutação em nenhuma das linhas de ancestralidade mostradas acima. O processo deve ocorrer assim:



O primeiro processo, diz que para encontrar a cópia ancestral na geração imediatamente anterior, **nenhuma das duas cópias alélicas pode sofrer mutação**. Portanto, a chance de ocorrer um homozigoto IBD dessa primeira forma será:

$$\frac{1}{2N}(1 - \mu)(1 - \mu)$$

Multiplicamos  $(1 - \mu)$  pois a chance é  $1/2N$  e não poderá mutar nem na linhagem paterna e na materna.

Algo semelhante ocorre na segunda parte da fórmula.

$$\left(1 - \frac{1}{2N}\right)(1 - \mu)(1 - \mu)$$

Pois não encontrará a molécula ancestral na geração imediatamente anterior,  $\left(1 - \frac{1}{2N}\right)$ , e não poderá mutar nem na linhagem paterna e na materna.



Desta forma, ao inserir a taxa de mutação  $\mu$  por alelo a cada geração, a formulação da homozigidade fica:

$$F_t = \frac{1}{2N}(1 - \mu)^2 + \left(1 - \frac{1}{2N}\right)(1 - \mu)^2 F_{t-1}$$

O termo recursivo  $F_{t-1}$  não precisa ser multiplicado por  $(1 - \mu)^2$  **pois isso já foi feito na geração anterior ( $t - 1$ ).**

Para encontrar o valor da heterozigidade  $H$  com a presença de mutação  $\mu$ , usamos a mesma estratégia anterior:

$$H_t = 1 - F_t$$

Portanto,

$$H_t = 1 - \left[ \frac{1}{2N}(1 - \mu)^2 + \left(1 - \frac{1}{2N}\right)(1 - \mu)^2 F_{t-1} \right]$$

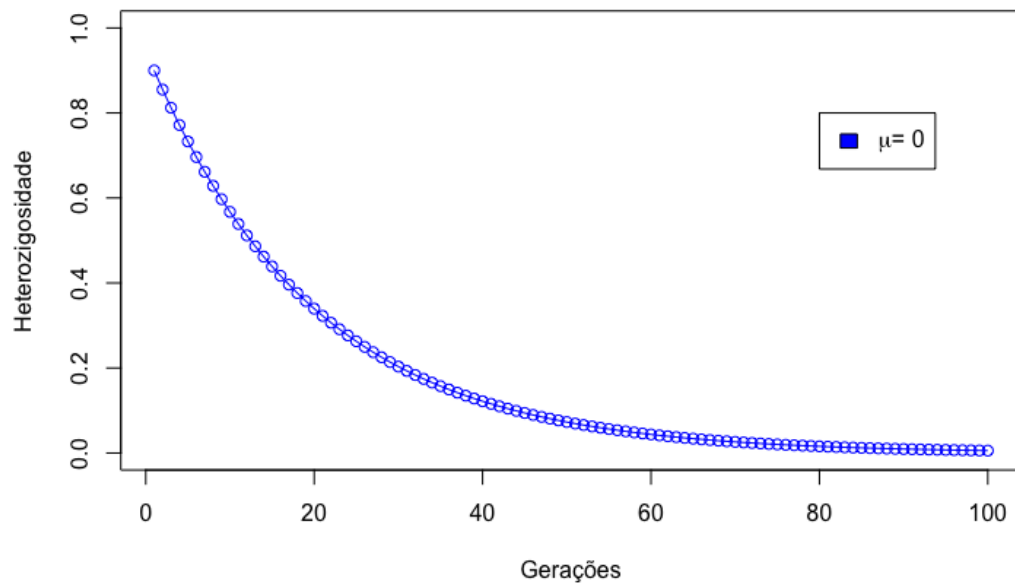
Que é igual a

$$H_t = 1 - \left[ \frac{1}{2N}(1 - \mu)^2 + \left(1 - \frac{1}{2N}\right)(1 - \mu)^2(1 - H_{t-1}) \right]$$

Podemos agora plotar o gráfico do decaimento da heterozigidade com a presença de mutação. Usemos  $N = 10$ ,  $H$  inicial de 90% e em princípio, taxa de mutação  $\mu = 0$ . Ao longo de 100 gerações temos:

```
N = 10
m = 0
H = 0.9
g = 100
for(i in c(2:g)){
  H = c(H, 1 - ( (1/(2*N)) * ((1-m)^2) + (1-(1/(2*N)))*((1-m)^2)*(1 - H[length(H)]) ) )
}

plot(H, col="blue", xlab="Gerações", ylab=c("Heterozigidade"), ylim=c(0,1))
lines(H, col="blue")
legend(80,0.8, legend=c(expression(paste(mu,"= 0"))) ), fill="blue" )
```



Evidentemente, temos um gráfico idêntico ao mostrado no início do estudo (quando a mutação não era considerada).

Ainda com  $N = 10$ , vamos testar a influência das taxas de mutação 0.001, 0.01 e 0.1 (além de 0) na heterozigosidade. Agora,  $H$  será uma matriz com 4 linhas (para cada taxa de mutação) e 100 colunas (gerações):

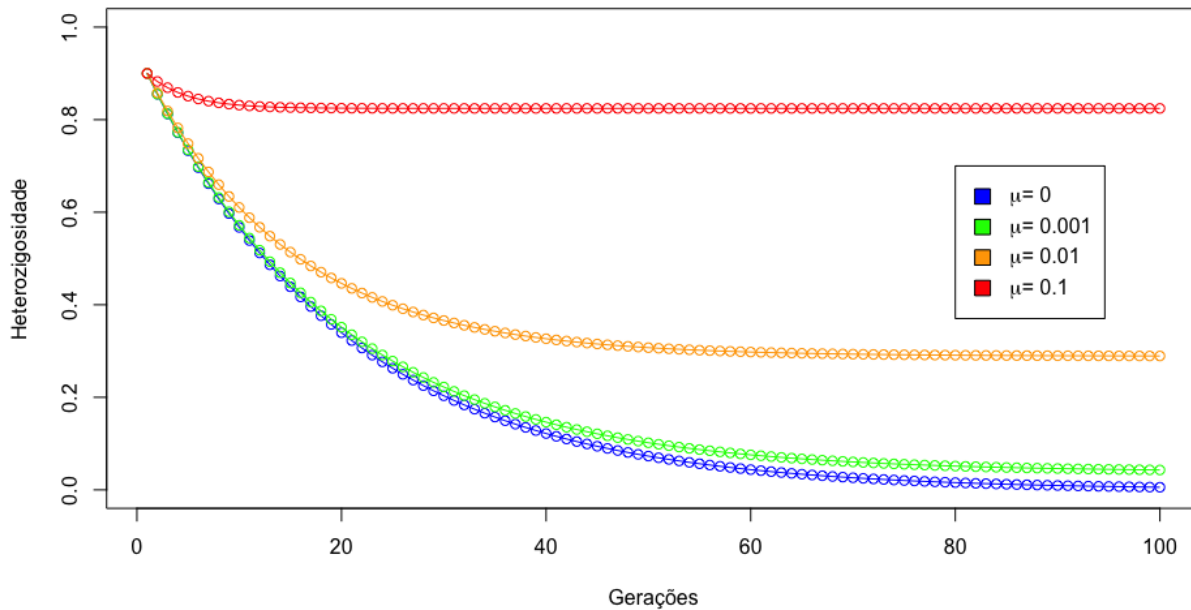
```
N = 10
m = c(0, 0.001, 0.01, 0.1)
g = 100
H = matrix(nrow=4, ncol=g)

H[1,1] = 0.9
H[2,1] = 0.9
H[3,1] = 0.9
H[4,1] = 0.9

for(i in c(2:100)){
  H[1,i] = 1 - ( (1/(2*N)) * ((1-m[1])^2) + (1-(1/(2*N))) * ((1-m[1])^2) * (1 - H[1,i-1]) )
  H[2,i] = 1 - ( (1/(2*N)) * ((1-m[2])^2) + (1-(1/(2*N))) * ((1-m[2])^2) * (1 - H[2,i-1]) )
  H[3,i] = 1 - ( (1/(2*N)) * ((1-m[3])^2) + (1-(1/(2*N))) * ((1-m[3])^2) * (1 - H[3,i-1]) )
  H[4,i] = 1 - ( (1/(2*N)) * ((1-m[4])^2) + (1-(1/(2*N))) * ((1-m[4])^2) * (1 - H[4,i-1]) )
}

plot(H[1,], col="blue", xlab="Gerações", ylab=c("Heterozigosidade"), ylim=c(0,1))
lines(H[1,], col="blue")
points(H[2,], col="green")
lines(H[2,], col="green")
points(H[3,], col="orange")
lines(H[3,], col="orange")
points(H[4,], col="red")
lines(H[4,], col="red")

legend(80,0.7,legend=c(expression(paste(mu,"= 0")), expression(paste(mu,"= 0.001")),
expression(paste(mu,"= 0.01")), expression(paste(mu,"= 0.1")) ), fill=c("blue", "green",
"orange", "red"))
```



Verificamos que, conforme a taxa de mutação aumenta, o valor da frequência de heterozigotos se afasta cada vez mais do valor nulo.

Também é notável que a partir de determinada geração existe uma **estabilidade do valor da frequência de heterozigotos**. Este ponto é chamado de ponto de **equilíbrio mutação-deriva**.

No ponto de equilíbrio entre mutação e deriva (os dois únicos fenômenos inseridos nas equações acima), a frequência de heterozigotos não muda, pois, **a taxa de eliminação de alelos por deriva é equivalente à taxa de inserção de novos alelos por mutação**.

Como podemos encontrar o valor da heterozigosidade referente ao ponto de equilíbrio entre mutação e deriva?

Por definição, no ponto de equilíbrio, a heterozigosidade não varia entre as gerações. Portanto:

$$H_t = H_{t-1}$$

Desta forma, no ponto de equilíbrio, a fórmula

$$H_t = 1 - \left[ \frac{1}{2N} (1 - \mu)^2 + \left( 1 - \frac{1}{2N} \right) (1 - \mu)^2 (1 - H_{t-1}) \right]$$

Será igual a

$$H_t = 1 - \left[ \frac{1}{2N} (1 - \mu)^2 + \left( 1 - \frac{1}{2N} \right) (1 - \mu)^2 (1 - H_t) \right]$$

Podemos agora resolver a equação para  $H_t$ :

$$\left\{ 1 - \left[ \frac{1}{2N} (1 - \mu)^2 + \left( 1 - \frac{1}{2N} \right) (1 - \mu)^2 (1 - H_t) \right] \right\} - H_t = 0$$



Você pode resolver essa equação no [WolframAlpha](https://www.wolframalpha.com):

The screenshot shows the WolframAlpha interface. The search bar contains the equation:  $Solve[1 - ((1/(2*N)) * ((1 - m)^2)) + ((1 - (1/(2*N)))) * ((1 - m)^2 * (1 - H)) - H == 0, H]$ . Below the search bar, there are options for 'NATURAL LANGUAGE' and 'MATH INPUT'. The 'Input interpretation' section shows the equation in a more readable form:  $solve \left( 1 - \left( \frac{1}{2N} (1 - m)^2 + \left( 1 - \frac{1}{2N} \right) (1 - m)^2 (1 - H) \right) - H = 0 \right)$  for  $H$ . The 'Result' section shows the solution:  $H = \frac{2(m - 2)mN}{m^2(2N - 1) + m(2 - 4N) - 1}$  and  $(m - 2)m(2N - 1) \neq 1$  and  $N \neq 0$ . There is also a 'Step-by-step solution' button and a 'Download Page' link.

Vamos tentar simplificar um pouco mais:

$$H = \frac{2(\mu - 2)\mu N}{\mu^2(2N - 1) + \mu(2 - 4N) - 1} = \frac{2\mu - 4N\mu}{2\mu^2N - \mu^2 - 4N\mu + 2\mu - 1}$$

Como a taxa  $\mu$  de mutação por alelo a cada geração é geralmente muito próxima a zero, é comum **ignorar os termos que contém  $\mu^2$** , ou seja,  $(2\mu^2N)$  e  $\mu^2$ . Ao fazer isso, evidentemente, teremos uma **aproximação do valor do ponto de equilíbrio mutação-deriva**:

$$H = \frac{2\mu - 4N\mu}{-4N\mu + 2\mu - 1}$$

Para simplificar ainda mais a fórmula, vamos agora **ignorar os termos  $(2\mu)$** , novamente admitindo que eles são muito pequenos **quando comparados aos termos que contém  $N$** , o tamanho da população:

$$H = \frac{-4N\mu}{-4N\mu - 1} = \frac{-1(4N\mu)}{-1(4N\mu + 1)} = \frac{4N\mu}{4N\mu + 1}$$

Desta forma, é o **valor de heterozigosidade equivalente ao ponto de equilíbrio entre mutação e deriva é comumente reportado como**:

$$H \approx \frac{4N\mu}{4N\mu + 1}$$

Conforme veremos adiante, o termo  $4N\mu$  tem uma importância grande em genética de populações. **Ele é frequentemente representado pela letra grega  $\theta$** . Desta forma, é comum encontrar a fórmula do ponto de equilíbrio mutação-deriva assim:

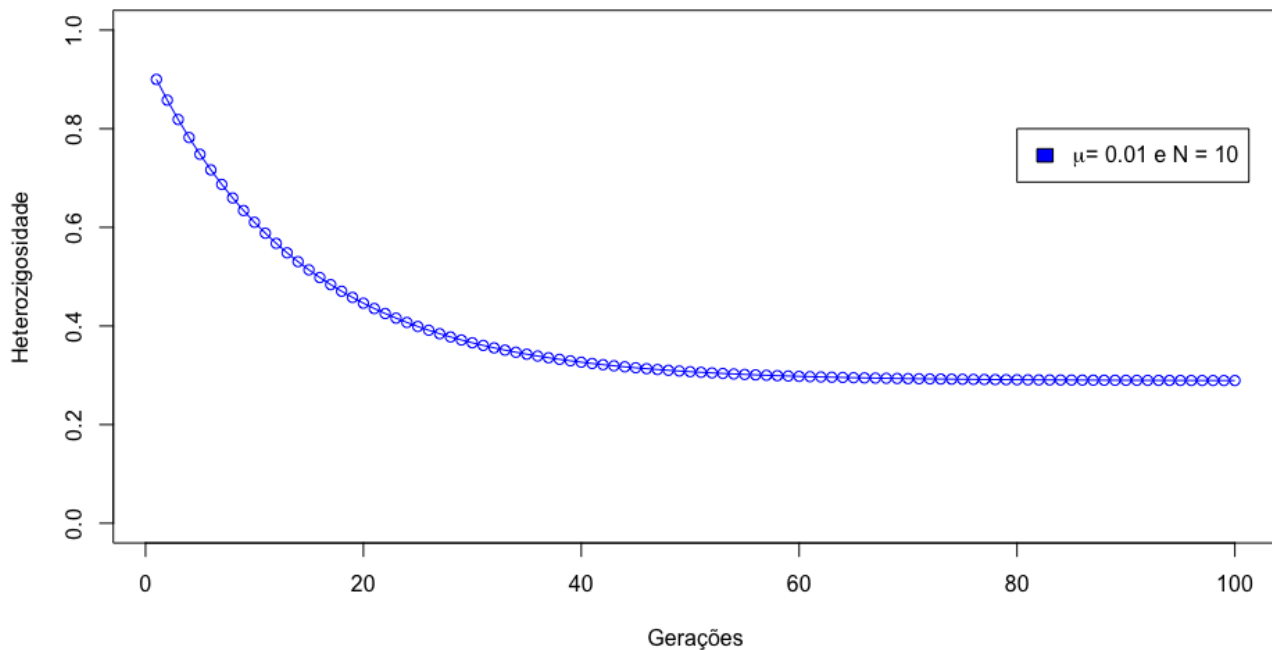
$$H = \frac{\theta}{\theta + 1}$$

Note que o símbolo de igualdade usado na fórmula é, na verdade, uma aproximação.

Podemos verificar quão boa é esta aproximação comparando os valores de  $H$  no equilíbrio com a fórmula com e sem a eliminação dos termos que contém  $\mu^2$  e  $2\mu$ . Vejamos o caso de  $N = 10$ ,  $\mu = 0.01$  e  $H$  inicial = 0.9:

```
N = 10
m = 0.01
H = 0.9
g = 100
for(i in c(2:g)){
  H = c(H, 1 - ( 1/(2*N) ) * ((1-m)^2) + (1-(1/(2*N)))*((1-m)^2)*(1 - H[length(H)] ) )
}

plot(H, col="blue", xlab="Gerações", ylab=c("Heterozigosidade"), ylim=c(0,1))
lines(H, col="blue")
legend(80,0.8, legend=c(expression(paste(mu,"= 0.01 e N = 10")) ), fill="blue" )
```



O ponto de  $H$  de equilíbrio parece estar entre 0.2 e 0.4. Vamos encontrar o valor com e sem a aproximação:

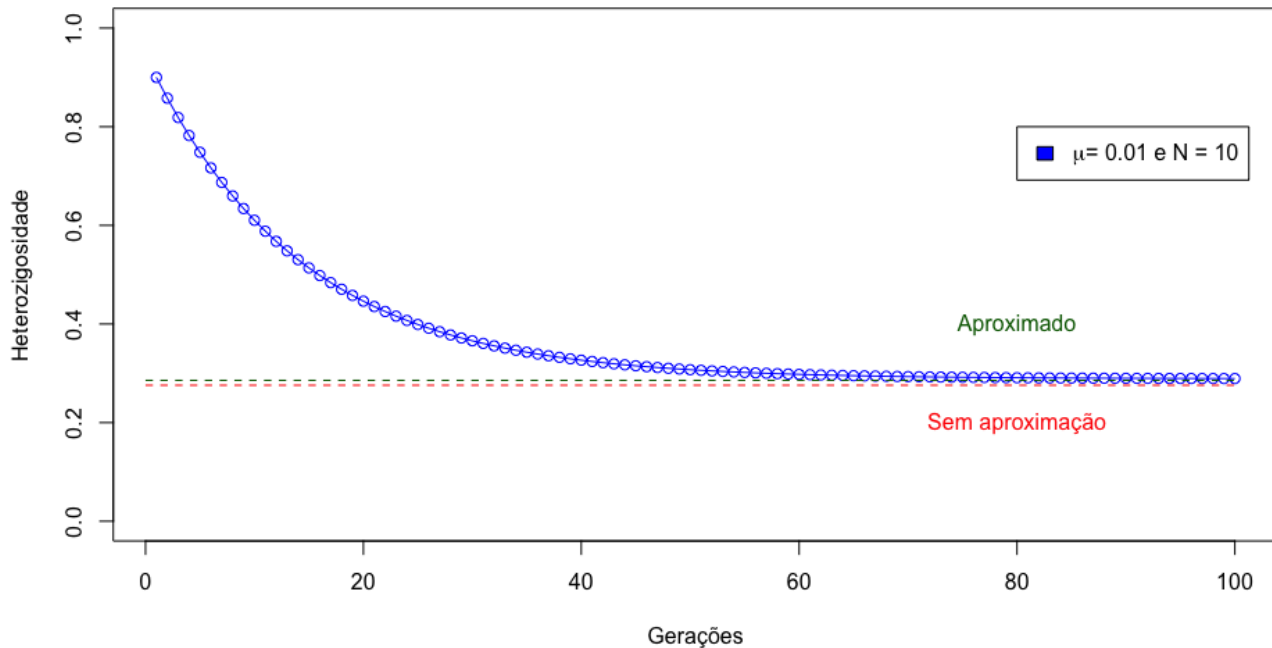
```
N = 10
m = 0.01
# Fórmula sem eliminação de  $m^2$  e  $2m$ :
H_equilibrio = ((2*m) - (4*N*m)) / ((2*(m^2)*N) - (m^2) - (4*N*m) + (2*m) - 1)

# Fórmula com aproximação:
H_equilibrio_aprox = (4*N*m) / ((4*N*m) + 1)
```

Sem aproximação, a frequência de heterozigotos do ponto de equilíbrio é 0.275742, enquanto o valor aproximado é igual a 0.2857143. Na prática, essa diferença é pouco relevante para fins biológicos. Portanto, a fórmula aproximada é largamente utilizada.

Finalmente, por curiosidade, podemos plotar no gráfico acima os valores de  $H$  de equilíbrio:

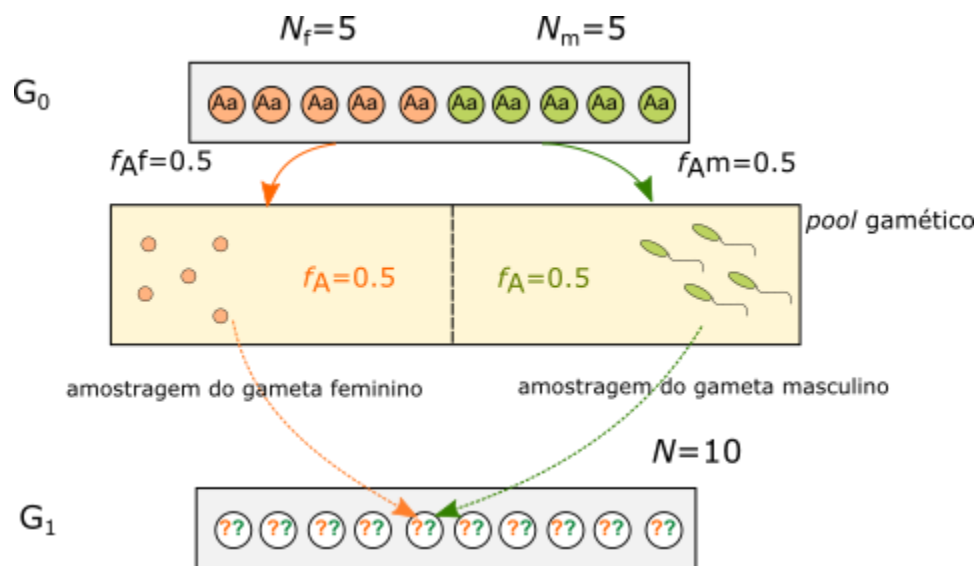
```
segments(0,0.275742,100,0.275742, col="red", lty=2)  
text(80,0.2, "Sem aproximação", col="red")  
  
segments(0,0.2857143,100,0.2857143, col="darkgreen", lty=2)  
text(80,0.4, "Aproximado", col="darkgreen")
```



## Efeito da variação do número de fêmeas e machos na diversidade genética e tamanho efetivo

Nesta prática, investigaremos uma outra situação muito comum nas populações naturais que também afetará no decaimento da frequência de homozigotos. Até agora, nossas simulações Wright-Fisher amostravam o *pool* gamético independente do gameta ser feminino ou masculino. Aqui, faremos uma distinção explícita entre gametas femininos e masculinos. Isso será importante para estudarmos o efeito de **sistemas diferenciais de cruzamento** no decaimento da frequência de heterozigotos por deriva.

Começaremos com o modelo mais simples de uma população que apresenta 100% de homozigotos e número igual de fêmeas (laranja) e machos (verde):

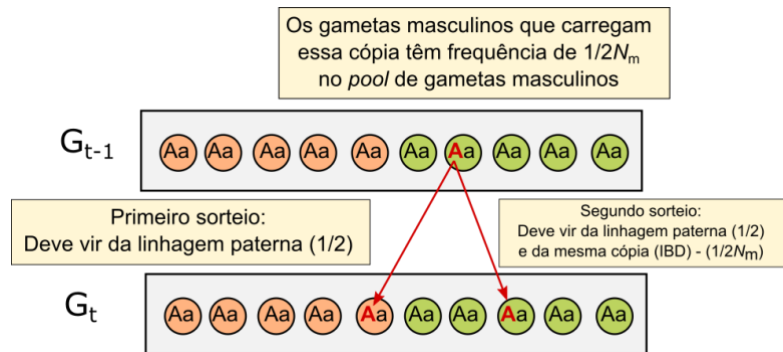


Neste exemplo, cada uma das fêmeas irá **contribuir igualmente** para o pool de gametas femininos. Desta forma, **a frequência dos gametas femininos que carregam o alelo A será idêntica à frequência deste alelo na população de fêmeas** (não há desvio meiótico). O mesmo ocorre para os machos. Com contribuição igual de cada indivíduo macho para o *pool* de gametas masculinos, não haverá diferença na frequência dos alelos entre os gametas masculinos e a população de machos.

Na geração seguinte, para formar o genótipo de um novo indivíduo, devemos amostrar um gameta feminino e um gameta masculino. A amostragem respeitará obviamente a frequência de cada um dos alelos nos respectivos *pools*. É importante notar a determinação genética do sexo deste novo indivíduo será feito pelo sexo heterogamético (por exemplo, o XY em humanos). **Assumindo que também não há viés para formação de um tipo específico de gameta no sexo heterogamético** (e.g., X ou Y nos humanos), a chance de formar uma fêmea (XX em humanos) ou um macho (XY em humanos) será idêntica. Desta forma, **esperamos que o número de fêmeas e machos seja equivalente** em  $G_1$ .

Agora podemos calcular um valor importante para estudar o decaimento da diversidade genética por deriva nesse modelo: **a probabilidade de amostrar, ao acaso, duas cópias alélicas que vieram de uma cópia alélica da mesma fêmea (ou de um mesmo macho) na geração anterior**. Em outras palavras, queremos calcular a probabilidade de as **duas cópias serem IBD de uma mesma fêmea**

**(ou macho) em uma geração (coalescência em 1 geração).** Vamos representar essa situação num esquema:



A lógica para encontrar essa probabilidade é semelhante àquela usada para calcular a probabilidade de encontrarmos um homocigoto IBD. Sabemos que a chance de ser IBD se define **apenas no segundo sorteio**. Desta forma, no **primeiro sorteio necessitamos apenas que a cópia venha do pool de gametas femininos**. Os gametas femininos compõem metade dos gametas do pool. Portanto, um gameta feminino será sorteado com probabilidade  $1/2$ . No segundo sorteio, queremos capturar um outro gameta feminino que possua uma cópia alélica que se originou da mesma molécula de DNA que o primeiro. Ora, inicialmente devemos sortear a cópia do alelo que o segundo indivíduo **herdou de sua mãe** (probabilidade =  $1/2$ ). Além disso, essa cópia alélica tem frequência no pool gamético de **1 entre total de cópias alélicas presentes na população de fêmeas**:  $1/2N_f$ .

Combinando os três termos usando multiplicação ( $1/2$  e  $1/2$  e  $1/2N_f$ ) temos que esse valor é  $1/8N_f$ .

Note que, se ignorarmos se o alelo vem da população de machos ou fêmeas, a probabilidade de sortear dois cópias alélicas IBD numa geração é  $1/2N$ , onde  $N$  é a população total, ou seja,  $N_f + N_m$ . Essa probabilidade é igual à chance de encontrar um indivíduo homocigoto IBD. Isso faz sentido pois, na prática, como os genótipos são formados pelo sorteio do *pool*, portanto é irrelevante se eles vão ser aferidos ao mesmo indivíduo ou a indivíduos diferentes.

Desta forma, podemos afirmar então que a chance de sorteio de duas cópias IBD é:

$$\frac{1}{8N_f} + \frac{1}{8N_m} = \frac{1}{2N}$$

Pois  $N = N_f + N_m$ . Usaremos essa formulação para definir o **tamanho populacional efetivo ( $N_e$ ) quando há variação na contribuição de machos e fêmeas para o *pool***:

$$\frac{1}{8N_f} + \frac{1}{8N_m} = \frac{1}{2N_e}$$

Que pode ser [revolvido para  \$N\_e\$](#) , resultando em:

$$N_e = \frac{4N_fN_m}{N_f + N_m}$$



Essa fórmula permite calcular qual seria a taxa de queda de heterozigiosidade na população idealizada de Wright-Fisher equivalente.

Por exemplo, admitamos que no sistema de reprodução do exemplo inicial exista uma única fêmea reprodutora (uma fêmea alfa ou a rainha de uma colméia). Desta forma, apenas uma fêmea contribui para o *pool* gamético, resultando em  $N_f = 1$ . **Apesar do tamanho populacional de censo seja 5, apenas 1 fêmea efetivamente contribui para o pool gamético.** Desta forma, temos que o tamanho efetivo desta população, ou seja, aquele que possibilitará aproximar o decaimento da heterozigiosidade será:

$$N_e = \frac{4(1)(5)}{1 + 5} = 3.333$$

Esse valor é bem menor que os  $N=10$  da população de censo. Desta forma, é evidente que num sistema reprodutivo de fêmea alfa existirá uma **atuação mais intensa da deriva genética.**

Podemos agora comparar o decaimento da diversidade genética em sistemas de reprodução onde nem todos os machos e fêmeas participam do pool gamético. Ou seja, existe uma **diferença entre o tamanho populacional de censo e o tamanho efetivo:**

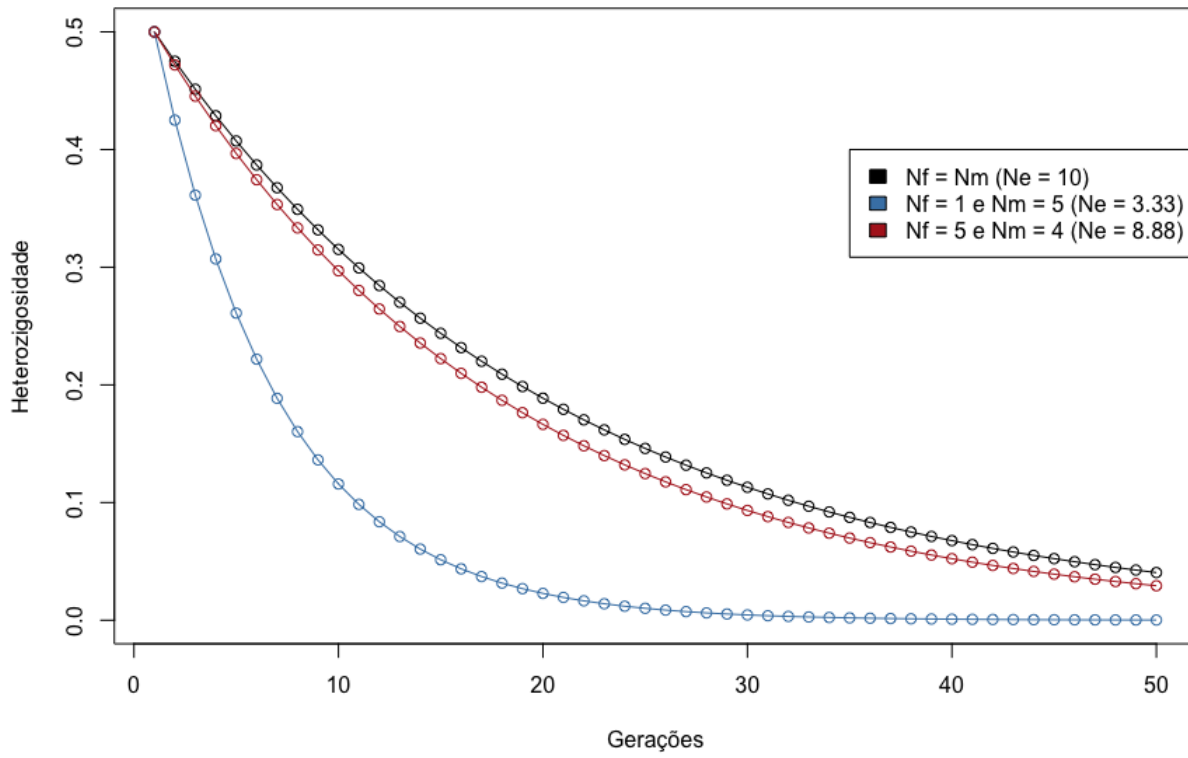
```
# Tamanho efetivo -- machos e fêmeas:
Nf = 5
Nm = 5

Ne = (4 * Nf * Nm) / (Nf + Nm)

# Equação do decaimento de H. Vamos admitir que H0 = 0.5

H = vector()
H[1] = 0.5

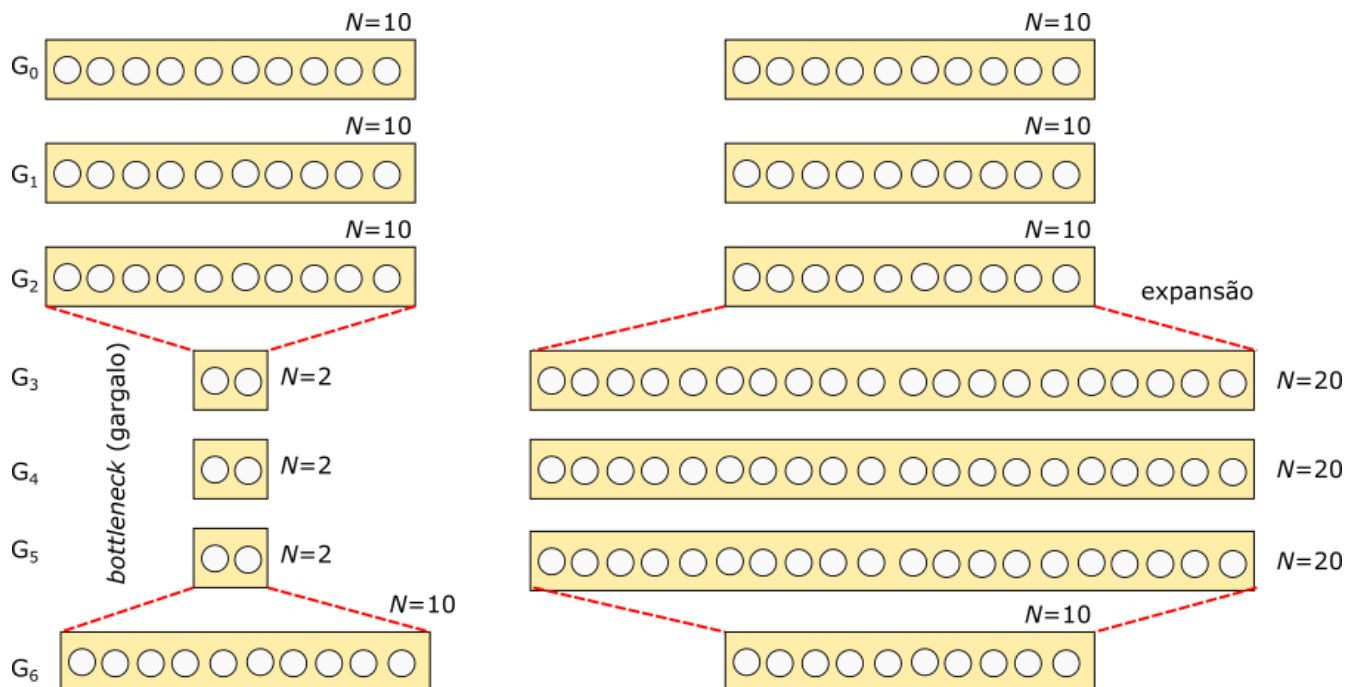
for(i in c(2:50)){
  H[i] = H[i -1] - (1/(2*Ne)) * H[i -1]
}
```



## Efeito da variação do tamanho da população na diversidade genética e tamanho efetivo

Nesta prática, iremos investigar os efeitos do *bottleneck* (gargalo) na heterozigosidade (= diversidade genética). Aplicaremos o mesmo modelo básico de Wright-Fisher, ou seja, a cada geração iremos compor um *pool* gamético com frequências alélicas idênticas às da população dos pais. Os genótipos da geração seguinte também serão compostos por amostragem do pool gamético. Entretanto, agora, **o tamanho da população da geração seguinte irá variar.**

Estudaremos dois exemplos que estão esquematizados abaixo:



No primeiro caso, em  $G_3$  ocorre uma redução do tamanho da população de  $N = 10$  para apenas 2 indivíduos. Essa população com tamanho reduzido continua até  $G_5$  quando ela retorna ao valor de  $N = 10$ . **Essa redução é chamada de *bottleneck* ou gargalo.** Neste caso, ele tem uma duração de três gerações. No segundo caso, ocorre uma expansão do tamanho da população para  $N = 20$  indivíduos. O período em que ocorre a expansão e sua duração são iguais ao primeiro caso.

Vejamos como essas dinâmicas populacionais alternativas ao tamanho constante afetam a evolução da frequência do alelo A. Podemos adaptar a simulação Wright-Fisher tradicional para acomodar essa variação do tamanho por três gerações:

```
# Bottleneck:
g = 50

N = rep(10, g) # vetor com os tamanhos populacionais a cada g geração (N = 10)
N[c(4:6)] = 2 # atribuição de N = 2 na 4ª, 5ª e 6ª geração

fA = nA = vector()

nA[1] = 10 # número de alelos na geração G0
```





```
fA[1] = nA[1]/(2*N[1]) # frequência do alelo A em G0

for(i in c(2:length(N))) {

  possibilidades = c(0:(2*N[i]))

  nA[i] = sample(possibilidades,1,prob =dbinom(possibilidades, (2*N[i]), prob=fA[i-1]) )

  fA[i] = nA[i]/(2*N[i])

}
```

No código acima, a simulação será rodada por  $g = 50$  gerações. Inicialmente, a frequência do alelo A é igual a 50% (10 cópias num total de  $2N = 20$ ). Agora, os tamanhos populacionais a cada geração estão guardados no vetor  $N$ , pois queremos informar explicitamente em quais gerações ocorre a diminuição ou aumento de  $N$ . Essa variação do número de  $N$  ao longo das gerações faz que precisemos fazer um novo vetor *possibilidades* de estados de número de alelos A a cada geração no *loop*. Por exemplo, se  $N = 10$ , existem 21 possibilidades de números de alelos A (de 0 até 20). Entretanto, quando  $N = 2$ , existem apenas 5 possibilidades (de 0 até 4). Finalmente, se  $N = 20$ , existem 41 possibilidades de estados de números de alelos A (de 0 até 40). A amostragem binomial, portanto, também deve ser calculada levando em conta esse novo vetor *possibilidades*.

O cenário de expansão pode, portanto, ser facilmente definida modificando o início do código. O mesmo se aplica caso queiramos comparar com o tamanho constante, onde basta não atribuir valores diferentes de tamanho populacional nas posições 4, 5 e 6 do vetor  $N$ .

```
# Expansão:
g = 50

N = rep(10, g) # vetor com os tamanhos populacionais a cada g geração (N = 10)
N[c(4:6)] = 20 # atribuição de N = 20 na 4ª, 5ª e 6ª geração
```

Ou

```
# Constante:
g = 50

N = rep(10, g) # vetor com os tamanhos populacionais a cada g geração (N = 10)
```

Sabemos que, ao rodar apenas uma vez cada um dos três cenários (*bottleneck*, expansão e constante), teremos uma enorme variação estocástica, pois o comportamento da frequência do alelo A ao longo das gerações é influenciado pelo erro amostral do *pool* gamético. Desta forma, assim como fizemos anteriormente no estudo da dinâmica Wright-Fisher clássica, podemos repetir a dinâmica da evolução da frequência alélica em cada caso um número grande de vezes independentes (*replicatas*). Assim, veremos, em média, a tendência do comportamento de  $fA$  em cada cenário. Lembre-se que essa é a lógica básica das cadeias de Markov.

Abaixo, usaremos um número de *replicatas* = 1000 para registrar a tendência geral do cenário de *bottleneck*. Acompanharemos a dinâmica da  $fA$  ao longo de  $g = 50$  gerações. Cada replicata independente (partindo do mesmo estado inicial) será registrada nas linhas da matriz  $fA$ . Desta forma,  $fA$  possui 1000 linhas e 50 colunas.

```
# Bottleneck:
g = 50
N = rep(10, g)
N[c(4:6)] = 2
```

```
replicatas = 1000
fA = nA = matrix(nrow=replicatas, ncol=g)

for(r in c(1:replicatas)){
  nA[r,1] = 10
  fA[r,1] = nA[r,1]/(2*N[1])

  for(i in c(2:g)){
    possibilidades = c(0:(2*N[i]))
    nA[r,i] = sample(possibilidades,1,prob =dbinom(possibilidades, (2*N[i]), prob=fA[r,i-1]))
    fA[r,i] = nA[r,i]/(2*N[i])
  }
}
```

Podemos calcular a frequência de replicatas em que A foi fixado com o comando:

```
length(fA[which(fA[,g] == 1) ]) / replicatas
```

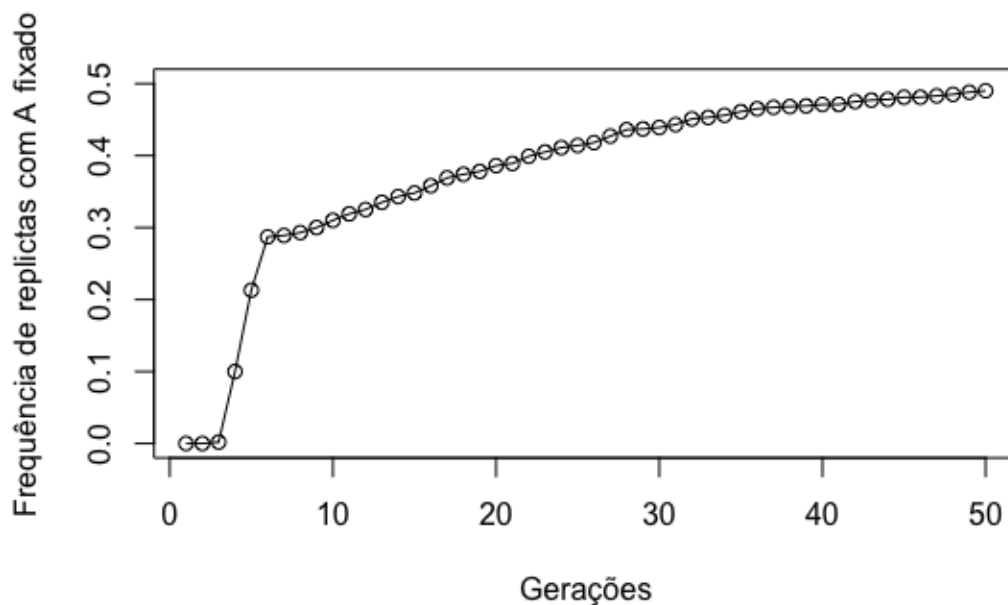
No meu experimento, esse valor foi igual a 49%.

Finalmente, iremos acompanhar ao longo das gerações a frequência de replicatas em que o alelo A foi fixado:

```
freq_fixados = vector()
for(i in c(1:g)){
  freq_fixados[i] = length(fA[which(fA[,i] == 1) ]) / replicatas
}
```

Podemos então plotar o gráfico da frequência de replicatas com fixação ao longo das gerações

```
plot(freq_fixados, xlab="Gerações", ylab="Frequência de replicatas com A fixado")
lines(freq_fixados)
```

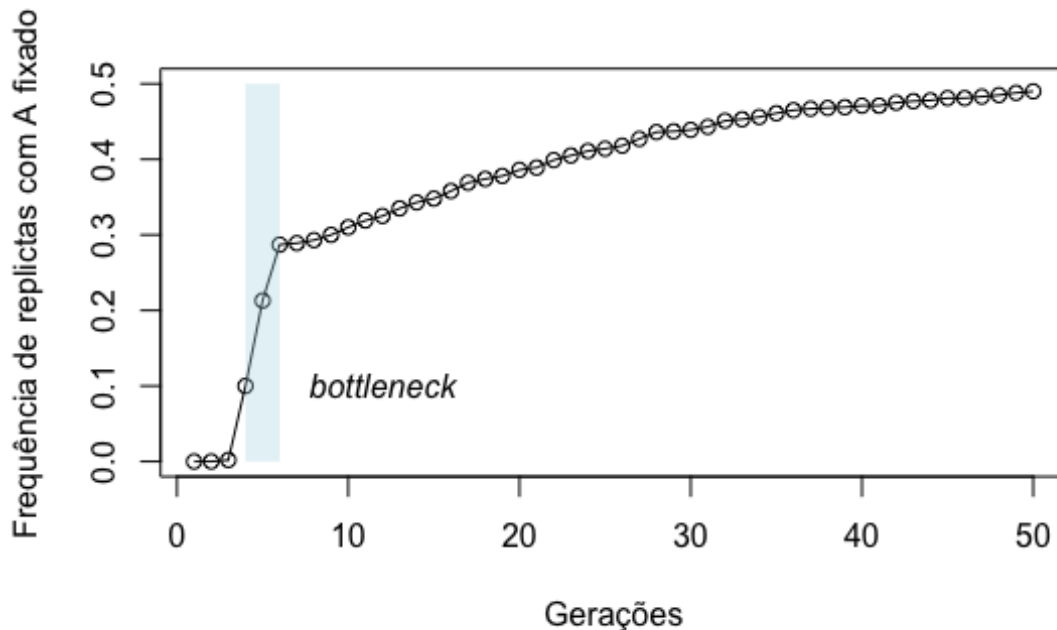


Enfim, vamos marcar neste gráfico as gerações em que o tamanho da população foi reduzido para  $N = 2$  (gerações 4, 5 e 6):

```

polygon(x=c(4,4,6,6),y=c(0,0.5,0.5,0), col=rgb(173, 216, 230, max = 255, alpha = 80), border=NA)
text(12,0.1, substitute(paste(italic("bottleneck"))))

```



Neste gráfico acima usamos duas funções novas. Na função **polygon** definimos os pontos dos vértices de um polígono –  $x=c(4,4,6,6)$ ,  $y=c(0,0.5,0.5,0)$  – os quatro vértices  $x$  e os quatro  $y$  respectivamente. Este polígono não tem bordas – `border=NA`, e a cor de preenchimento é um azul claro transparente – `col=rgb(173, 216, 230, max = 255, alpha = 80)`. Para definir a cor de preenchimento, usamos uma outra função, **rgb**,

Na função **rgb**, definimos os valores de tons red, green e blue da cor “lightblue” (173, 216, 230), o azul claro, numa escala que vai até 255 – `rgb(173, 216, 230, max = 255, alpha = 80)`. Por fim, `alpha = 80` indicou o nível de transparência da cor. Para saber o valor dos tons de “lightblue”, podemos usar o comando **col2rgb**, que retornará:

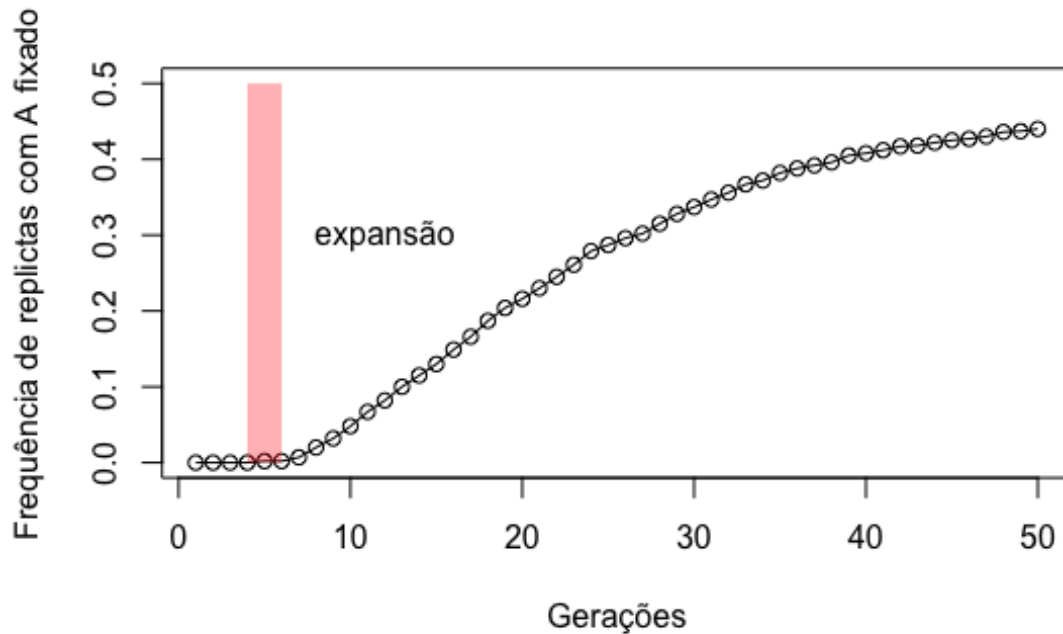
```

> col2rgb("lightblue")
      [,1]
red      173
green    216
blue     230

```

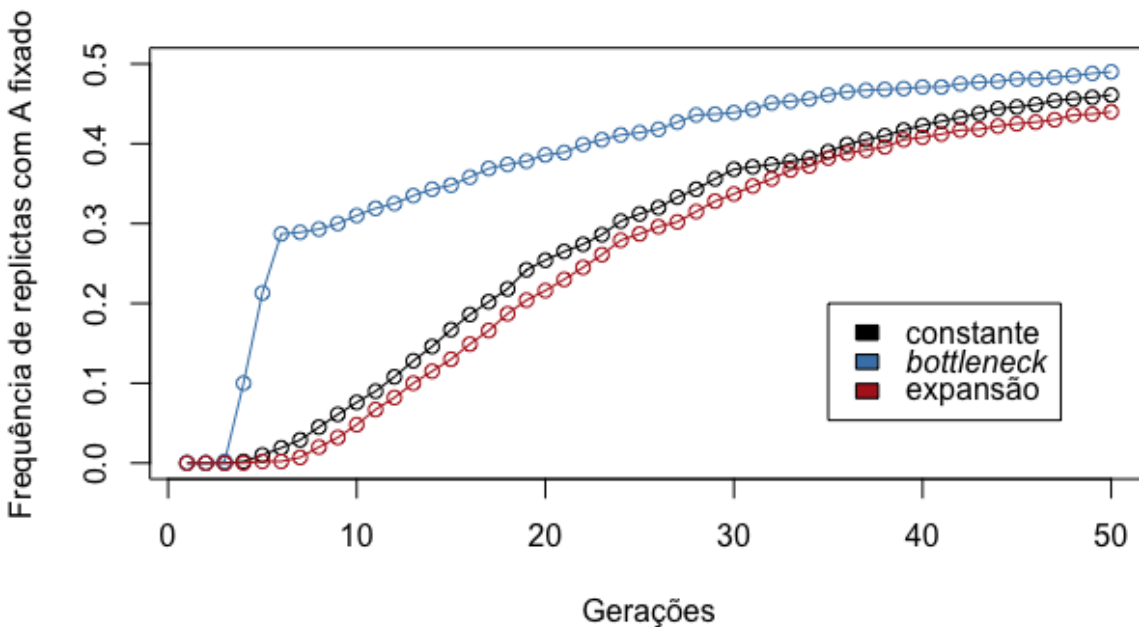
No gráfico acima, fica evidente que, durante o período de gargalo, a frequência de replicatas em que houve fixação do alelo A aumentou bastante. Lembremos que, em genética de populações, frequência de ocorrência é igual à probabilidade. Desta forma, podemos afirmar que **o bottleneck aumentou a probabilidade de fixação do alelo A**. Isso aconteceu porque houve **aumento da intensidade da deriva genética**. Consequentemente, o *bottleneck* resultou numa maior perda de diversidade (frequência de heterozigotos).

Façamos agora o cenário de expansão, usando os mesmos comandos aplicados anteriormente. No meu experimento, após  $g = 50$  gerações, 44% das replicatas estavam fixadas para o alelo A. O gráfico de evolução da frequência de replicatas com o alelo A fixado foi o seguinte:



Comparado com o gráfico anterior, é evidente que o período de expansão **atenuou a probabilidade do alelo A ser fixado**, pois a frequência de replicatas com este alelo fixado é bem próxima de zero. Desta forma, podemos afirmar que a expansão populacional diminuiu a probabilidade de fixação, **atenuando o efeito da deriva**. Consequentemente, a expansão aumentou a probabilidade de encontrarmos heterozigotos, atuando favoravelmente à **manutenção da diversidade genética**.

Vamos agora fazer um gráfico conjunto comparando os três cenários:



Comparado com o tamanho populacional constante ao longo do tempo, o **gargalo acentuou** a intensidade da deriva, enquanto a **expansão atenuou** a deriva.

Enfim, podemos também verificar o decaimento da heterozigosidade em cada um dos cenários. Para fins práticos, fazer esse cálculo, a cada geração vamos admitir que os gametas **amostrados do pool gamético** se encontrarão ao acaso. Desta forma, esperamos que ao amostrar  $p$  gametas  $A$  e  $q$  gametas  $a$ , encontraremos  $2pq$  heterozigotos na geração seguinte. Note que essa estratégia **não é equivalente** ao modelo Hardy-Weinberg, pois o cálculo de  $2pq$  foi feito **após a amostragem** dos  $2N$  gametas do *pool*. Portanto, temos a atuação da deriva.

Para ter acesso ao decaimento aproximado o máximo possível a curva teórica, podemos repetir o procedimento para um número muito grande de replicatas. Abaixo, usaremos *replicatas* = 100000. Esse código, portanto, será um pouco mais demorado para finalizar. Vamos guardar todas as heterozigosidades calculadas numa matriz  $H$  com um número *replicatas* de linhas e  $g$  de colunas:

```
# Decaimento da heterozigosidade:
g = 50
N = rep(10, g)
N[c(4:6)] = 20
replicatas = 100000
fA = nA = H = matrix(nrow=replicatas, ncol=g)

for(r in c(1:replicatas)){

  nA[r,1] = 10
  fA[r,1] = nA[r,1]/(2*N[1])
  H[r,1] = 2 * fA[r,1] * (1 - fA[r,1]) # 2pq

  for(i in c(2:g)){

    possibilidades = c(0:(2*N[i]))
```



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.

Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

```
nA[r,i] = sample(possibilidades,1,prob =dbinom(possibilidades, (2*N[i]), prob=fA[r,i]-
1))
fA[r,i] = nA[r,i]/(2*N[i])
H[r,i] = 2 * fA[r,i] * (1 - fA[r,i]) # Assumimos que os gametas do pool se encontram ao
acaso (2pq)

}

}
```

Após rodar o código, podemos calcular a heterozigosidade média entre as replicatas de cada geração com o comando:

```
apply(H,2,mean)
```

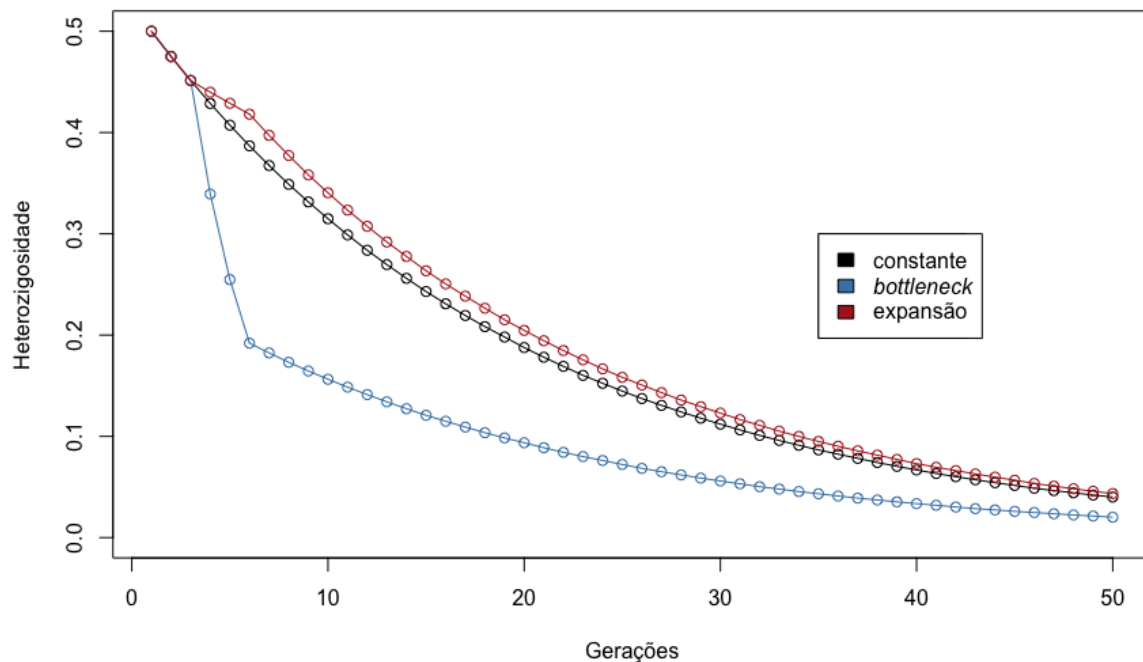
O comando **apply** indica que queremos a média (mean) das colunas (o “2” do comando – se usássemos “1” teríamos a média das linhas). Ou seja, esse comando retornará um vetor de tamanho  $g = 50$  com as médias das replicatas para cada geração. Evidentemente, o primeiro valor é igual a 50%, pois iniciamos com todas as replicatas com  $fA = 0.5$  e, conseqüentemente,  $H = 0.5$ .

```
> apply(H,2,mean)
 [1] 0.50000000 0.47501820 0.45145880 0.42870105 0.40725990 0.38681525
 [7] 0.36740070 0.34882695 0.33145870 0.31489090 0.29893745 0.28376275
[13] 0.26954370 0.25602560 0.24296010 0.23089270 0.21933775 0.20834155
[19] 0.19809200 0.18769655 0.17814100 0.16912915 0.16024175 0.15239815
[25] 0.14468825 0.13731945 0.13059040 0.12410635 0.11793065 0.11207015
[31] 0.10635645 0.10098955 0.09594950 0.09120430 0.08663070 0.08245845
[37] 0.07827110 0.07417120 0.07042130 0.06683160 0.06334475 0.06017700
[43] 0.05712110 0.05426060 0.05154650 0.04878885 0.04646200 0.04429425
[49] 0.04203510 0.04006915
```

Para plotar o gráfico do decaimento, basta usarmos:

```
plot(apply(H,2,mean), xlab="Gerações", ylab="Heterozigosidade",ylim=c(0,0.5))
lines(apply(H,2,mean))
```

Abaixo mostraremos as linhas de decaimento dos três cenários num único gráfico:



Novamente, observamos como o *bottleneck* diminui a heterozigosidade, enquanto a expansão impediu um decaimento mais intenso da diversidade genética.

Evidenciamos, portanto, que a variação do tamanho da população ao longo das gerações pode influenciar a probabilidade de fixação dos alelos. Obviamente, essa variação adiciona complexidade ao estudo da deriva genética, pois como podemos fazer aplicar o processo de Markov quando existe essa variação de  $N$  ao longo do tempo de forma a prever o comportamento do sistema?

Essa é uma das principais motivações para a utilização do conceito de **tamanho populacional efetivo**, representado pelo símbolo  $N_e$ .

O tamanho populacional efetivo é a medida do tamanho de uma população idealizada Wright-Fisher ( $N$  constante) que sofre uma quantidade de deriva equivalente àquela observada numa população com uma dinâmica qualquer.

Por exemplo, o efeito da deriva no cenário de *bottleneck* é **equivalente** ao efeito da deriva numa população Wright-Fisher constante com qual tamanho? A mesma situação se aplica ao cenário de expansão. Pelo gráfico comparativo acima, é intuitivo afirmar que no cenário de *bottleneck*, o valor do tamanho da população Wright-Fisher (WF) equivalente seria menor que  $N = 10$ , pois a deriva neste cenário é mais intensa daquela encontrada numa população WF de  $N = 10$  (linha preta do gráfico). O oposto se aplica ao cenário de expansão. A população WF equivalente deve ser maior, pois a deriva é mais atenuada.



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

Para calcular os valores de equivalência com populações WF, ou seja, o tamanho efetivo  $N_e$ , usamos o seguinte raciocínio para o caso de **variação do tamanho populacional no tempo**:

Lembremos que, com a atuação da deriva genética, a heterozigiosidade tende a cair na taxa de  $\frac{1}{2N}$  a cada geração:

$$H_t = H_{t-1} - \frac{1}{2N} H_{t-1}$$

Essa relação pode também ser expressa da forma:

$$H_t = \left(1 - \frac{1}{2N}\right) H_{t-1}$$

Assim, por exemplo, na 2ª geração, temos o seguinte valor de heterozigiosidade:

$$H_2 = \left(1 - \frac{1}{2N}\right) H_1$$

Se a frequência de heterozigotos inicial é  $H_0$ , podemos afirmar que  $H_1 = \left(1 - \frac{1}{2N}\right) H_0$ . Desta forma, podemos escrever que  $H_2$  é equivalente a:

$$H_2 = \left(1 - \frac{1}{2N}\right) \left(1 - \frac{1}{2N}\right) H_0$$

É evidente pela nova formulação acima que, após  $t$  gerações de decaimento com a mesma taxa, teremos o seguinte valor de heterozigiosidade:

$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0$$

Ou mesmo:

$$\frac{H_t}{H_0} = \left(1 - \frac{1}{2N}\right)^t$$

Se, entretanto, o valor  $N$  variar a cada geração **teremos que especificar isso na fórmula acima**. Por exemplo, suponhamos que em quatro ( $t = 4$ ) gerações temos os valores  $N_0$ ,  $N_1$ ,  $N_2$  e  $N_3$ . O decaimento da heterozigiosidade comparada a  $H_0$  será:

$$\frac{H_4}{H_0} = \left(1 - \frac{1}{2N_0}\right) \left(1 - \frac{1}{2N_1}\right) \left(1 - \frac{1}{2N_2}\right) \left(1 - \frac{1}{2N_3}\right)$$

Ou seja, temos o produto (simbolizado pela letra grega maiúscula  $\prod$ ) dos termos  $\left(1 - \frac{1}{2N}\right)$  para cada valor de  $N$  da geração correspondente. Isso pode ser escrito de forma compacta como:

$$\frac{H_t}{H_0} = \prod_{i=1}^{t-1} \left(1 - \frac{1}{2N_i}\right)$$





O que nos interessa, entretanto, é que demos encontrar um novo valor de  $N$ , chamado de tamanho populacional efetivo ( $N_e$ ) **que seja a solução da seguinte equação:**

$$\left(1 - \frac{1}{2N_0}\right)\left(1 - \frac{1}{2N_1}\right)\left(1 - \frac{1}{2N_2}\right)\left(1 - \frac{1}{2N_3}\right) = \left(1 - \frac{1}{2N_e}\right)^4$$

Por quê? Pois sabemos que, **se o tamanho da população fosse constante**, a taxa de redução na geração  $t = 4$  seria  $\left(1 - \frac{1}{2N}\right)^4$ .

Desta forma, ao resolver essa equação, iremos encontrar um novo valor de  $N$  (chamado de  $N_e$ ) que corresponderá **ao mesmo decaimento da heterozigiosidade ocorrido em  $N_0, N_1, N_2$  e  $N_3$ , entretanto, numa população idealizada com tamanho constante.**

De forma geral, devemos resolver, para  $N_e$ , a equação:

$$\prod_{i=1}^{t-1} \left(1 - \frac{1}{2N_i}\right) = \left(1 - \frac{1}{2N_e}\right)^t$$

Para resolver essa equação, inicialmente chamemos o termo  $\prod_{i=1}^{t-1} \left(1 - \frac{1}{2N_i}\right) = a$ . Assim,  $a = \left(1 - \frac{1}{2N_e}\right)^t$ . Resolvendo para  $N_e$ , **temos:**

$$N_e = -\frac{1}{2(\sqrt[t]{a} - 1)} = \frac{1}{2\left\{1 - \left[\prod_{i=1}^{t-1} \left(1 - \frac{1}{2N_i}\right)\right]^{1/t}\right\}}$$

Cuja **aproximação** resulta em:

$$N_e = \frac{t}{\sum_{i=1}^{t-1} \frac{1}{N_i}}$$

A fórmula indica que o tamanho efetivo é dado pela **média harmônica dos tamanhos das populações a cada geração**. Pesquise sobre a [média harmônica](#). Essa fórmula também é frequentemente representada nos livros didáticos assim:

$$\frac{1}{N_e} = \frac{1}{t} \left( \frac{1}{N_0} + \frac{1}{N_1} + \dots + \frac{1}{N_{t-1}} \right)$$

Uma propriedade importante da média harmônica é que **os valores menores têm um peso maior** (pois os valores se encontram no denominador). Assim, quando existe uma redução do tamanho da população, o impacto na redução da heterozigiosidade é significativo. Esse impacto será capturado no cálculo do tamanho populacional efetivo –  $N_e$ . Uma forma fácil de enxergar essa propriedade é trabalharmos com um caso muito simples: Calcular o  $N_e$  para  $t = 2$  gerações, com  $N_0 = 1$  e  $N_1 = 10$ .

Se

$$\prod_{i=1}^{t-1} \left(1 - \frac{1}{2N_i}\right) = \left(1 - \frac{1}{2N_e}\right)^t$$

Temos



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
 Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

$$\left(1 - \frac{1}{2N_e}\right) \left(1 - \frac{1}{2N_e}\right) = \left(1 - \frac{1}{2N_0}\right) \left(1 - \frac{1}{2N_1}\right)$$

Com  $N_0 = 1$  e  $N_1 = 10$ :

$$\left(1 - \frac{1}{2N_e}\right) \left(1 - \frac{1}{2N_e}\right) = \left(1 - \frac{1}{2(1)}\right) \left(1 - \frac{1}{2(10)}\right)$$

Essa formulação deixa evidente que o valor de  $N_e$  que resolve essa igualdade deve ser mais próximo de 1 do que 10, pois o termo  $\left(1 - \frac{1}{2(1)}\right) = 0.5$  é mais impactante para a multiplicação do que  $\left(1 - \frac{1}{2(10)}\right) = 0.95$ . Como é um produto, quanto mais próximo de 1, menos impactante será. Essa discrepância é lavada em consideração pela média harmônica.

Para calcular a média harmônica dos valores do vetor  $N$ , que armazena os valores de tamanho populacional a cada geração, basta usarmos o comando:

```
Ne = length(N) / sum(1/N)
```

Os tamanhos efetivos, ao longo das 50 gerações em que a simulação foi rodada serão:

$$N_e = 8.064516 \text{ (cenário de } bottleneck)$$

$$N_e = 10.30928 \text{ (cenário de expansão)}$$

Verificamos, portanto, que o *bottleneck* aplicado resultou numa taxa de queda de heterozigosidade **correspondente ao de uma população Wright-Fisher** de tamanho igual a 8.064516. Por outro lado, a expansão do tamanho populacional resultou numa atenuação do efeito da deriva na heterozigosidade correspondente ao que acontece numa população Wright-Fisher com 10.30928 indivíduos.

Podemos comparar os gráficos do decaimento observado com aquele previsto usando  $N_e$  com intuito de verificar se o decaimento da população Wright-Fisher idealizada é uma aproximação razoável:

```
# curva do decaimento usando Ne (bottleneck)
Ne = 8.064516
g = 50

H_Ne = vector()
H_Ne[1] = 0.5

for(i in c(2:g) ){
  H_Ne[i] = H_Ne[i-1] - (1/(2*Ne)) * H_Ne[i-1]
}

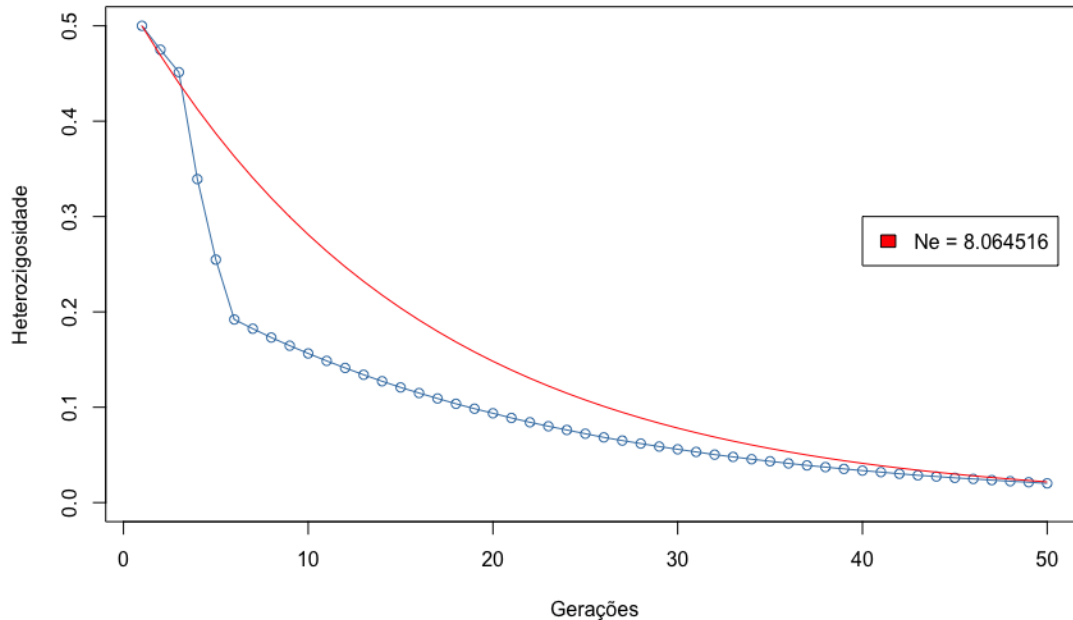
plot(apply(H,2,mean), xlab="Gerações", ylab="Heterozigosidade",ylim=c(0,0.5), col="steelblue",
main="Aproximação do gargalo por Ne")

lines(apply(H,2,mean), col="steelblue")

lines(H_Ne, col="red")

legend(40,0.3,legend="Ne = 8.064516", fill="red")
```

### Aproximação do gargalo por Ne



Embora a curva vermelha aparentemente não represente uma boa aproximação do decaimento do *bottleneck*, após a geração 45, o decaimento usando  $Ne = 8.064516$  **prediz muito bem o valor da heterozigidade**.

Vamos agora para o cenário de expansão:

```
# curva do decaimento usando Ne (expansão)
Ne = 10.30928
g = 50

H_Ne = vector()
H_Ne[1] = 0.5

for(i in c(2:g) ){
  H_Ne[i] = H_Ne[i-1] - (1/(2*Ne)) * H_Ne[i-1]
}

plot(apply(H,2,mean), xlab="Gerações", ylab="Heterozigidade",ylim=c(0,0.5), col="firebrick",
main="Aproximação da expansão por Ne")

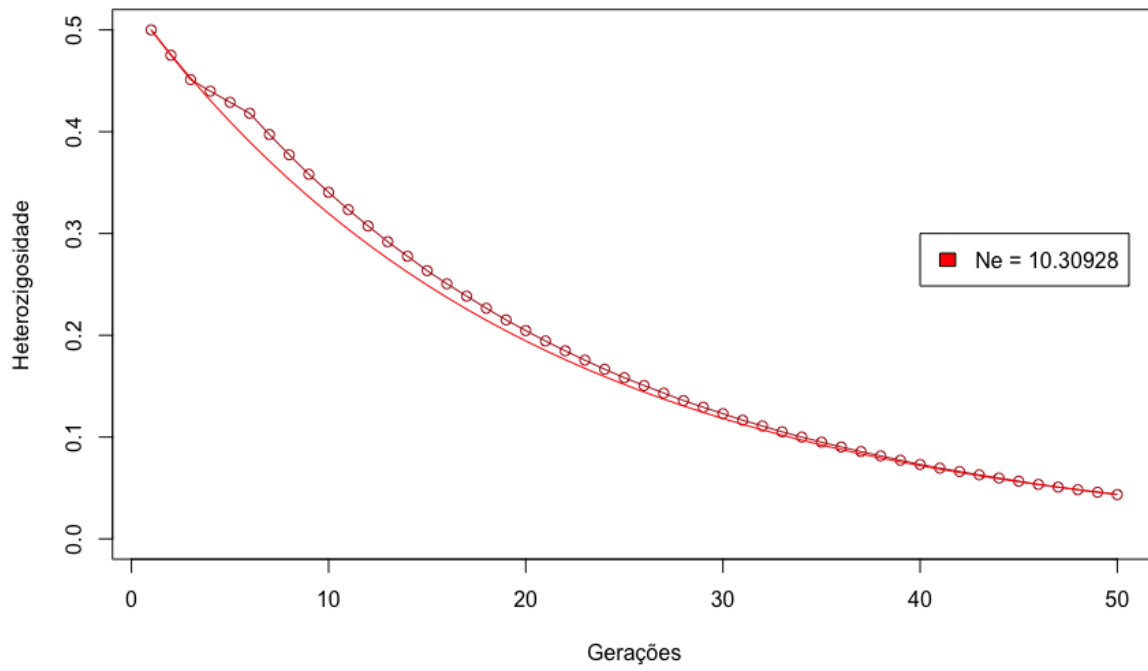
lines(apply(H,2,mean), col="firebrick")

lines(H_Ne, col="red")

legend(40,0.3,legend="Ne = 10.30928", fill="red")
```



### Aproximação da expansão por $N_e$

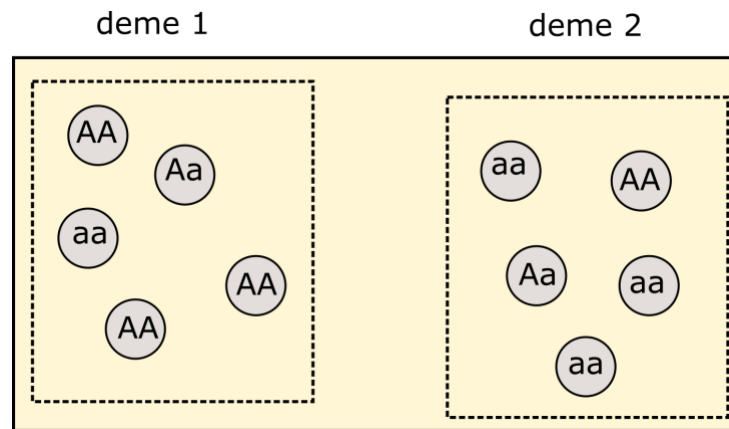


## Redução da heterozigidade devido à estruturação populacional

As populações biológicas raramente apresentam homogeneidade no espaço geográfico. Isso significa que existe uma chance diferenciada de formação de pares reprodutivos devido à proximidade espacial: indivíduos mais próximos no espaço frequentemente possuem maior chance de realizarem cruzamento que indivíduos distantes. Em casos mais drásticos, duas ou mais subunidades de uma população (**demes** ou **metapopulações**) podem ficar algumas gerações sem trocar nenhum migrante (fluxo gênico nulo). A consequência dessa propriedade é que a diversidade genética não será espalhada no espaço de forma homogênea.

Essa compartimentalização da diversidade genética é chamada de estruturação populacional. Na prática, devido ao isolamento das populações locais, a deriva genética e endocruzamento ficam intensificados nas diversas demes e, como resultado, temos uma redução da frequência esperada de heterozigotos.

Vamos verificar esse fenômeno com um caso simples. Suponhamos que existam duas demes/metapopulações (chamadas de 1 e 2) em que a composição dos genótipos em cada uma delas é a seguinte:



A partir deste cenário, podemos medir as frequências alélicas  $fA$  e a heterozigidade  $HI$  de cada deme. Vamos inserir esses valores em dois vetores, onde cada posição equivale a uma deme:

`HI = c(1/5 , 1/5) # heterozigosidades da deme1 e deme2`

`HI = mean(HI) # media das heterozigosidades observadas`

`fA = c(6/10 , 3/10) # frequencias alélicas da deme1 e deme2`

Agora podemos calcular as heterozigosidades  $H_S$  e  $H_T$  conforme as fórmulas

$$H_S = \frac{2p_1q_1 + 2p_2q_2}{2}$$

Assim,  $H_S$  nada mais é do que a média entre as heterozigosidades em cada deme assumindo o equilíbrio Hardy-Weinberg (ou seja, sem deriva e endocruzamento):

`HS = 2 * fA * (1-fA) # faz 2pq da deme1 e deme2`

`HS = mean(HS) # faz a média dos 2pq`



Por outro lado,  $H_T$  é calculada como:

$$H_T = 2\bar{p}\bar{q}$$

Ou seja, em  $H_T$ , assumimos que **não existem as demes** 1 e 2 (portanto, usamos as frequências médias globais  $\bar{p}$  e  $\bar{q}$ ). Desta forma, a população como um todo está em EHW:

`HT = 2 * mean(fA) * (1 - mean(fA)) # faz 2pq da media dos alelos p e q (=1-p).`

Evidentemente, se as demes não resultassem numa interrupção da panmixia global, ou seja, não prejudicassem o fluxo de genes entre as áreas 1 e 2,  **$H_S$  seria igual a  $H_T$** . Qualquer desvio dessa igualdade, indica que efetivamente os indivíduos das áreas 1 e 2 possuem uma chance maior de cruzarem com aqueles pares localizados dentro de suas respectivas demes. Essa regionalização da probabilidade de cruzamento tem como consequência a redução da heterozigosidade  $H_S$  quando comparada a  $H_T$ . Essa redução pode ser capturada da seguinte forma:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

Note que o  $F_{ST}$  compara dois cenários em que o EHW é explicitamente assumido ( $2pq$ ). Desta forma, as diferenças entre os valores de  $H_S$  e  $H_T$  são exclusivamente oriundos da estruturação da diversidade genética nas demes 1 e 2:

`FST = (HT - HS) / HT`

Além do  $F_{ST}$ , podemos conceber mais duas comparações entre as heterozigosidades:

$$F_{IS} = \frac{H_S - H_I}{H_S}$$

E

$$F_{IT} = \frac{H_T - H_I}{H_T}$$

Note que  $F_{IS}$  é uma medida que compara a média das heterozigosidades observadas com a média do esperado pelo EHW. O desvio entre os dois valores pode ser causado pelos fatores que, portanto, alteram o EHW, como ausência de panmixia (cruzamento preferencial). Por fim,  $F_{IT}$  é uma medida global da redução da heterozigosidade causada por ausência de panmixia e estruturação populacional, ou seja, variação da frequência dos alelos entre as demes.

`FIS = (HS - HI) / HS`

`FIT = (HT - HI) / HT`

Para o cenário acima, os valores calculados para cada uma das estatísticas são:

$$\begin{aligned} F_{ST} &= 0.09 \\ F_{IS} &= 0.56 \\ F_{IT} &= 0.60 \end{aligned}$$

A leitura que fazemos deste cenário é que 9% da redução da heterozigosidade entre as demes é oriunda de estruturação populacional, enquanto 56% ocorre por desvio das condições de EHW dentro

de cada deme. Ao todo, temos uma redução de 60% da heterozigidade esperada caso as demes 1 e 2 se comportassem como uma única unidade panmítica.

```
N = 10
g = 50
estados = c(0:(2*N))
replicatas = 100000
fA = nA = matrix(nrow=replicatas, ncol=g)
Fst = vector()
Fst[1] = 0

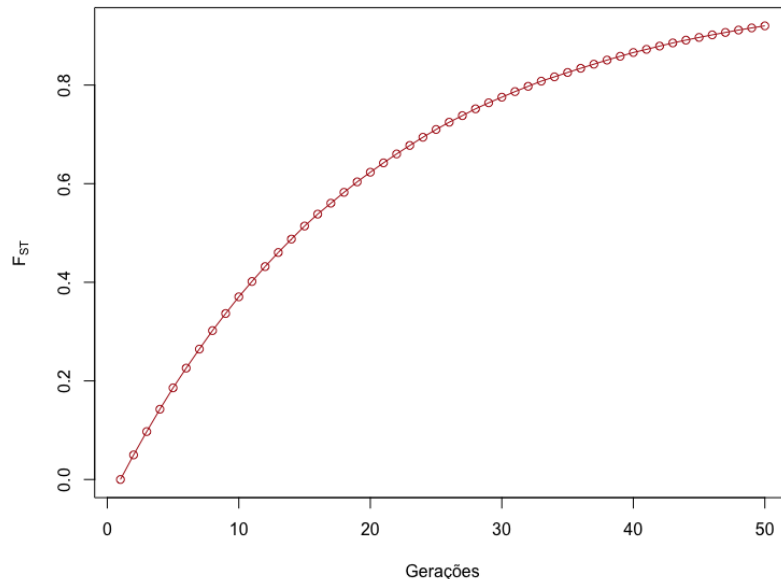
for(r in c(1:replicatas)){
  nA[r,1] = N
  fA[r,1] = nA[r,1] / (2*N)

  for(i in c(2:g)){
    nA[r,i] = sample(estados, 1, prob=dbinom(estados, (2*N), prob = fA[r, (i-1)]))
    fA[r,i] = nA[r,i] / (2*N)
  }
}

for(i in c(1:g)){
  HT = 2 * mean(fA[,i]) * (1 - mean(fA[,i]))
  HS = mean( 2 * fA[,i] * (1 - fA[,i]) )
  Fst[i] = (HT - HS)/HT
}
```

Podemos notar claramente que o valor de  $F_{ST}$  aumenta entre as *replicatas* ao longo das gerações:

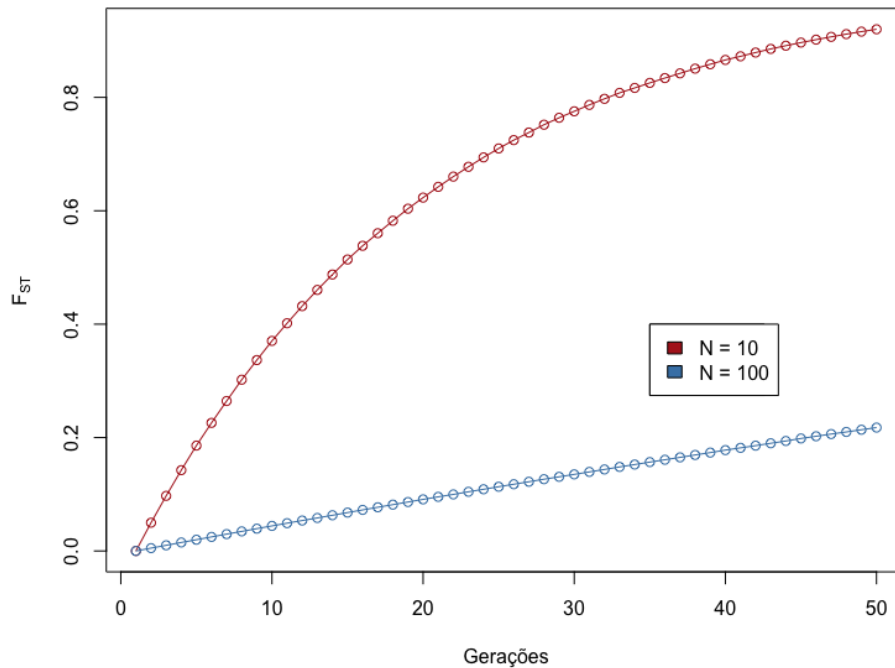
```
plot(Fst, ylab=expression(F[ST]), xlab="Gerações", col="firebrick")
lines(Fst, col="firebrick")
```



Isso indica que a diferenciação entre as frequências alélicas em cada uma das 100000 replicatas aumenta com o tempo. Isso ocorre pois em cada replicata teremos a **ação da deriva eliminando os heterozigotos** e, conseqüentemente, aumentando a discrepância entre a frequência esperada de heterozigotos na população total e aquela de cada replicata. Para comprovar que o tamanho da

população é um fator decisivo na diferenciação genética entre as populações, vamos rodar a mesma simulação para  $N = 100$ :

$N = 10$



Note que o comportamento de aumento do  $F_{ST}$  ao longo das gerações está intimamente relacionado ao aumento da homozigosidade IBF ( $F$ ) calculada usando a equação:

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_{t-1}$$

Ambos são considerados **índices de fixação** ( $F$ ) de alelos, pois a fixação de alelos tem como consequência direta o aumento da frequência de homozigotos.

Quando populações ficam separadas em demes por algum período, existe a possibilidade de elas voltarem a compor uma única unidade panmítica. Quando isso acontece, dizemos que ocorreu **quebra de isolamento**. Um dos efeitos da união das demes isoladas é o **aumento da heterozigosidade comparada ao esperado pelo equilíbrio**.

Suponhamos um caso simples onde existe fixação de alelos alternativos nas demes 1 e 2:

```
fA = c(1, 0) # a deme1 tem o alelo A fixado e a deme2 tem o alelo a fixado
HT = 2 * mean(fA) * (1 - mean(fA))
```

A frequência de heterozigotos esperada é 50%. Entretanto, o cruzamento dos indivíduos AA da deme 1 com os indivíduos aa da deme 2 resultará em 100% de heterozigotos Aa na primeira geração de filial.

Desta forma teremos uma discrepância entre a frequência observada (100%) e a esperada de heterozigotos (50%):





$$F_{IT} = \frac{0.5 - 1}{0.5} = -1$$

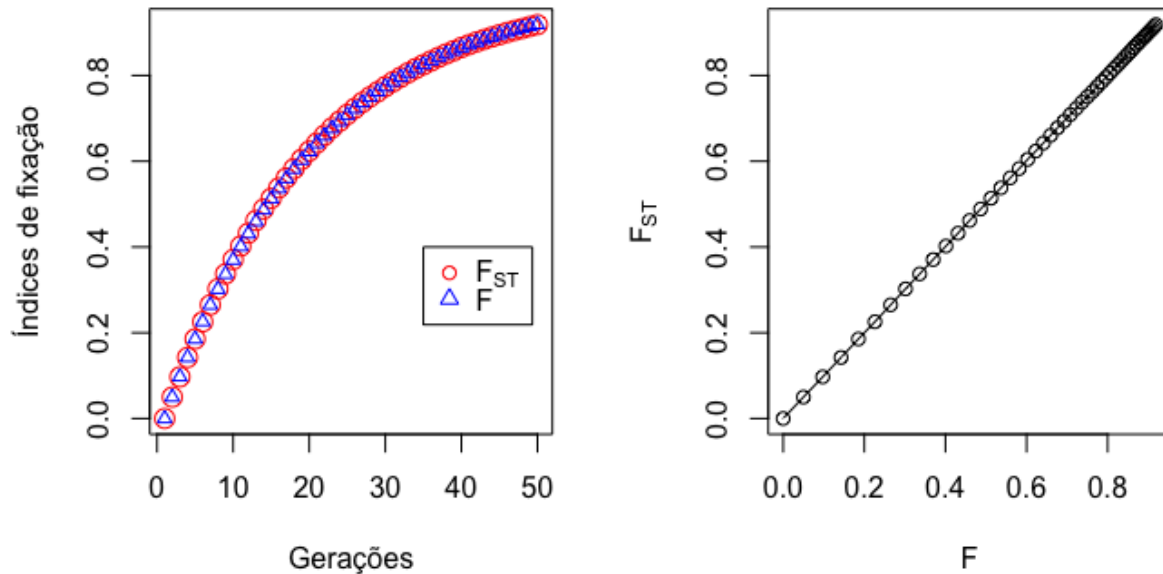
**Um valor de  $F$  negativo indica excesso de heterozigotos.** Pense em outras situações em que isso pode ocorrer.

Esses fenômenos associados à redução da frequência de heterozigotos da população total e, conseqüentemente, aumento da frequência de homozigotos nas demes associado ao aumento da frequência de heterozigotos quando demes isoladas voltam a funcionar como uma unidade panmítica foram reportados inicialmente em 1928 pelo geneticista sueco Sten Whalund. Desta forma, eles são denominados de **efeito Whalund**.

É relevante mencionar que todos os índices de fixação  $F$  são também medidas de probabilidade de amostragem de alelos IBD numa população. Quando não existe fluxo gênico algum entre as demes, o aumento da homozigosidade  **$F$  em cada uma das demes será numericamente equivalente** ao valor  $F_{ST}$ . Isso pode ser facilmente verificado plotando o valor do aumento da homozigosidade IBD no gráfico anterior que mostra o aumento do  $F_{ST}$  em demes com  $N = 10$ :

```
N = 10
g = 50
F = vector()
F[1] = 0
for(i in c(2:g)){
  F[i] = (1/(2*N)) + (1 - (1/(2*N))) * F[i-1]
}

par(mfcol=c(1,2)) # prepara área de plot para 1 linha com 2 plots
# primeiro plot
plot(Fst, pch=1, col="red", cex=1.5, ylab="Índices de fixação", xlab="Gerações")
points(F, pch=2, col="blue", cex=0.8)
legend(35,0.4,pch=c(1,2), col=c("red", "blue"), legend=c(expression(F[ST], "F")))
# segundo plot
plot(Fst ~ F, ylab=expression(F[ST]), xlab="F")
lines(Fst ~ F)
```



Por fim, o  $F_{ST}$ , uma medida de diferenciação genética entre as demes, é também equivalente a razão entre a variância das frequências alélicas entre as demes – no nosso caso  $fA = p$ , portanto  $\text{var}(p)$  – e a variância da frequência alélica esperada se a população se comportasse como uma única unidade panmítica ( $\bar{p}\bar{q}$ ), que é a variância esperada pela distribuição binomial:

$$F_{ST} = \frac{\text{var}(p)}{\bar{p}\bar{q}}$$

Essa fórmula se aplica apenas quando o número de demes analisadas é grande. Por exemplo, no experimento em que `replicatas = 100000` avaliamos 100000 demes independentes que começavam com  $fA = p = 0,5$ . Podemos calcular a  $\text{var}(p)$  a cada geração com o comando:

```
apply(fA, 2, var)
```

A nova fórmula de  $F_{ST}$  seria:

```
Fst_var = apply(fA, 2, var) / ( apply(fA, 2, mean) * ( 1 - apply(fA, 2, mean) ) )
```

Podemos plotar os valores de  $F_{ST}$  calculados das duas formas diferentes para verificar sua equivalência:

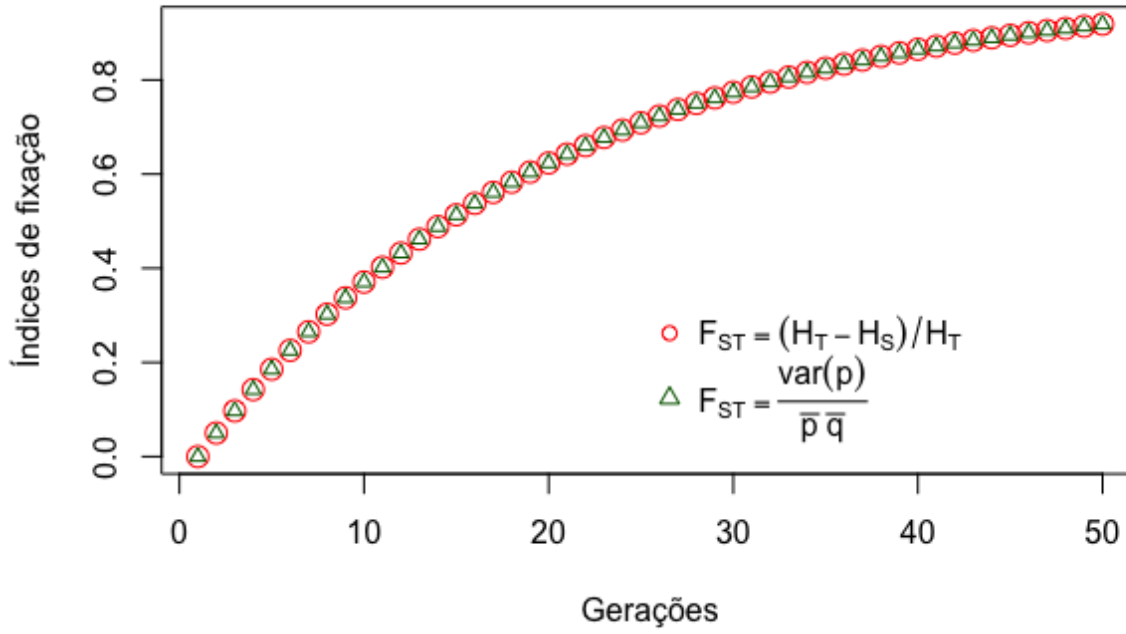
```
plot(Fst, pch=1, col="red", cex=1.5, ylab="Índices de fixação", xlab="Gerações")
```

```
points(Fst_var, pch=2, col="darkgreen", cex=0.8)
```

```
legend(25,0.4,pch=c(1,2), col=c("red", "darkgreen"), legend=c(expression(F[ST] == (H[T]-H[S])/H[T]),  
expression(F[ST] == frac(var(p),bar(p)~ bar(q))), box.lwd = 0,box.col = "white",bg = "white")
```



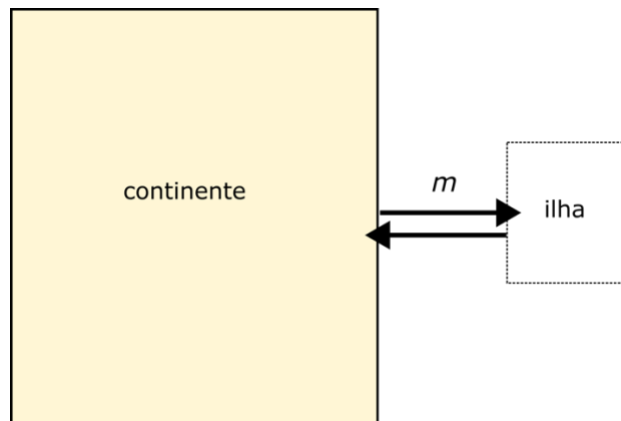
*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)



## Consequências do fluxo gênico nas frequências alélicas e diversidade genética

Nesta prática analisaremos alguns modelos simples de fluxo gênico entre subpopulações (demes) para entender seu efeito nas frequências alélicas ao longo das gerações. Em princípio, é relevante mencionar que existem diversos modelos de fluxo gênico entre subunidades populacionais. Todos são claramente simplificações do fluxo real de **indivíduos reprodutores** entre as diversas demes. Obviamente, os geneticistas de populações estão interessados em calcular o fluxo migratório de indivíduos que de fato **contribuíram para o pool gamético**.

Iniciaremos nossa análise com o caso mais simples possível: o **modelo ilha-continente**. Neste modelo, existe uma deme relativamente isolada em uma ilha. Também existe uma deme significativamente maior habitando o continente. Esta deme continental é tão maior que aquela da ilha que se assume que os migrantes oriundos da ilha **não afetam as frequências alélicas da deme continental**. Assim, vamos admitir que exista um fluxo gênico de migração  $m$  entre ilha e continente. Inicialmente, na geração 0, a frequência do alelo  $A$  ( $f_A = p$ ) na ilha é  $p_0$ . No continente, vamos assumir que a frequência do alelo  $A$  será  $p_{\text{continente}}$ :



Na próxima geração, a frequência do alelo  $A$  na ilha será:

$$p_1 = p_0(1 - m) + p_{\text{continente}}m$$

Ou seja, a frequência do alelo  $A$  na geração seguinte será igual à frequência na geração anterior ( $p_0$ ) multiplicada (**e**) pela taxa de não-migração ( $1 - m$ ) somada ao ganho de alelos  $A$  do continente ( $p_{\text{continente}}m$ ).

Perceba que, com esta formulação, o parâmetro que mede o fluxo gênico ( $m$ ) tem um significado não muito simples. Matematicamente, ele é **a parcela da frequência do alelo  $A$  nas demes que foi modificada por migração**. Ele **não é igual** ao número de indivíduos migrantes. Alternativamente, podemos afirmar que  $m$  é a **chance de receber um alelo de uma outra deme**. Neste modelo simples, a parcela da frequência alélica da ilha cedida ao continente ( $p_0m$ ) é ignorada, pois a população do continente não é afetada pela migração de reprodutores ilha  $\rightarrow$  continente.

A dinâmica da frequência do alelo  $A$  na ilha ao longo das gerações pode ser facilmente verificada no R:

```
## Modelo continente-ilha:

f_cont = 0.7 # freq do alelo A no continente
f_ilha = vector()
f_ilha[1] = 0.05 # freq do alelo A na ilha

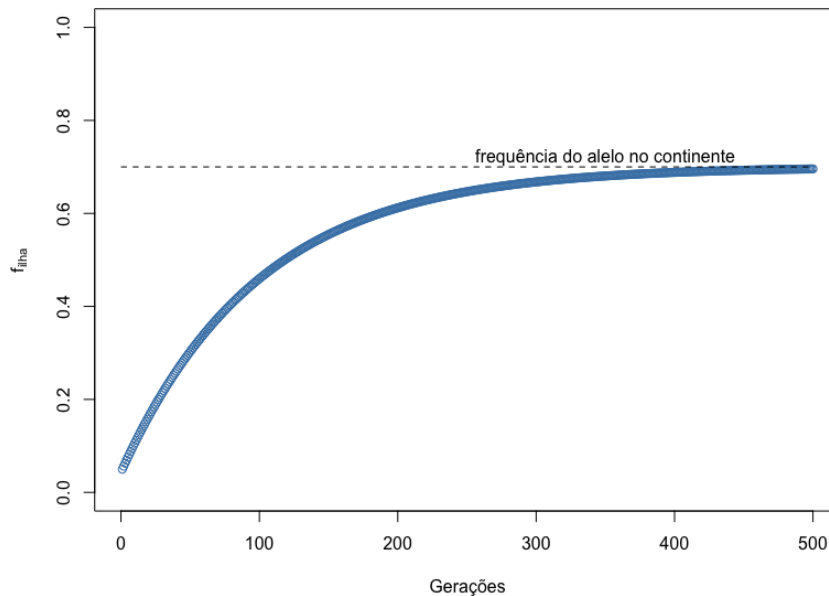
g = 500
m = 0.01 # taxa de fluxo gênico

for(i in c(2:g)){

  f_ilha[i] = f_ilha[i-1] * (1-m) + f_cont*m

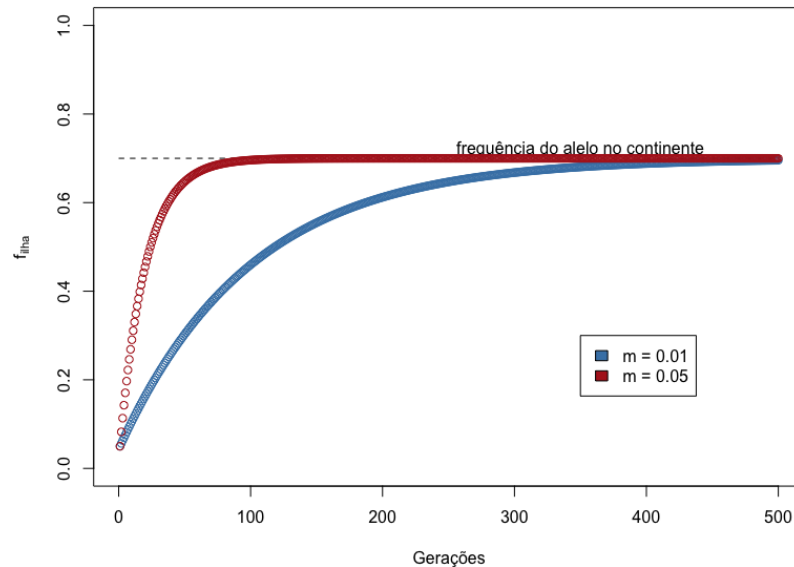
}

plot(f_ilha, col="steelblue", ylab=expression(f[ilha]), xlab="Gerações", ylim=c(0,1))
segments(0,f_cont,g,f_cont, lty=2 )
text(350,0.72,"frequência do alelo no continente")
```

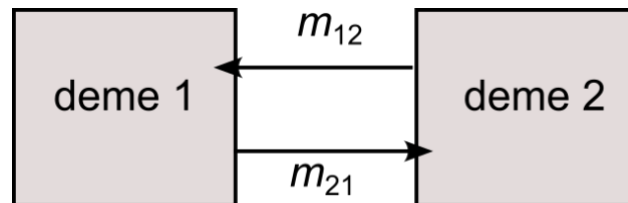


É claro que a frequência do alelo *A* na ilha, que inicialmente é de apenas 5%, aumenta até ficar equivalente à frequência da deme continental. Neste exemplo, a chance de receber um alelo do continente foi  $m = 0.01$ . Evidentemente, quanto maior for o fluxo gênico, mais rapidamente a frequência da ilha ficará igual a do continente:

```
m = 0.05 # taxa de fluxo gênico
```



O modelo continente-ilha pode ser expandido para um modelo ilha-ilha, onde a população as frequências de ambas as demes são modificadas, segundo o seguinte esquema:



A quantificação desse esquema é realizada através da utilização de uma **matriz de taxas de migração** definida por:

$$M = \begin{bmatrix} (1 - m_{1 \leftarrow 2}) & m_{1 \leftarrow 2} \\ m_{2 \leftarrow 1} & (1 - m_{2 \leftarrow 1}) \end{bmatrix}$$

Os termos  $m_{i \leftarrow j}$  da matrix M indicam a probabilidade de um indivíduo nascido na ilha  $i$  (linhas) ter pais oriundos da ilha  $j$  (colunas). Desta forma, a direção da migração ocorre de  $j \rightarrow i$ . Note que **as linhas da matriz devem somar 1**, pois um os indivíduos serão filhos de migrantes ou não. Após uma geração de migração, as frequências alélicas serão dadas por:

$$p_i^{(t+1)} = \sum_{j=1}^k p_j^{(t)} m_{i \leftarrow j}$$

No exemplo da figura acima, existem  $k = 2$  demes. Assim, a frequência alélica na deme 1 após uma geração de migração será:

$$p_1^{(t+1)} = p_1^{(t)}(1 - m_{1 \leftarrow 2}) + p_2^{(t)} m_{1 \leftarrow 2}$$



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

Cuja interpretação é a seguinte: a frequência do alelo *A* na deme 1 na geração (*t*+1) é igual à frequência deste alelo na geração anterior nesta mesma deme  $p_1^{(t)}$  multiplicada pela probabilidade de um indivíduo em (*t* + 1) ter pais oriundos da própria deme ( $1 - m_{1\leftarrow 2}$ ) somado à probabilidade de um indivíduo ter pais oriundos da deme 2 ( $m_{1\leftarrow 2}$ ) multiplicado pela frequência de *A* na deme 2 ( $p_2^{(t)}$ ) na geração anterior. Em outras palavras, **um indivíduo nascido na deme 1 tem um pai oriundo desta mesma deme,  $p_1^{(t)}(1 - m_{1\leftarrow 2})$ , ou (+) da deme 2:  $p_2^{(t)} m_{1\leftarrow 2}$ .**

O seguinte código aplica essas equações em R:

```
## Modelo ilha-ilha:

m12 = 0.1 # probabilidade de um indivíduo em 1 ter pais oriundos de 2
m21 = 0.1 # probabilidade de um indivíduo em 2 ter pais oriundos de 1

g = 200

M = matrix(nrow=2,ncol=2) # matriz M de taxas de migração
p = matrix(nrow=g,ncol=2) # frequência do alelo A

M[1,] = c((1-m12), m12) # preenchimento da primeira linha da matriz M
M[2,] = c(m21, (1-m21)) # preenchimento da segunda linha da matriz M

p[1,] = c(0.9,0.2) # frequências iniciais do alelo A na deme 1 e na deme 2

for(i in c(2:g)){

  p[i,1] = p[i-1,1] * M[1,1] + p[i-1,2] * M[1,2]
  p[i,2] = p[i-1,1] * M[2,1] + p[i-1,2] * M[2,2]

}

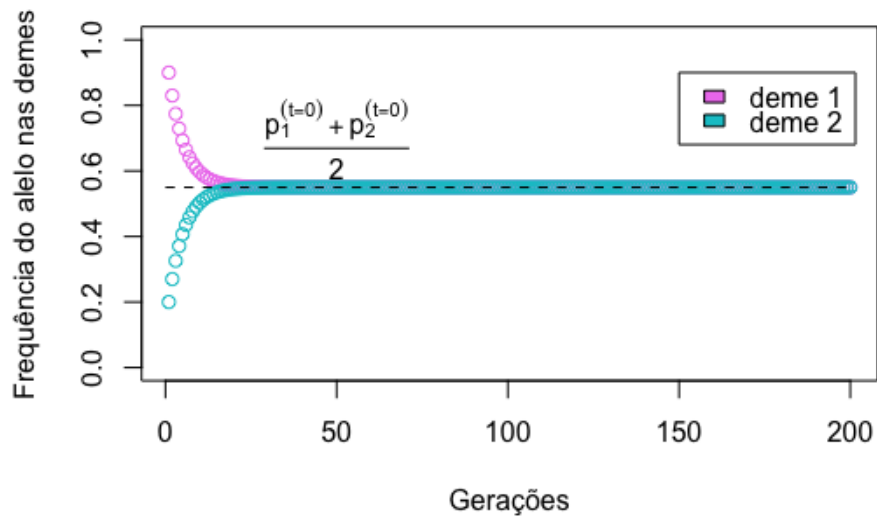
# Plotar o gráfico:
plot(p[,1], ylim=c(0,1), col="violet", ylab="Frequência do alelo nas demes",
xlab="Gerações")

points(p[,2], col="turquoise3")

legend(150,0.9,legend=c("deme 1", "deme 2"), fill=c("violet", "turquoise3"))

segments(0,mean(p[1,]), 200,mean(p[1,]), lty=2)

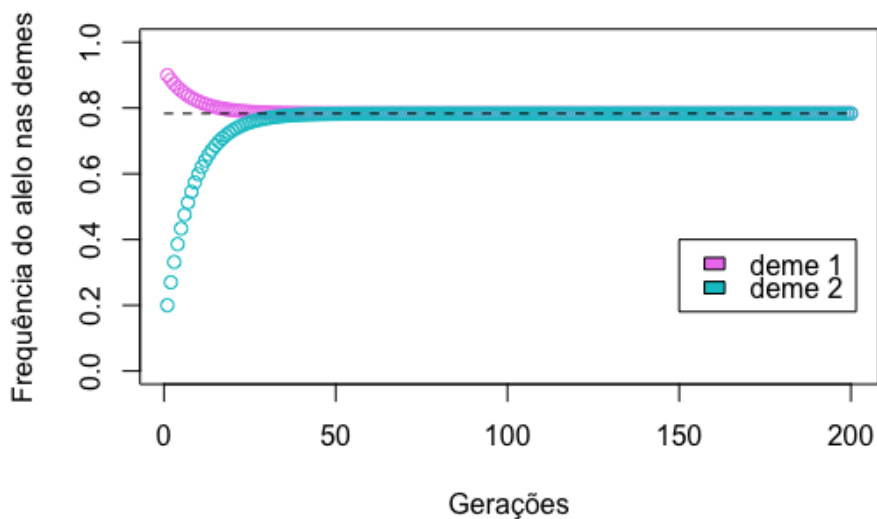
text(50,0.7, expression(frac(p[1]^(t==0)+p[2]^(t==0),2 ) ) )
```



Podemos observar que a frequência do alelo *A* nas duas ilhas converge para um único valor comum, mostrando o efeito **de homogeneização da diversidade genética** que o fluxo gênico possui. Este valor é igual à **média entre as frequências alélicas iniciais** ( $p_1^{(t=0)}$  e  $p_2^{(t=0)}$ ) em cada uma das ilhas (demes). No nosso caso, esse valor é igual a  $(0.9 + 0.2)/2 = 0.55$ .

Podemos agora verificar a dinâmica das frequências alélicas nas demes quando a taxa de migração entre elas é **assimétrica**. Por exemplo:

```
m12 = 0.02 # probabilidade de um indivíduo em 1 ter pais oriundos de 2
m21 = 0.1  # probabilidade de um indivíduo em 2 ter pais oriundos de 1
```





Verificamos que também existe uma frequência de equilíbrio. Entretanto, como a deme 2 recebe uma frequência maior de migrantes que a deme 1, a frequência alélica de equilíbrio não é igual à média aritmética das frequências iniciais. Esta pode ser encontrada pela média ponderada:

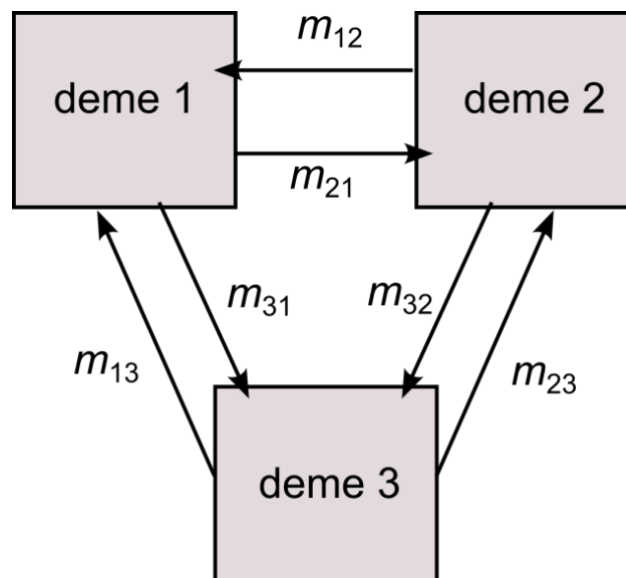
$$\bar{p} = \frac{p_1^{(t=0)} m_{2 \leftarrow 1} + p_2^{(t=0)} m_{1 \leftarrow 2}}{m_{1 \leftarrow 2} + m_{2 \leftarrow 1}}$$

No exemplo dado, temos:

$$\bar{p} = \frac{(0.9)(0.1) + (0.2)(0.02)}{0.02 + 0.1} = 0.7833333$$

Note que essa generalização do modelo de ilhas permite que estudemos a dinâmica de qualquer cenário de taxa de migração entre demes, devemos assumir, no entanto, que **as taxas de migração são constantes entre das gerações**.

Por exemplo, estudemos um caso mais complexo com três demes com taxas de mutação assimétricas, incluindo casos em que não há fluxo gênico ( $m = 0$ ):



A matriz de taxas de migração será:

$$M = \begin{bmatrix} (1 - (m_{1 \leftarrow 2} + m_{1 \leftarrow 3})) & m_{1 \leftarrow 2} & m_{1 \leftarrow 3} \\ m_{2 \leftarrow 1} & (1 - (m_{2 \leftarrow 1} + m_{2 \leftarrow 3})) & m_{2 \leftarrow 3} \\ m_{3 \leftarrow 1} & m_{3 \leftarrow 2} & (1 - (m_{3 \leftarrow 1} + m_{3 \leftarrow 2})) \end{bmatrix}$$

Segue o código em R:

```
# Taxas de migração:
m12 = 0.02
m13 = 0 # não há fluxo gênico da deme 3 para a deme 1
m21 = 0.08
m23 = 0.1
m31 = 0.01
```

```

m32 = 0 # não há fluxo gênico da deme 2 para a deme 3

g = 500

M = matrix(nrow=3,ncol=3)
p = matrix(nrow=g,ncol=3)

M[1,] = c( (1 - (m12+m13)), m12, m13)
M[2,] = c(m21, (1 - (m21+m23)), m23 )
M[3,] = c(m31, m32, (1 - (m31+m32)) )

p[1,] = c(0.95,0.4,0.01)

for(i in c(2:g)){

  p[i,1] = p[i-1,1] * M[1,1] + p[i-1,2] * M[1,2] + p[i-1,3] * M[1,3]
  p[i,2] = p[i-1,1] * M[2,1] + p[i-1,2] * M[2,2] + p[i-1,3] * M[2,3]
  p[i,3] = p[i-1,1] * M[3,1] + p[i-1,2] * M[3,2] + p[i-1,3] * M[3,3]

}

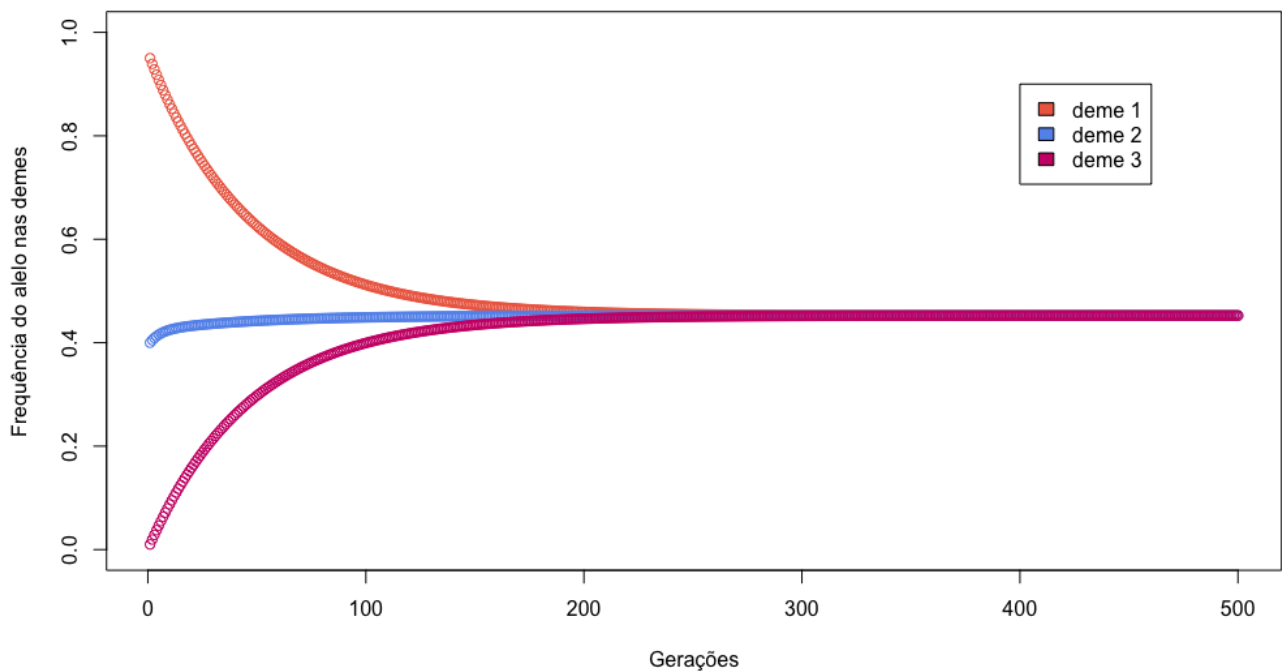
# Plotar o gráfico:
plot(p[,1], ylim=c(0,1), col="coral2", ylab="Frequência do alelo nas demes", xlab="Gerações")

points(p[,2], col="cornflowerblue")

points(p[,3], col="deeppink3")

legend(400,0.9,legend=c("deme 1", "deme 2", "deme3"), fill=c("coral2", "cornflowerblue","deeppink3" ))

```



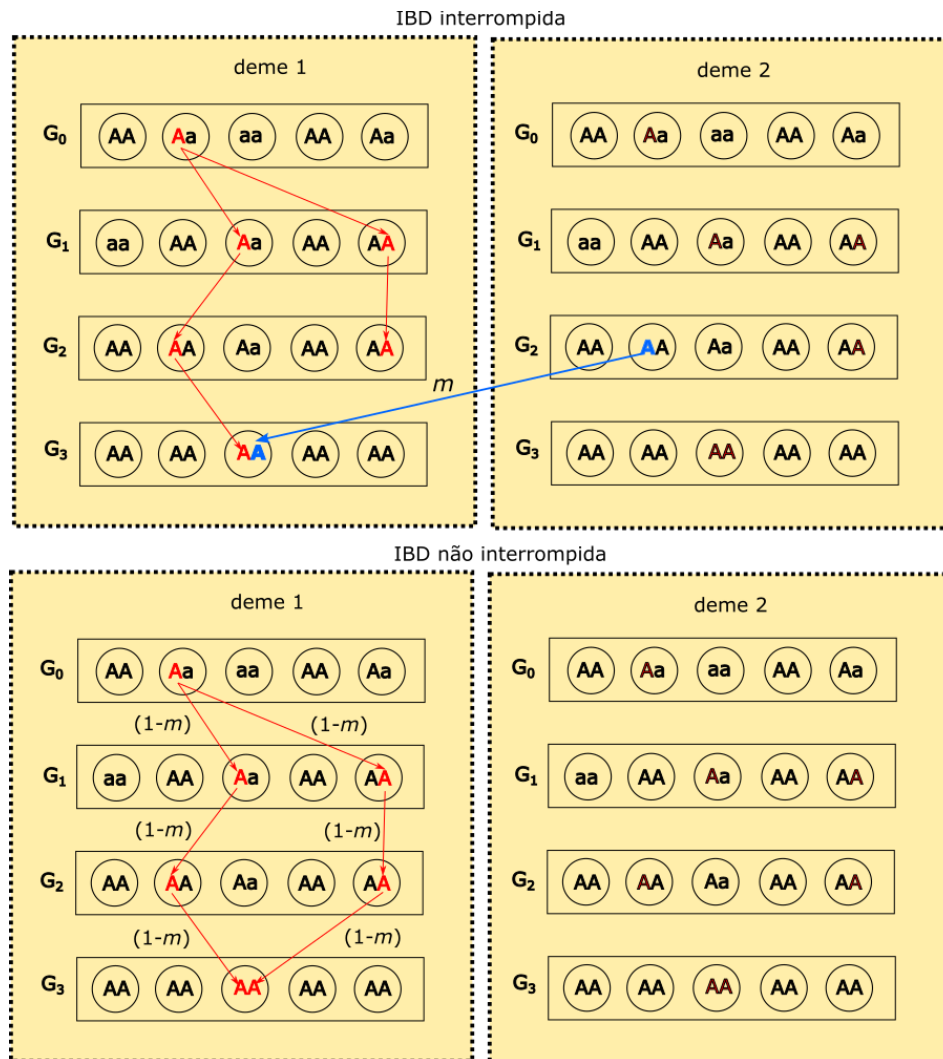
As frequências alélicas das três demes estabilizam em 0.4525. Neste momento, não há mais variação de frequências alélicas entre as demes.

Tente generalizar o código acima para um número qualquer de demes.

Na primeira metade do século 20, Sewall Wright utilizou uma generalização do modelo de ilhas para estimar o **efeito da migração na deriva genética**. A generalização de Wright é muito similar ao

modelo de Markov utilizado anteriormente para estudar a variação da frequência alélica ao longo das gerações por deriva. Wright imaginou um **número infinito de demes com o mesmo tamanho populacional  $N$** . Essas demes trocam migrantes a cada geração com **uma mesma taxa  $m$**  (modelo simétrico).

Ao estabelecer esta generalização, é possível calcular a frequência de ocorrência de homozigotos IBD **da mesma forma que fizemos para o modelo de mutação-deriva**. Um homozigoto IBD apenas será formado se suas cópias alélicas **coalescerem na mesma deme**. A cópia alélica, portanto, não pode vir de uma outra deme. A chance de a cópia vir de outra deme é, justamente, a taxa de migração  $m$ . Verifique como o esquema se assemelha à modelagem da taxa de mutação:



Seguindo esse raciocínio, conseqüentemente podemos definir que a probabilidade de encontrar um homozigoto IBD numa deme na geração  $t$  será:

$$F_t = \frac{1}{2N} (1 - m)^2 \left(1 - \frac{1}{2N}\right) (1 - m)^2 F_{t-1}$$



Na condição de equilíbrio (ver a derivação usada anteriormente no estudo sobre mutação-deriva), teremos que a homozigosidade IBD será **aproximadamente**:

$$F = \frac{1}{4Nm + 1}$$

Esse valor será **numericamente equivalente ao índice de fixação  $F_{ST}$** .

Como a formulação da homozigosidade IBD na interação entre migração e deriva é idêntica à formulação feita na interação entre mutação e deriva, podemos afirmar que, matematicamente, **o efeito da migração é equivalente ao efeito da mutação na diversidade genética (heterozigosidade)**.

## Calculando as probabilidades IBD em 1 geração de duas cópias amostradas da população

Inicialmente avaliaremos um caso trivial, o lançamento de moedas, para confirmar que o **conceito de probabilidade está intimamente associado à frequência de ocorrência de um dado evento**. Suponhamos que possuímos uma moeda que não é viciada, ou seja, é justa. Desta forma, as probabilidades de tirar cara ( $p_K$ ) ou coroa ( $p_C$ ) são iguais (= 50%). Consideremos agora que nós **não sabemos** se essa moeda é justa. Como poderíamos obter os valores de  $p_K$  e  $p_C$ ?

A estratégia que parece óbvia é **jogarmos a moeda um número muito grande de vezes** e contarmos as frequências de caras (K) e coroas (C). Esses valores serão estimativas de  $p_K$  e  $p_C$ , que são representados com um acento circunflexo (“^”) em estatística –  $\widehat{p}_K$  e  $\widehat{p}_C$ .

Façamos isso no R:

```
## Lançamento de uma moeda

pK = 0.5 ## não sabemos esse valor!
pC = 0.5 ## não sabemos esse valor!

tentativas = 100000 # lançaremos a moeda esse número de vezes

nK = 0 # número de vezes que caiu cara (inicia com 0)

for(i in c(1:tentativas)){
  x = sample(c("K","C"), 1, prob=c(pK,pC)) # joga a moeda 1 vez e põe o resultado em x
  if(x == "K"){ # caiu cara!
    nK = nK + 1 # soma ao número de caras
  }
}

# Frequência (=probabilidade) de caras:
nK/tentativas
```

No meu experimento, eu obtive que frequência de ocorrência de caras ( $K$ ) nas  $tentativas = 100000$  lançamentos foi 0.50242. Desta forma,  $\widehat{p}_K = 0.50424$ . Esse valor é muito próximo do valor verdadeiro do parâmetro  $p_K = 0.5$  (50%). Isso confirma que a moeda é justa.

Aumentemos o número de lançamentos ( $tentativas$ ) para 1000000.

tentativas = 100000 # lançaremos a moeda esse número de vezes

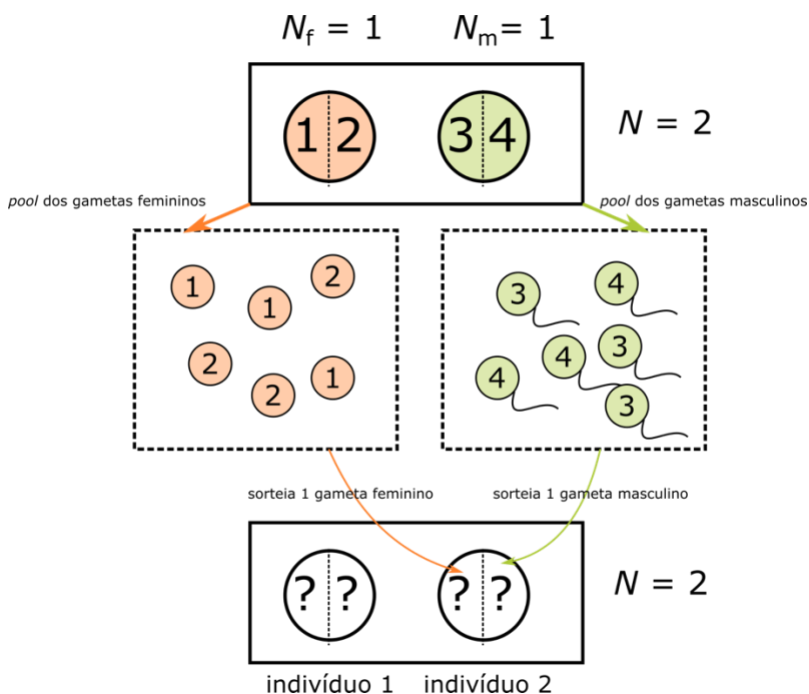
Agora, no meu experimento, obtive o valor de probabilidade de cara igual a 0.500807. Fica evidente que **quanto maior o número de tentativas (replicatas), mais próximo do valor real de  $p_K$  a minha estimativa  $\hat{p}_K$  será.**

Passemos agora para o problema do cálculo das probabilidades IBD:

Examinaremos o caso mais simples. Uma população de apenas 1 fêmea e 1 macho ( $N = 2$ ). A fêmea será heterozigota e possuirá no locus examinado alelos chamados de 1 e 2. O macho também é heterozigoto e, neste mesmo locus, possui alelos chamados de 3 e 4.

O *pool* gamético das fêmeas, portanto, será composto por óvulos contendo o alelo 1 ou o alelo 2 em igual frequência (não há desvio meiótico). Da mesma forma, o *pool* gamético dos machos será composto de espermatozoides com o alelo 3 ou o alelo 4 em igual frequência.

Na próxima geração, devemos compor os genótipos de  $N = 2$  indivíduos. Cada indivíduo será composto pelo sorteio de um óvulo e um espermatozoide.

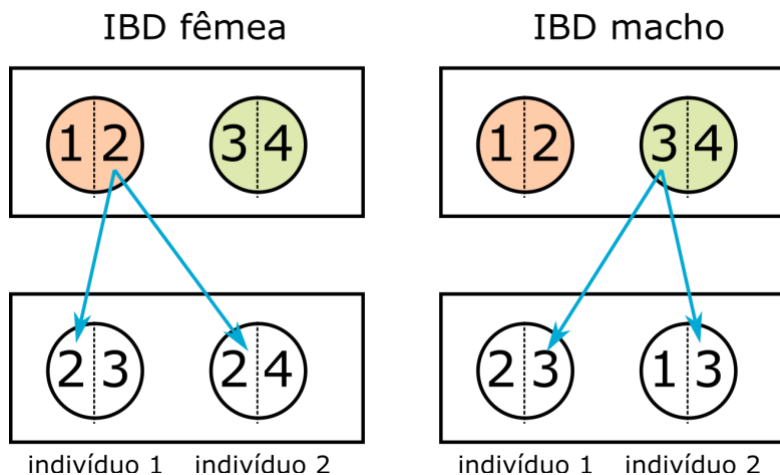


Após compor os genótipos. Podemos verificar se ocorreu identidade por descendência (*identity by descent*, IBD) nesta população.

Para tal, queremos calcular qual seria a **chance de obter alelos IBD ao sortear dois indivíduos e depois sortear uma cópia alélica de cada**. Isso é equivalente às seguintes grandezas:

- A probabilidade de se coletar duas cópias alélicas IBD nesta população ao acaso, ou
- A probabilidade de duas cópias coletadas ao acaso coalescerem na mesma cópia na geração imediatamente anterior

Inclusive, podemos saber se essa identidade por descendência ocorreu na população de fêmeas ou de machos:



Estamos interessados em saber **qual é a chance de ocorrência de cada um desses eventos** (IBDf e IBDm). Obviamente, se fizermos apenas 1 tentativa de composição dos  $N = 2$  indivíduos a partir dos *pools* feminino e masculino não teremos uma estimativa desse valor. Devemos, portanto, repetir esse processo um número grande de vezes e contar a ocorrência de cada um desses eventos para conseguir calcular a sua frequência.

Segue o código R que faz isso. Faremos 1000000 procedimentos independentes:

```
Individuo1 = vector()
Individuo2 = vector()
IBD = IBDf = IBDm = 0
tentativas = 1000000

for(i in c(1:tentativas)){

  Individuo1[1] = sample(c(1,2),1) # sorteia um gameta feminino do pool
  Individuo1[2] = sample(c(3,4),1) # sorteia um gameta masculino do pool

  Individuo2[1] = sample(c(1,2),1) # sorteia um gameta feminino do pool
  Individuo2[2] = sample(c(3,4),1) # sorteia um gameta masculino do pool

  s1 = sample(Individuo1, 1) # sorteia uma cópia do individuo 1
  s2 = sample(Individuo2, 1) # sorteia uma cópia do individuo 2

  if(s1 == s2){ # alelos são IBD (macho OU fêmea)
    IBD = IBD + 1
    if(s1 == 1 | s1 == 2){ # são IBD nas fêmeas
      IBDf = IBDf + 1
    }
    else{ # IBD nos machos
      IBDm = IBDm + 1
    }
  }
}

IBD/tentativas # probabilidade IBD macho OU fêmea
IBDf/tentativas # probabilidade IBD fêmea
IBDm/tentativas # probabilidade macho
```



No meu experimento, obtive os seguintes valores:

IBD = 0.249493  
 IBDf = 0.124944  
 IBDm = 0.124549

Lembremos que a probabilidade teórica do sorteio de 2 cópias IBD (independente de ser vir de um macho **ou** de uma fêmea) é  $1/2N$ , ou seja,  $1/2(2) = 0.25$ . No meu experimento, das *tentativas* = 1000000 vezes que o processo foi realizado, em 0.249493 delas foram sorteadas cópias IBD. É uma aproximação muito boa.

No caso do IBD das fêmeas, a expectativa teórica é que o valor seja  $1/8N_f$ , ou seja,  $1/8(1) = 0.125$ . O mesmo se aplica para os machos:  $1/8N_m = 0.125$ . Os valores obtidos foram, também, muito próximos de 0.125.

Abaixo, você encontrará um código geral para qualquer número de machos e fêmeas. Lembre-se que neste caso o IBD total deverá ser igual a  $\frac{4N_f N_m}{N_f + N_m}$ .

```
Nf = 1 #numero de femeas
Nm = 10 #numero de machos
N = Nf + Nm #numero total

alelof = c(1:(2*Nf)) #série de alelos das femeas
alelosm = c( ((2*Nf)+1) : ( (2*Nf)+(2*Nm) ) ) #série de alelos dos machos

tentativas = 1000000

IBD = IBDf = IBDm = 0

for(i in c(1:tentativas)){

  popf = popm = vector()
  pop = matrix(nrow=N, ncol=2)

  #amostra N alelos do pool de gametas femininos:
  popf = sample(alelof, N, replace=T)
  #amostra N alelos do pool de gametas masculinos:
  popm = sample(alelosm, N, replace=T)

  #composicao da população de N individuos com alelo F e alelo M
  for(j in c(1:N) ){
    pop[j,1] = popf[j]
    pop[j,2] = popm[j]
  }

  x = sample(c(1:N),2) # sorteia 2 individuos da população
  y1 = sample(c(1,2),1) # sorteia um dos alelos do primeiro individuo
  y2 = sample(c(1,2),1) # sorteia um dos alelos do segundo individuo

  if(pop[x[1], y1] == pop[x[2], y2]){ # verifica se os alelos são iguais (IBD)
    IBD = IBD + 1
    if((pop[x[1], y1] %in% popf) == TRUE){ #verfica se veio de uma femea (IBD femea)
      IBDf = IBDf + 1
    }
  }
}
```



```
}  
else{ #caso contrário, veio de um macho (IBD macho)  
  IBDm = IBDm + 1  
}  
}  
  
IBD/tentativas # probabilidade IBD macho OU fêmea  
IBDf/tentativas # probabilidade IBD fêmea  
IBDm/tentativas # probabilidade macho
```





## A ideia de seleção natural

Seleção natural é um fenômeno descrito no século 19 por Charles Darwin e Alfred R. Wallace. Entende-se que a seleção seja um fenômeno majoritariamente populacional, ocorrendo entre indivíduos de uma mesma espécie.

Tradicionalmente, a existência das seguintes propriedades resulta em modificação associada à seleção natural:

- Existe variação entre os indivíduos.
- Essa variação é herdada.
- Essa variação afeta o sucesso dos indivíduos em acessar recursos e a reprodução.
- Os recursos do ambiente são finitos.
- A cada geração, nascem mais indivíduos do que a capacidade suporte do ambiente.

Do ponto de vista biológico, há evidências empíricas de que todas as propriedades acima são verdadeiras. Desta forma, a ocorrência da seleção natural é uma conclusão lógica.

Embora não existam grandes debates sobre a ocorrência de seleção nas populações naturais, o impacto que este processo tem/teve na diversificação dos seres vivos como um todo é alvo de intensa discussão desde o século 19.

Comumente, a ideia de seleção natural é apresentada usando como exemplo o crescimento de dois genótipos de um organismo com reprodução assexuada (a.g., *A* e *B*). Existe **sucesso reprodutivo diferencial** entre os genótipos, de forma que a taxa de crescimento das populações de cada genótipo é diferente. Para cada um dos genótipos, o número de indivíduos na próxima geração é dado por uma simples equação recursiva:

$$N_A^{(t+1)} = N_A^{(t)} R_A$$

e

$$N_B^{(t+1)} = N_B^{(t)} R_B$$

Ou seja, o número de indivíduos em  $(t+1)$  é igual ao número da geração anterior  $(t)$  multiplicado por  $R$ , o **número médio de descendentes que cada indivíduo deixa**. Evidentemente, se  $R = 1$ , o tamanho da população ficará constante. Vejamos como se comportam as curvas de crescimento de cada genótipo quando  $R_A = 1.03$  e  $R_B = 1.01$ :

```
# exponencial
Ra = 1.03
Rb = 1.01

Na = vector()
Nb = vector()

Na[1] = 1
Nb[1] = 1

g = 100

for(t in c(2:g)){
  Na[t] = Na[t-1] * la
  Nb[t] = Nb[t-1] * lb
}

# gráfico:
```

*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.

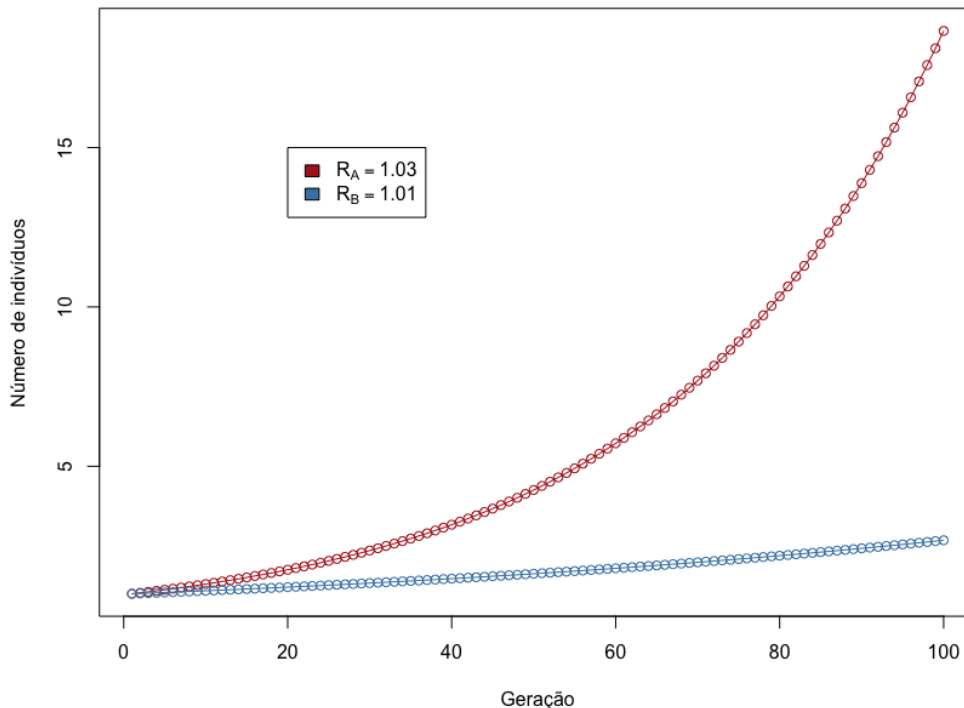
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

```
plot(Na, col="firebrick", ylab="Número de indivíduos", xlab="Geração" )
lines(Na, col="firebrick")

points(Nb,col="steelblue")

lines(Nb, col="steelblue")

legend(20,15,legend=c(expression(N[A] == 1.03),expression(N[B] == 1.01) ),
fill=c("firebrick","steelblue"))
```



Ao longo das gerações, a frequência relativa do genótipo *A* aumenta exponencialmente. Essa propriedade é usada para definir a **fitness relativa (*w*)** do genótipo *A* em relação ao *B*. Desta forma  $w = R_A/R_B$  ou do genótipo *B* em relação ao *A*,  $w = R_B/R_A$ .

Neste modelo simples, as fitness relativas se mantêm constantes ao longo do tempo. Além disso, embora a frequência relativa do genótipo *B* seja cada vez menor quando comparada ao *A*, **essas equações não preveem que o genótipo *B* será eliminado.**

Não existe, neste modelo exponencial, um limite para o número de indivíduos que o ambiente suporta. Efetivamente, **não existe seleção natural**, pois quantidade de recursos do ambiente é infinita. Precisamos ajustar o nosso modelo para contemplar essa outra propriedade necessária para ocorrência de seleção.

Uma alternativa famosa ao crescimento exponencial é o **crescimento logístico**. Neste modelo, existe explicitamente um número máximo de indivíduos que o ambiente suporta. Esse valor é controlado pelo parâmetro de **capacidade suporte *K***.

No modelo logístico, o número médio de descendentes por indivíduo é dado pela equação:



Licença de utilização: Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
 Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

$$R_A = 1 + r_A \left( 1 - \frac{N^{(t-1)}}{K} \right)$$

e

$$R_B = 1 + r_B \left( 1 - \frac{N^{(t-1)}}{K} \right)$$

Onde  $N = N_A + N_B$ , o tamanho total da população.

Assim, diferentemente do modelo exponencial simples, no modelo logístico o número de descendentes por indivíduo **muda ao longo das gerações** conforme as equações acima. O parâmetro  $r$  controla a taxa intrínseca dessa mudança. Se  $r > 0$ , as populações tendem a aumentar de tamanho.

Suponhamos que  $r_A = 0.3$  e  $r_B = 0.1$  e que a capacidade suporte do ambiente seja de  $K = 150$  indivíduos no total.

```
# logístico:
Na = vector()
Nb = vector()

Ra = vector()
Rb = vector()

K = 150

Na[1] = 1
Nb[1] = 1

ra = 0.3
rb = 0.1

g = 100

for(t in c(2:g)){

  N = Na[t-1] + Nb[t-1]

  Ra[t] = (1 + ra * (1 - N/K) )
  Rb[t] = (1 + rb * (1 - N/K) )

  Na[t] = Ra[t] * Na[t-1]
  Nb[t] = Rb[t] * Nb[t-1]

}

# gráfico:
plot(Na, col="firebrick", ylab="Número de indivíduos", xlab="Geração",ylim=c(0,151) )

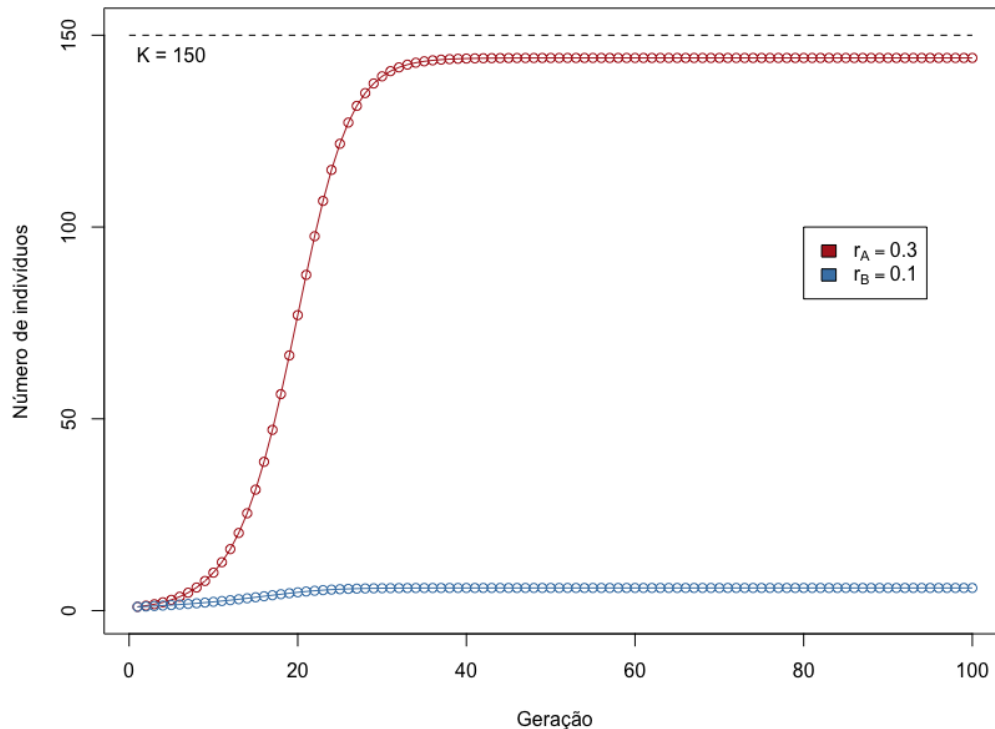
lines(Na, col="firebrick")

points(Nb,col="steelblue")

lines(Nb, col="steelblue") segments(0,150,100,150,lty=2)

text(5,145,"K = 150")

legend(80,100,legend=c(expression(r[A] == 0.3),expression(r[B] == 0.1) ),
fill=c("firebrick","steelblue"))
```



No modelo logístico, as taxas de reprodução de ambos os genótipos são modificadas para que se ajustem à capacidade suporte do ambiente  $K = 150$ . Entretanto, como tanto  $r_A$  e  $r_B$  são  $> 0$ , as duas populações tendem a crescer. Quando a capacidade suporte é atingida, ainda existem indivíduos do genótipo azul. Ou seja, o azul também não foi eliminado. Matematicamente, o azul só será eliminado se  $r_B < 0$ .

Será que está acontecendo seleção neste cenário? A resposta é um sim parcial. Certamente, neste modelo, o número de indivíduos não é infinito. Mas por que o genótipo azul não foi eliminado se a taxa de reprodução dele é menor que a do genótipo vermelho?

Isso ocorreu porque o ambiente (através do parâmetro  $K$ ) modificou as taxas reprodutivas dos genótipos até que ambas se adequassem ao número limite de 150. Neste sentido, ocorreu seleção do ambiente no ajuste das taxas. Entretanto, **não houve nenhuma interação (competição etc) entre os genótipos azul e vermelho**. O genótipo vermelho foi selecionado para se adequar ao ambiente, mas seu crescimento nas gerações iniciais não competiu diretamente com o genótipo azul. Apenas adequou  $r_B$  indiretamente através do multiplicador  $\left(1 - \frac{N^{(t-1)}}{K}\right)$  que é **idêntico para ambos os genótipos**.

A proposição de um modelo mais realista de crescimento populacional foi feita na década de 1920 por Alfred Lotka e Vito Volterra. No **modelo Lotka-Volterra**, o número de descendentes de cada genótipo também será ajustado, entretanto, agora existe um fator de competição  $C$  assimétrico que também irá afetar o número de descendentes, definido por:

$$R_A = 1 + r_A \left( 1 - \frac{N_A^{(t-1)} + C_{A \leftarrow B} N_B^{(t-1)}}{K} \right)$$



E

$$R_B = 1 + r_B \left( 1 - \frac{C_{B \leftarrow A} N_A^{(t-1)} + N_B^{(t-1)}}{K} \right)$$

O novo parâmetro  $C$  determina o efeito que cada indivíduo do genótipo  $B$  tem no  $A$  ( $C_{A \leftarrow B}$ ), assim como cada indivíduo  $A$  tem em  $B$  ( $C_{B \leftarrow A}$ ). Portanto,  $C_{A \leftarrow B}$  é um multiplicador de  $N_B$ , e  $C_{B \leftarrow A}$  é um multiplicador de  $N_A$ .

Suponhamos que o efeito de  $B$  em  $A$  seja menor que o de  $A$  em  $B$ . Como exemplo, façamos  $C_{A \leftarrow B} = 1$  e  $C_{B \leftarrow A} = 2$ . Vejamos a dinâmica dos tamanhos populacionais:

```
# Dinâmica Lotka-Volterra:
Na = vector()
Nb = vector()

Ra = vector()
Rb = vector()

K = 150

Na[1] = 1
Nb[1] = 1

ra = 0.3
rb = 0.1

g = 150

# efeito de competição:
Cab = 1 # efeito de B em A
Cba = 2 # efeito de A em B

for(t in c(2:g)){

  Ra[t] = (1 + ra * (1 - (Na[t-1] + Cab*Nb[t-1])/K) )
  Rb[t] = (1 + rb * (1 - (Cba*Na[t-1] + Nb[t-1])/K) )

  Na[t] = Ra[t] * Na[t-1]
  Nb[t] = Rb[t] * Nb[t-1]
}

# gráfico:
plot(Na, col="firebrick", ylab="Número de indivíduos", xlab="Geração",ylim=c(0,151) )

lines(Na, col="firebrick")

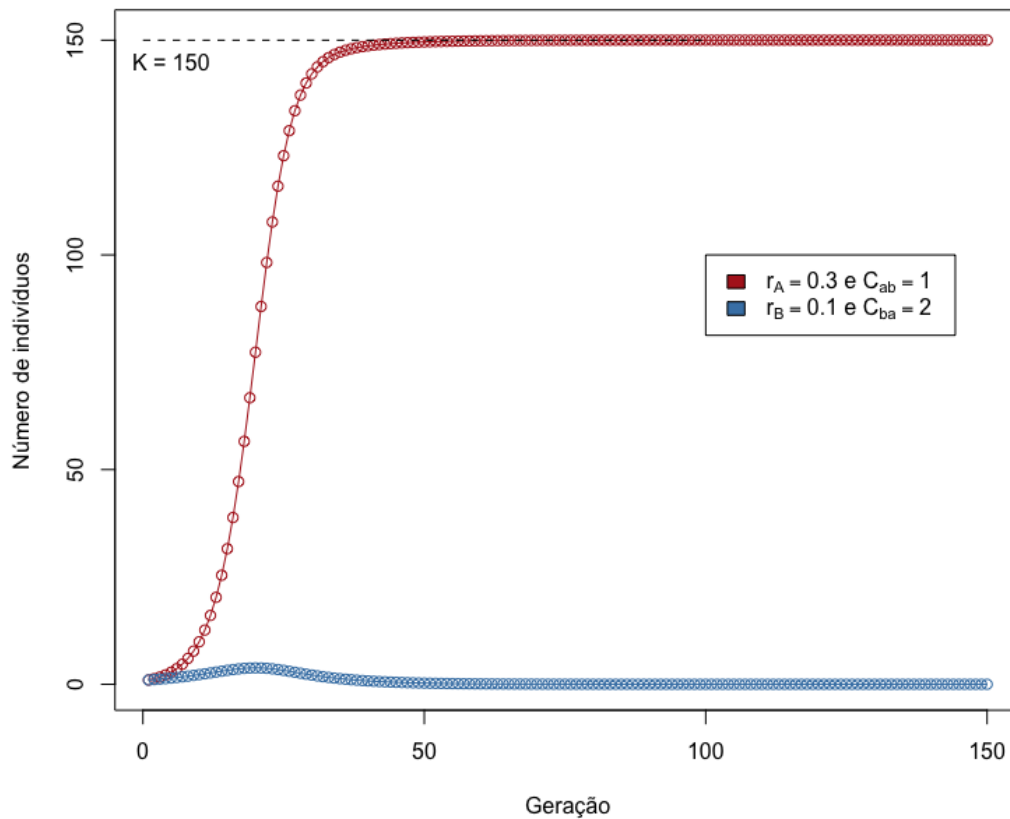
points(Nb,col="steelblue")

lines(Nb, col="steelblue")

segments(0,150,100,150,lty=2)

text(5,145,"K = 150")

legend(100,100,legend=c(expression(paste(r[A] == 0.3, " e ", C[ab] == 1)),expression(paste(r[B] == 0.1,
" e ", C[ba] == 2 ) ) ), fill=c("firebrick","steelblue"))
```



Nesta dinâmica de Lotka-Volterra, observamos que os indivíduos do genótipo *B* são efetivamente eliminados por competição com os indivíduos do genótipo *A*.

Conforme vimos, não é trivial modelar um simples cenário de seleção natural de uma forma em relação a outra. Além de sucesso reprodutivo diferencial e limitação do número de indivíduos que o ambiente suporta, tivemos que adotar um fator de interação entre os indivíduos que possuem genótipos diferentes (competição). Apenas com essas condições matemáticas (Lotka-Volterra) foi que um dos genótipos foi totalmente eliminado.

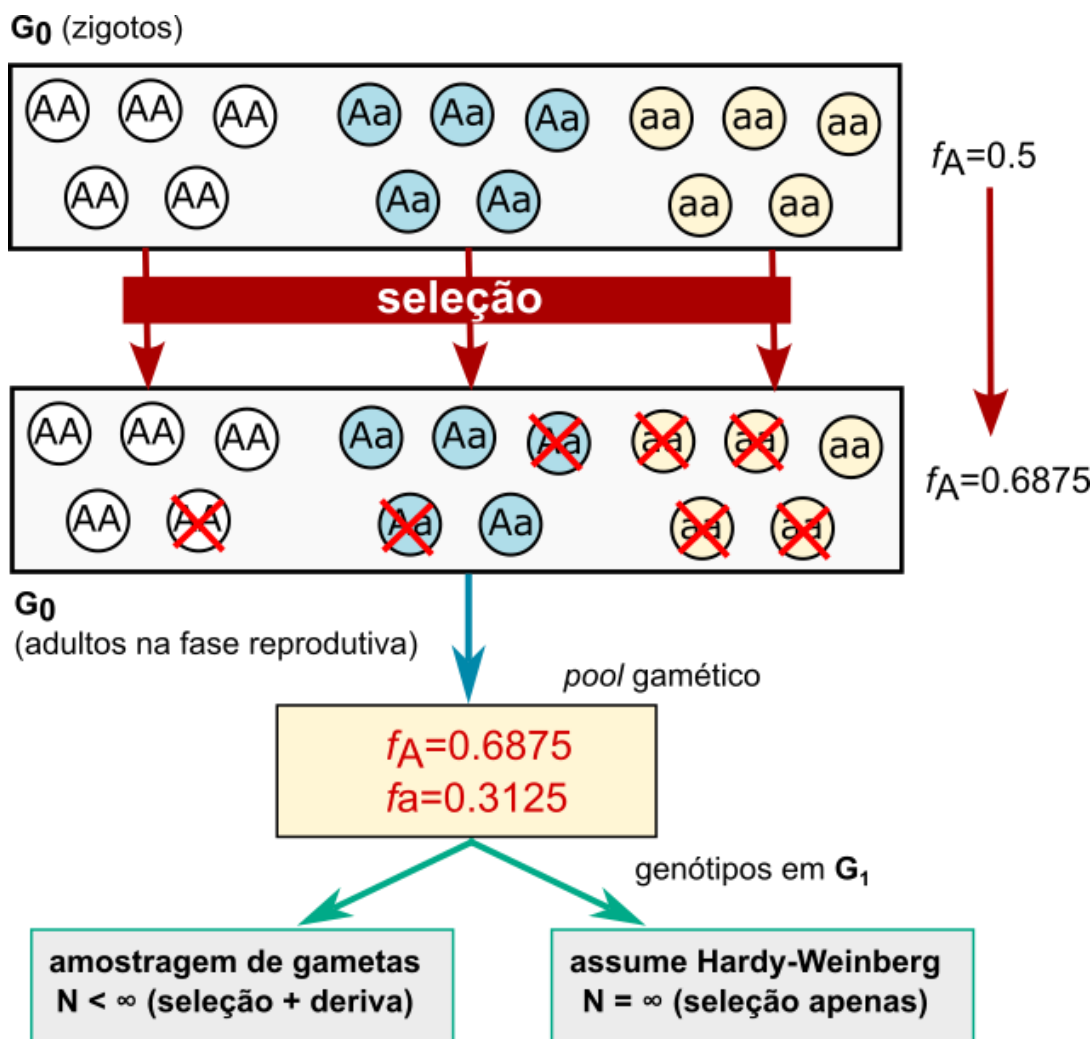
## Quantificação da ação da seleção natural nas populações

Durante o processo de desenvolvimento dos indivíduos até a fase reprodutiva podem ocorrer uma série de fatores que impedem que eles efetivamente participem do *pool* gamético que será usado para compor a geração seguinte. Se esse acesso diferencial ao *pool* gamético ocorrer devido a características associadas à expressão do genótipo dos indivíduos (características fenotípicas, comportamentais etc.), existirá **uma relação de desigualdade entre os diversos genótipos** na participação do *pool* gamético.

Note que esse processo **não ocorre por erro amostral do *pool***, a participação diferencial no conjunto de gametas **tem efetivamente uma razão biológica**. Por exemplo, um genótipo ser mais eficiente que outro na obtenção recursos alimentares ou de parceiros reprodutivos. Quando ocorre esse processo de **participação diferencial na reprodução**, dizemos que existe **seleção natural**.

A quantificação da ação da seleção natural é feita usando uma variável central: o **valor adaptativo relativo** dos genótipos, também chamada de **fitness relativo**, simbolizada pela letra **w**.

Para compreendermos o que é o valor adaptativo, vamos estudar o seguinte exemplo:





No exemplo acima, a população tem  $N = 15$  indivíduos na fase dos zigotos. Entretanto, ao longo do processo de desenvolvimento até a fase adulta reprodutiva, alguns morreram ou simplesmente não conseguiram participar do *pool* gamético que será usado para compor a próxima geração. Esses indivíduos estão marcados com uma cruz vermelha no esquema. Na composição do *pool* gamético, portanto, **a frequência dos alelos nos gametas será diferente da frequência dos alelos na população de zigotos.**

A partir do *pool* gamético existem duas possibilidades básicas de modelagem para compor os genótipos dos zigotos da geração seguinte: (i) **assumir o equilíbrio Hardy-Weinberg** (consequentemente admitindo não existir erro amostral do *pool*, ou seja,  $N = \infty$ ) ou (ii) **fazer amostragem do *pool* semelhante àquele realizada no modelo Wright-Fisher.** No caso *i*, verificaremos a ação da **seleção natural apenas.** Já no caso *ii* ocorrerá **também a presença de variação estocástica da frequência dos alelos associada à deriva.** O caso *ii*, portanto, será um modelo em que **seleção e deriva estarão atuando em conjunto.**

Para iniciarmos o estudo da mecânica da seleção, vamos então definir uma medida chamada **valor adaptativo absoluto.** Ela medirá a **eficiência absoluta de um genótipo em participar do *pool* de gametas.** Assim como o fitness relativo, esta medida também é simbolizada por **w.** O uso da mesma simbologia é inadequado, pois gera confusão.

Valores adaptativos absolutos:

$$w_{AA} = \frac{4}{5} = 0.8$$

$$w_{Aa} = \frac{3}{5} = 0.6$$

$$w_{aa} = \frac{1}{5} = 0.2$$

A interpretação dessas medidas é simples: 80% dos zigotos AA chegam à fase reprodutiva, assim como 60% dos Aa e 20% dos aa.

Vamos agora estabelecer uma **relação comparativa** entre os fitness dos três genótipos. Para facilitar a discussão, escolheremos como **referência** para a comparação, o genótipo com o maior fitness absoluto, o AA. O fitness absoluto desse genótipo servirá como uma unidade para quantificar todos os outros genótipos **de forma relativa.** Desta forma:

Valores adaptativos relativos:

$$w_{AA} = \frac{0.8}{0.8} = 1$$

$$w_{Aa} = \frac{0.6}{0.8} = 0.75$$

$$w_{aa} = \frac{0.2}{0.8} = 0.25$$

O significado dos valores relativos de fitness são: A **eficiência relativa** do genótipo Aa comparado ao AA é de 75%. A eficiência relativa do aa é 25%. Ou seja, para cada 1 indivíduo AA que consegue chegar na fase reprodutiva, 0.75 indivíduos Aa chegam e apenas 0.25 indivíduos aa chegam.

Os valores adaptativos relativos podem então ser usados para quantificar a **mecânica de atuação** da seleção natural.



Essa mecânica medirá a frequência dos genótipos **antes** e **após** seleção. Ela pode ser resumida no seguinte quadro:

|                               | AA                             | Aa                             | aa                             |
|-------------------------------|--------------------------------|--------------------------------|--------------------------------|
| Antes da seleção              | $f_{AA}$                       | $f_{Aa}$                       | $f_{aa}$                       |
| Valores adaptativos relativos | $w_{AA}$                       | $w_{Aa}$                       | $w_{aa}$                       |
| Após seleção                  | $\frac{f_{AA}w_{AA}}{\bar{w}}$ | $\frac{f_{Aa}w_{Aa}}{\bar{w}}$ | $\frac{f_{aa}w_{aa}}{\bar{w}}$ |

Desta forma, para calcular as frequências dos genótipos após a seleção, basta multiplicar a frequência do genótipo antes da seleção por seu valor adaptativo relativo. Entretanto, existe um termo novo na fórmula que mede a frequência após a seleção:  $\bar{w}$ .

Este valor é definido como:

$$\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa}$$

Ou seja, o somatório dos produtos dos três numeradores. Este valor serve para **normalizar as frequências genotípicas após a seleção entre os valores de 0 e 1**. Além disso, veremos que  $\bar{w}$  também é uma medida importante para compreender a dinâmica da atuação da seleção natural na população.

Façamos agora os cálculos baseados no cenário estabelecido no início da prática:  $w_{AA} = 1$ ,  $w_{Aa} = 0.75$  e  $w_{aa} = 0.25$ . As frequências iniciais de cada um dos genótipos são de 5 indivíduos em 15, ou seja, 33.3%. Usaremos o R para calcular os valores das frequências genotípicas após a seleção:

```
# Valores adaptativos relativos:
wAA = 1
wAa = 0.75
waa = 0.25

# Frequencias iniciais ANTES:
fAA = 5/15
fAa = 5/15
faa = 5/15

w_medio = (fAA*wAA) + (fAa*wAa) + (faa*waa)

# Frequências APÓS:

fAA = (fAA*wAA) / w_medio
fAa = (fAa*wAa) / w_medio
faa = (faa*waa) / w_medio
```

Após seleção as frequências genotípicas serão:  $f_{AA} = 0.5$ ,  $f_{Aa} = 0.375$  e  $f_{aa} = 0.125$ . Comparado às frequências antes da seleção (33.3% cada), notamos que os genótipos que possuem o alelo A (AA e Aa) subiram de frequência.

É simples calcular a variação da frequência do alelo A causada pela seleção. Antes da seleção, existiam 15 cópias do alelo A num total de 30 cópias ( $2N$ ), ou seja,  $f_A = 0.5$ . Após a seleção, podemos calcular a frequência do alelo na população de reprodutores a partir dos genótipos:



$$f_A = f_{AA} + \frac{f_{Aa}}{2} = 0.5 + \frac{0.375}{2} = 0.6875$$

Após apenas uma rodada de seleção a frequência do A subiu de 50% para 68.75%. Evidentemente, a frequência do a caiu. **O pool gamético, portanto, terá 68.75% dos gametas contendo o alelo A.**

Outro fato interessante que ocorre após a seleção é a mudança do valor de  $\bar{w}$ :

**Antes** da seleção:

$$\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa} = (0.333 \times 1) + (0.333 \times 0.75) + (0.333 \times 0.25) = 0.66667$$

**Depois** da seleção, os genótipos serão formados pelo modelo Hardy-Weinberg (**apenas** seleção atuando) a partir das novas frequências alélicas do pool gamético:

$$\begin{aligned} f_{AA} &= f_A^2 = 0.6875^2 = 0.4726562 \\ f_{Aa} &= 2f_Af_a = 2 \times 0.6875 \times 0.3125 = 0.4296875 \\ f_{aa} &= f_a^2 = 0.3125^2 = 0.09765625 \end{aligned}$$

O valor do  $\bar{w}$  da próxima geração será:

$$\begin{aligned} \bar{w} &= f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa} = (0.4726562 \times 1) + (0.4296875 \times 0.75) + (0.09765625 \times 0.25) \\ &= 0.8193359 \end{aligned}$$

Após uma rodada de seleção,  $\bar{w}$  subiu de 0.66 para 0.82. **Guarde esta propriedade na memória.**

Podemos agora construir um *loop* para estudar todo o processo de seleção natural ao longo das gerações e seu impacto nas frequências genotípicas e alélicas. Usaremos os mesmos valores de frequência inicial dos genótipos e valores adaptativos médios. Armazenaremos as três frequências genotípicas, a frequência do alelo A, e o  $\bar{w}$  em vetores ( $f_{AA}$ ,  $f_{Aa}$ ,  $f_{aa}$ ,  $f_A$  e  $w_{\text{medio}}$ ). O *loop* será rodado por 50 gerações:

```
# Valores adaptativos relativos:
wAA = 1
wAa = 0.75
waa = 0.25

fAA = fAa = faa = fA = w_medio = vector()

# Frequencias iniciais ANTES:
fAA[1] = 5/15
fAa[1] = 5/15
faa[1] = 5/15

g = 50

fA[1] = fAA[1] + fAa[1]/2

w_medio[1] = fAA[1]*wAA + fAa[1]*wAa + faa[1]*waa

for(i in c(2:g)){

  # tem que trabalhar com a frequência alélica por que é o pool gamético que é sorteado!
  # Pool gamético da geração i-1:
  fA[i] = ( fA[i-1] * ( (fA[i-1]*wAA) + ( (1-fA[i-1])*wAa) ) ) / w_medio[i-1]
```

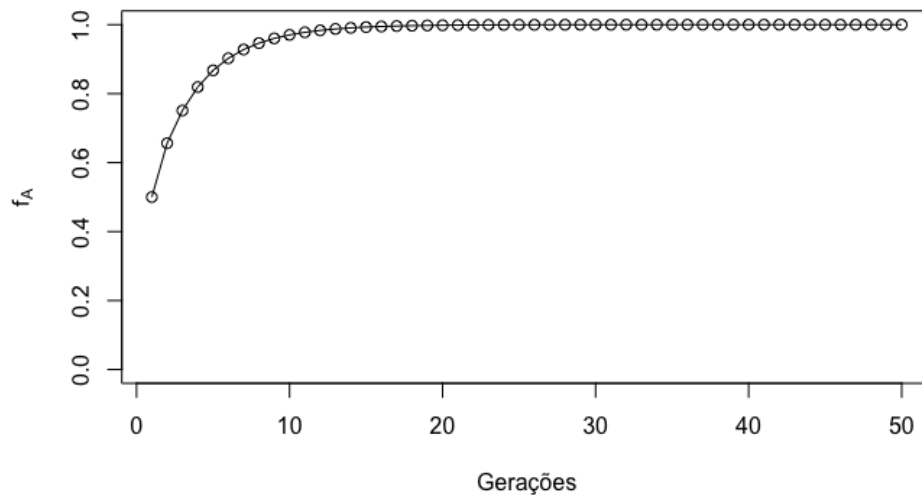
```
fAA[i] = (fA[i])^2
fAa[i] = 2 * fA[i] * (1 - fA[i])
faa[i] = (1 - fA[i])^2

w_medio[i] = fAA[i]*wAA + fAa[i]*wAa + faa[i]*waa

}
```

Podemos plotar então o gráfico da evolução da frequência do alelo A ao longo das 50 gerações:

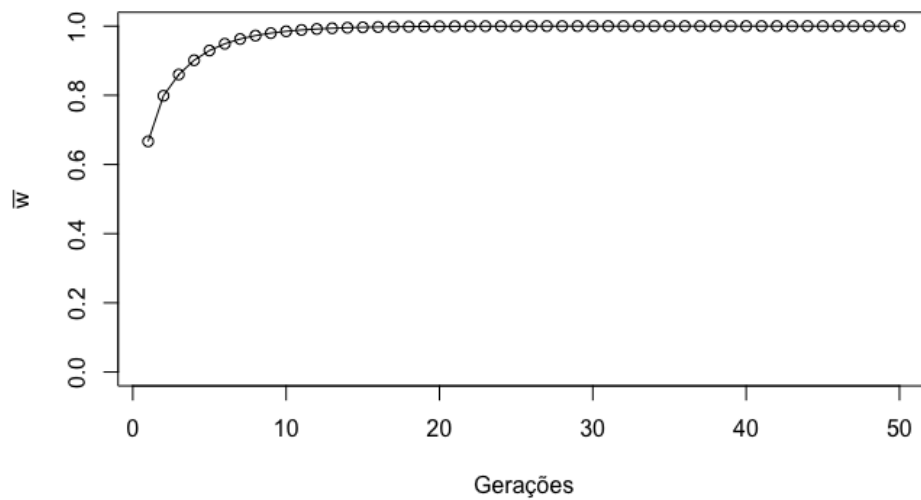
```
plot(fA,ylim=c(0,1), xlab="Gerações", ylab=expression(f[A]))
lines(fA)
```



Conforme observado no gráfico, a frequência do alelo A, inicialmente em 50%, sobe até atingir 100%. Indicando que a população possui 100% de homocigotos Aa. O alelo foi, portanto, fixado pela ação da seleção natural. Os genótipos que continham o alelo A eram aqueles com os maiores valores adaptativos relativos (AA e Aa).

É interessante também observar o gráfico do  $\bar{w}$  ao longo das gerações:

```
plot(w_medio ,ylim=c(0,1), xlab="Gerações", ylab=expression(bar(w)))
lines(w_medio)
```



O valor de  $\bar{w}$  também sobe até atingir o valor de 1. Esta é uma das principais propriedades da ação da seleção natural. **A seleção natural é a única força evolutiva que atua no sentido de sempre aumentar o valor de  $\bar{w}$ .**



## Tipos de seleção natural

Nesta prática estudaremos a dinâmica evolutiva das frequências alélicas e a paisagem adaptativa em diversos regimes de seleção natural. Esses regimes são definidos pelas diversas **relações de desigualdade** que podem ocorrer entre os valores adaptativos relativos dos genótipos. A mecânica básica do processo é a mesma aplicada na prática de quantificação da seleção.

No caso de um locus com dois alelos em organismos diplóides, podemos conceber três relações possíveis de desigualdade entre os valores adaptativos relativos, que definem os tipos de seleção:

| Tipo de seleção | Relação de desigualdade  |
|-----------------|--|
| Direcional      | $w_{AA} > w_{Aa} > w_{aa}$ OU $w_{AA} = w_{Aa} > w_{aa}$ OU $w_{AA} > w_{Aa} = w_{aa}$ |
| Balanceada      | $w_{AA} < w_{Aa} > w_{aa}$   |
| Disruptiva      | $w_{AA} > w_{Aa} < w_{aa}$   |

Cada um desses tipos de seleção define dinâmicas de mudança da frequência dos alelos e do valor adaptativo médio  $\bar{w}$  ao longo das gerações. Vejamos exemplos para cada um dos casos usando a quantificação que usada na prática anterior.

### Seleção direcional

Em todos os casos de seleção direcional, começaremos com frequências baixas para os genótipos com maior valor adaptativo relativo para visualizar melhor a evolução da frequência alélica.

#### **Caso $w_{AA} > w_{Aa} > w_{aa}$ (direcional):**

```
# Valores adaptativos relativos:
wAA = 1
wAa = 0.8
waa = 0.3

fAA = fAa = faa = fA = w_medio = vector()

# Frequência inicial do alelo A (valor escolhido):
fA[1] = 0.1

g = 30

# Vamos assumir que a população inicia com frequências genotípicas em Hardy-Weinberg (embora não seja
# necessário):

fAA[1] = (fA[1])^2
fAa[1] = 2 * fA[1] * (1 - fA[1])
faa[1] = (1 - fA[1])^2

w_medio[1] = fAA[1]*wAA + fAa[1]*wAa + faa[1]*waa

for(i in c(2:g)){

  # Composição do pool gamético de (i-1) usando a frequência alélica
  # dos indivíduos reprodutores da geração anterior (i-1), ou seja, após seleção:

  fA[i] = ( fA[i-1] * ( (fA[i-1]*wAA) + ( (1-fA[i-1])*wAa) ) ) / w_medio[i-1]

  # O pool gamético é sorteado infinitas vezes para não ter erro amostral
```

*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
 Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

# Neste modelo, APENAS a seleção está atuando. Não há deriva:

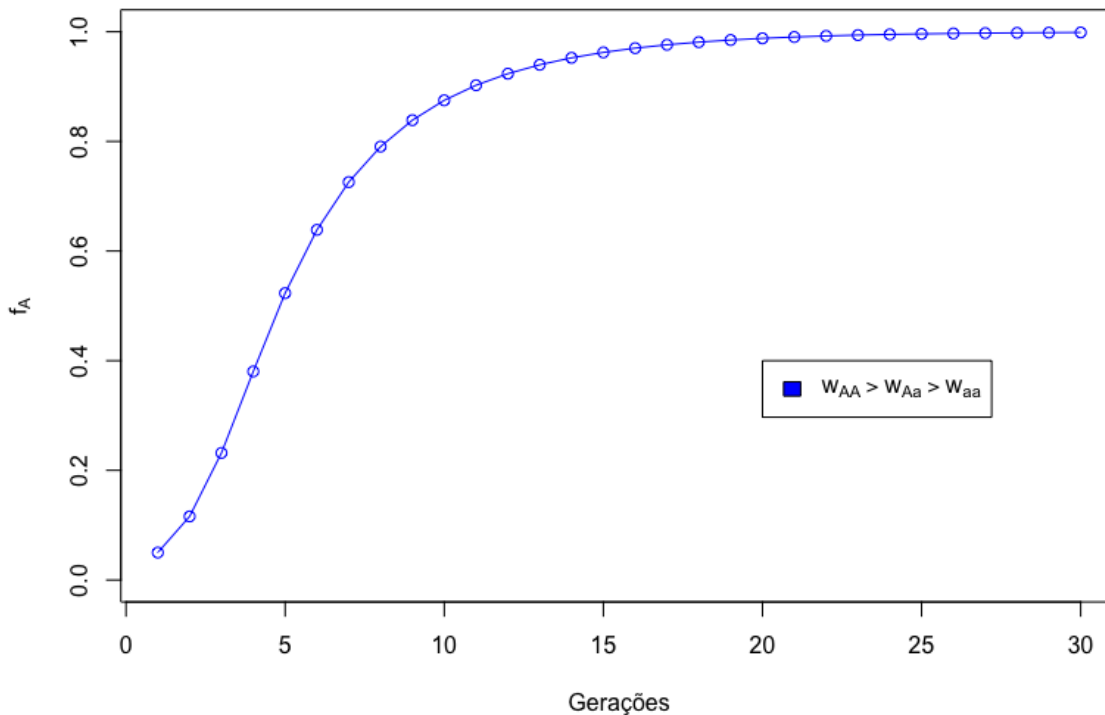
```
fAA[i] = (fA[i])^2
fAa[i] = 2 * fA[i] * (1 - fA[i])
faa[i] = (1 - fA[i])^2

# Cálculo do w_medio dos zigotos da geração i:
w_medio[i] = fAA[i]*wAA + fAa[i]*wAa + faa[i]*waa
```

}

Podemos plotar então o gráfico da evolução da frequência do alelo A ao longo de 30 gerações:

```
plot(fA,ylim=c(0,1), xlab="Gerações", ylab=expression(f[A]), col="blue")
lines(fA, col="blue")
legend(20,0.4,legend=c(expression(paste(w[AA] > w[Aa])>w[aa] )), fill=c("blue"))
```



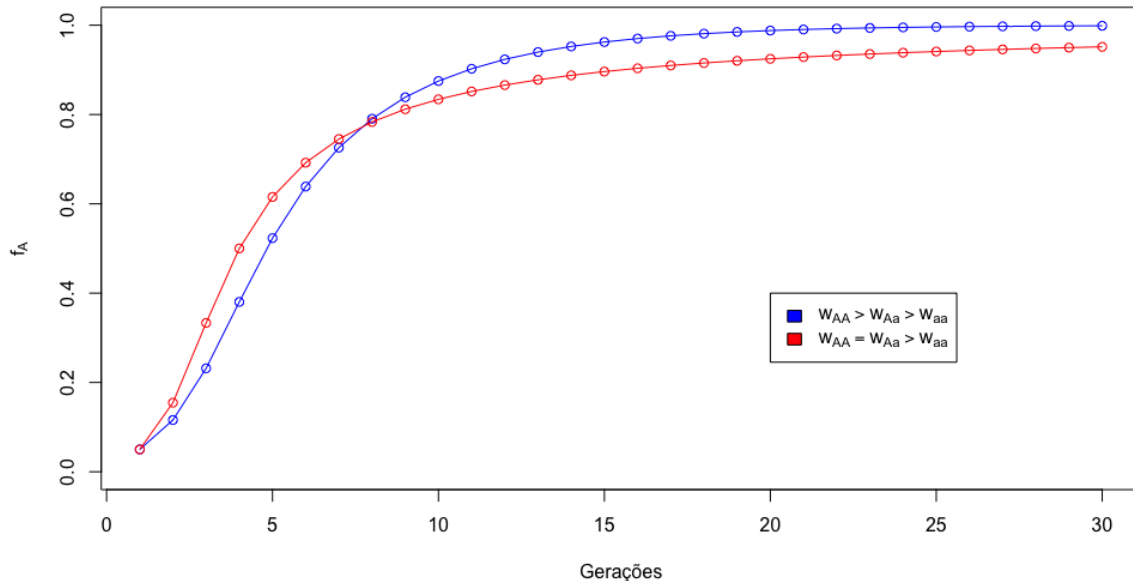
O gráfico mostra que, embora inicialmente com uma frequência muito próxima a zero, como os genótipos que possuem o alelo A são selecionados (maiores valores adaptativos relativos), em poucas gerações ocorre a fixação do A na população, ou seja, 100% dos indivíduos têm genótipo AA.

**Caso  $w_{AA} = w_{Aa} > w_{aa}$  (direcional):**

Use o mesmo código acima para estudar esse segundo caso, apenas modifique os valores adaptativos relativos para:

```
# Valores adaptativos relativos:
wAA = 1
wAa = 1
waa = 0.25
```

Vamos plotar o gráfico comparando ambos os casos:



Neste segundo caso, é evidente que a frequência do alelo A **não atinge a fixação**. Isso ocorre, pois a relação  $w_{AA} = w_{Aa} > w_{aa}$  indica um cenário de **dominância completa** do alelo A em relação ao a. Para os olhos da seleção natural, os genótipos AA e Aa são igualmente bons, pois possuem o mesmo fenótipo (dominância A sobre a).

O gráfico acima evidencia também que, nas primeiras gerações de seleção, a frequência do A sobe com mais intensidade. Isso ocorre pois os alelos A do heterozigoto também estão sendo selecionados (além do homozigoto AA). O alelo A não consegue ser fixado, pois os alelos a não são eliminados por estarem “**escondidos**” no heterozigoto Aa.

### Caso $w_{AA} > w_{Aa} = w_{aa}$ (direcional):

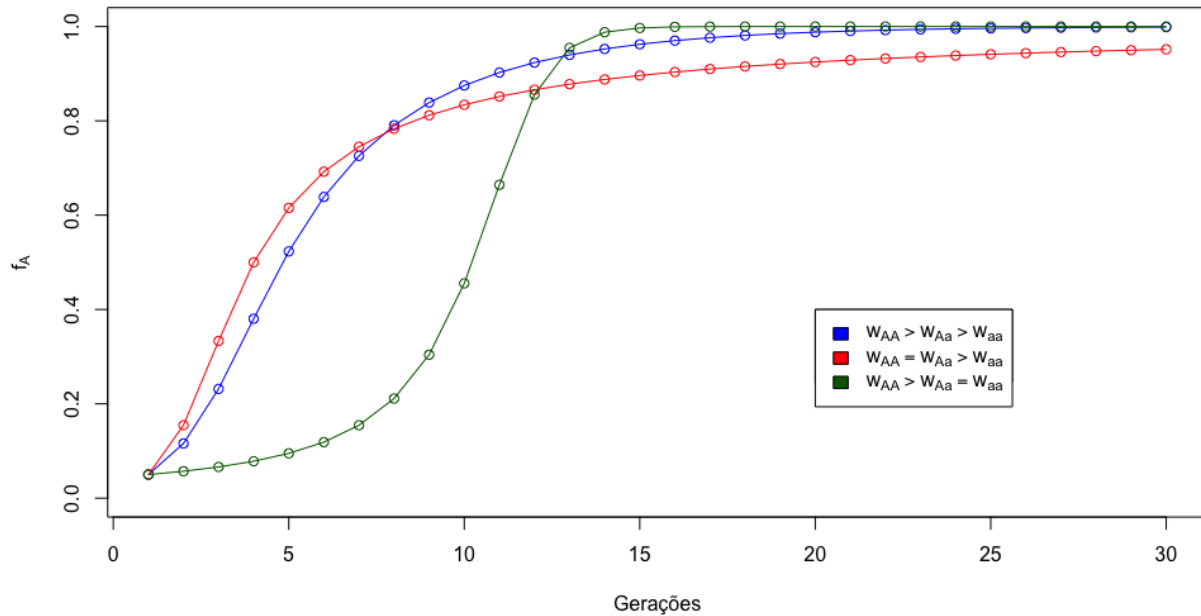
Por fim, veremos um caso em que também existe dominância completa. Entretanto, agora o alelo dominante é o a e não o A. Portanto,  $w_{Aa} = w_{aa}$ , pois ambos são idênticos aos olhos da seleção natural. Embora a seja dominante sobre A, o genótipo com o maior valor adaptativo médio é o homozigoto AA:

# Valores adaptativos relativos:

```
wAA = 1
wAa = 0.25
waa = 0.25
```

```
fA[1] = 0.05
```

Comparemos os três casos de seleção direcional:



Verificamos que, embora inicialmente a frequência do alelo *A* aumente mais lentamente (linha verde), que os outros casos, após aproximadamente a 10ª geração, a taxa de aumento torna-se a mais veloz. Isso ocorre, pois, a frequência de indivíduos *AA*, que começa inicialmente baixa ( $0.05 \times 0.05 = 0.0025$ ), atinge um ponto que permite a seleção natural acelerar o processo de eliminação dos genótipos que possuem o alelo *a* (*Aa* e *aa*) ambos muito com valores adaptativos médios bem menores que *AA*. Após este momento, a frequência do alelo *A* rapidamente atinge a fixação.

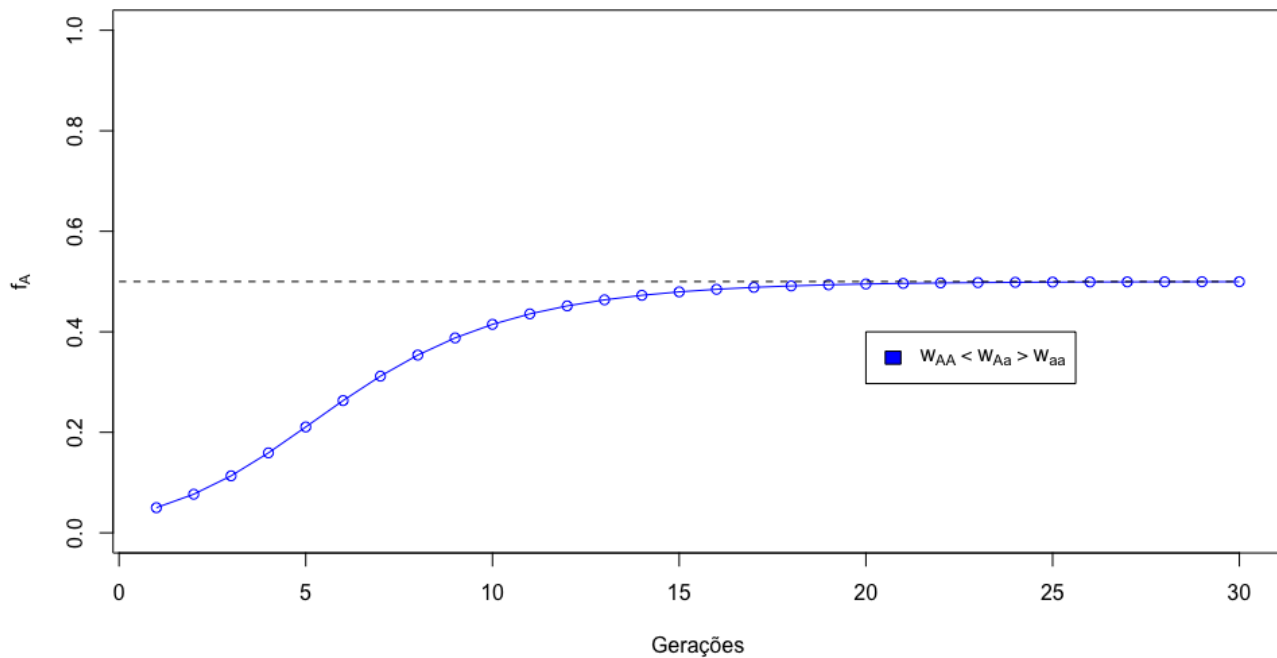
**Caso  $w_{AA} < w_{Aa} > w_{aa}$  (balanceada):**

Neste caso, o heterozigoto possui o maior valor adaptativo relativo. Estudemos os seguintes valores:

```
# Valores adaptativos relativos:
wAA = 0.6
wAa = 1
waa = 0.6
```

A frequência inicial do alelo *A* também é 0.05. O gráfico da evolução da frequência do alelo *A* é (use o mesmo código do início da prática para obter os novos valores de  $f_A$ ):



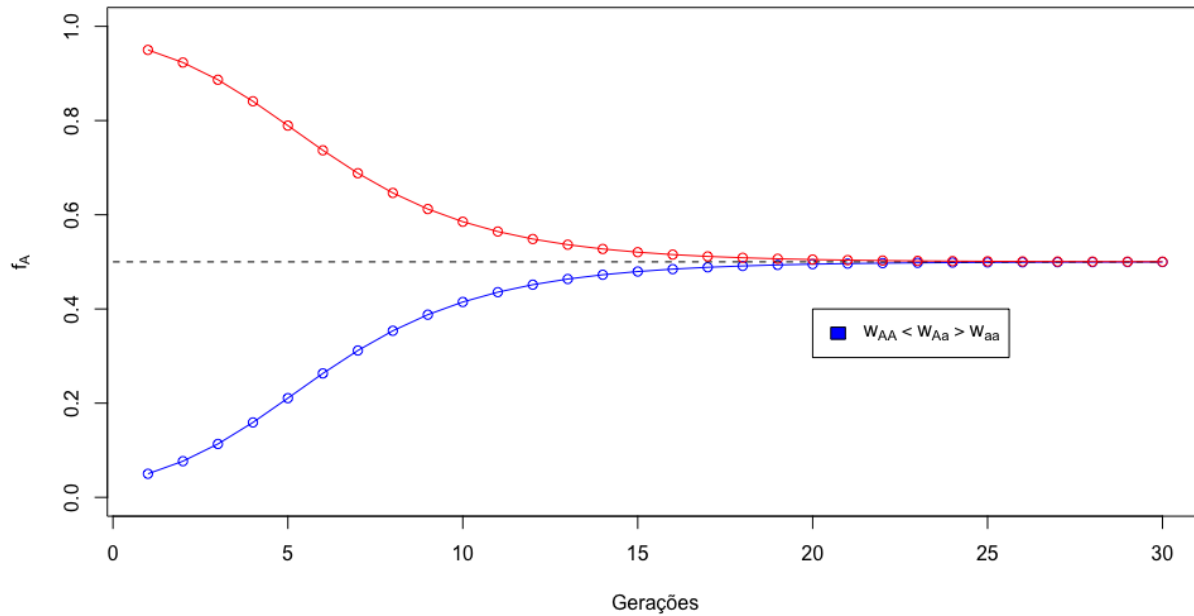


Neste caso, não há fixação do alelo *A*, pois o heterozigoto *Aa* é o melhor genótipo. Evidentemente, o heterozigoto possui ambos os alelos. A frequência do alelo *A* aumenta até atingir o valor de 50% (linha pontilhada). Isso ocorre pois os dois homozigotos *AA* e *aa* são **igualmente** piores que o heterozigoto ( $w_{AA} = w_{aa}$ ).

Estudemos o seguinte caso agora, onde a frequência do alelo *A* começa em 95%:

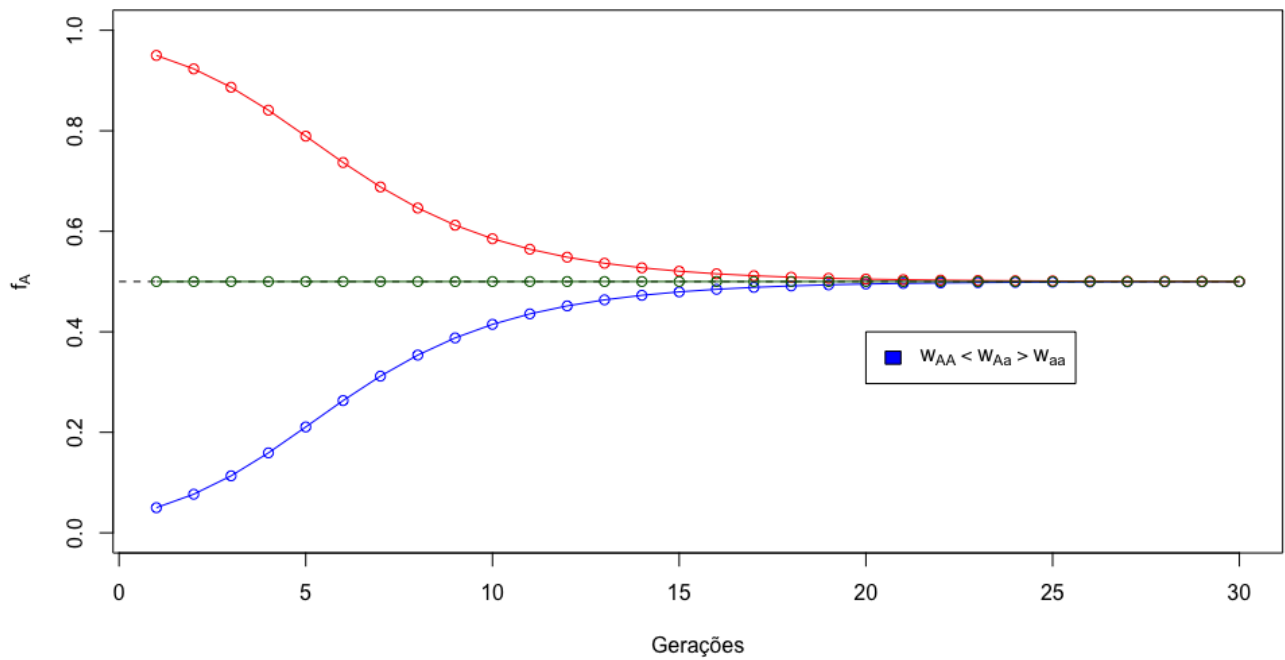
```
# Valores adaptativos relativos:
wAA = 0.6
wAa = 1
waa = 0.6

fA[1] = 0.95
```



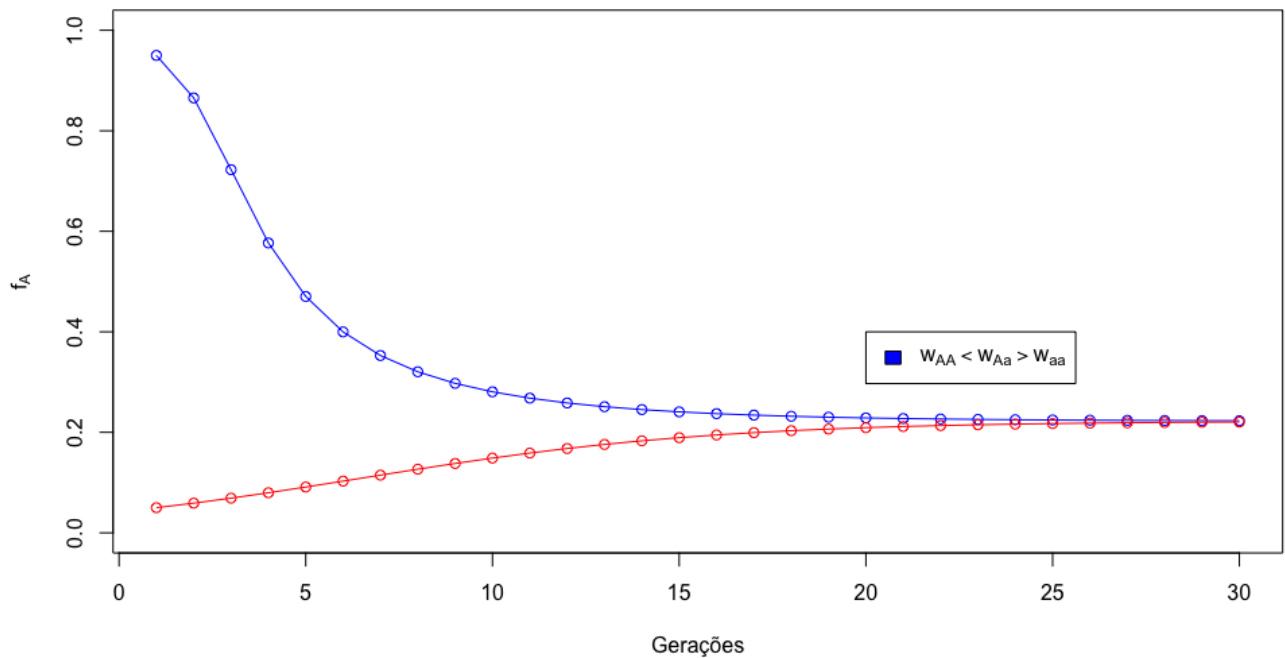
Podemos notar que a frequência do alelo A diminui até o valor de 50%, quando atinge a estabilidade. Trata-se de um **ponto de equilíbrio estável**. Evidentemente, se a frequência de A começar em 50%, ela não irá ser alterada:

```
# Valores adaptativos relativos:  
wAA = 0.6  
wAa = 1  
waa = 0.6  
  
fA[1] = 0.5
```



Lembremos que o ponto de equilíbrio de 50% é atingido apenas quando  $w_{AA} = w_{aa}$ . Estudemos agora este último caso de seleção balanceada em que  $w_{AA} < w_{aa}$ , embora o heterozigoto continue a ser o melhor genótipo. Vamos adotar  $f_A[1] = 0.95$  (linha azul) e também  $f_A[1] = 0.05$  (linha vermelha).

```
# Valores adaptativos relativos:  
wAA = 0.3  
wAa = 1  
waa = 0.8
```



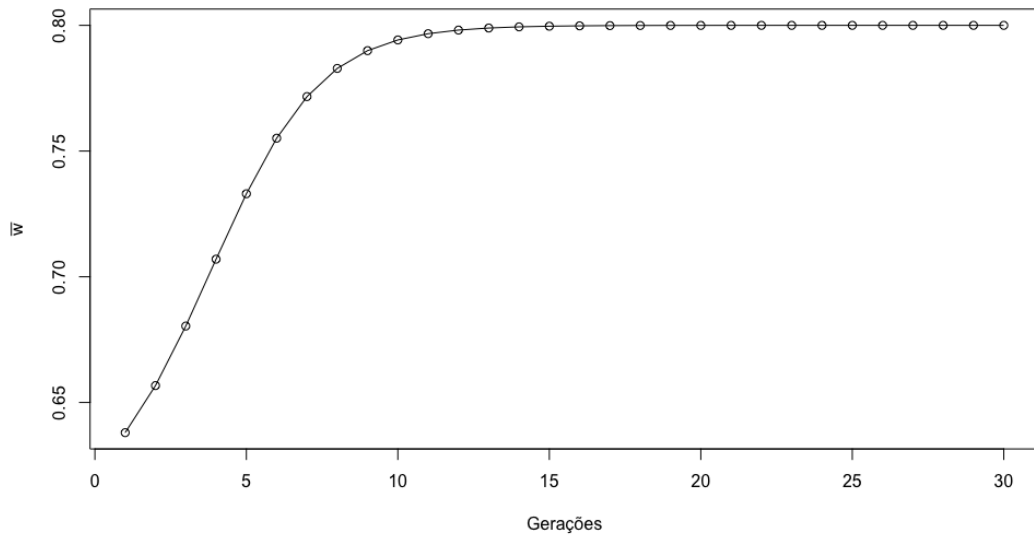
Verificamos que existe um novo ponto de equilíbrio diferente de 50%. Isso ocorre, pois, os dois homocigotos não são igualmente piores que o heterocigoto. Embora o heterocigoto *Aa* continue a ser o melhor genótipo, o genótipo *AA* possui valor adaptativo relativo pior que o *aa*.

Para calcular o novo ponto de equilíbrio, se  $w_{AA} = 1 - s$ ,  $w_{Aa} = 1$  e  $w_{aa} = 1 - t$ , o valor da frequência de equilíbrio do alelo *A* será dado por:

$$f_A = \frac{t}{s + t}$$

No exemplo dado,  $w_{AA} = 0.3$ ,  $w_{Aa} = 1$  e  $w_{aa} = 0.8$ , portanto  $s = 0.7$  e  $t = 0.2$ . Assim, o ponto de equilíbrio do gráfico acima é  $\frac{0.2}{0.7+0.2} = 0.222$ .

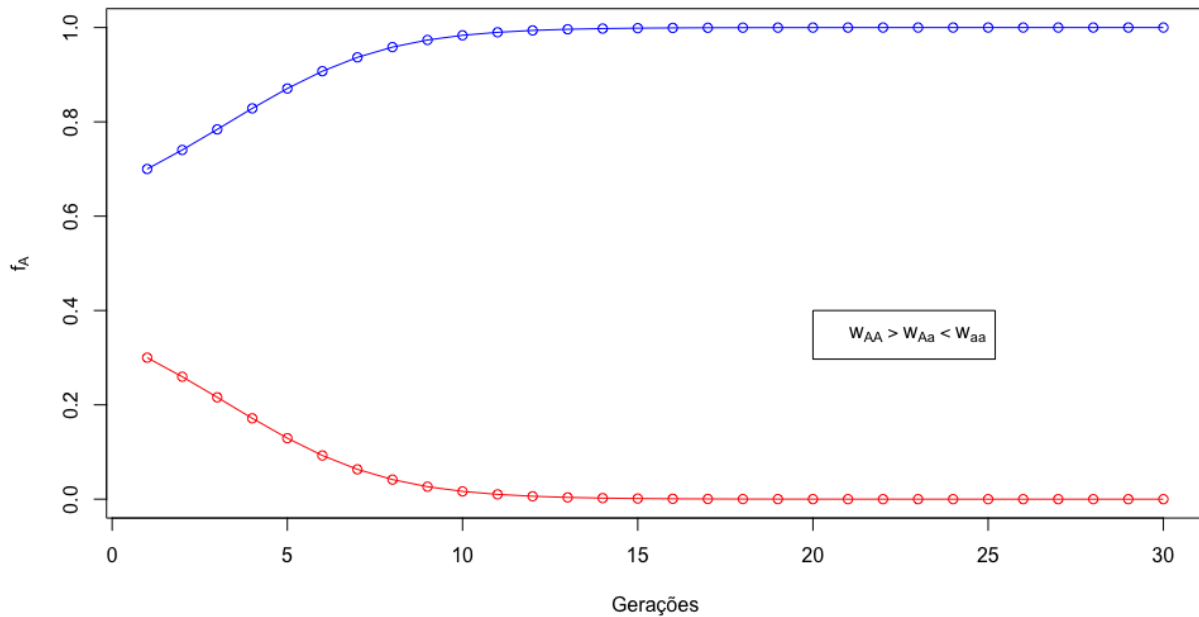
Finalmente, é interessante observar que, como o cruzamento de heterocigotos formam uma prole com 25% de *AA* e 25% de *aa*, a população não conseguirá levar a população até o valor de  $\bar{w} = 1$ , conforme acontece na seleção direcional.



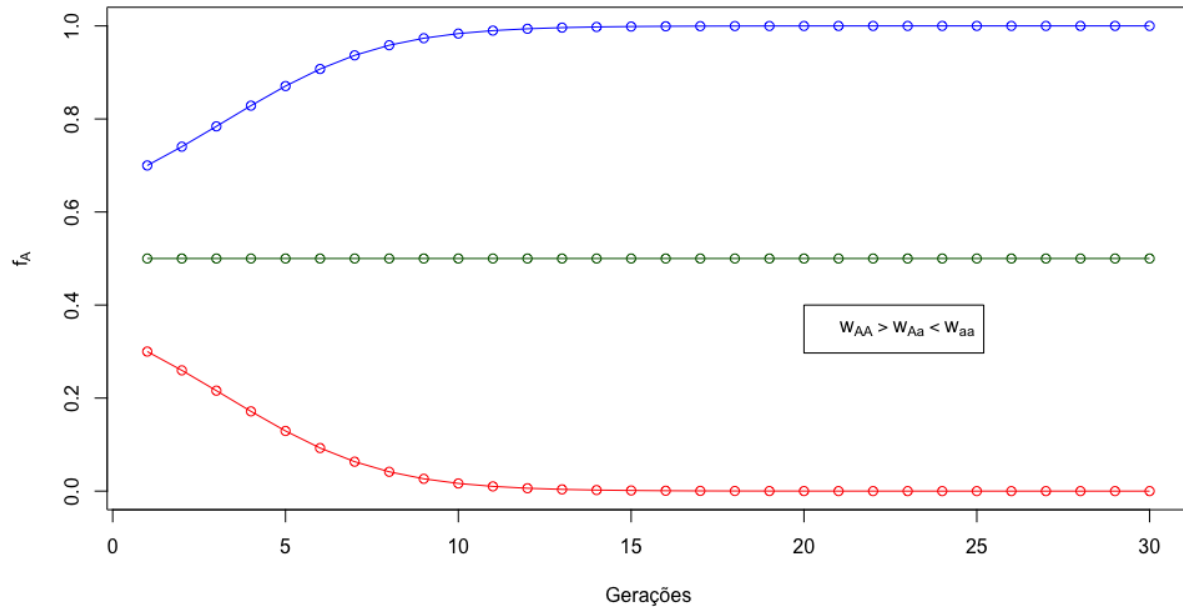
### Caso $w_{AA} > w_{Aa} < w_{aa}$ (disruptiva):

Neste último caso, temos os genótipos homozigotos com valores adaptativos superiores ao heterozigoto. Vamos analisar dois casos com  $f_A[1] = 0.7$  (linha azul) e também  $f_A[1] = 0.3$  (linha vermelha):

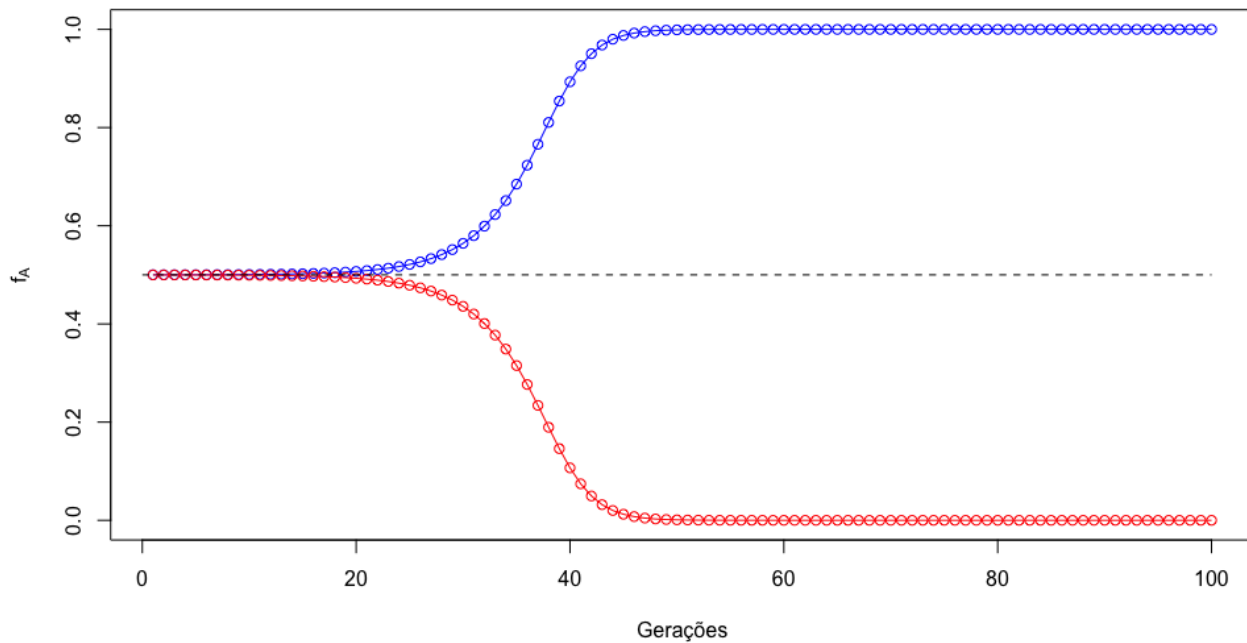
```
# Valores adaptativos relativos:  
wAA = 1  
wAa = 0.6  
waa = 1
```



Podemos verificar que, dependendo do valor inicial da frequência do alelo *A*, ele será fixado ou perdido. O valor de frequência alélica inicial que define se *A* será fixado ou perdido é justamente 50% se **ambos os homozigotos forem igualmente melhores que o heterozigoto, ou seja,  $w_{AA} = w_{aa}$** . Caso  $f_A[1] = 0.5$ , temos o seguinte cenário:



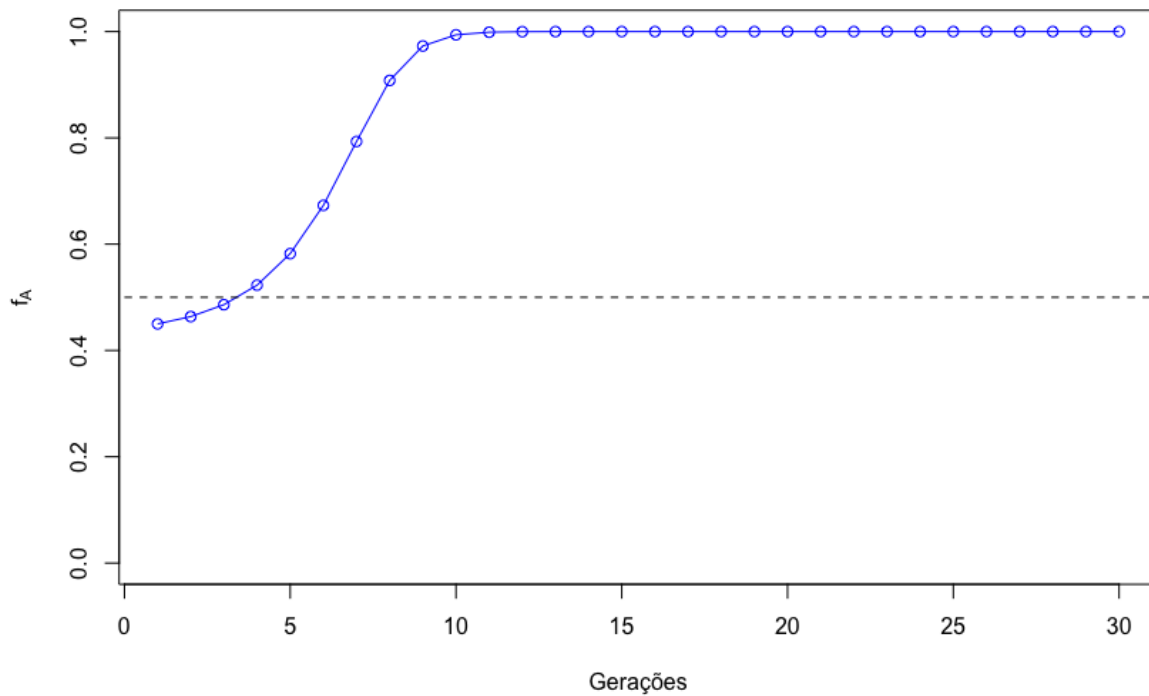
O ponto de 50% será um **equilíbrio instável**, pois ele existe apenas caso a frequência inicial seja exatamente 0.5 e **não existe nenhuma variação estocástica desse valor ao longo das gerações** (ou seja, não há deriva). Qualquer valor diferente de 0.5 levará o alelo à fixação (>0.5) ou perda (<0.5). Vamos analisar dois casos com  $f_A[1] = 0.5001$  (linha azul) e também  $f_A[1] = 0.4999$  (linha vermelha):



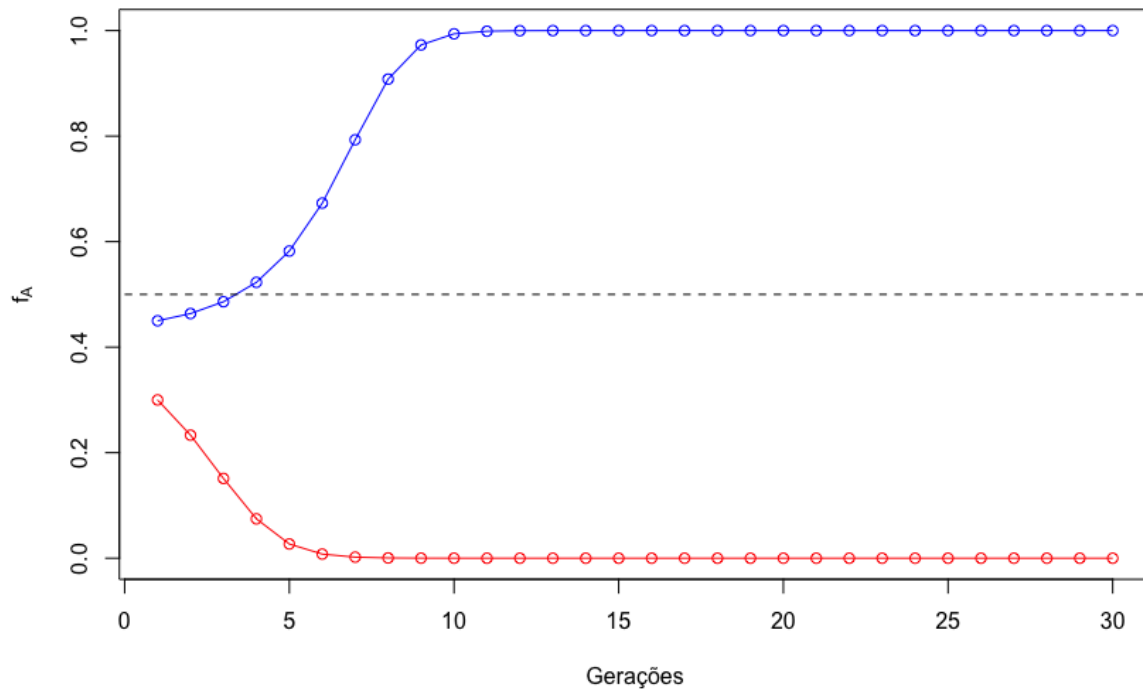
Embora necessitemos aguardar até aproximadamente a geração 30 para que as frequências se afastem notavelmente do equilíbrio instável, antes mesmo da 60ª geração ocorrerá fixação ou perda.

Analisemos agora um caso em que o heterozigoto é o genótipo com a pior performance, entretanto,  $w_{AA} > w_{aa}$ :

```
# Valores adaptativos relativos:  
wAA = 1  
wAa = 0.2  
waa = 0.8  
  
fA[1] = 0.45
```



Agora, mesmo começando abaixo de 50% (0.45), o alelo  $A$  é fixado. Isso ocorreu porque o homocigoto  $aa$  é um pouco menos eficiente que o homocigoto  $AA$ . Entretanto, se iniciarmos com frequência de  $A = 0.3$  (linha vermelha), temos:

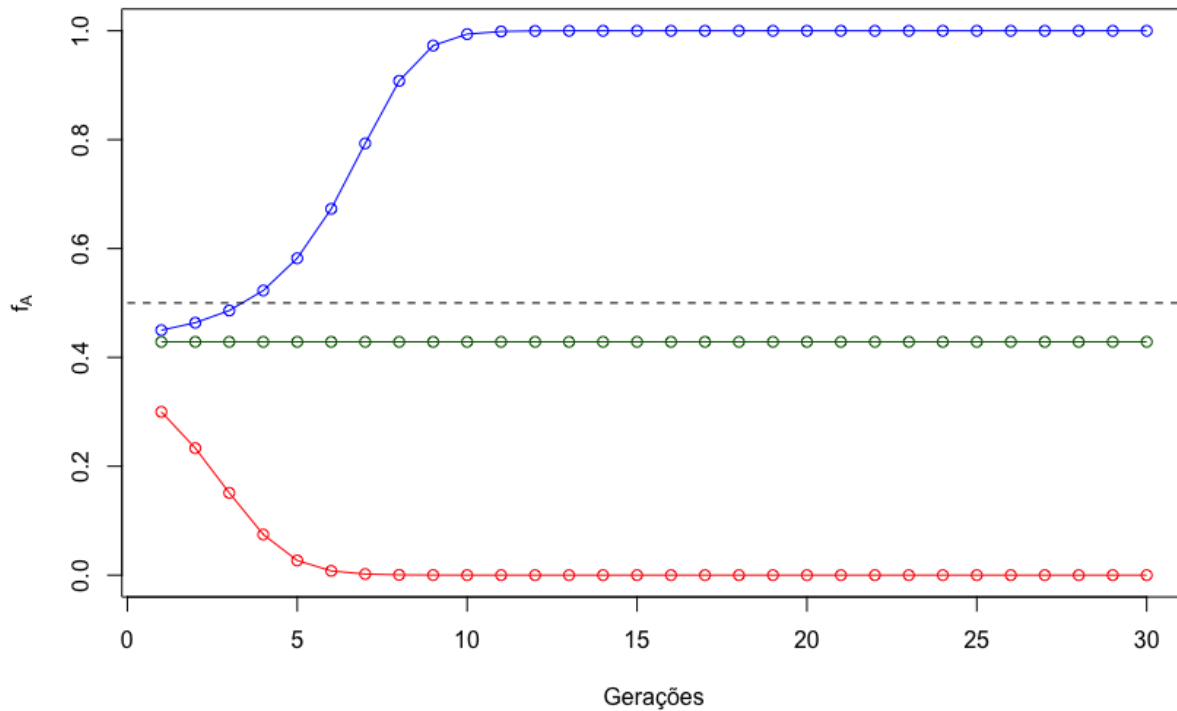




Abaixo de um determinado valor, o *A* é perdido. Esse valor é definido por:

$$f_A = \frac{w_{Aa} - w_{aa}}{2w_{Aa} - w_{AA} - w_{aa}}$$

Portanto, temos que o novo ponto de **equilíbrio instável** será 0.4285714 (linha verde):



Para finalizar a análise dos tipos de seleção, é necessário investigar como o valor adaptativo médio  $\bar{w}$  responde a cada tipo de seleção. O gráfico que estuda a dinâmica do  $\bar{w}$  para cada valor de frequência do alelo *A* é chamado de **paisagem adaptativa**.

Vamos estudar qual será o valor de  $\bar{w}$  para valores de frequência do alelo *A* que variam de 0 até 1, com intervalos de 0.1. Isso é obtido usando o comando `fA = seq(from=0, to=1, by=0.1)`.

Iniciemos com os três casos de seleção **direcional** que estudamos acima.

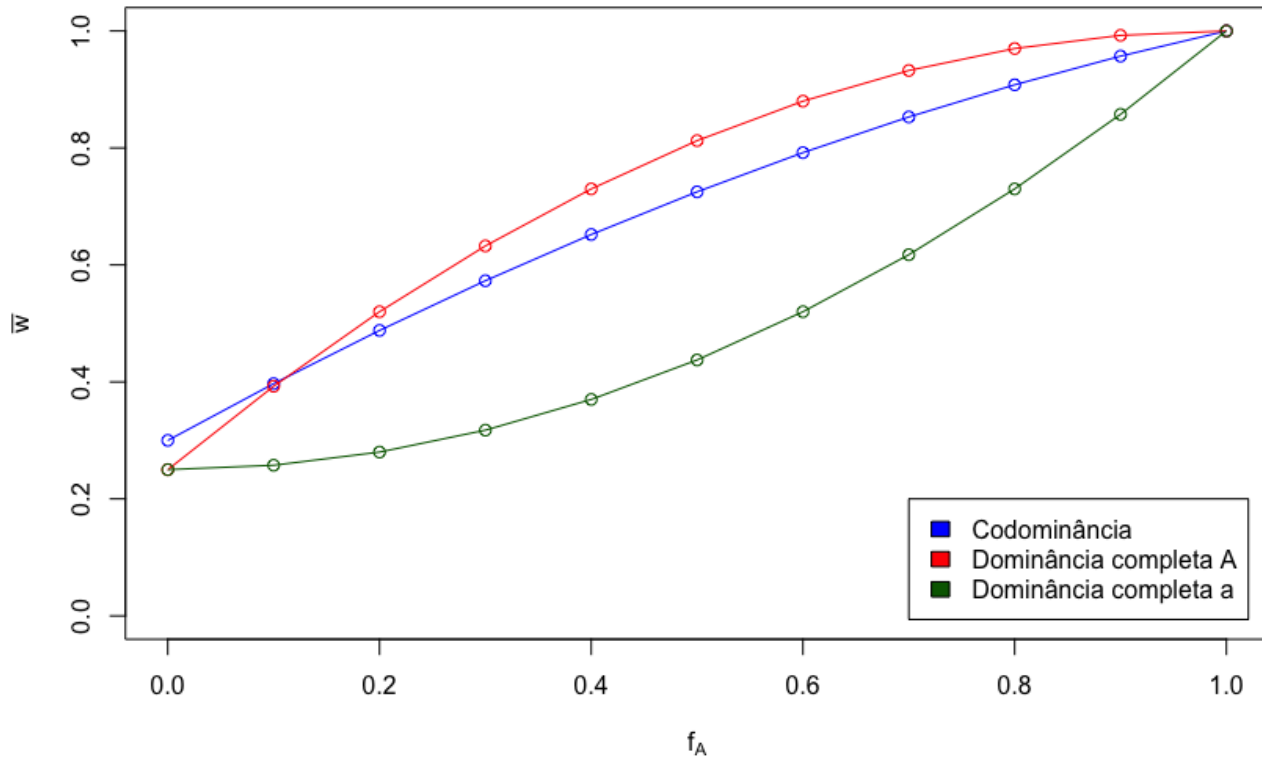
```
wAA = 1
wAa = 0.8
waa = 0.3

fA = seq(from=0, to=1, by=0.1)

w_medio = NULL

for(i in fA){
  w_medio = c(w_medio, ((i^2)*wAA) + ((2*i*(1-i))*wAa) + (((1-i)^2)*waa))
}
```

```
}
plot(w_medio ~ fA, col="blue")
lines(w_medio ~ fA, col="blue")
```



As curvas vermelha e verde podem ser obtidas modificando os valores de  $w$  no início do código pelos valores usados acima na seção de seleção direcional. Para plotar, basta usar:

```
points(w_medio ~ fA, col="red")
lines(w_medio ~ fA, col="red")

points(w_medio ~ fA, col="darkgreen")
lines(w_medio ~ fA, col="darkgreen")

legend(0.7,0.2,legend=c("Codominância", "Dominância completa A", "Dominância completa a"),
fill=c("blue","red", "darkgreen" ))
```

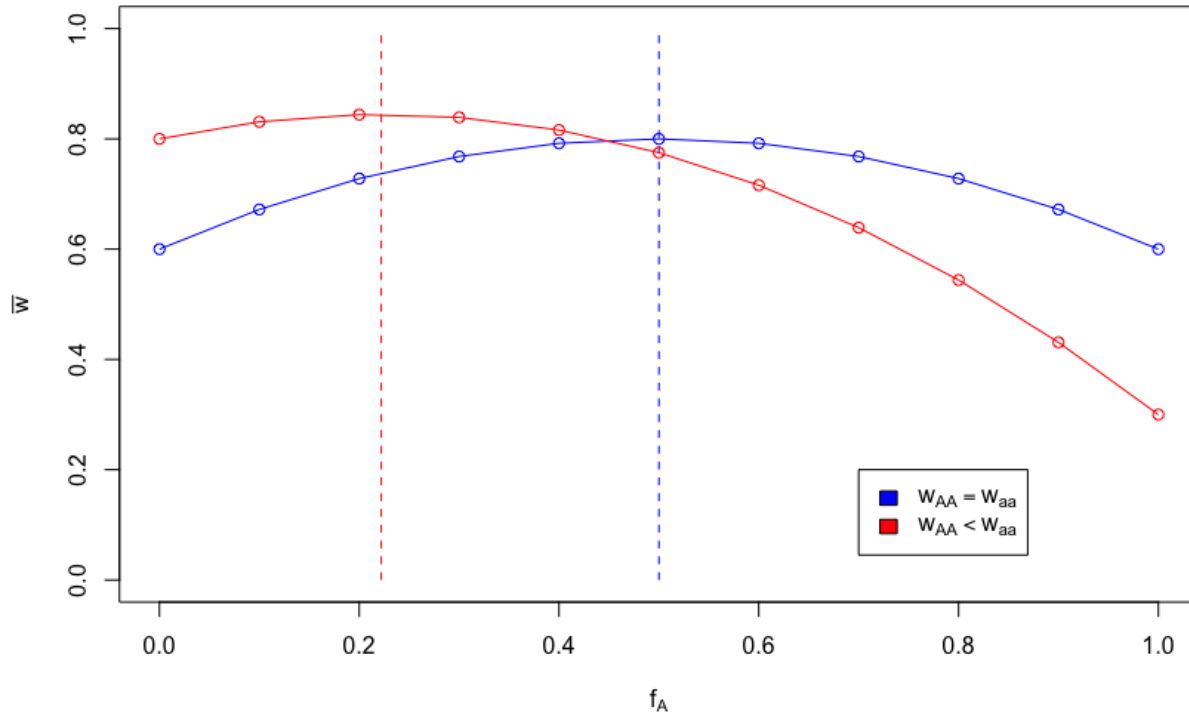
Embora a **taxa de ação da seleção natural** seja diferente nos três casos, como a seleção natural atua no **sentido de aumentar o valor de  $\bar{w}$** , quando a frequência do alelo A for 100%, teremos o valor mais alto de  $\bar{w}$ . **Entretanto, sabemos que no caso de dominância completa do A, a fixação não ocorrerá.** Portanto, o gráfico da paisagem adaptativa representa a trajetória teórica de  $\bar{w}$ .

Vejamos agora a paisagem adaptativa da seleção **balanceada**. Novamente, estudaremos os casos analisados acima:

```
# Valores adaptativos relativos (linha azul)
wAA = 0.6
wAa = 1
waa = 0.6
```

e

```
# Valores adaptativos relativos (linha vermelha)
wAA = 0.3
wAa = 1
waa = 0.8
```



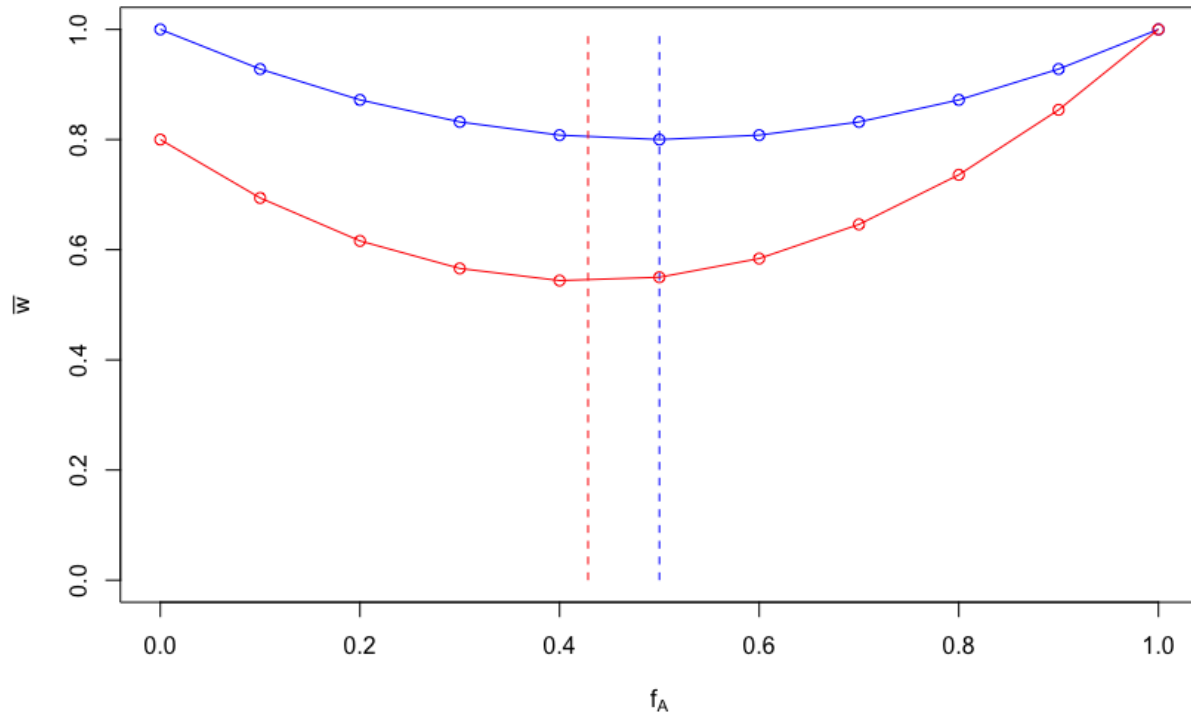
Podemos visualizar os **pontos de equilíbrio estáveis**, marcados pelas linhas tracejadas em cada um dos casos. A **seleção natural irá atuar no sentido de levar as populações até o valor de frequência de equilíbrio**. Pelas razões expostas anteriormente, **o valor máximo de  $\bar{w}$  não será igual a 1**.

Por fim, vejamos a paisagem adaptativa dos exemplos estudados de seleção **disruptiva**:

```
# Valores adaptativos relativos (linha azul)
wAA = 1
wAa = 0.6
waa = 1
```

e

```
# Valores adaptativos relativos (linha vermelha)
wAA = 1
wAa = 0.2
waa = 0.8
```



As linhas tracejadas indicam os **pontos de equilíbrio instável** igual a 0.5 (azul) e 0.4285714 (vermelha). Note que, se  $w_{AA} = w_{aa}$ , caso o alelo A seja perdido ou fixado,  $\bar{w} = 1$ , pois os homozigotos possuem valores adaptativos relativos idênticos. Entretanto, se  $w_{AA} > w_{aa}$ , apenas quando o A for fixado temos  $\bar{w} = 1$ .



## Formalização da seleção e variação das frequências alélicas

Nesta prática iremos apresentar uma forma comumente usada para investigação da ação da seleção natural nas populações. Ela é baseada no mesmo mecanismo usado anteriormente, entretanto, os valores adaptativos relativos são escritos em função de duas variáveis: o **coeficiente de seleção** ( $s$ ) e o **grau de dominância** ( $h$ ). Desta forma, ao invés de escrevermos

|                   |           |           |           |
|-------------------|-----------|-----------|-----------|
|                   | <b>AA</b> | <b>Aa</b> | <b>aa</b> |
| Fitness relativos | $w_{AA}$  | $w_{Aa}$  | $w_{aa}$  |

Estabeleceremos valores aos fitness usando a seguinte formalização:

| Tipos de seleção                | $w_{A_1A_1}$ | $w_{A_1A_2}$ | $w_{A_2A_2}$ |
|---------------------------------|--------------|--------------|--------------|
| a. Recessivo letal              | 1            | 1            | 0            |
| b. Alelo deletério recessivo    | 1            | 1            | $1 - s$      |
| c. Alelo deletério aditivo      | 1            | $1 - s/2$    | $1 - s$      |
| d. Seleção purificadora         | 1            | $1 - hs$     | $1 - s$      |
| e. Seleção positiva (do $A_1$ ) | $1 + s$      | $1 + hs$     | 1            |
| f. Vantagem do heterozigoto     | $1 - s_1$    | 1            | $1 - s_2$    |
| g. Desvantagem do heterozigoto  | $1 + s_1$    | 1            | $1 + s_2$    |

O uso das variáveis  $s$  e  $h$  oferece um detalhamento maior aos três tipos básicos de seleção (direcional, balanceada e disruptiva), embora a dinâmica geral do processo permaneça inalterada. Conforme podemos observar no quadro acima, o coeficiente de seleção mede a **redução da fitness** de um genótipo comparado ao melhor genótipo (casos b,c,d,f) ou mesmo o **aumento da fitness** em alguns casos (casos e,g). Desta forma, **o coeficiente de seleção pode ser favorável ou contrário ao genótipo numa parcela  $s$** , ou seja, o quanto a fitness difere (para + ou -) comparado com um genótipo de referência.

Já o grau de dominância  $h$  irá atuar no genótipo heterozigoto indicando o nível de dominância de um alelo. Se  $h = 0$ , temos dominância completa do alelo  $A_1$  (caso d) ou do  $A_2$  (caso e). Por outro lado, se  $h = 1$ , temos o cenário inverso ao anterior. Se  $h = 0,5$  estaremos num caso aditivo conforme o exposto em c. Assim como  $s$ , **é importante ler a descrição do modelo de seleção para saber exatamente a qual genótipo ( $A_1A_1$  ou  $A_2A_2$ ) o valor de  $h$  se refere.**

A formalização anterior é útil para estudar **como as diferenças entre os valores adaptativos influenciam a variação da frequência alélica**. Façamos a derivação básica da variação da frequência após a ação da seleção natural usando o modelo mais geral. Analisemos a frequência do alelo  $A_1 = p$ .

Frequência alélica após seleção:

$$p' = f_{A_1A_1}' + \frac{1}{2}f_{A_1A_2}'$$

Substituindo pelas frequências genotípicas após seleção:

$$p' = \left( \frac{p^2 w_{A_1A_1}}{\bar{w}} \right) + \frac{1}{2} \left( \frac{(2pq) w_{A_1A_2}}{\bar{w}} \right)$$

$$p' = \frac{p(pw_{A_1A_1} + qw_{A_1A_2})}{\bar{w}}$$



Antes da seleção a frequência alélica era  $p$ . Desta forma, a variação da frequência alélica é dada por:

$$\Delta p = p' - p$$

$$\Delta p = \frac{p(pw_{A_1A_1} + qw_{A_1A_2})}{\bar{w}} - p$$

Estudemos agora um caso de seleção purificadora contra os genótipos que possuem o alelo  $A_2$ , aumentando a frequência do alelo  $A_1$ . Neste caso, temos:

| Tipos de seleção     | $w_{A_1A_1}$ | $w_{A_1A_2}$ | $w_{A_2A_2}$ |
|----------------------|--------------|--------------|--------------|
| Seleção purificadora | 1            | $1 - hs$     | $1 - s$      |

Vamos admitir que se trata de um caso de herança aditiva com  $h = 0.5$  e que o coeficiente de seleção, ou seja, a diferença entre os fitness seja de 0,1. Assim,

| Tipos de seleção     | $w_{A_1A_1}$ | $w_{A_1A_2}$     | $w_{A_2A_2}$ |
|----------------------|--------------|------------------|--------------|
| Seleção purificadora | 1            | $1 - (0.5)(0.1)$ | $1 - 0.1$    |

Calculemos agora a variação da frequência de  $A_1$ ,  $\Delta p$ , ao longo das gerações:

```
### Variação da frequência alélica por seleção apenas

# Valores adaptativos relativos A1 = A e A2 = a:
s = 0.1
h = 0.5

wAA = 1
wAa = 1 - (h*s)
waa = 1 - s

fAA = fAa = faa = fA = w_medio = vector()
Dp = vector()

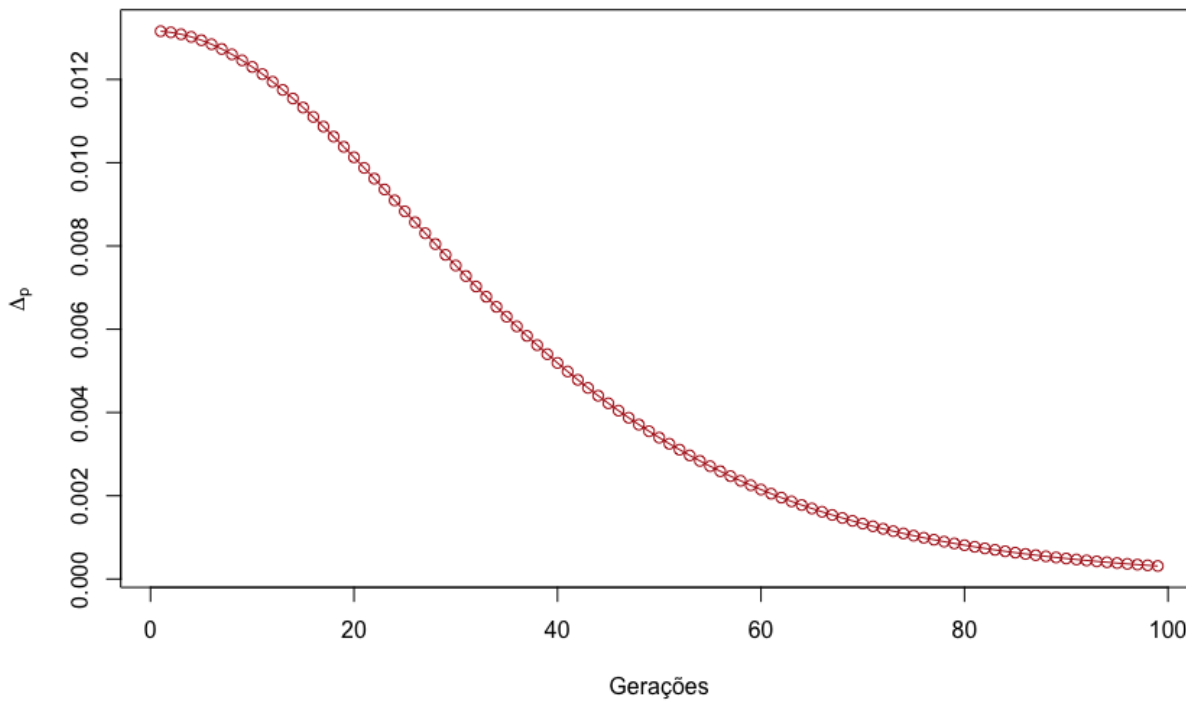
# Frequência inicial do alelo A (valor escolhido):
fA[1] = 0.5
g = 100
# Vamos assumir que a população inicia com frequências genotípicas em Hardy-Weinberg (embora não seja necessário):
fAA[1] = (fA[1])^2
fAa[1] = 2 * fA[1] * (1 - fA[1])
faa[1] = (1 - fA[1])^2
w_medio[1] = fAA[1]*wAA + fAa[1]*wAa + faa[1]*waa

for(i in c(2:g)){
  # frequência após seleção do alelo A no pool:
  fA[i] = ( fA[i-1] * ( (fA[i-1]*wAA) + ( (1-fA[i-1])*wAa) ) ) / w_medio[i-1]

  # Variação da frequência do alelo A:
  Dp[i - 1] = fA[i] - fA[i-1]

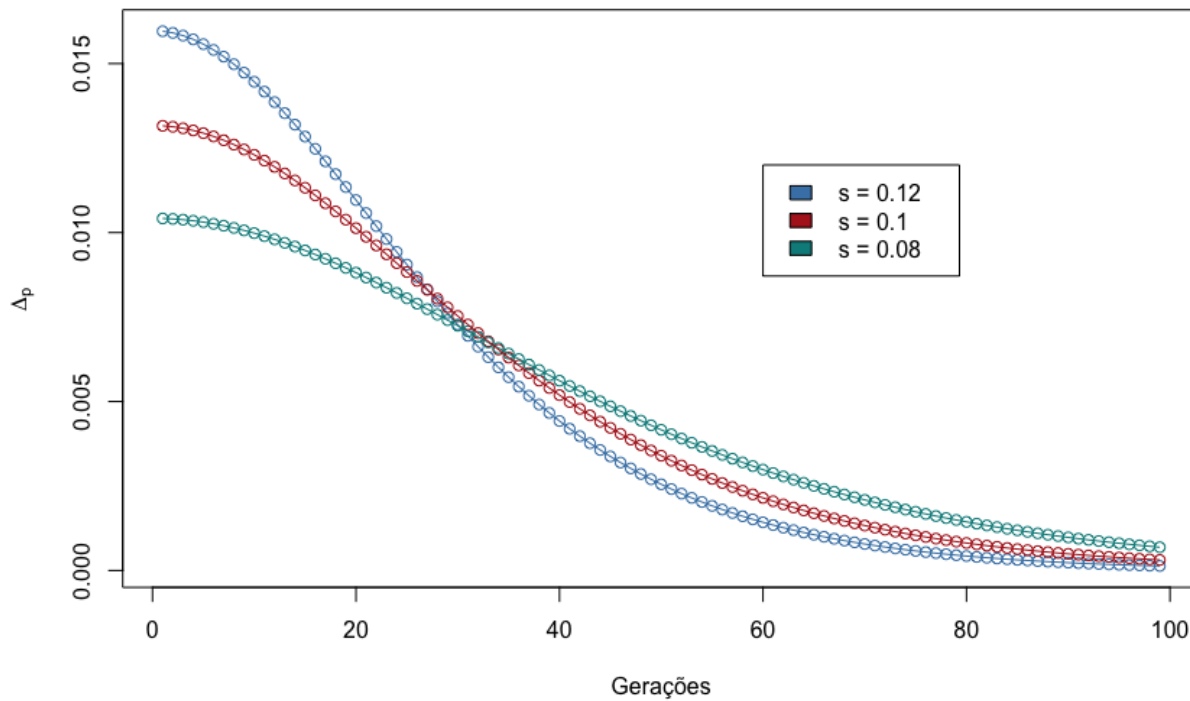
  # frequências genotípicas e w_medio da próxima geração
  fAA[i] = (fA[i])^2
  fAa[i] = 2 * fA[i] * (1 - fA[i])
  faa[i] = (1 - fA[i])^2
  w_medio[i] = fAA[i]*wAA + fAa[i]*wAa + faa[i]*waa
}
```

```
}  
  
# gráfico:  
plot(Dp, xlab="Gerações", ylab=expression(Delta[p]), col="firebrick" )  
lines(Dp, col="firebrick")
```



Podemos observar que **nas primeiras gerações de seleção a variação da frequência alélica é mais intensa**. Indicando que parte significativa da mudança da frequência alélica ocorre, ou seja, o aumento da frequência do alelo  $A_1$  ocorre nas primeiras gerações. Após algumas gerações de mudança drástica, a velocidade da variação da frequência cai.

Comparemos agora o mesmo regime de seleção com  $s = 0.1$ ,  $s = 0.12$  e  $s = 0.08$ . Perceba que a diferença entre os coeficientes de seleção é de apenas 0.02. Vejamos os gráficos:



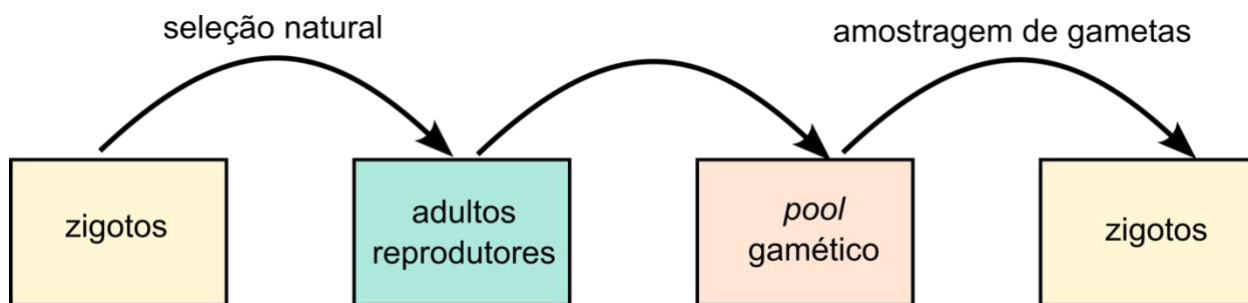
Embora a diferença entre os coeficientes seja pequena, é evidente que **a capacidade da seleção em alterar a frequência alélica é muito eficiente**. Esses resultados impressionaram muito os biólogos na primeira metade do século 20 e foi uma das principais razões para o **resgate da seleção natural como uma força evolutiva importante**, diminuindo o apoio da comunidade científica para processos alternativos, como o neolamarckismo e a ortogênese. Entretanto, notemos que **a dinâmica exposta acima ocorre com populações idealizadas com tamanho infinito** (uso das proporções Hardy-Weinberg).



## Seleção natural em populações finitas

No modelo básico, após a atuação da seleção natural no sucesso dos genótipos em chegarem à fase de adultos reprodutores que contribuem para *pool* gamético, calculamos as frequências dos genótipos na geração seguinte utilizando as frequências de Hardy-Weinberg. Desta forma, **ignoramos qualquer erro amostral dos gametas na composição da geração seguinte**. Na prática, esse procedimento fez que admitíssemos um tamanho populacional infinito.

Neste estudo, adotaremos uma estratégia diferente. Ao invés de usas as frequências genotípicas de HW para compor a população de zigotos a partir do *pool* gamético, iremos **amostrar os gametas do *pool***. Ao fazer isso, estamos admitindo que **não há seleção natural na população de gametas**, de forma que a formação dos genótipos dos zigotos acontecerá de forma estocástica:



Nesta nova modelagem, para fins de simplificação, admitiremos que a população de zigotos e adultos reprodutores terão tamanho constante ( $N$ ).

O código R para implementar esse modelo requer que modifiquemos o código do modelo básico de seleção apresentado anteriormente para incluir a amostragem probabilística que seguirá uma distribuição binomial, exatamente como fizemos no modelo básico de Wright-Fisher para estudar o efeito da deriva. Esse novo código é um híbrido dos códigos R de seleção (em populações infinitas) e deriva, ou seja, permite a análise de **interação entre seleção e deriva**.

Ao estudar abaixo, algumas características devem ser mencionadas. Inicialmente, para tornar a dinâmica mais didática, fizemos uma separação entre a população de zigotos (com frequências genotípicas  $fAAz$ ,  $fAaz$  e  $faaz$  e frequência alélica  $fAz$ ) e de adultos reprodutores (com frequências genotípicas  $fAAr$ ,  $fAar$  e  $faar$  e frequência alélica  $fAr$ ). Além disso, como a atuação da deriva genética é realizada pela função **dbinom**, que retorna o número de alelos  $A$  amostrados do *pool* ao invés da frequência dos genótipos. Dessa forma, adotamos as proporções Hardy-Weinberg (quadro de Punnett) apenas para compor os genótipos próxima geração de zigotos **após amostragem do *pool***. Dessa maneira, o efeito da deriva é contabilizado.

Primeiramente, estudemos um caso em que o coeficiente de seleção contra os genótipos com alelo  $a$  (ou  $A_2$ ) seja  $s = 0.1$  e grau de dominância do  $A$  sobre o  $a$  de  $h = 0.5$ . O tamanho da população será  $N = 10$  indivíduos (a deriva genética será forte). A frequência inicial do alelo  $A$  na população de zigotos será de 10%.

### **$s = 0.1$ , $h = 0.5$ , $N = 10$ :**

```
### Interação entre seleção e deriva
```

```
# Valores adaptativos relativos  $A_1 = A$  e  $A_2 = a$ :
```

```
s = 0.1 # coeficiente de seleção contra os genótipos com o alelo  $a$  (=  $A_2$ )
```



```
h = 0.5 # grau de dominância do A (A1) em relação ao a (A2)

wAA = 1
wAa = 1 - (h*s)
waa = 1 - s

fAAz = fAaz = faaz = fAz = vector() # frequências dos zigotos
fAaR = fAaR = faaR = fAr = vector() # frequências dos reprodutores
w_medio = vector()

g = 50 # número de gerações que a simulação rodará
N = 10 # tamanho efetivo da população

# Frequências genotípicas iniciais dos zigotos:
fAAz[1] = 0.05
fAaz[1] = 0.1
faaz[1] = 0.85

fAz[1] = fAAz[1] + fAaz[1]/2

# Primeiro round de seleção e frequências genotípicas dos reprodutores:
w_medio[1] = fAAz[1]*wAA + fAaz[1]*wAa + faaz[1]*waa

fAaR[1] = (fAAz[1]*wAA) / w_medio[1]
fAaR[1] = (fAaz[1]*wAa) / w_medio[1]
faaR[1] = (faaz[1]*waa) / w_medio[1]

# Frequência do alelo A no pool gamético dos reprodutores:
fAr[1] = fAaR[1] + faaR[1]/2

for(i in c(2:g)){

  # amostragem de alelos A do pool para compor N zigotos da próxima geração:
  fAz[i] = ( sample(c(0:(2*N)), 1, prob= dbinom(c(0:(2*N)), (2*N), fAr[i-1] ) ) )/(2*N)

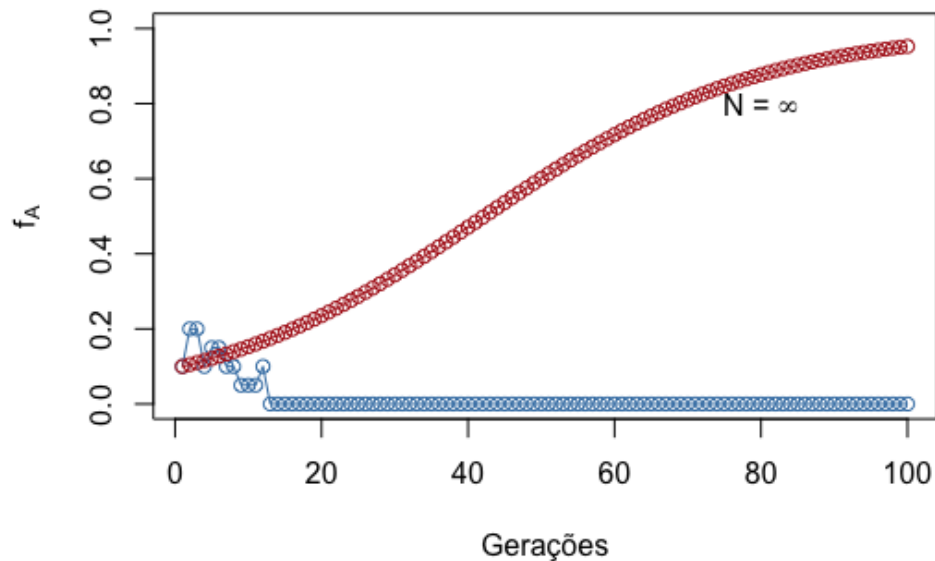
  # frequências genotípicas da população de zigotos (quadro de Punnett):
  fAAz[i] = ( fAz[i])^2
  fAaz[i] = 2 * fAz[i] * (1 - fAz[i])
  faaz[i] = (1 - fAz[i])^2

  # frequências genotípicas dos reprodutores após seleção:
  w_medio[i] = fAAz[i]*wAA + fAaz[i]*wAa + faaz[i]*waa

  fAaR[i] = (fAAz[i]*wAA) / w_medio[i]
  fAaR[i] = (fAaz[i]*wAa) / w_medio[i]
  faaR[i] = (faaz[i]*waa) / w_medio[i]

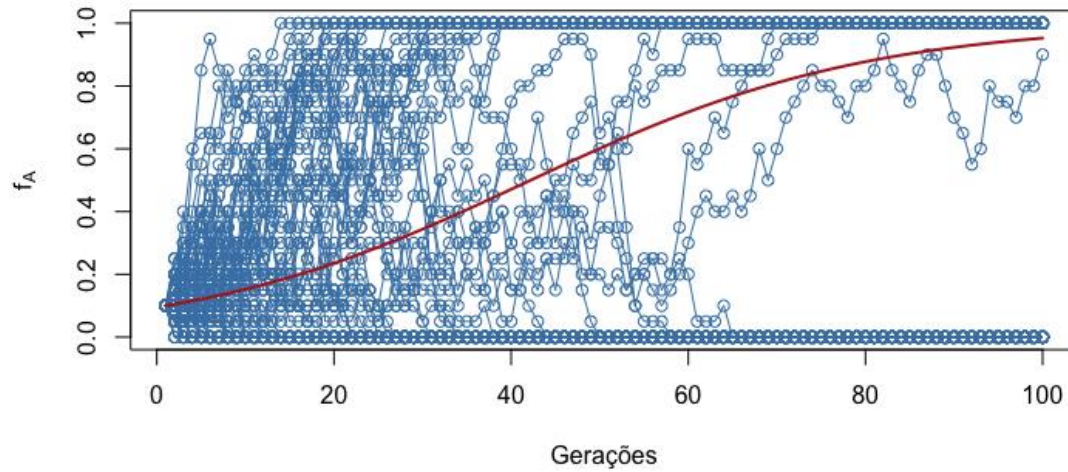
  # Frequência do alelo A no pool gamético dos reprodutores:
  fAr[i] = fAaR[i] + faaR[i]/2
}

# gráfico:
plot(fAz, col="steelblue", ylab=expression(f[A]), xlab="Gerações", ylim=c(0,1))
lines(fAz,col="steelblue")
```



No gráfico acima, a linha azul indica a evolução da frequência do alelo  $A$  quando numa população com  $N = 10$  e seleção natural favorável ao genótipo  $AA$ . Em princípio, esperaríamos que o alelo  $A$  fosse fixado, entretanto, no meu experimento, **mesmo com seleção favorecendo esse alelo, ele foi perdido**. Para fins de comparação, também foi plotada a evolução da frequência do alelo  $A$  nas mesmas condições de seleção ( $s = 0.1$  e  $h = 0.5$ ) numa população com tamanho infinito (linha vermelha). Ela mostra que, de fato, **com a atuação apenas da seleção natural, o alelo  $A$  deveria ser fixado**.

Como a deriva é um processo estocástico, cada vez que a simulação acima for rodada, teremos um caminho diferente da evolução das frequências alélicas. Façamos 100 replicatas com as mesmas condições ( $s = 0.1$ ,  $h = 0.5$  e  $N = 10$ ). O código para a simulação com as replicatas está disponibilizado ao final desta prática (tente fazer você mesmo! Veja o Apêndice).

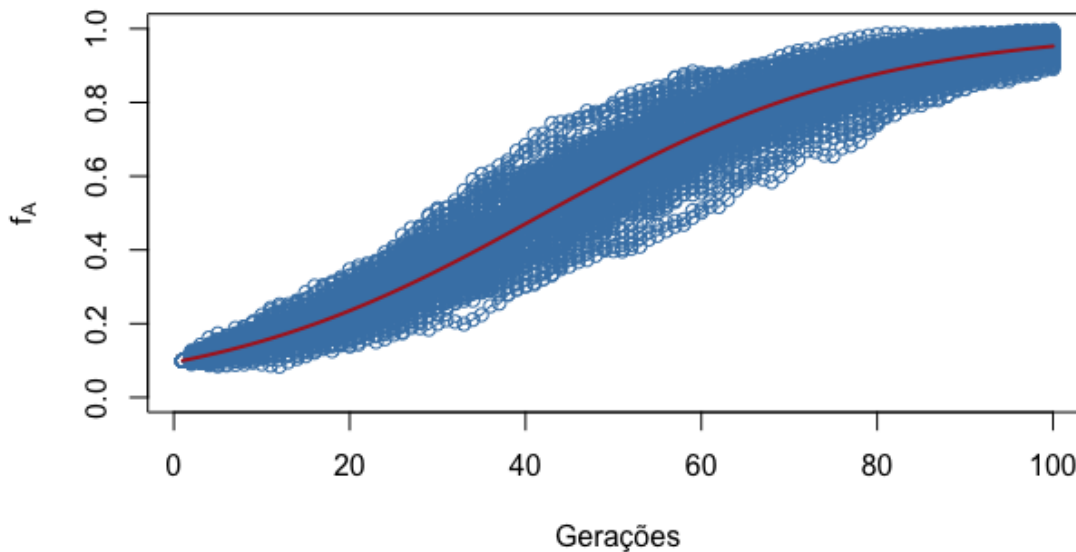


A linha vermelha é a dinâmica com o tamanho populacional infinito. No meu experimento, após 100 gerações, o alelo A tinha sido fixado em apenas 24% das replicatas. Um resultado completamente contrário ao esperado pela ação pura da seleção natural (linha vermelha).

Vejamos agora o que acontece se o tamanho populacional for aumentado para  $N = 1000$  indivíduos, usando as mesmas condições de seleção ( $s = 0.1$ ,  $h = 0.5$ ).

$N = 1000$  # tamanho efetivo da população

**$s = 0.1$ ,  $h = 0.5$ ,  $N = 1000$ :**

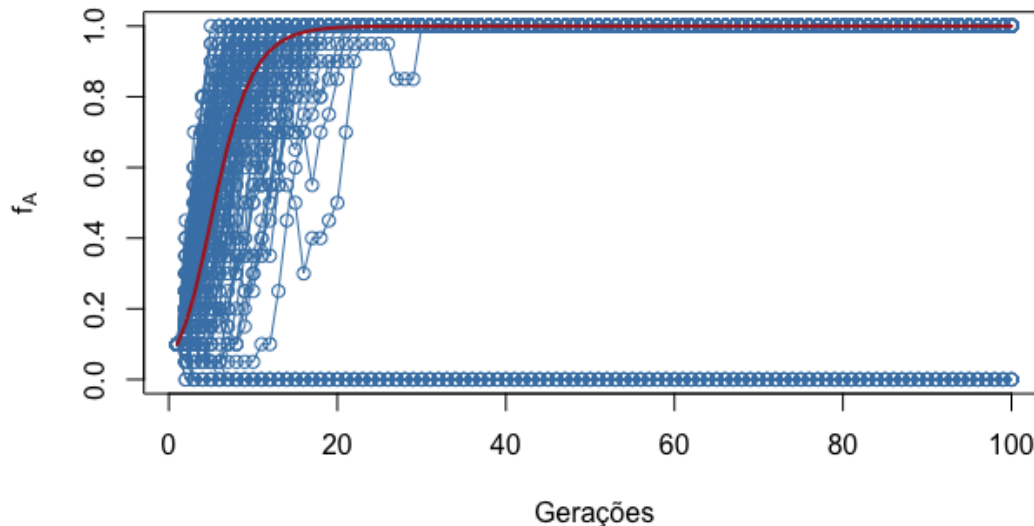


Agora, com a deriva reduzida (maior tamanho populacional), **a seleção natural conseguiu sobrepujar o efeito estocástico da deriva**. Neste segundo experimento, nenhuma replicata teve o

alelo A perdido. Todas as replicatas estão tendendo a fixar o alelo, exatamente conforme previsto por seleção.

Vamos agora verificar se, ao aumentar a intensidade da seleção ( $s = 0.6$ ), mesmo com uma população pequena ( $N = 10$ ), a deriva irá conseguir sobrepujar a seleção:

**$s = 0.6, h = 0.5, N = 10:$**



Com o aumento da intensidade da seleção, no meu experimento, o alelo A foi fixado em 93% das replicatas. Apenas em 7% das replicatas, tivemos a perda deste alelo. Isso indica que, agora, a seleção conseguiu vencer o efeito estocástico da deriva em 93% das vezes. Note que, mesmo com uma seleção natural muito intensa, a deriva sobrepujou a seleção em 7% das replicatas!

### **Mas como saberemos qual das forças irá guiar a dinâmica das frequências alélicas? O**

geneticista japonês Motoo Kimura sugeriu uma regra simples para entender a interação entre seleção natural e deriva genética. Se o produto  $4N_e s \ll 1$ , a deriva genética irá sobrepujar a seleção. Por outro lado, se  $4N_e s \gg 1$ , a seleção irá sobrepujar a deriva na maioria das replicatas. Quando  $4N_e s \approx 1$ , teremos uma dinâmica em que aproximadamente 50% das replicatas serão fixadas, ou seja, não há como fazer uma previsão sobre as forças.

### **Apêndice**

Código para obtenção das replicatas:

```
### Interação entre seleção e deriva

replicatas = 100
g = 100

# Valores adaptativos relativos A1 = A e A2 = a:
s = 0.6 # coeficiente de seleção contra os genótipos com o alelo a (A2)
h = 0.5 # modelo aditivo

wAA = 1
```



```
wAa = 1 - (h*s)
waa = 1 - s

fAAz = fAaz = faaz = fAz = matrix(ncol=g, nrow=replicatas) # frequencias dos zigotos
fAAr = fAar = faar = fAr = matrix(ncol=g, nrow=replicatas) # frequencias dos reprodutores
w_medio = matrix(ncol=g, nrow=replicatas)

N = 10 # tamanho efetivo da população

# Frequências genotípicas iniciais dos zigotos:
for(i in c(1:replicatas)){
  fAAz[i,1] = 0.05
  fAaz[i,1] = 0.1
  faaz[i,1] = 0.85

  fAz[i,1] = fAAz[i,1] + fAaz[i,1]/2
}

# Primeiro round de seleção e frequencias genotípicas dos reprodutores:
for(i in c(1:replicatas)){
  w_medio[i,1] = fAAz[i,1]*wAA + fAaz[i,1]*wAa + faaz[i,1]*waa
  fAAr[i,1] = (fAAz[i,1]*wAA) / w_medio[i,1]
  fAar[i,1] = (fAaz[i,1]*wAa) / w_medio[i,1]
  faar[i,1] = (faaz[i,1]*waa) / w_medio[i,1]

  fAr[i,1] = fAAr[i,1] + fAar[i,1]/2
}

for(i in c(1:replicatas)){
  for(j in c(2:g)){

    # amostragem de alelos A do pool para compor N zigotos da proxima geração:
    fAz[i,j] = ( sample(c(0:(2*N)), 1, prob= dbinom(c(0:(2*N)), (2*N), fAr[i,j-1] ) ) ) / (2*N)

    # frequências genotípicas da população de zigotos (quadro de Punnett):
    fAAz[i,j] = ( fAz[i,j] )^2
    fAaz[i,j] = 2 * fAz[i,j] * (1 - fAz[i,j])
    faaz[i,j] = (1 - fAz[i,j])^2

    # frequências genotípicas dos reprodutores após seleção:
    w_medio[i,j] = fAAz[i,j]*wAA + fAaz[i,j]*wAa + faaz[i,j]*waa

    fAAr[i,j] = (fAAz[i,j]*wAA) / w_medio[i,j]
    fAar[i,j] = (fAaz[i,j]*wAa) / w_medio[i,j]
    faar[i,j] = (faaz[i,j]*waa) / w_medio[i,j]

    # Frequência do alelo A no pool gamético dos reprodutores:
    fAr[i,j] = fAAr[i,j] + fAar[i,j]/2

  }
}
```



## Custo da seleção natural

Em 1957, o biólogo inglês John Haldane propôs um conceito que seria bastante influente em análises subsequentes da ação da seleção natural. Haldane argumentou que após a atuação da seleção natural, com a eliminação dos indivíduos antes da fase de adultos reprodutores, o tamanho total da população irá diminuir. Desta forma, os genótipos com maior valor adaptativo deverão **umentar sua taxa de fertilidade** para “repor” os indivíduos perdidos pela seleção e manter a população com tamanho constante. Caso contrário, após algumas gerações, a **população poderá atingir um número de indivíduos que a tornem inviável** e, conseqüentemente, será extinta. Assim, a seleção natural possui um custo: **o número de indivíduos eliminados por seleção deve ser repostado pelo aumento de fertilidade dos genótipos com maior fitness.**

Desta forma, após o surgimento de um novo alelo mutante vantajoso, o processo de seleção até a sua fixação por seleção custará um total de excesso de fertilidade (ou mortes genéticas) na população. Definimos então o **custo da seleção** como o **total de aumento de fertilidade** que a população teve que se impor para continuar com o mesmo tamanho ao longo de todo o processo de fixação do novo alelo vantajoso.

A ideia de Haldane é relativamente fácil de ser quantificada. Analisemos, para fins de simplificação, um modelo haplóide com dois genótipos no locus A:  $A_1$  e  $A_2$ , com os seguintes valores de *fitness*:

|                | $W_{A1}$ | $W_{A2}$ |
|----------------|----------|----------|
| <i>Fitness</i> | 1        | $1 - s$  |

Com coeficiente de seleção  $s$ , fica evidente que uma parcela  $f_{A_2}s$  de indivíduos  $A_2$  será eliminada por seleção a cada geração.

O valor adaptativo médio desta população será:

$$\bar{w} = f_{A_1}w_{A_1} + f_{A_2}w_{A_2} = f_{A_1}(1) + (1 - f_{A_1})(1 - s) = 1 - f_{A_2}s$$

O custo da seleção  $C$  em uma geração  $t$  pode ser definido pelo número de mortes genéticas do genótipo  $A_2$  comparado com o fitness médio da população:

$$C_t = \frac{f_{A_2}s}{1 - f_{A_2}s}$$

O custo total da seleção para a fixação do alelo será dado pelo somatório dos custos por geração até a fixação do alelo com maior fitness, o  $A_1$ . Nas palavras de Haldane, o custo é, portanto, **o excesso de fertilidade necessário para a seleção natural operar.**

Haldane demonstrou que o **custo total** da seleção de um alelo vantajoso até a sua fixação por seleção seria equivalente a

$$C = -\ln f_0$$

Onde  $f_0$  é a frequência inicial do alelo vantajoso (ver Apêndice para derivação da fórmula).

A derivação de Haldane, portanto, mostra que **apenas a frequência inicial** do alelo vantajoso é relevante para calcular o excesso de fertilidade total necessário para sua fixação. O coeficiente de seleção ( $s$ ) não é determinante para o custo total. Embora contraintuitivo, essa derivação é fácil de



entender – como o custo é um somatório, o valor de  $s$  apenas influenciará o tempo necessário até a fixação. Uma seleção contra o alelo desvantajoso mais intensa fará que o alelo vantajoso seja fixado mais rápido, enquanto uma seleção mais tênue resultará num tempo maior até a fixação do alelo vantajoso. Entretanto, ao final do processo, **o excesso de fertilidade total será o mesmo em ambos os casos.**

Imaginemos um caso em que um novo alelo vantajoso surge por mutação na população haplóide. Como sua frequência inicial é  $1/N$ , o custo total da seleção em uma população com o  $N = 10$  será de  $C = -\ln(1/10) = 2.302585$ . Se a população tiver tamanho de  $N = 1000$ , teremos  $C = -\ln(1/1000) = 6.907755$ . Ou seja, o total de mortes genéticas, que devem ser repostas pelo excesso de fertilidade, será de  $2.302585 \times 10 = 23.02585$  indivíduos no primeiro caso e  $6.907755 \times 1000 = 6907.755$  no segundo. **Quanto maior o tamanho da população, maior será o custo da seleção.**

Vejamos agora um código R para o cálculo do custo da seleção, ou seja, o total de excesso de fertilidade até a fixação do alelo.

Neste código, inicialmente num caso de uma população de haplóides, o número de descendentes por indivíduos reprodutores ( $r$ ) **será atualizado a cada geração para manter o tamanho da população constante.**

```
# Custo da seleção no modelo haplóide:

# fitness absolutos - existem dois alelos num locus: A (=A1) e B (=A2)
wA = 1
wB = 0.8

r = 1 # inicialmente, temos fertilidade de 1 descendente/indivíduo
g = 100

fA = fert = vector()
N = 1000 # tamanho da população

fA[1] = 1/N # vamos admitir que o alelo vantajoso acabou de surgir por mutação

# primeiro round de seleção:
nrep[1] = (fA[1] * wA * N) + ((1-fA[1])*wB * N) # número de reprodutores após seleção
fA[2] = (fA[1] * wA * N) / nrep[1] # frequência do alelo A na população de reprodutores
r = N/nrep[1] # excesso de fertilidade necessário para manter a população com N constante
fert = c(fert,r)

for(i in c(3:g)){
  nrep[i - 1] = (fA[i-1] * wA * N) + ((1-fA[i-1])*wB * N)

  fA[i] = (fA[i-1] * wA * N) / nrep[i-1]

  r = N/nrep[i-1]

  fert = c(fert,r)
}
```

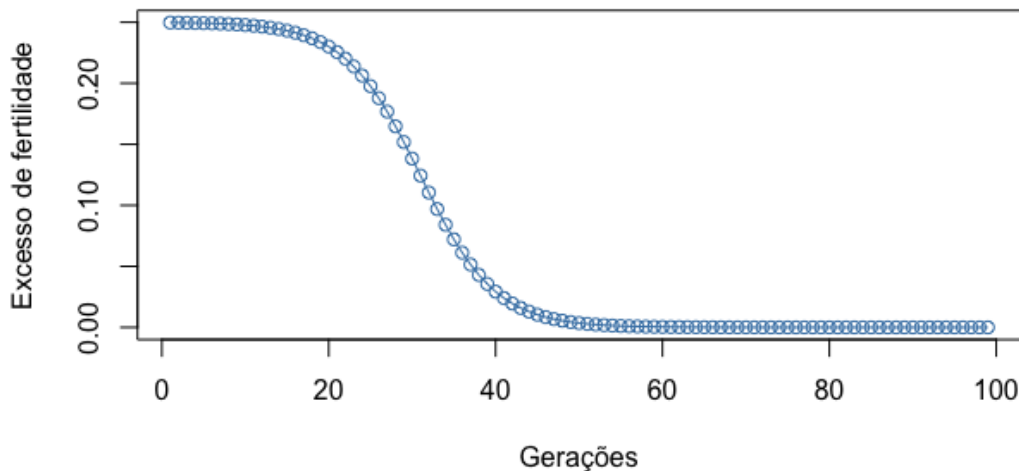
No código acima, a cada geração, a razão  $r$  será maior ou igual a 1, de forma a manter o tamanho da população constante. Por exemplo, quando  $r = 1.23$ , o excesso de fertilidade será de 0.23 descendentes por indivíduo. Os valores guardados no vetor *fert* podem ser facilmente transformados



no excesso de fertilidade calculando ( $fert - 1$ ). Fazemos o gráfico do excesso de fertilidade ao longo das gerações:

```
# gráfico:
plot(fert - 1, ylab="Excesso de fertilidade", xlab="Gerações", col="steelblue")

lines(fert -1, col="steelblue" )
```



Inicialmente, o excesso de fertilidade foi de aproximadamente 0.25 descendentes por indivíduo. Ou seja, cada indivíduo da geração parental teve que deixar 1.25 descendentes para que a população continuasse com o mesmo  $N$  após a seleção.

O custo total da seleção pode ser calculado somando todos os custos por geração:

```
sum( fert - 1 )
> 7.614268
```

Portanto, ao longo de todo o processo de fixação do novo mutante vantajoso, a população teve que produzir um excesso de 7.62 x  $N$  indivíduos, ou seja, 7614.268 indivíduos foram eliminados por seleção e tiveram que ser “repostos”.

Usando a derivação de Haldane, o valor do custo total é  $C = -\ln f_0 = -\ln\left(\frac{1}{1000}\right) = 6.907755$ .

A razão do valor da simulação ter sido diferente do valor derivado por Haldane é devido a **simulação ter sido feita em tempo discreto, enquanto a fórmula de Haldane foi calculada para um tempo contínuo** (ver Apêndice). Entretanto, os valores são razoavelmente semelhantes.

Para o caso diplóide, basta fazermos uma pequena modificação no código. Entretanto, uma simplificação deve ser feita: as frequências genotípicas a cada geração serão calculadas assumindo panmixia, ou seja, as proporções de Hardy-Weinberg. Conforme vimos anteriormente, é possível transformar esse pressuposto num modelo com amostragem de gametas (interação deriva e seleção natural).



## Vejamos o código para o modelo diplóide:

```
# Custo da seleção - modelo diplóide:

# fitness absolutos
wAA = 1
wAa = 0.8
waa = 0.6

r = 1
g = 100

fA = fert = vector()
N = 1000

# frequências genotípicas iniciais
fAA = 0
fAa = 1/N # um novo alelo vantajoso mutante (A) acabou de surgir. Inicialmente, ele está no heterozigoto.
faa = 1 - (1/N)

# primeiro round de seleção:

nrep = (N*fAA*wAA) + (N*fAa*wAa) + (N*faa*waa) # numero de reprodutores

fA[1] = ( (N*fAA*wAA)/nrep ) + ( (N*fAa*wAa)/nrep )/2 # freq A no pool gamético

fAA = fA[1]^2
fAa = 2 * fA[1] * (1-fA[1])
faa = (1-fA[1])^2

r = N/( (fAA*nrep) + (fAa*nrep) + (faa*nrep) ) ## valor de r para manter a pop estável

for(i in c(3:g)){

  nrep = (N*fAA*wAA) + (N*fAa*wAa) + (N*faa*waa) # numero de reprodutores

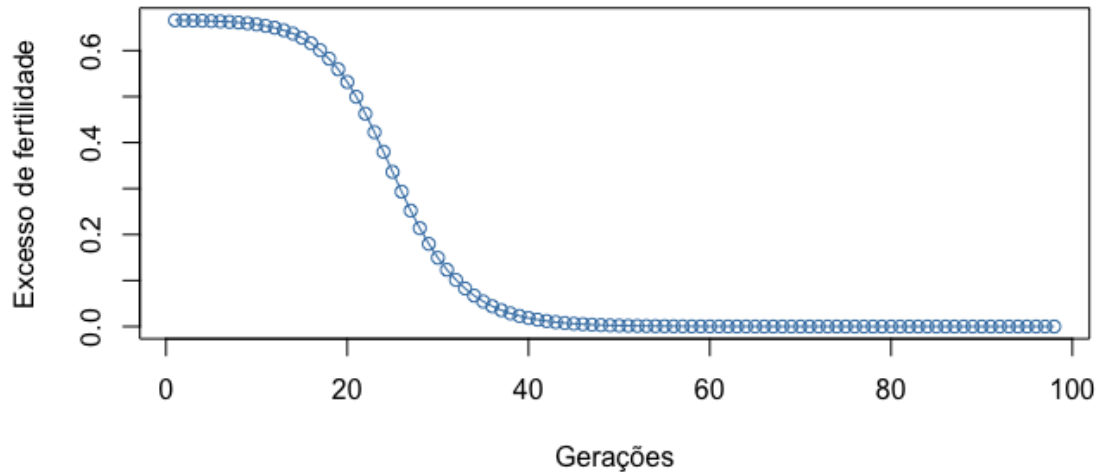
  fA[i-1] = ( (N*fAA*wAA)/nrep ) + ( (N*fAa*wAa)/nrep )/2 # frequencia de A no pool gamético

  fAA = fA[i-1]^2
  fAa = 2 * fA[i-1] * (1-fA[i-1])
  faa = (1-fA[i-1])^2

  r = N/( (fAA*nrep) + (fAa*nrep) + (faa*nrep) ) ## valor de r para manter o tamanho da pop estável

  fert = c(fert,r)
}

# gráfico
plot(fert - 1, ylab="Excesso de fertilidade", xlab="Gerações", col="steelblue")
lines(fert -1, col="steelblue" )
```



Notamos que o excesso de fertilidade para o modelo diplóide foi maior. Na primeira geração de seleção, os indivíduos com maior fitness tiveram que ter, em média 1.665927 filhos para manter a população com o mesmo  $N$ . O custo total da seleção foi de 16.5756, indicando que até a fixação do novo alelo vantajoso  $A$ ,  $16.5756 \times N$  (16575.6) indivíduos foram eliminados por seleção.

### Referência

Haldane, J.B.S. The cost of natural selection. *J Genet* 55, 511–524 (1957).  
<https://doi.org/10.1007/BF02984069>

### Apêndice

#### Derivação do custo total da seleção no modelo haplóide

Suponhamos que existem dois genótipos  $A_1$  e  $A_2$  na população, com os respectivos fitness  $w_{A_1}$  e  $w_{A_2}$ , com  $w_{A_1} > w_{A_2}$ . Desta forma, a seleção favorecerá  $A_1$ . O custo total da seleção até a fixação do melhor genótipo é obtido a seguir.

A frequência do genótipo  $A_1$  após um round de seleção é:

$$f_{A_1}^{(t+1)} = \frac{f_{A_1}^{(t)} w_{A_1}}{\bar{w}}$$

Desta forma, a variação da frequência de  $A_1$  é dada por:

$$\Delta_{f_{A_1}} = f_{A_1}^{(t+1)} - f_{A_1}^{(t)} = \frac{f_{A_1}^{(t)} w_{A_1}}{\bar{w}} - f_{A_1}^{(t)} = \frac{f_{A_1}^{(t)} (w_{A_1} - \bar{w})}{\bar{w}}$$

Assumindo que  $\bar{w} = 1 - f_{A_2} s$  e  $w_{A_1} = 1$ , a expressão pode ser simplificada em:



$$\Delta_{fA_1} = \frac{fA_1(1 - (1 - fA_2s))}{1 - fA_2s} = \frac{fA_1fA_2s}{1 - fA_2s}$$

Esta equação de diferença pode ser transformada em tempo contínuo pela seguinte diferencial quando  $\Delta_{fA_1} \rightarrow 0$ :

$$\Delta_{fA_1} = \frac{dfA_1}{dt} = \frac{fA_1fA_2s}{1 - fA_2s}$$

Guardemos esta relação acima.

Paralelamente, sabemos que o custo da seleção em uma geração é dado por:

$$C_t = \frac{fA_2s}{1 - fA_2s}$$

Como o custo total é o somatório dos custos por geração, em tempo discreto, temos:

$$C = \sum_{t=0}^{\infty} \frac{fA_2s}{1 - fA_2s}$$

Isso equivale à soma dos custos da geração 0 até o infinito (sabemos que, na prática, após a fixação o custo será zero, portanto, podemos considerar o total até o infinito). Essa mesma formulação em tempo contínuo equivale a:

$$C = \int_0^{\infty} \frac{fA_2s}{1 - fA_2s} dt$$

Anteriormente, concluímos que  $\frac{dfA_1}{dt} = \frac{fA_1fA_2s}{1 - fA_2s}$ . Desta forma,

$$dt = \frac{1 - fA_2s}{fA_1fA_2s} dfA_1$$

Substituindo  $dt$  na formulação contínua do custo, temos:

$$C = \int_{fA_1}^1 \frac{dfA_1}{fA_1} = -\ln fA_1$$

Nesta última formulação,  $fA_1$  indica a frequência inicial do alelo vantajoso.



## Carga genética, manutenção do polimorfismo genético e o debate balanceado-clássico

Na década de 1950, após a consolidação da chamada síntese evolutiva, uma parcela grande de biólogos considerava que as regiões do genoma que abrigavam loci polimórficos (com mais de um alelo) eram muito numerosas. Além disso, esses alelos eram mantidos nas populações por seleção favorável aos genótipos heterozigotos (**seleção balanceada**). A principal voz defensora desta ideia era o biólogo ucraniano-estadunidense Theodosius Dobzhansky.

Uma outra parcela dos biólogos entendia, entretanto, que os loci polimórficos eram raros, ou seja, a maioria do genoma era composto de regiões homozigotas. Quando uma região polimórfica existia, esse heterozigoto provavelmente consistia em um **alelo recessivo deletério** “escondido”. Como no caso de dominância completa  $AA$  e  $Aa$  possuem o mesmo fitness, é possível que um recessivo deletério ( $a$ ) continue segregando na população escondido no heterozigoto. Desta forma, a manutenção do polimorfismo genético estaria associada escape da **seleção negativa (purificadora)** contra o homozigoto recessivo deletério. Essa ideia era defendida por biólogos como Herman Muller.

Dobzhansky caracterizou essas duas ideias acima de escolas **balanceada e clássica**, respectivamente. Naquele momento, a estrutura química da molécula de DNA acabara de ser descoberta (1953) e não havia técnicas de sequenciamento de nucleotídeos. Desta forma, o debate entre as duas “escolas” foi realizado majoritariamente no campo teórico.

Um conceito frequentemente associado ao debate clássico vs. balanceado é o de **carga genética**. A **carga genética ( $L$ ) de uma população é quanto o fitness médio ( $\bar{w}$ ) desta se desvia do genótipo com o maior fitness**. Matematicamente:

$$L = \frac{(w_{max} - \bar{w})}{w_{max}}$$

Por exemplo, suponhamos a seguinte relação de valores adaptativos entre os genótipos:  $w_{AA} = 0.8$ ,  $w_{Aa} = 1$  e  $w_{aa} = 0.8$ . Esse é um caso de seleção balanceada. Se a população possuir 100% de indivíduos  $Aa$ , seu  $\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa} = (0)(0.8) + (1)(1) + (0)(0.8) = 1$ . O fitness do melhor genótipo ( $Aa$ ) é 1. Portanto,  $w_{max} = 1$ . Desta forma, a carga genética será:

$$L = \frac{(1 - 1)}{1} = 0$$

Evidenciando que a composição genotípica desta população está no valor máximo que a seleção natural poderá levar.

Entretanto, se existem 100% de indivíduos heterozigotos  $Aa$  na população, todos os pares reprodutivos consistirão de  $Aa \times Aa$ . Na próxima geração, teremos 25% dos filhos com genótipo  $AA$  e outros 25% com genótipo  $aa$ . Na geração dos filhos,  $\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa} = (0.25)(0.8) + (0.5)(1) + (0.25)(0.8) = 0.9$ , e

$$L = \frac{(1 - 0.9)}{1} = 0.1$$



Ne geração dos filhos, a população terá uma carga genética de 10%. Como a redução do  $\bar{w}$  ocorreu pela segregação dos alelos para formar os genótipos homozigotos, denominamos esta carga genética de **segregacional**.

Para compreendermos melhor a importância desse valor, lembremos do conceito de custo da seleção natural. Para manter o tamanho da população constante, após a eliminação de 20% dos AA (o  $w_{AA} = 0.8$ ) e 20% dos aa (o  $w_{aa} = 0.8$ ) por seleção, a fertilidade média de cada casal deverá ser aumentada para repor os indivíduos perdidos. O valor da fertilidade média necessária para repor os indivíduos perdidos será:

$$fertilidade = \frac{1}{\bar{w}}$$

Assim, quanto menor for o  $\bar{w}$ , maior será a fertilidade média necessária para manter a população com tamanho constante.

Desta forma, quando  $\bar{w}$  se afasta do valor máximo (carga genética  $L > 0$ ), há necessidade de aumento da fertilidade média.

O código R usado para calcular o custo da seleção pode ser usado para o exemplo descrito acima:

```
# Carga genética segregacional

# fitness absolutos dos genótipos (seleção balanceada)
wAA = 0.8
wAa = 1
waa = 0.8

r = 1
g = 100

fA = fert = vector()
N = 1000

# frequências genotípicas iniciais
fAA = 0.25
fAa = 0.5
faa = 0.25

# primeiro round de seleção:

nrep = (N*fAA*wAA) + (N*fAa*wAa) + (N*faa*waa) # numero de reprodutores

fA[1] = ( (N*fAA*wAA)/nrep ) + ( (N*fAa*wAa)/nrep )/2 # freq A no pool gamético

fAA = fA[1]^2
fAa = 2 * fA[1] * (1-fA[1])
faa = (1-fA[1])^2

for(i in c(3:g)){

  nrep = (N*fAA*wAA) + (N*fAa*wAa) + (N*faa*waa) # numero de reprodutores

  fA[i-1] = ( (N*fAA*wAA)/nrep ) + ( (N*fAa*wAa)/nrep )/2 # frequencia de A no pool gamético

  # frequências sem erro amostral (H-W):
  fAA = fA[i-1]^2
  fAa = 2 * fA[i-1] * (1-fA[i-1])
}
```

```

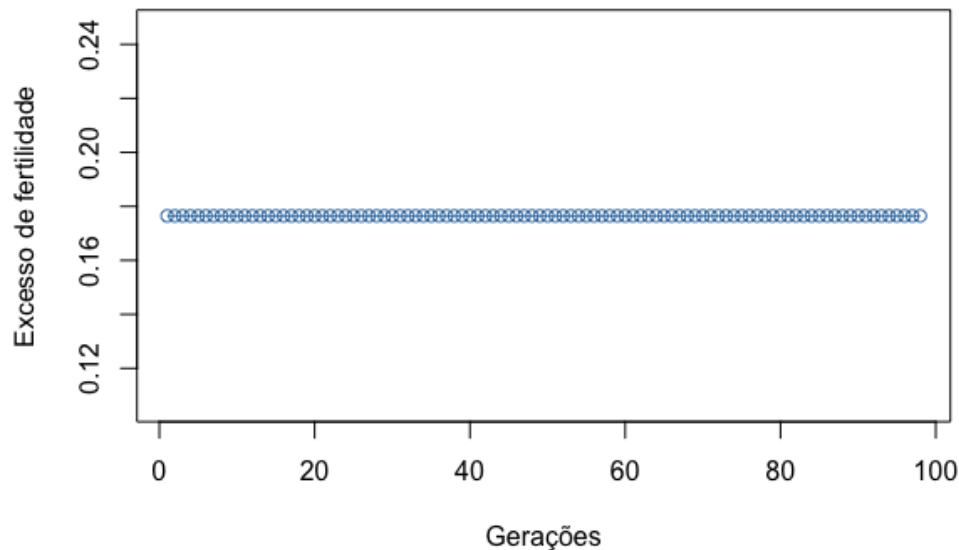
faa = (1-fA[i-1])^2

r = N/( (fAA*nrep) + (fAa*nrep) + (faa*nrep) ) ## valor de r para manter a pop estável

fert = c(fert,r)
}

# gráfico:
plot(fert - 1, ylab="Excesso de fertilidade", xlab="Gerações", col="steelblue")
lines(fert -1, col="steelblue" )

```



Notemos que, ao contrário da seleção direcional ou da seleção disruptiva, onde o excesso de fertilidade cai até atingir o valor nulo (quando o alelo vantajoso é fixado), **na seleção balanceada o excesso de fertilidade é estável ao longo das gerações.**

Essa observação fez que muitos biólogos criticassem a noção de que a maioria do polimorfismo genético seria mantido por seleção balanceada. Se isso fosse verdade, a carga genética associada às populações seria muito elevada e, conseqüentemente, o excesso de fertilidade necessário seria alto e constante.

Por outro lado, se o polimorfismo está associado a alelos deletérios escondidos nos genótipos heterozigotos, temos a seguinte relação de fitness entre os genótipos:  $W_{AA} = W_{Aa} = 1$  e  $w_{aa} = 1 - s$ . Sabemos que essa relação estabelece um regime de seleção direcional favorecendo ao alelo A. Entretanto, este não irá se fixar. Ao rodar a seleção natural por várias gerações usando a quantificação que aprendemos anteriormente, veremos que o alelo A irá ficar muito próximo da frequência de 1 (fixação). Isso indica que  $f_{AA}$  irá se aproximar de 100% com o passar das gerações. Desta forma, se  $f_{AA} \approx 1$ , temos que  $f_{Aa} \approx 0$  e  $f_{aa} \approx 0$ . Desta forma,  $\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa} = (1)(1) + (0)(1) + (0)(1 - s) \approx 1$ , indicando que **a carga genética próxima de zero**, não havendo necessidade de aumento significativo de fertilidade.

Entretanto, muitos observaram que essa situação não seria realista, pois existiria uma taxa de mutação  $\mu$  do alelo A se transformar num alelo deletério, gerando a chamada **carga genética mutacional**.



Vamos implementar um código R em que, a cada geração, no pool gamético, uma frequência  $\mu$  de alelos *A* tornem-se alelos recessivos deletérios. Vejamos como se comporta o excesso de fertilidade necessário para manter a população constante neste caso:

```
# Carga genética mutacional

# fitness absolutos dos genótipos (seleção direcional + mutação)

s = 0.6

wAA = 1
wAa = 1
waa = 1 - s

mu = 0.0001

r = 1
g = 1000

fA = fert = vector()
N = 1000

# frequências genotípicas iniciais
fAA = 0.25
fAa = 0.5
faa = 0.25

# primeiro round de seleção:

nrep = (N*fAA*wAA) + (N*fAa*wAa) + (N*faa*waa) # numero de reprodutores

fA[1] = ( ( (N*fAA*wAA)/nrep ) + ( (N*fAa*wAa)/nrep )/2 ) # frequencia de A no pool
gamético
fA[1] = fA[1] - (fA[1] * mu) # freq. Do alelo A que não mutou

fAA = fA[1]^2
fAa = 2 * fA[1] * (1-fA[1])
faa = (1-fA[1])^2

for(i in c(3:g)){

  nrep = (N*fAA*wAA) + (N*fAa*wAa) + (N*faa*waa) # numero de reprodutores

  fA[i-1] = ( ( (N*fAA*wAA)/nrep ) + ( (N*fAa*wAa)/nrep )/2 ) # frequencia de A no pool
gamético
  fA[i-1] = fA[i-1] - (fA[i-1] * mu) # freq. Do alelo A que não mutou

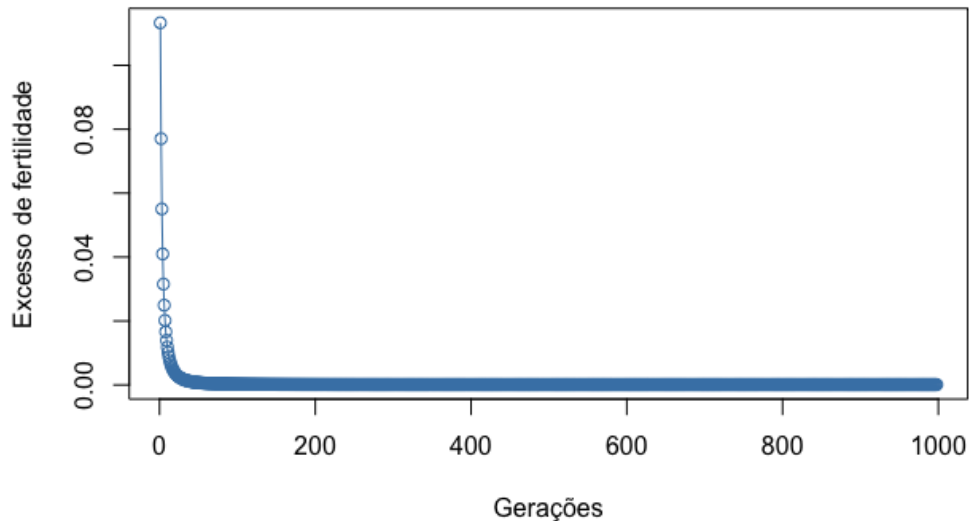
  # frequências sem erro amostral (H-W):
  fAA = fA[i-1]^2
  fAa = 2 * fA[i-1] * (1-fA[i-1])
  faa = (1-fA[i-1])^2

  r = N/( (fAA*nrep) + (fAa*nrep) + (faa*nrep) ) ## valor de r para manter a pop estável

  fert = c(fert,r)
}

# gráfico
plot(fert - 1, ylab="Excesso de fertilidade", xlab="Gerações", col="steelblue")
lines(fert -1, col="steelblue" )
```





Podemos verificar que o excesso de fertilidade necessário ficou muito próximo de zero, ele foi estabilizado no valor de 0.0001000101, indicando que a carga genética mutacional neste locus seria muito baixa.

Notamos também que a frequência do alelo *A* atinge um ponto de equilíbrio de 0.9870901. Este ponto é denominado de **equilíbrio mutação-seleção**, pois **os alelos *A* perdidos por mutação são repostos por seleção natural favorecendo o genótipo *AA***.

Este ponto pode ser calculado usando a frequência de equilíbrio do alelo deletério *a*, que é dada aproximadamente por:

$$\hat{f}_a = \sqrt{\frac{\mu}{s}}$$

Desta forma, [temos](#):

$$\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa} = (1 - \hat{f}_a)^2(1) + 2(1 - \hat{f}_a)(\hat{f}_a)(1) + \hat{f}_a^2(1 - s) = 1 - s\hat{f}_a^2$$

A carga genética mutacional para o locus com um alelo deletério recessivo será:

$$L = \frac{(w_{max} - \bar{w})}{w_{max}} = \frac{(1 - (1 - s\hat{f}_a^2))}{1} = s \left( \sqrt{\frac{\mu}{s}} \right)^2 = \mu$$

**A carga genética mutacional seria, portanto, igual a taxa de mutação deletéria.** Esse valor é menor que a carga segregacional. Desta forma, os partidários da escola clássica defendiam que seria mais realista que os polimorfismos existentes nas populações naturais fossem relacionados ao equilíbrio mutação-seleção ao invés de seleção favorável ao heterozigoto.

### Considerações históricas

A discussão teórica entre as escolas balanceada e clássica foi travada antes da possibilidade de sequenciar genomas. Na década de 1960, o estudo de isoenzimas possibilitou o início de estudos moleculares de genética de populações. Verificou-se que o polimorfismo genético em populações naturais era maior que o esperado pela escola clássica (Lewontin e Hubby, 1966). Entretanto, ele era grande demais para ser explicado completamente por seleção balanceada. Isso motivou a proposição da teoria neutra (ou neutralista) da evolução molecular por Kimura (1968) e King e Jukes (1969). Evidentemente, hoje, esse debate ganhou outra dimensão com acesso à diversidade genômica.

É importante mencionar que a carga genética mutacional será igual à taxa de mutação deletéria **apenas em casos de dominância completa**, ou seja,  $w_{AA} = w_{Aa} = 1$  e  $w_{aa} = 1 - s$ . Quando existe dominância incompleta, o cálculo deve ser feito para o caso  $w_{AA} = w_{Aa} = 1 - hs$  e  $w_{aa} = 1 - s$ . Quando  $h$  é baixo, o valor da carga genética é aproximadamente  $2\mu$ . Na prática, o valor de  $h$  fica em torno de 0.05 para diversos loci estudados (Simmons e Crow, 1977). Desta forma, é comum usar a equivalência de  $L = 2\mu$  por locus.

O valor da carga genética para todos os loci do genoma será o somatório do valor para cada locus. Se existem  $n$  loci no genoma, temos

$$L_{genoma} = \sum_{i=1}^n 2\mu_i$$

Consequentemente, temos que o fitness médio da população para todo o genoma será

$$\bar{w} = \sum_{i=1}^n (1 - 2\mu_i) \approx e^{-L}$$

Na década de 1950, Muller usou esse cálculo da carga genética mutacional para estimar o número de genes nos humanos *antes do sequenciamento de genomas*. Neste momento, se achava que a taxa de mutação deletéria era aproximadamente  $\mu = 0.00001$ . Muller assumiu que a fertilidade de excesso por casal dos humanos estaria entre 0.8 e 1 no máximo. Cada indivíduo geraria 1.8 ou 2 descendentes no máximo (~4 filhos/casal). Assim:

$$\begin{aligned} \text{fertilidade} &= \frac{1}{\bar{w}} \\ 1.8 &= \frac{1}{\bar{w}} \rightarrow \bar{w} = 0.55 \end{aligned}$$

Substituindo esse valor na formulação acima e assumindo que a taxa de mutação deletéria é idêntica para cada locus, temos:

$$\begin{aligned} \bar{w} &\approx e^{-L} \rightarrow \bar{w} \approx e^{-n2\mu} \\ 0.55 &\approx e^{-n2(0.00001)} \end{aligned}$$

O número de loci estaria por volta de 15 a 30 mil. Esse valor é uma aproximação razoável do valor que conhecemos hoje.

Por fim, devemos ressaltar que os defensores da escola balanceada argumentavam que se o número de alelos por locus for muito grande, a carga genética associada seria muito baixa. No exemplo que estudamos acima, onde apenas dois alelos ( $A$  e  $a$ ) existiam no locus, se as frequências de cada um dos alelos forem idênticas ( $f_A = f_a = 0.5$ ), temos  $(0.5^2 + 0.5^2) = 50\%$  homozigotos com reduzido



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

formados. Quando o número de alelos sobe para 5 (A1, ..., A5) temos um total de  $(5 \times (1/5)^2) = 20\%$  de homozigotos com fitness reduzido. Com 20 alelos segregando num locus, apenas 5% dos indivíduos formados terão fitness reduzido (os heterozigotos possuem o maior fitness). Desta forma, a carga genética da seleção balanceada seria atenuada. Esse é um dos mecanismos que supostamente mantém a alta diversidade genética nos loci MHC. Em 1964, entretanto, os geneticistas teóricos Motoo Kimura e James Crow demonstraram que, **em populações finitas, o número de alelos que podem ser mantidos num locus não poderia ser tão alto conforme defendido pelos teóricos da escola balanceada**, notavelmente Dobzhansky.

## Referências

- Dobzhansky T. (1955) A review of some fundamental concepts and problems of population genetics. Cold Spring Harb Symp Quant Biol. 20:1-15.
- Lewontin RC, Hubby JL. (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics. 54(2):595-609.
- Kimura, M (1968). "Evolutionary rate at the molecular level". Nature. 217 (5129): 624–6
- Kimura M, Crow JF. (1964) The Number Of Alleles That Can Be Maintained In A Finite Population. Genetics. 49(4):725-38.
- King, J. L.; Jukes, T. H. (1969). "Non-Darwinian evolution". Science. 164 (3881): 788–98



## Destino das mutações I: Tempos médios para fixação e perda de novos alelos mutantes

Uma pergunta importante em genética de populações é o tempo que leva para um novo alelo mutante ser fixado ou perdido. Esses valores foram calculados por Kimura e Ohta em 1969. Através da aplicação da modelagem da dinâmica estocástica dos alelos usando cadeia de Markov em tempo contínuo conforme descrita pelas [equações de Kolmogorov](#), após o trabalho do matemático russo [Andrey Kolmogorov](#).

Em genética de populações, as [equações progressivas \(forward\) de Kolmogorov](#), que são também conhecidas como equações Fokker-Planck em Física, foram inicialmente estudadas por Sewall Wright em 1945 e posteriormente por Kimura em 1955. As [equações regressivas \(backward\) de Kolmogorov](#) foram introduzidas em genética de populações por Motoo Kimura em 1957.

Aplicando esse arcabouço teórico, Kimura e Ohta (1969) demonstraram que o tempo **médio** até a fixação de um novo alelo mutante **neutro** é dado aproximadamente por:

$$\bar{t}_{fixação} \approx 4N$$

Por outro lado, o tempo médio até a perda de um novo alelo mutante neutro é dado aproximadamente por:

$$\bar{t}_{perda} \approx 2\log(2N)$$

Em ambas as equações, assumimos que  $N = N_e$ . Por exemplo, se a população possui tamanho efetivo de  $N = 10$ , o tempo médio que levará para o novo alelo mutante neutro ser fixado será de 40 gerações. Por outro lado, o tempo médio até a sua perda será de aproximadamente 6 gerações.

No código abaixo, estudaremos esse processo 10000 vezes independentemente (replicatas). Em cada replicata registraremos na variável *tfix* o tempo (em gerações) caso o novo alelo mutante neutro ter sido fixado ou em *tloss* o tempo que levou caso ele tenha sido perdido. Posteriormente, faremos um histograma de cada uma dessas variáveis que são vetores. Usaremos a função **break** para parar o *loop* das gerações caso tenha ocorrido fixação ou perda. A simulação é **essencialmente a mesma** daquela que estudamos nas populações Wright-Fisher.

```
# Tempos de fixação e perda de alelos neutros:

N = 10
g = 500
replicates = 10000

states = c(0:(2*N))
n_m = f_m = matrix(nrow = replicates, ncol=g) # matrizes que registram o número de cópias
do alelo mutante e sua frequência respectivamente

tfix = tloss = vector()

for(r in c(1:replicates)){
  n_m[r,1] = 1 # inicialmente, todas as replicatas começam com apenas 1 cópia do alelo
mutante neutro
  f_m[r,1] = n_m[r,1]/(2*N) # frequência inicial do novo alelo mutante neutro - 1/2N
  stopfix = stoploss = 0
```

```
for(i in c(2:g)){
  n_m[r,i] = sample(c(0:(2*N)), 1, prob=dbinom(states, (2*N),prob = f_m[r,(i-1)] ))
  if(n_m[r,i] == 0 & stoploss == 0){
    tloss = c(tloss, i)
    stoploss = 1
  }
  if(n_m[r,i] == (2*N) & stopfix == 0){
    tfix = c(tfix,i)
    stopfix = 1
  }
  f_m[r,i] = n_m[r,i]/(2*N)
  if(stopfix == 1 | stoploss == 1){
    break # se ocorreu fixação ou perda, interromper o loop das gerações nesta replicata e
    passar para a próxima
  }
}
}
```

Calculemos agora o tempo médio até a fixação no novo alelo usando todas as replicatas:

```
> mean(tfix)
[1] 39.29741
```

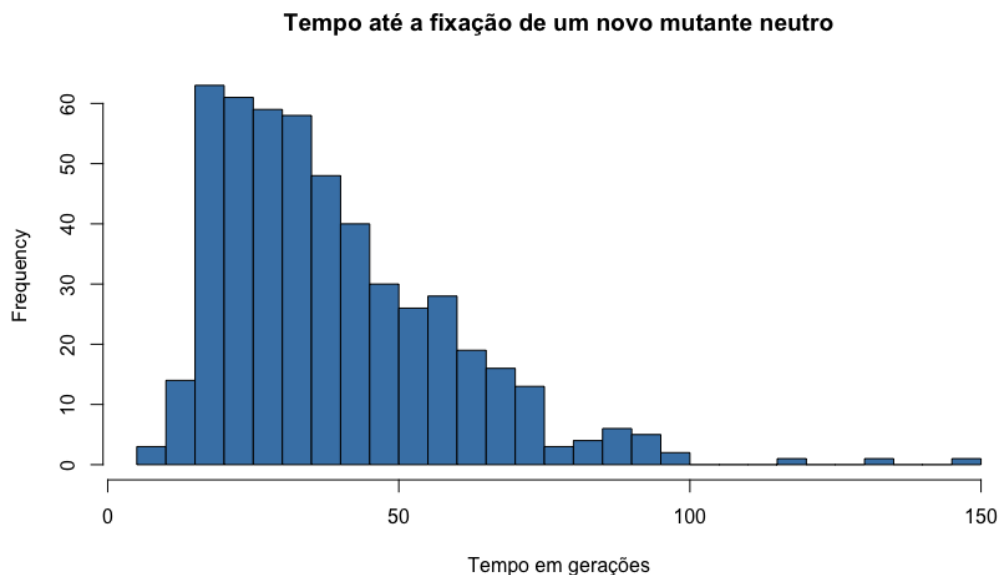
E o tempo médio até a perda:

```
> mean(tloss)
[1] 6.535846
```

Note que esses valores são muito próximos das expectativas teóricas, de 40 gerações para fixação e 6 gerações para a perda.

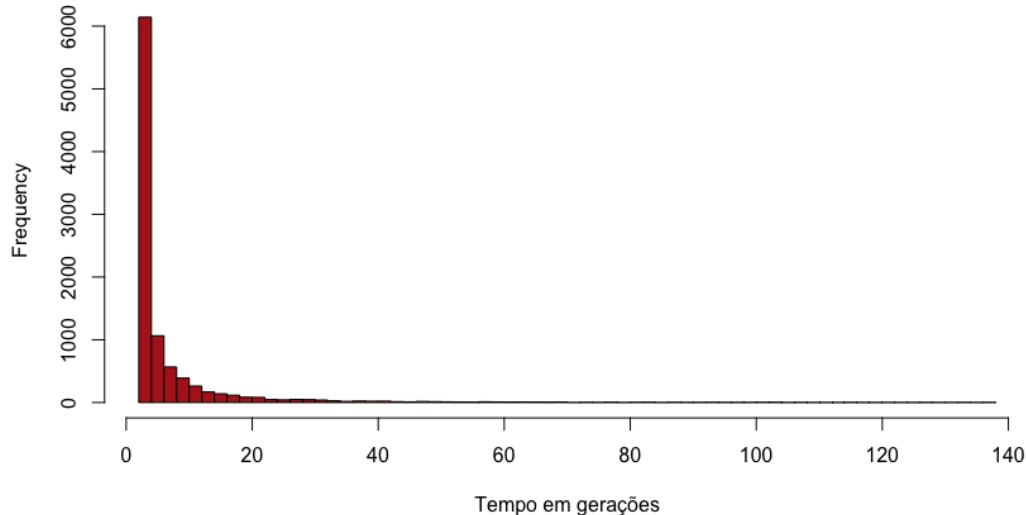
Façamos agora os histogramas *tfix* e *tloss*:

```
hist(tfix, nclass=50, main = "Tempo até a fixação de um novo mutante neutro", xlab="Tempo em
gerações", col="steelblue")
```



```
hist(tloss, nclass=50, main = "Tempo até a perda de um novo mutante neutro", xlab="Tempo em gerações", col="firebrick")
```

### Tempo até a perda de um novo mutante neutro



Algumas propriedades importantes podem ser observadas nos gráficos acima. Primeiramente, embora existam expectativas teóricas para a média do tempo de fixação e a média do tempo de perda, **a variância desses valores é grande**. Isso é particularmente verdadeiro para o tempo de fixação. Em média, dizemos que ele é aproximadamente  $4N$ . Entretanto, valores menores ou maiores que  $4N$  (40 no nosso caso) são comumente encontrados.

Por outro lado, o tempo até a perda do novo alelo mutante neutro possui uma distribuição **aproximadamente exponencial**. A maior concentração de valores está próxima de 1 geração, enquanto valores grandes de tempo ( $> 40$  gerações) ficam **cada vez mais improváveis**. Desta forma, é mais fácil prever o tempo de perda, embora exista uma longa cauda na curva característica da distribuição exponencial.

Para fins práticos, é seguro afirmar que **a maior parte dos novos mutantes neutros segregam na população por poucas gerações até ser perdido**.

### Referências

- Kimura, M., 1957. Some Problems of Stochastic Processes in Genetics. *The Annals of Mathematical Statistics* 28, 882–901.
- Kimura, M., 1955. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences* 41, 144–150.
- Kimura, M., Ohta, T., 1969. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics* 61, 763–771.
- Wright, S., 1945. The Differential Equation of the Distribution of Gene Frequencies. *Proceedings of the National Academy of Sciences* 31, 382–389.

## Destino das mutações II: Origem do polimorfismo entre espécies e probabilidades de fixação

Quando regiões homólogas do genoma de indivíduos de uma espécie ou de diferentes espécies são sequenciadas, é comum observarmos algum polimorfismo. Anteriormente, verificamos que, dentro de uma espécie, uma das métricas usadas para quantificar o grau de polimorfismo é frequência de heterozigotos. Sabemos que, em equilíbrio mutação-deriva, e assumindo alelos neutros, esse valor será  $\hat{H} = \frac{4N_e\mu}{4N_e\mu+1}$ .

Entretanto, quando comparamos espécies diferentes ou mesmo populações isoladas geneticamente há algumas gerações, a quantificação do polimorfismo deve ser outra, pois é comum que as espécies comparadas **não compartilham alelos em comum**. Desta forma, a frequência de heterozigotos em cada uma delas é irrelevante para compreendermos o quanto de variação genética existe **entre** elas.

Por exemplo, analisemos uma região do gene *FAT4*, responsável pela expressão de uma proteína de membrana, em três de mamíferos (cão, homem e chimpanzé):

|                  |  |
|------------------|--|
| Canis_familiaris | GCGGGACCTCAACGACAACGCGCCGGTCTTCCCCGATCCCTCCATCGTGG |
| Homo_sapiens     | GCGGGACCTCAATGACAACGCCCCCGTTTTCCCGGACCCCTCTATCGTGG |
| Pan_troglodytes  | GCGGGACCTCAATGACAACGCTCCCGTTTTCCCGGACCCCTCTATCGTGG |

Como podemos observar, cada uma dessas espécies possui alelos **exclusivos** no locus *FAT4*. A maioria dos sítios da sequência de nucleotídeos desse gene é idêntica entre as três espécies. Entretanto, observamos alguns sítios que apresentam polimorfismo. Destes, alguns são compartilhados entre as espécies, enquanto outros são exclusivos de cada espécie.

Mas como caracterizar esse polimorfismo entre espécies?

Em genética evolutiva e evolução molecular, entendemos que estas diferenças entre as espécies são oriundas de processos de **substituição de uma forma alélica por outra ao longo do processo evolutivo ocorrendo nas diversas populações**.

Essa substituição de uma forma alélica por outra em uma população indica que **uma das formas alélicas foi fixada**, enquanto as outras foram perdidas.

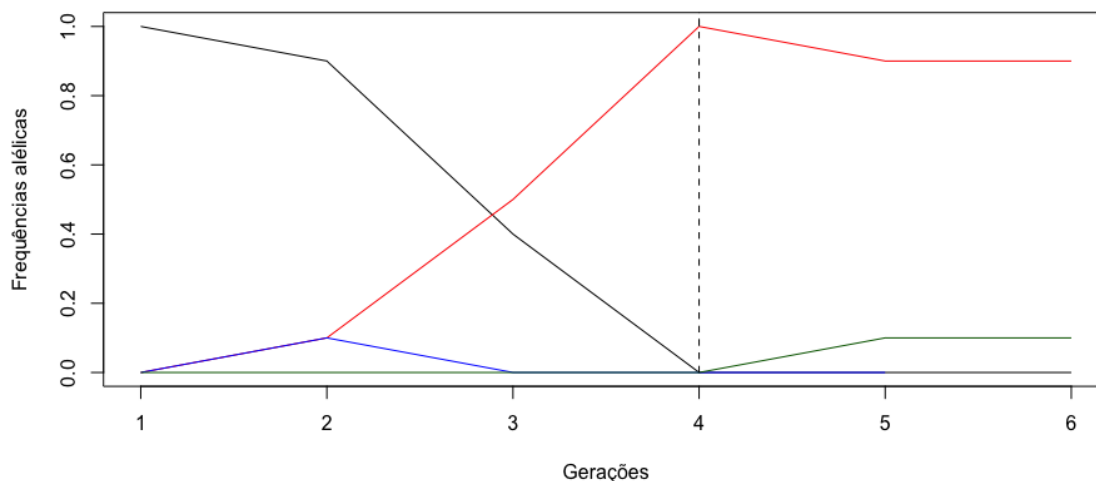
Assim, fazemos uma equivalência entre **substituição** e **fixação**.

Vamos esse processo no seguinte esquema a seguir, onde mostramos a composição alélica de uma população da geração 0 até a geração 5. Inicialmente, todos os 10 indivíduos da população possuíam a forma alélica com a sequência "ATGCCGT". Em G1, ocorreu uma **mutação** de C para G (C>G) no quarto sítio originando o alelo "ATG**G**CGT". Esse alelo aumentou de frequência e, G2. Em G2, também ocorreu uma mutação de A>T no primeiro sítio de uma das cópias alélicas ("CTGCCGT"). Esse novo alelo, entretanto, foi perdido na geração seguinte, G3. Nesta mesma geração, o alelo "ATG**G**CGT" foi **fixado**. Portanto, ocorreu a **substituição** do alelo "ATGCCGT" pelo alelo "ATG**G**CGT". Em G4, uma outra mutação aconteceu de G>A no sexto sítio de uma das cópias alélicas, resultando no alelo "ATG**G**CAT". Esse alelo ainda está segregando em G5.

*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
 Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)



Esse processo todo pode ser sumarizado usando o gráfico de frequências alélicas abaixo. Na quarta geração (G3), marcada com linha pontilhada, **ocorreu a substituição alélica**, pois o alelo vermelho foi **fixado**.



Pela dinâmica exposta acima, fica evidente que **mutação e substituição são processos diferentes** (embora frequentemente confundidos). Também é evidente que, na maior parte do tempo, a população analisada possui mais de um alelo segregando, resultando numa variação genética persistente (**standing genetic variation**).

Desta forma, ao comparar as sequências dos alelos derivados ainda segregando com o alelo ancestral, temos o seguinte alinhamento de sequências:





```
Alelo_ancestral      ATGCCGT
Alelo_derivado_1    ATGCGT
Alelo_derivado_2    ATGCAT
```

Neste alinhamento, o G é resultado de uma substituição, enquanto o A é **uma mutação que ainda não passou por substituição**. Quando comparamos alelos em diferentes espécies, como é o caso da figura acima, **frequentemente assumimos que todas as mudanças são substituições**, pois o tempo de separação das espécies geralmente é grande, diminuindo a probabilidade de que mutações que ainda não foram fixadas sejam observadas.

### Probabilidade de fixação de um novo alelo mutante

Conforme estudamos anteriormente, sabemos que um novo alelo mutante neutro tem probabilidade de fixação igual a  $1/2N$ . Entretanto, se o novo alelo mutante é vantajoso, conferindo àqueles que o possuem uma vantagem  $s$ , a sua probabilidade de fixação será maior.

Em 1927, Haldane, e posteriormente Fisher (1930) e Wright (1931), demonstraram que a probabilidade de fixação de um novo alelo mutante vantajoso seria aproximadamente

$$P_{fixação} \approx 2s$$

Note que, curiosamente, a probabilidade de fixação de um novo mutante vantajoso é **independente do tamanho da população**. Essa derivação foi subsequentemente detalhada por Kimura (1962), que encontrou uma fórmula geral para a probabilidade de fixação de alelos vantajosos numa frequência  $p$  **qualquer** na população. Se o alelo vantajoso confere uma vantagem  $s$  ao homocigoto e  $s/2$  ao heterocigoto (modelo aditivo), temos:

$$P_{fixação} \approx \frac{1 - e^{-4spN}}{1 - e^{-4sN}}$$

Uma vez que a frequência inicial de um novo alelo mutante sempre será  $1/2N$ , se usarmos  $p = 1/2N$  na fórmula acima:

$$P_{fixação} \approx \frac{1 - e^{-4s(\frac{1}{2N})N}}{1 - e^{-4sN}} = \frac{1 - e^{-2s}}{1 - e^{-4sN}}$$

Para fins práticos, tanto a fórmula de Haldane (1927) e Kimura (1962) são usadas, pois ambas resultam em valores semelhantes. Nas palavras de Kimura (1962):

“For a positive  $s$  and very large  $N$  we obtain the known result that the probability of ultimate survival of an advantageous mutant gene is approximately twice the selection coefficient (Haldane 1927). On the other hand, if we let  $s \rightarrow 0$ , we obtain  $U = 1/2N$ , the result known for a neutral gene.”

Comparemos as duas formulações:

```
# Haldane (1927) vs. Kimura (1962):
```

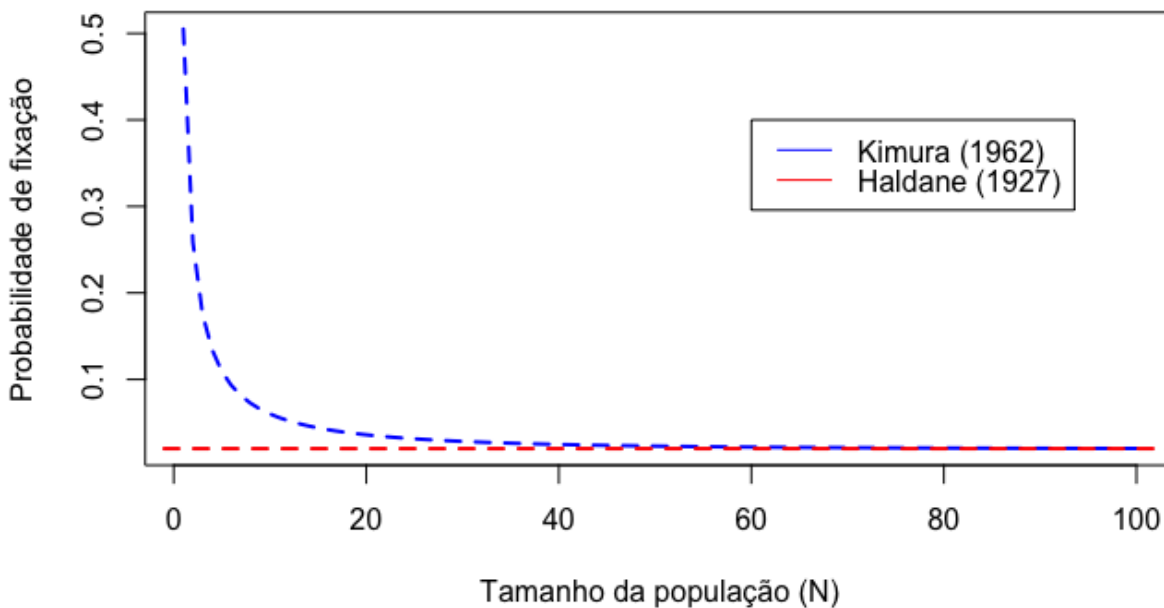
```
s = 0.01
```

```
prob_fix_haldane = 2*s # prob fix de um novo mutante vantajoso por Haldane 27
```

```
# definição da função de prob fix de um novo mutante vantajoso por Kimura 62:
prob_fix_kimura = function(s,N){
  return( (1 - exp(-2*s))/(1 - exp(-4*s*N)) )
}

# gráfico
curve(prob_fix_kimura(s=s, N=x), xlim=c(1,100), col="blue", lty=2, ylab="Probabilidade de
fixação", xlab="Gerações", lwd=2)
segments(-1,prob_fix_haldane,120, prob_fix_haldane, col="red", lty=2, lwd=2)

legend(60,0.4,legend=c("Kimura (1962)", "Haldane (1927)"), col=c("blue", "red"), lty=1)
```



O gráfico deixa evidente que, **conforme o tamanho da população aumenta, a fórmula de Kimura se aproxima do valor  $2s$ , calculado por Haldane.** Entretanto, para populações muito pequenas (aprox. <20), existe uma diferença significativa entre as duas predições.

### Taxa de substituição alélica

A diferença entre mutações e substituições apresentada acima motivou que os geneticistas evolutivos tentassem calcular a **taxa de substituição alélica**, pois esse valor é importante para entender a **taxa de divergência genética entre as espécies**.

Kimura (1968) e King e Jukes (1969) utilizaram o seguinte raciocínio para obtenção desta taxa. Se taxa de mutação por geração é  $\mu$ , a cada geração uma população de indivíduos diplóides com  $N$  indivíduos produz um total de  $2N\mu$  novos alelos.

Assumindo que esses  $2N\mu$  novos alelos são neutros quando comparados aos que estão segregando na população ( $s = 0$ ), cada um deles terá chance de fixação de  $1/2N$ , pois eles se fixarão por ação de deriva genética.



Desta forma, a cada geração, a taxa de fixação (=substituição alélica) será:

$$k = 2N\mu \times \frac{1}{2N} = \mu$$

Ou seja, **se os novos mutantes alelos forem neutros, a taxa de substituição será igual à taxa de mutação**. Desta forma, a taxa de substituição, que em última análise determinará a divergência genética entre as espécies **não depende do tamanho populacional**.

Esse raciocínio foi um dos fundamentos teóricos usados para explicar o **relógio molecular**, a observação originalmente de Zuckerkandl e Pauling (1962) de que **a divergência genética entre as espécies é acumulada de forma aproximadamente linear**. A explicação teórica do relógio molecular foi um dos primeiros resultados da **teoria neutra (ou neutralista) da evolução molecular** (Kimura, 1983).

No início da década de 1970, a geneticista japonesa Tomoko Ohta, que trabalhava na época sob orientação de Motoo Kimura, investigou situações em que um novo alelo vantajoso se comportaria efetivamente como um alelo neutro. Essa situação pode ser matematicamente estabelecida igualando as probabilidades de fixação de novos alelos mutantes neutros com a probabilidade de fixação de novos alelos vantajosos:

$$2s = \frac{1}{2N}$$

Consequentemente, quando o produto  $4Ns = 1$ , temos que a dinâmica evolutiva desse novo alelo mutante que, embora não seja neutro, seja fundamentalmente a mesma da dinâmica de um alelo neutro (Kimura e Ohta, 1971).

Posteriormente, Ohta usou essa observação como uma forma de resolver o chamado **paradoxo de Lewontin**, que diz respeito sobre a pequena diferença entre heterozigosidades de espécies que possuem tamanhos populacionais muito diferentes (lembre-se da fórmula que prevê a heterozigosidade de equilíbrio dentro das espécies).

Ohta observou que a **taxa de mutação de novos alelos levemente deletérios** poderia ser maior que o considerado originalmente por Kimura. Ela mostrou que **alelos levemente deletérios poderiam segregar como se fossem alelos neutros (escapando da seleção natural) em populações pequenas**. Por exemplo, se o novo alelo é levemente desvantajoso ( $s = 0.001$ ), uma população com  $N = 10$ , terá o produto  $4Ns = 0.04$ . Esse valor é próximo de 1 e esse alelo poderá eventualmente se fixar por deriva por se comportar semelhante a um alelo neutro. Entretanto, se a população tem  $N = 1000$ ,  $4Ns = 4$ , indicando que ocorrerá forte seleção negativa contra o alelo. A formulação de Ohta é conhecida como **nearly neutral theory of molecular evolution (teoria quasi neutra)** (Ohta, 1973).

## Referencias

- Haldane, J.B.S., 1927. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society* 23, 838–844.
- Kimura, M. (1962) On the Probability of Fixation of Mutant Genes in a Population. *Genetics*, 47, 713–719.
- Kimura, M. and Ohta, T. (1971) Protein Polymorphism as a Phase of Molecular Evolution. *Nature*, 229, 467–469.



- Kimura, M., 1968. Evolutionary Rate at the Molecular Level. *Nature* 217, 624–626.
- King, J.L., Jukes, T.H., 1969. Non-Darwinian Evolution. *Science* 164, 788–798.
- Ohta, T., 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* 246, 96–98.
- Zuckerkandl E, Pauling (1962). "Molecular disease, evolution, and genic heterogeneity". In Kasha, M., Pullman, B (eds.). *Horizons in Biochemistry*. Academic Press, New York. pp. 189–225.

## Apêndice

```
# Código para simular a dinâmica de alelos mutantes em uma população.
# Carlos G. Schrago, 2022.

alelos = c(1) # começamos com apenas 1 alelo, denominado de "alelo 1"
N = 50 # tamanho populacional efetivo
g = 2000 # numero de gerações
mu = 0.01 # taxa de mutação

freq = matrix(nrow=length(alelos), ncol=g) # matriz de frequências - as colunas são as
# gerações e as linhas são os alelos.

freq[length(alelos),1] = 1 # frequência do "alelo 1" = 100% na primeira geração

if(runif(1) <= mu){ # vai ganhar um novo alelo mutante
  freq = rbind(freq, c(1/(2*N), rep(0, (g - 1))))
  freq[(length(alelos)), 1] = 1 - (1/(2*N))
  alelos = c(alelos, (alelos[length(alelos)] + 1))
}

freq[is.na(freq)] = 0

for(i in c(2:g)){
  # amostra os alelos do pool:
  x = sample(alelos, 2*N, replace=T, prob=freq[, (i-1)])

  # pergunta se cada um dos alelos irá mutar
  for(j in c(1:length(x))){
    if(runif(1) <= mu){ # esse gameta sofrerá mutação
      alelos = c(alelos, (alelos[length(alelos)] + 1))
      x[j] = alelos[length(alelos)]
      freq = rbind(freq, rep(0,g))
    }
  }

  # calcula as frequências da geração i
  for(j in c(1:length(alelos))){
    freq[j,i] = length(which(x == j))/(2*N)
  }
}

plot(c(-1),xlim=c(1,g), ylim=c(0,1), xlab="Frequência dos alelos", ylab="Gerações")
cores = terrain.colors(length(alelos))
```



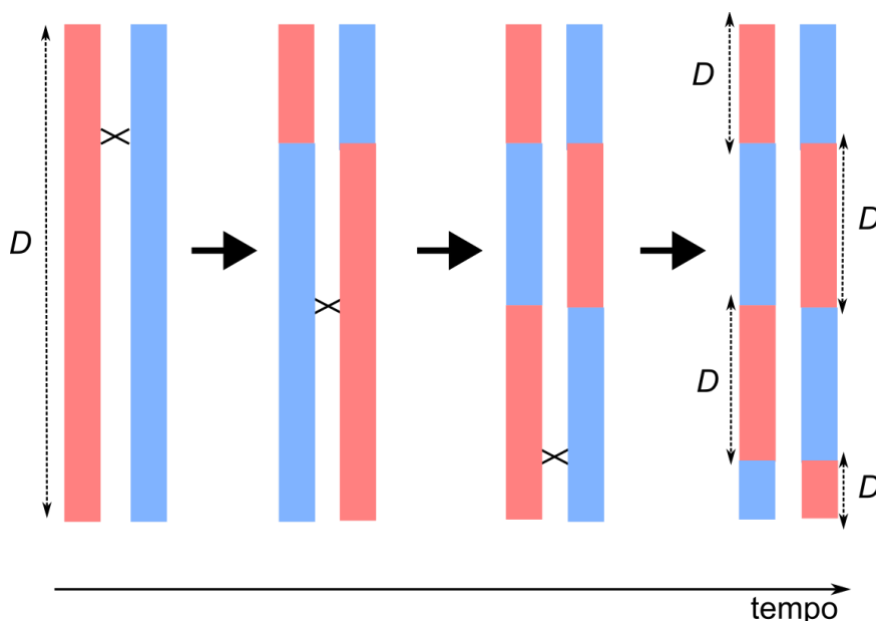
```
for(i in c(1:length(alelos))){  
  lines(freq[i,], col=cores[i])  
}  
  
## numero de fixações:  
if(length(alelos) > 1){  
  n = 0  
  for(i in c(2:length(alelos))){  
    if( (length(which(freq[i,] == 1)) != 0 ) == TRUE){  
      n = n + 1  
    }  
  }  
}
```

## Evolução multi-loci: O relógio da recombinação

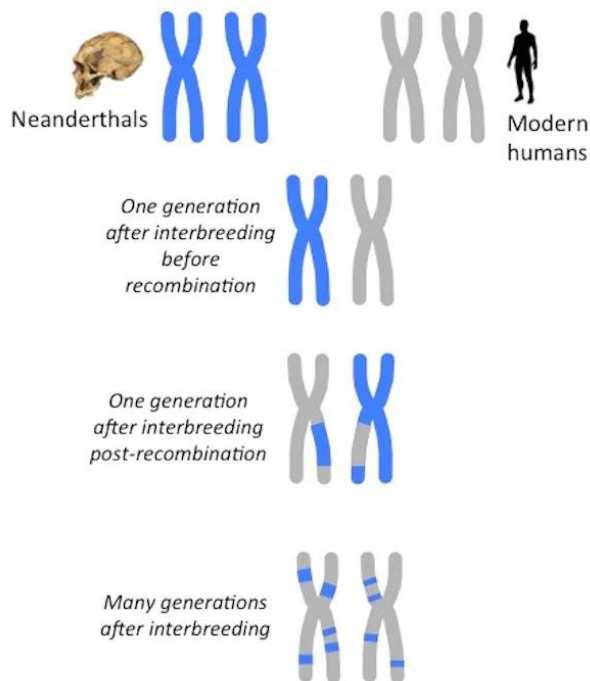
Em genética de populações, quando analisamos mais de um locus do genoma, devemos considerar que eles podem estar fisicamente localizados em regiões próximas a ponto de ocorrer **desequilíbrio de ligação** entre os alelos que segregam em cada um desses loci. Neste caso, é comum o uso do termo **haplótipo** para descrever cada uma das possibilidades de combinação de alelos nos loci estudados.

O **desequilíbrio de ligação** é em essência uma medida de associação não randômica entre os alelos segregando nos loci no momento da formação dos gametas que vão compor *pool* gamético. Existem diversas métricas que, de maneiras diferentes, tentam quantificar o grau desta associação não randômica em escala genômica.

Como os eventos de crossing-over quase sempre ocorrem entre pelos menos um dos braços de pares de cromossomos homólogos durante meiose, **a recombinação ocorre numa taxa aproximadamente constante ao longo das gerações**. Isso resulta numa propriedade interessante: o tamanho dos blocos cromossômicos que compartilham a mesma origem vai diminuindo ao longo do tempo. Essa propriedade é chamada de relógio de recombinação. Na figura abaixo, observamos que como o tamanho dos blocos cromossomos de origem única ( $D$ ) diminuem a cada evento de recombinação. Os blocos haplotípicos originais ficarão cada vez menores:



As aplicações desta propriedade são várias. Por exemplo, usando esta propriedade, [Moorjani et al. \(2016\)](#) calcularam que o evento de introgressão entre humanos e neandertais. Conforme esquematizado na figura abaixo, os blocos cromossômicos azuis (de origem neandertal) irão diminuir de tamanho com o passar das gerações:



Fonte: <https://tinyurl.com/4wt83db3>

Vejamos agora ver a dinâmica deste decaimento usando um código R. Neste código, a população começa com  $N_c$  cromossomos diferentes. Cada cromossomo tem tamanho de  $l$  de sítios. A cada geração, metade dos  $N_c$  cromossomos são escolhidos para sofrer recombinação. A recombinação ocorrerá com uma taxa  $c$  numa posição cromossômica aleatoriamente escolhida. Após os eventos de recombinação, verificaremos os tamanhos (em número de sítios) dos cromossomos originais no total de cromossomos da população. Esses valores ficarão registrados num vetor que ocupará uma posição (correspondente à geração  $g$ ) da lista `len_bloco`. Ao final, plotaremos a média dos tamanhos relativos dos blocos originais em cada uma das gerações. O tamanho relativo simplesmente indica o quanto existe do tamanho original do bloco. Evidentemente, o valor inicial é igual a 1, pois os cromossomos estão 100% sem recombinação.

```
# Decaimento do tamanho dos blocos de ligação:

N = 100 # numero de individuos
Nc = 2*N # número de cromossomos
c = 0.5 # taxa de recombinação entre os cromossomos
l = 100 # tamanho (número de sítios) no cromossomo
g = 100 # número de gerações

len_bloco = list()

cromossomos = matrix(nrow=Nc, ncol=1)
j = 1
for(i in c(1:Nc)){
  cromossomos[i,] = rep(j,l)
  j = j + 1
}

len_bloco[[1]] = 1

for(i in c(2:g)){
  # escolhe aleatoriamente Nc/2 cromossomos para recombinar:
  recombinantes = sample(c(1:Nc), Nc/2, replace = F)
  for(j in c(1 : (length(recombinantes) - 1) )){
```



*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

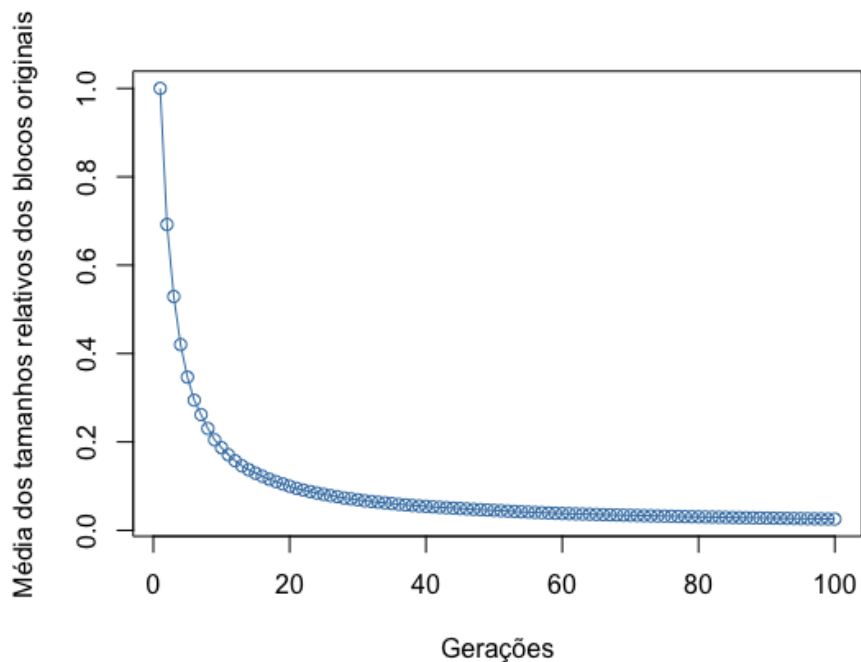
```
if(runif(1) < c){ # vai recombinar?
  # escolhe sítio de recombinação:
  x = round(runif(1,min=1, max=1) )
  y = cromossomos[recombinantes[j], c(1:x)]
  cromossomos[recombinantes[j], c(1:x) ] = cromossomos[recombinantes[(j + 1)],c(1:x) ]
  cromossomos[recombinantes[(j + 1)], c(1:x) ] = y
}
}
# verifica o tamanho dos blocos originais:
lens = vector()
for(j in c(1:Nc)){
  bloco = 0
  for(k in c(2:1)){

    if(cromossomos[j,k] == cromossomos[j, (k - 1)]){
      bloco = bloco + 1
    }
    else{
      lens = c(lens, (bloco + 1)/l )
      bloco = 0
    }
  }
  lens = c(lens, (bloco + 1)/l)
}
len_bloco[[i]] = lens
}

mean_len = vector()
for(i in c(1:g)){
  mean_len = c(mean_len, mean(len_bloco[[i]]))
}

#gráfico
plot(mean_len, xlab="Gerações", ylab="Média dos tamanhos relativos dos blocos originais",
col="steelblue")

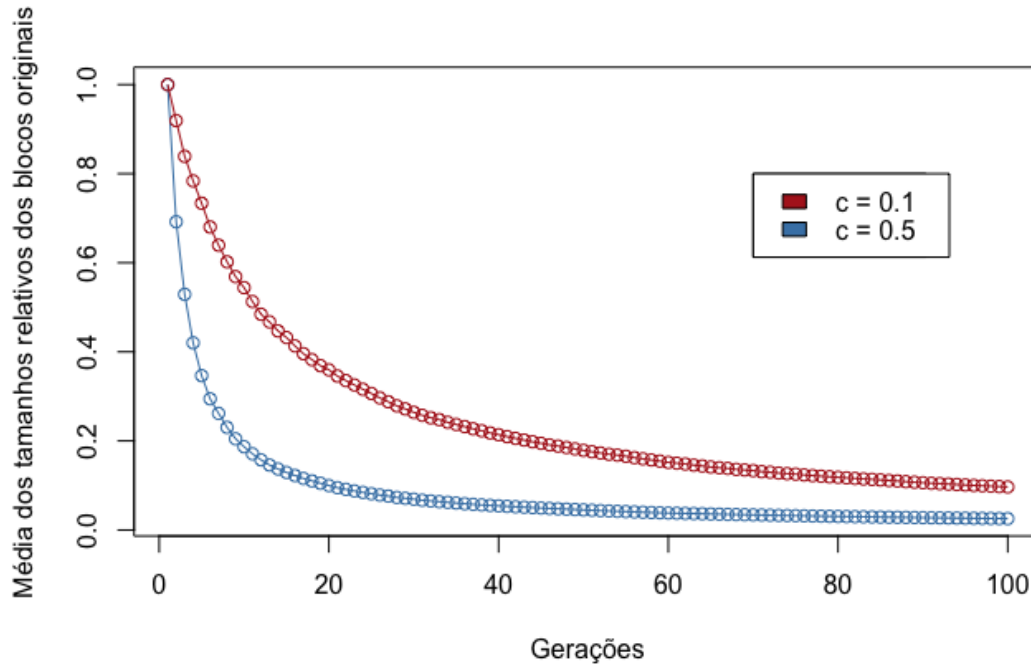
lines(mean_len, col="steelblue")
```





Conforme podemos verificar no gráfico acima, os tamanhos dos blocos cromossômicos originais decaem ao longo do tempo de forma aproximadamente exponencial. Desta forma, diferentemente do relógio molecular, o relógio de recombinação não é linear. Tentemos agora diminuir a taxa de recombinação de 0.5 para 0.1:

`c = 0.1 # taxa de recombinação entre os cromossomos`

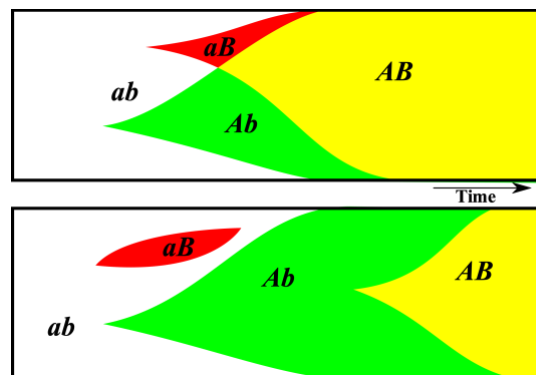


Conforme esperado, quanto menor for a taxa de recombinação, mais lento será o decaimento do tamanho médio dos blocos cromossômicos originais.

## Evolução multi-loci: Efeito da seleção no polimorfismo genômico com ligação genética

Quando um novo alelo mutante surge na população, ele surge numa região específica do genoma. **A dinâmica deste novo alelo mutante irá influenciar a dinâmica das regiões próximas a ele.** O quão próximo será justamente definido pela taxa de recombinação – quanto maior a taxa de recombinação, maior será a independência evolutiva das regiões genômicas. Num caso extremo, se a taxa de recombinação for nula, o cromossomo inteiro será influenciado pela dinâmica de um único mutante.

Em casos em que a taxa de recombinação é muito baixa (ou mesmo ausente), mutações vantajosas no cromossomo que surgem em indivíduos diferentes nunca serão combinadas por recombinação para compor um cromossomo que possua todo o conjunto de alelos vantajosos nos diferentes genes. Inclusive, os diferentes indivíduos que carregam esses alelos irão competir. Esse fenômeno é conhecido como **interferência Hill-Robertson** (1966). Portanto, a recombinação permite que mutações vantajosas se encontrem no mesmo cromossomo (quadro superior da figura abaixo). Caso contrário, será necessário que as mutações vantajosas ocorram independentemente no mesmo cromossomo, aumento o tempo necessário para que o genótipo superior surja (quadro inferior da figura abaixo):



Além disso, na ausência de recombinação, mutações levemente deletérias irão se acumular independentemente ao longo dos cromossomos presentes nos diferentes indivíduos. Se, em algum momento, esta população de indivíduos voltar a recombinar, os cromossomos recombinantes resultantes carregarão uma quantidade muito grande de mutações deletérias. Isso poderia explicar a razão de espécies que evoluem reprodução assexuada (sem recombinação) não retornarem à reprodução sexuada (com recombinação). Essa hipótese é chamada de **catraca de Muller** (1964).

### Selective sweeps vs. background selection

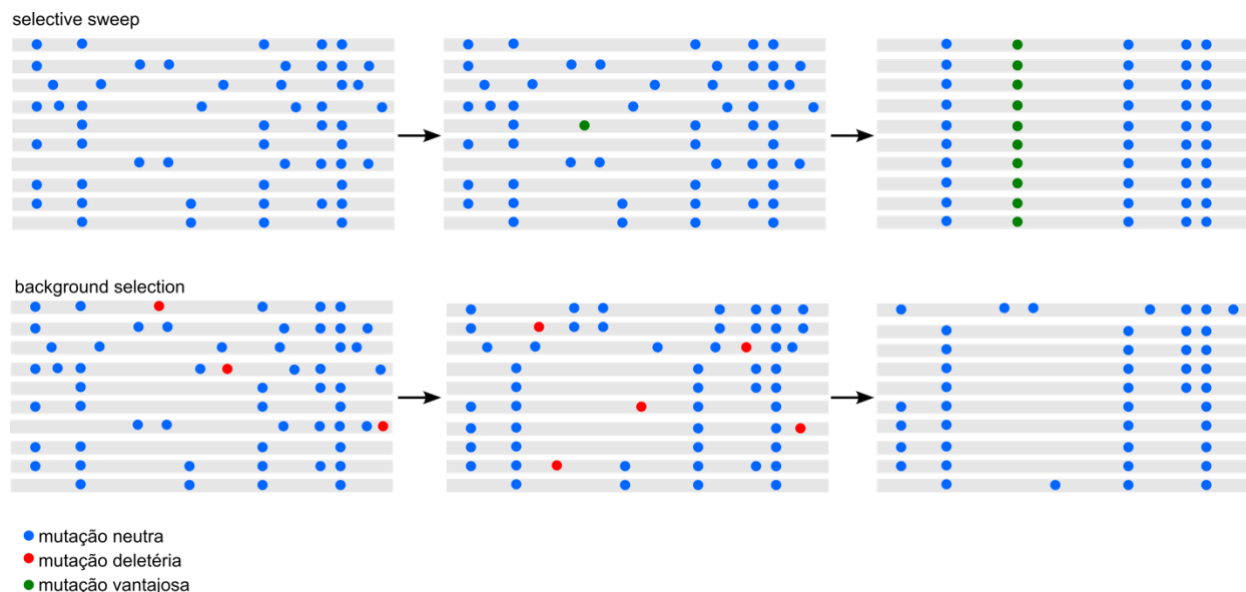
A partir da década de 1970, os geneticistas começaram a se questionar qual seria o efeito da ligação genética no polimorfismo genético próximo à região genômica em que uma mutação vantajosa ou deletéria acontece. Em 1974, Maynard-Smith e Haigh concluíram que, quando uma mutação vantajosa ocorre e ela começa a subir em frequência, devido à ligação genética, ocorre o **efeito hitch-hiking**, em que outras mutações próximas à mutação vantajosa também aumentarão de frequência, principalmente se estas forem neutras. Esse aumento rápido do haplótipo que contém a mutação vantajosa é chamado de **selective sweep** e resulta na diminuição da diversidade genética

nesta região genômica entre os indivíduos da população, pois todos terão o mesmo haplótipo. O tamanho da região sem diversidade vai depender do (1) **tempo em que a varredura de seleção ocorreu** e (2) a **taxa de recombinação**. Se a varredura ocorreu há pouco tempo, teremos uma grande região genômica com baixa diversidade. Quanto maior a taxa de recombinação, menor será o tamanho do bloco haplotípico com baixa diversidade. Observemos que a **seleção positiva** é o tipo de seleção que opera no processo de *selective sweep* (veja a figura abaixo, primeira linha).

Em 1993, o casal de geneticistas Brian e Deborah Charlesworth demonstraram que a seleção purificadora também poderia resultar em redução da diversidade genética em regiões do genoma. Como as mutações deletérias devem ser mais frequentes que as vantajosas, estas surgem nas populações com uma taxa maior. As regiões cromossômicas que carregam estas mutações serão, portanto, eliminadas por **seleção negativa** (= purificadora). Ao longo do tempo, também observaremos uma redução da diversidade genética nesta região genômica na população. Esse processo foi denominado de **background selection**. Assim como as varreduras de seleção, o impacto da background selection na redução da diversidade também depende do tempo em que a seleção negativa está eliminando as variações deletérias e da taxa de recombinação.

Na prática, **não é simples diferenciar qual dos dois processos ocorreu na população** observando os dados empíricos. Uma parcela significativa dos trabalhos teóricos recentes em genética evolutiva se dedica a essa questão.

Vejamos o esquema abaixo:



Nesse esquema, cada bloco cinza representa um cromossomo num indivíduo. A população possui, portanto, 10 cromossomos (10 linhas cinzas). Em ambos os casos, a maior parte da diversidade genética é oriunda de polimorfismos neutros (pontos azuis). Entretanto, essa variação neutra será influenciada por mutações vantajosas (pontos verdes) que podem surgir. O haplótipo que carrega a variação vantajosa subirá de frequência e, através de efeito *hitch-hiking* (= efeito carona), aumentará também a frequência das mutações neutras próximas à vantajosa. Alternativamente, temos um cenário onde mutações deletérias (pontos vermelhos) se acumulam constantemente nos cromossomos e são eliminadas por seleção purificadora constantemente. Ao longo do tempo, a diversidade desta região genômica vai diminuir por *background selection*.



## Referências

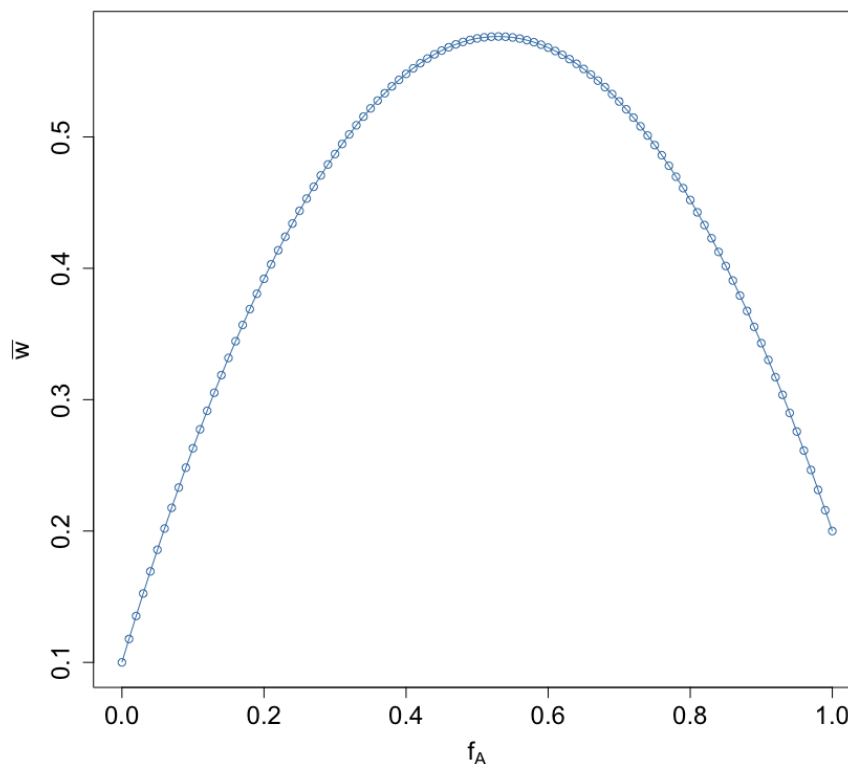
- Hill, W.G., Robertson, A., 1966. The effect of linkage on limits to artificial selection. *Genetics Research* 8, 269–294. <https://doi.org/10.1017/S0016672300010156>
- Muller, H.J., 1964. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 1, 2–9. [https://doi.org/10.1016/0027-5107\(64\)90047-8](https://doi.org/10.1016/0027-5107(64)90047-8)
- Maynard-Smith, J., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genet Res* 23, 23–35.
- Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303. <https://doi.org/10.1093/genetics/134.4.1289>

## Evolução multi-loci: Interação de múltiplos loci e a hipótese *shifting balance*

Existem diversos exemplos de características fenotípicas que variam nas populações naturais e que são resultado da interação de mais de um locus. São casos de interação gênica, genericamente chamadas de **epistasia**. Em 1932, Sewall Wright, que estudou a herança do padrão de cores dos pelos de roedores do gênero *Cavia* (porco do Índia), propôs que a interação entre genes poderia resultar numa paisagem adaptativa complexa, com vários máximos locais.

A partir desta ideia, Wright sugeriu que populações que apresentam subestruturação genética, ou seja, que não são panmíticas, seriam justamente aquelas em que o processo de especiação (ou de geração de raças) seria acelerado. Ele chamou essa hipótese de ***shifting balance***.

Para entendermos o conceito de Wright, lembremos que a paisagem adaptativa (*fitness landscape*) é o gráfico  $\bar{w} \times f_A$ , onde  $\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa}$ . Assim, num cenário de seleção favorável ao heterozigoto, temos uma paisagem adaptativa semelhante à da figura abaixo, onde foi assumido  $w_{AA} = 0.2$ ,  $w_{Aa} = 1$  e  $w_{aa} = 0.1$ . Note também como o valor máximo de  $\bar{w}$  está afastado de 1, indicando uma alta carga genética nesta população.



Imaginemos agora que, ao invés de apenas um locus, estudamos dois loci (*A* e *B*) com dois alelos segregando em cada um deles (*A*, *a* e *B*, *b*). Existem, portanto, nove genótipos possíveis nesta população. Assim como o caso mais simples acima, cada um desses genótipos terá um valor adaptativo  $w_{AABB}$ ,  $w_{AABb}$ ,  $w_{AAbb}$ , etc. Ao todo, teremos nove valores adaptativos, um para cada genótipo.

Assim como fizemos anteriormente, podemos então calcular o  $\bar{w}$  desta população com os valores das frequências genotípicas e os valores adaptativos dos genótipos. Apesar de longa, a fórmula é fácil de



entender: é o somatório dos nove produtos entre as nove frequências genotípica e seus respectivos valores adaptativos:

$$\bar{w} = f_{AABB}w_{AABB} + f_{AABb}w_{AABb} + f_{AAbb}w_{AAbb} + f_{AaBB}w_{AaBB} + f_{AaBb}w_{AaBb} + f_{Aabb}w_{Aabb} + f_{aaBB}w_{aaBB} + f_{aaBb}w_{aaBb} + f_{aabb}w_{aabb}$$

Podemos então conceber uma paisagem adaptativa para esse sistema. Entretanto, como agora estudamos não apenas a frequência do alelo *A*, pois também temos a do alelo *B*, devemos incluir mais um eixo no gráfico. Ao invés de um sistema (*x*,*y*) em que os valores de *x* são *f<sub>A</sub>* e os *y* são  $\bar{w}$ , temos um sistema (*x*, *y*, *z*), onde *x* é *f<sub>A</sub>*, *y* é *f<sub>B</sub>*, e *z* é  $\bar{w}$ .

Vamos estudar então um exemplo em que os valores adaptativos dos nove genótipos são dados pelo quadro abaixo:

|           | <b>AA</b>         | <b>Aa</b>        | <b>Aa</b>         |
|-----------|-------------------|------------------|-------------------|
| <b>BB</b> | $w_{AABB} = 0.87$ | $w_{AaBB} = 0.1$ | $w_{aaBB} = 0.99$ |
| <b>Bb</b> | $w_{AABb} = 0.56$ | $w_{AaBb} = 1$   | $w_{aaBb} = 0$    |
| <b>bb</b> | $w_{AAbb} = 0.89$ | $w_{Aabb} = 0$   | $w_{aabb} = 1$    |

Vamos admitir para fins de simplificação que, apesar dos loci *A* e *B* apresentarem interação gênica, não existe ligação física entre eles. Por exemplo, eles se encontram em cromossomos diferentes. Desta forma, as frequências genotípicas podem ser calculadas por um produto simples. Os valores de  $\bar{w}$  serão armazenados na matriz *w*. Segue o código em R:

```
# Paisagem adaptativa 2 loci:

# valores adaptativos dos genótipos
wAABB = 0.87
wAABb = 0.56
wAAbb = 0.89
wAaBB = 0.1
wAaBb = 1
wAabb = 0
waaBB = 0.99
waaBb = 0
waabb = 1

fA = seq(from=0, to=1, by=0.01)
fB = seq(from=0, to=1, by=0.01)

w = matrix(ncol=length(fA), nrow=length(fB))
rownames(w) = fA
colnames(w) = fB
x = 1
for(i in fA){

  # frequências HW para o locus A
  fAA = i^2
  fAa = 2 * i * (1 - i)
  faa = (1 - i)^2

  y = 1
  for(j in fB){

    # frequências HW para o locus B
    fBB = j^2
    fBb = 2 * j * (1 - j)
    fbb = (1 - j)^2
```

*Licença de utilização:* Distribuição e utilização livres, sem modificação do conteúdo e com referência à fonte original.  
 Atualizações disponíveis em [www.lbem.net.br/ensino](http://www.lbem.net.br/ensino)

*# frequências dos genótipos SEM ligação gênica - produto simples*

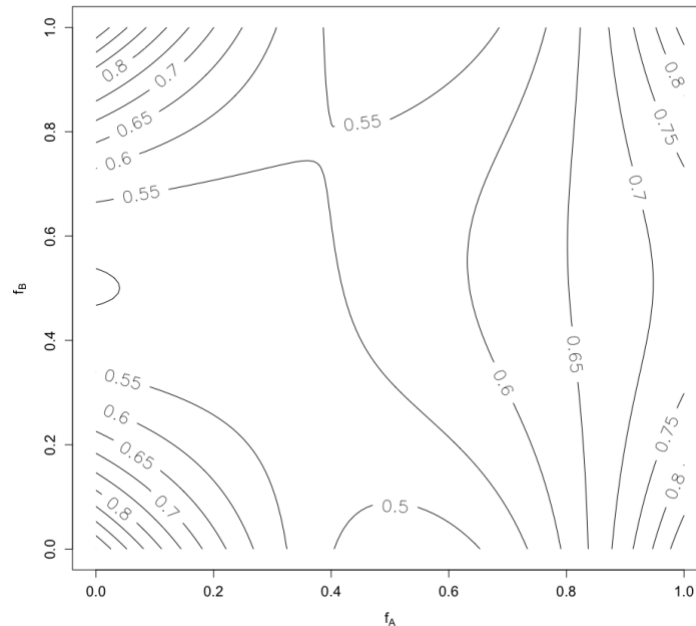
```
fAABB = fAA*fBB
fAABb = fAA*fBb
fAAbb = fAA*fbb
fAaBB = fAa*fBB
fAaBb = fAa*fBb
fAabb = fAa*fbb
faaBB = faa*fBB
faaBb = faa*fBb
faabb = faa*fbb
```

*# cálculo do w médio*

```
w[x,y] = fAABB*wAABB + fAABb*wAABb + fAAbb*wAAbb + fAaBB*wAaBB + fAaBb*wAaBb + fAabb*wAabb +
faaBB*waaBB + faaBb*waaBb + faabb*waaabb
  y = y + 1
}
  x = x + 1
}
```

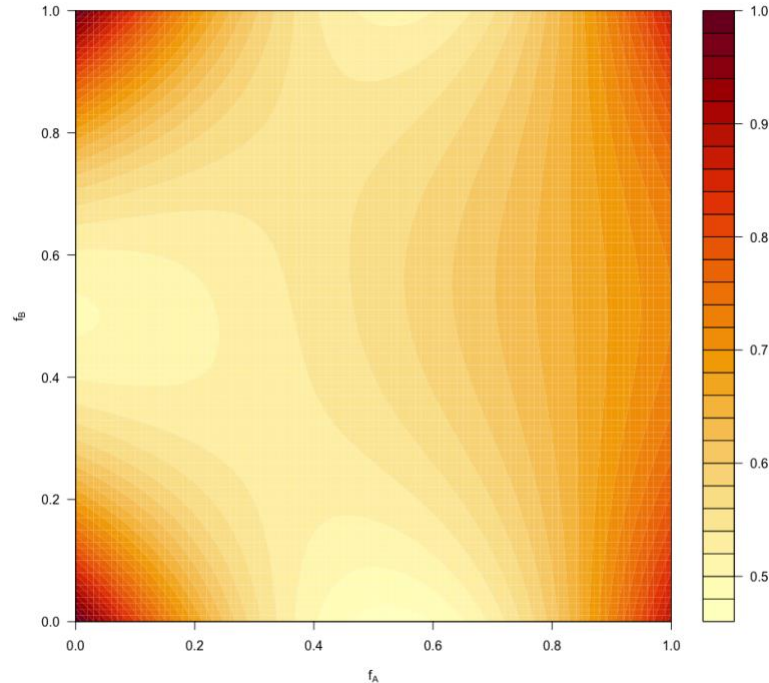
Vamos plotar o gráfico em três dimensões usando a função **contour**:

```
contour(fA,fB,w, labcex = 1.5, xlab=expression(f[A]), ylab=expression(f[B]))
```



Neste gráfico, o eixo z, correspondente aos valores de  $\bar{w}$ , está representado nas linhas. Uma outra forma de representar estes valores é com a função **filled.contour**:

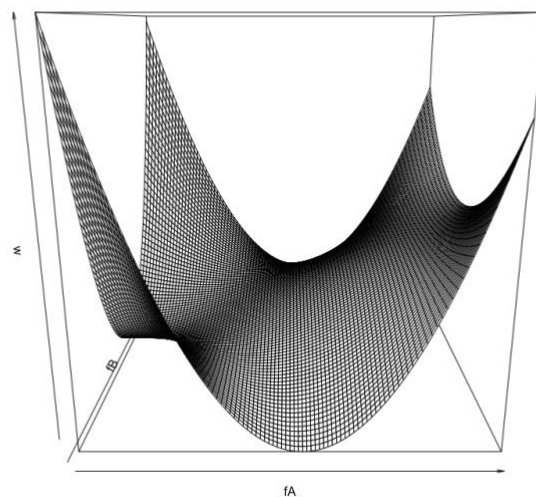
```
filled.contour(fA,fB,w, xlab=expression(f[A]), ylab=expression(f[B]))
```



Agora, os valores de  $\bar{w}$  podem ser mais facilmente visualizados usando a escala com gradação localizada à direita do gráfico. Fica evidente que as combinações de frequência que resultam no maior valor de  $\bar{w}$  são  $f_A = 0$  e  $f_B = 0$ , assim como  $f_A = 0$  e  $f_B = 1$ . São pontos de **picos adaptativos**. Também conseguimos visualizar algumas regiões de **vales adaptativos**.

Vejamos uma terceira forma de visualizar essa paisagem adaptativa usando a função **persp**:

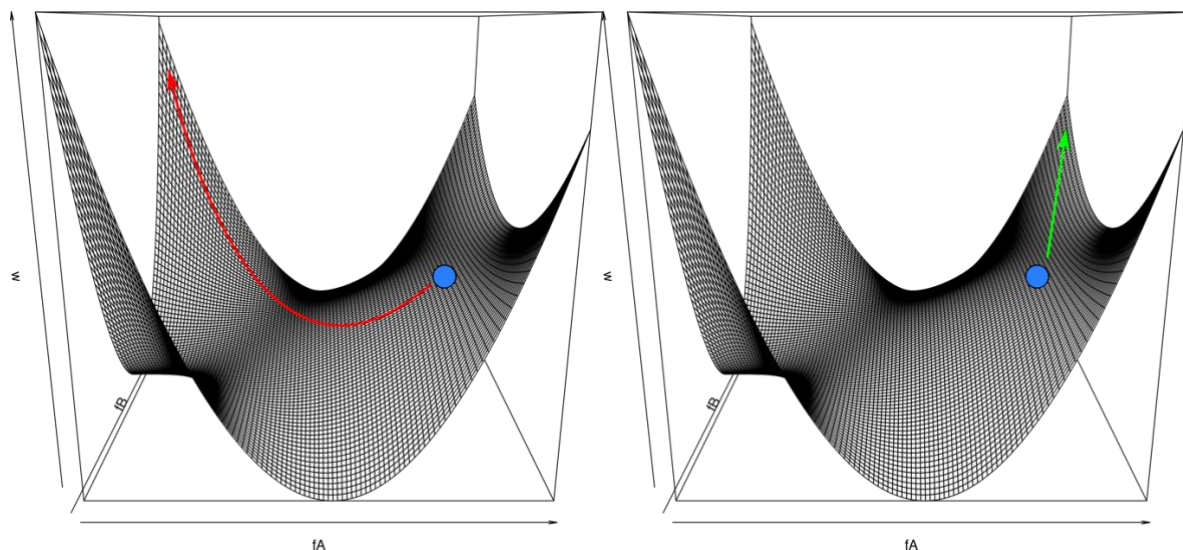
```
persp(fA, fB, w, xlab="fA", ylab="fB", zlab="w" )
```



Nesta representação, as combinações entre  $f_A$  e  $f_B$  que apresentam valores altos e baixos de  $\bar{w}$  ficam bastante evidentes. Também verificamos a presença de **máximos locais**, ou seja, pontos em que o valor de  $\bar{w}$  é alto **relativo aos valores da redondeza**.



Da mesma forma que o gráfico de paisagem com apenas um locus mostrado no início deste estudo, neste gráfico tridimensional também devemos entender que a seleção natural sempre irá direcionar a população para um aumento do  $\bar{w}$ . A **seleção não cruza vales adaptativos**.



Assim, se uma população se encontra numa combinação de frequências alélicas iguais a  $f_A = 0.8$  e  $f_B = 0.7$  (representado pelo ponto azul acima), a seleção natural **não poderá** alterar as frequências alélicas para levar esta população até o máximo pico adaptativo (caminho representado pela seta vermelha), pois será necessário cruzar um vale adaptativo e a seleção não diminui o valor de  $\bar{w}$ . A ação da seleção será, portanto, levar a população até um máximo local (seta verde), onde os alelos A e B serão fixados:  $f_A = 1$  e  $f_B = 1$ . Isso significa que a população possui 100% de indivíduos com o genótipo AABB, cujo  $w_{AABB} = 0.87$ , menor que o valor máximo possível.

Sewall Wright entendia que em populações que apresentam um alto nível de estruturação, **as diversas demes da população poderiam explorar melhor a paisagem adaptativa**. Nestas populações, **através da ação da deriva genética**, cada deme teria uma composição alélica distinta que permitiria uma melhor exploração das possibilidades da paisagem. Com isso, cada deme poderia estar **localmente adaptada**. Essa situação, para Wright, seria uma primeira etapa do processo de diferenciação genética que levaria à **especiação**.

## Referências

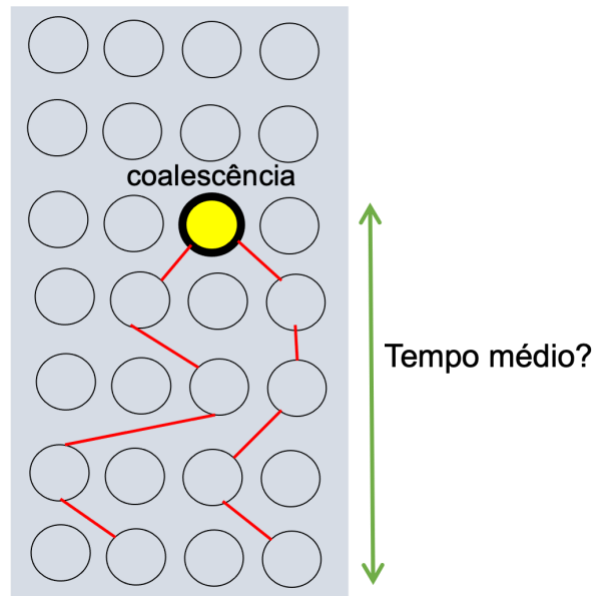
Wright, S. (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress on Genetics 1, 356–366.

<https://www.nature.com/scitable/topicpage/sewall-wright-and-the-development-of-shifting-30508/>

## Tempo de espera até a coalescência de um par de alelos

Quando estudamos o processo reprodutivo Wright-Fisher, verificamos que é possível acompanhar a ancestralidade dos alelos até encontrarmos a molécula de DNA mãe. Esse processo é denominado de **coalescência**.

Sewall Wright foi o primeiro a se questionar **qual o seria o tempo médio até o processo de coalescência de um par de alelos sorteados de uma população ocorrer?**



Conforme estudamos anteriormente, Wright sabia que, a cada geração, a chance de um par de alelos sorteados de uma população coalescer é  $1/2N$ . Desta forma, ele assumiu que esse tempo médio até a coalescência seria modelado por uma **distribuição geométrica**. Por quê?

Quando lançamos diversas vezes uma mesma moeda justa ( $Prob_{cara}=1/2$ ), o tempo médio (em número de lançamentos) que leva até observamos a primeira cara (ou coroa) é o recíproco da probabilidade do evento (cair cara ou coroa) a cada lançamento. Assim sendo, temos  $(1/2)^{-1} = 2$ . Ou seja, **em média**, teremos que lançar 2 vezes a moeda para obter a primeira cara (ou coroa).

Paralelamente, se a chance de coalescência por geração é  $1/2N$ , Wright pensou que levaria em média  $(1/2N)^{-1} = 2N$  gerações até a coalescência ocorrer usando a distribuição geométrica. Notemos que, na distribuição geométrica, se a chance de ocorrência do fenômeno é  $p$  por geração, o tempo médio até a primeira ocorrência será  $p^{-1}$ , ou seja,  $1/p$ . A fórmula da distribuição geométrica é dada por

$$\Pr(X = x) = (1 - p)^{x-1}p$$

Que retorna a probabilidade de um evento que tem chance  $p$  de acontecer a cada tentativa ocorrer após  $x$  tentativas (tempo em lançamento).

Wright sabia que a distribuição geométrica, que é em tempo discreto, pode ser aproximada por uma distribuição exponencial de tempo contínuo com taxa de decaimento  $\lambda = 1/2N$ , cuja fórmula é:

$$\Pr(x; \lambda) = \lambda e^{-\lambda x}$$

Comparemos ambas as distribuições no R para uma população com  $N = 10$ :

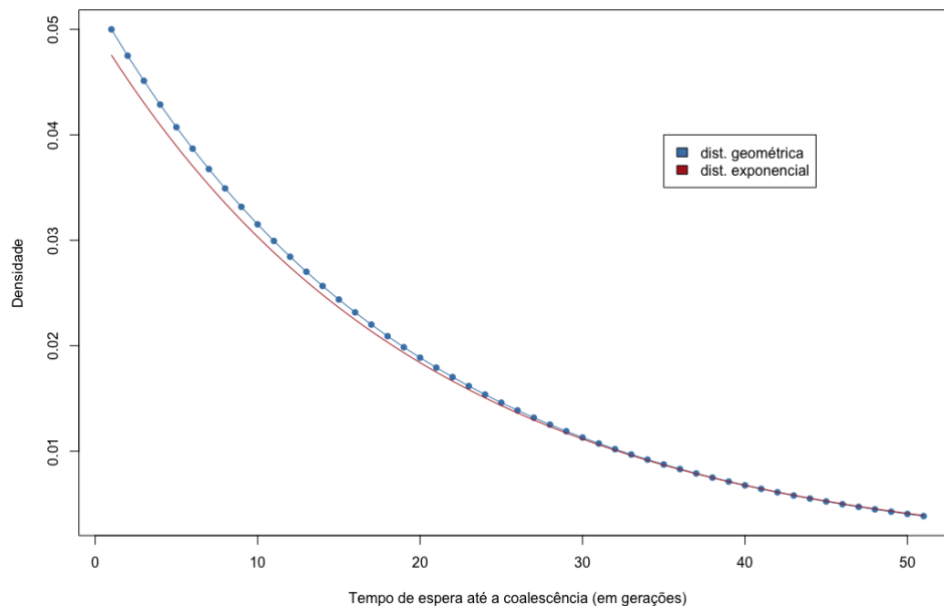
`N = 10`

```
# distribuição geométrica
plot(dgeom(seq(from=0,to=50, by=1), prob= 1/(2*N)), ylab="Densidade", xlab="Tempo de espera
até a coalescência (em gerações)", pch=16, col="steelblue" )

lines(dgeom(seq(from=0,to=50, by=1), prob= 1/(2*N)), col="steelblue")

# distribuição exponencial
curve(dexp(x, rate=1/(2*N)), col="firebrick", add=T)

legend(35, 0.04, legend=c("dist. geométrica", "dist. exponencial"), fill=c("steelblue",
"firebrick"))
```



A distribuição exponencial possui média  $\bar{x} = \lambda^{-1}$ . Desta forma, como  $\lambda = 1/2N$ , temos que o tempo médio até ocorrer a coalescência do par de alelos também será  $(1/2N)^{-1} = 2N$  gerações. Esse é o mesmo valor de média obtido na distribuição geométrica.

Em síntese, Wright derivou que, em média, **o tempo até ocorrer a coalescência de um par de alelos sorteados em uma população é  $2N$  gerações**. Observemos que, como esse tempo é exponencialmente distribuído, existe uma enorme variância associada a esse valor. A variância da distribuição exponencial é dada por

$$\text{var}(x) = \frac{1}{\lambda^2}$$

Assim, embora, em média, demore  $2N$  gerações para a coalescência ocorrer, a variância deste valor é igual a

$$\text{var}(x) = \frac{1}{\left(\frac{1}{2N}\right)^2} = 4N^2$$

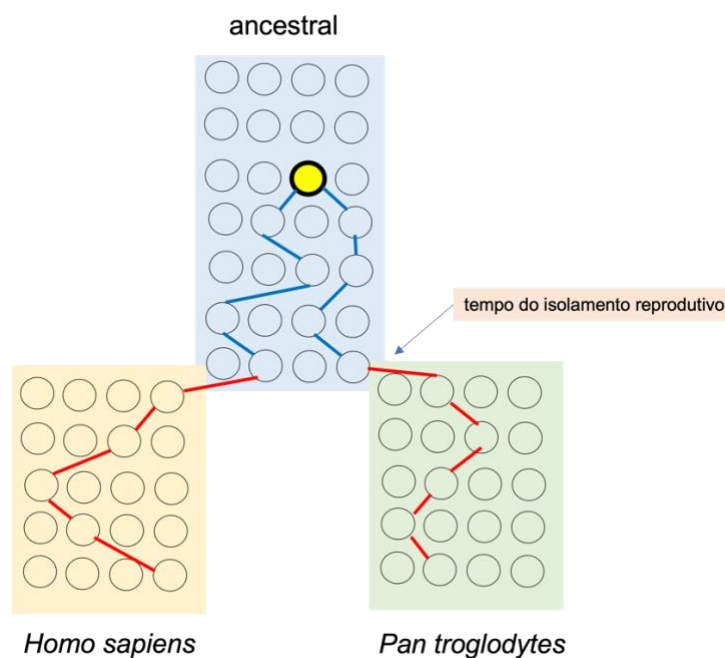
### Varição genética na população ancestral

A derivação de Wright é útil para entendermos o **quanto de variação genética é herdada da população ancestral** quando comparamos seqüências de genes homólogos em diferentes espécies. Por exemplo, ao compararmos as seqüências do gene *FAT4* entre humanos chimpanzés, observaremos algumas mudanças:

|                 |  |
|-----------------|--|
| Homo_sapiens    | GCGGGACCTCAATGACAACGCCCCCGTTTTCCCGGACCCCTCTATCGTGC |
| Pan_troglodytes | GCGGGACCTCAATGACAACGCTCCCGTTTTCCCGGACCCCTCTATCGTGC |

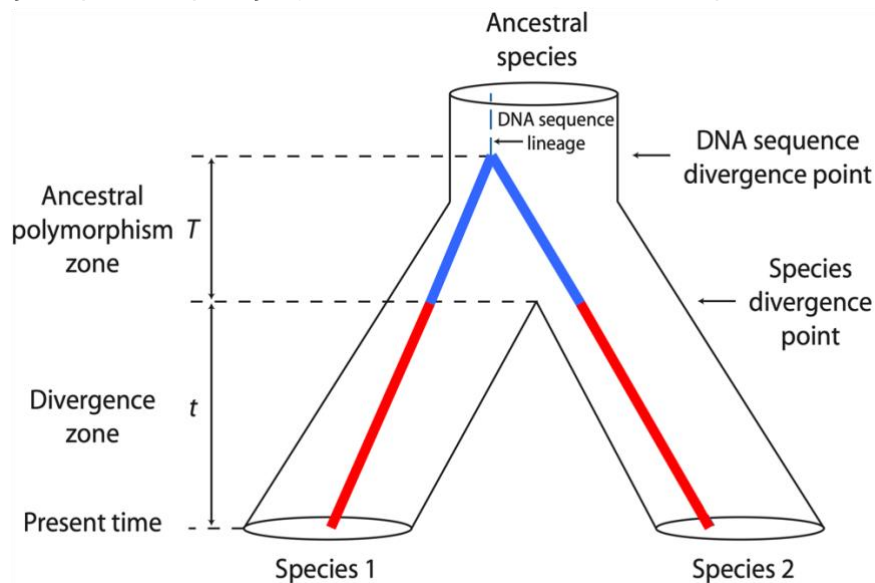
É comum achar que, no momento da divergência entre as duas espécies, na população ancestral, a seqüência desse gene era idêntica nos indivíduos dessa população. Entretanto, sabemos que existe polimorfismo (diferentes alelos) segregando nas populações, **inclusive na espécie ancestral de humanos e chimpanzés**. Essa situação pode ser estudada usando o processo de coalescência.

Podemos esquematizar esse problema assim:



Neste esquema mostramos que as cópias alélicas do gene *FAT4* dos humanos e chimpanzés encontram a sua molécula mãe (ou seja, coalescem) na população ancestral. As populações de cada uma das espécies estão marcadas com cores distintas. Desta forma, fica evidente que uma parte das mudanças encontradas no gene *FAT4* **ocorreu no ancestral** de humanos e chimpanzés (linhas azuis), enquanto outras mudanças ocorreram nas linhagens após o isolamento reprodutivo destas (linhas vermelhas).

Em síntese, o número total de mudanças acumuladas na comparação de sequências entre espécies (a **distância genética**) é composto de duas partes. A parte azul (mudanças no ancestral) e a parte vermelha (mudanças após a separação). Podemos visualizar o mesmo padrão assim:



O quanto ocorre de contribuição de cada uma das partes?

Se o tempo de divergência (= isolamento reprodutivo) entre espécies é  $t$  e assumindo uma taxa de mutação constante  $\mu$ , o total de mudanças nas linhas vermelhas será:

$$\mu \times t \times 2 = 2\mu t$$

Multiplicamos por dois, pois temos 2 linhas vermelhas (uma para cada espécie).

Na parte azul, usamos a mesma lógica. Entretanto, notemos que o tempo representado por  $T$  na figura acima é justamente o tempo de coalescência de um par de alelos na população ancestral (o par que, no futuro, será herdado um para os humanos e outro para os chimpanzés). Sabemos que esse tempo é, em média,  $2N$  gerações. Assim:

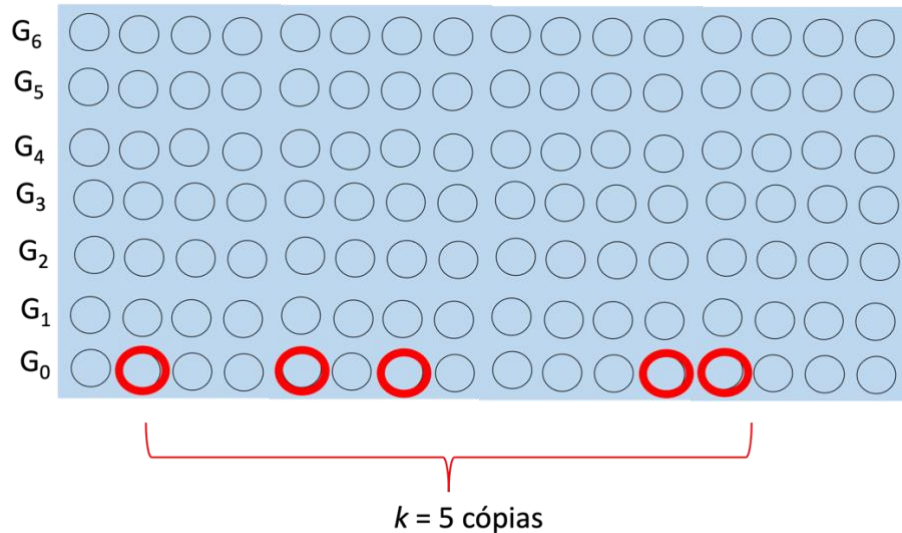
$$\mu \times T \times 2 = \mu \times 2N \times 2 = 4N\mu$$

Já vimos que, em genética de populações, o produto  $4N\mu$  é denominado de  $\theta$ . Desta forma, podemos dizer que **distância genética**  $d$  total (linhas azuis + vermelhas) entre sequências de amostradas de diferentes espécies separadas num tempo  $t$  será:

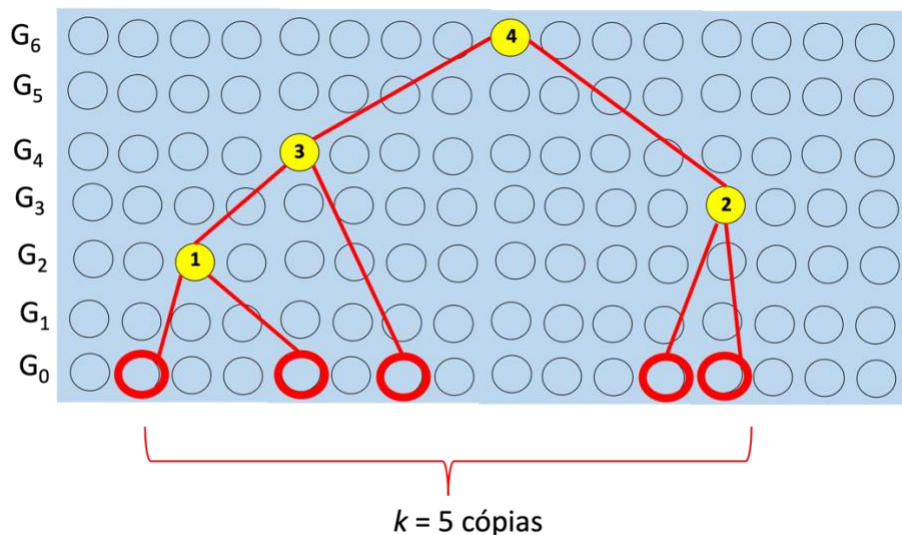
$$d = 2\mu t + \theta$$

## Tempos de espera até as diversas coalescências de vários alelos: a coalescência Kingman

No estudo anterior verificamos que o tempo médio que leva até um par de alelos amostrados ao acaso em uma população coalescerem é igual a  $2N$  gerações. Em 1982, o matemático britânico John Kingman explorou o problema da coalescência de alelos numa população quando mais de um par é amostrado. A situação pode ser esquematizada assim:

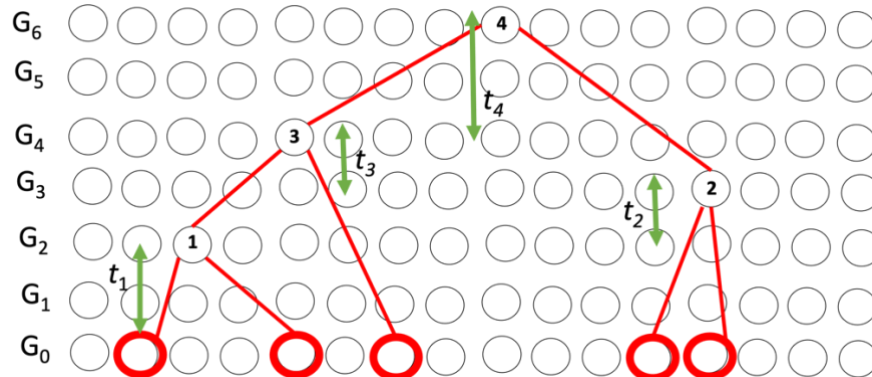


Nesta figura, no tempo presente (marcado com  $G_0$ ), foram amostrados  $k = 5$  cópias alélicas na população. Kingman se interessou em descrever **os tempos médios de todas as coalescências** que acontecem entre essas cópias:



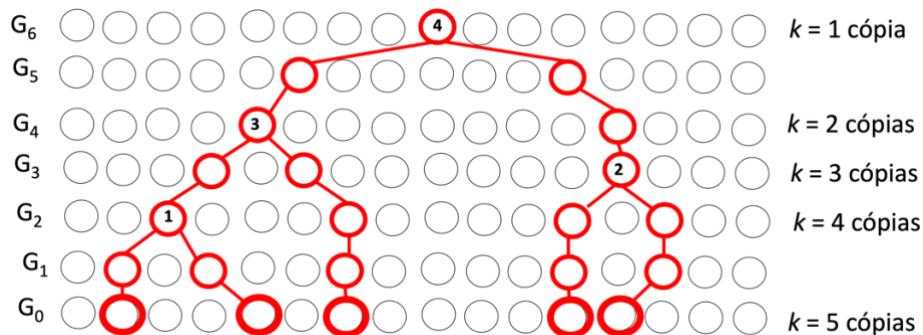
Na figura acima, verificamos que a primeira coalescência (1) ocorreu em  $G_2$ , ou seja, duas gerações atrás. A segunda (2) ocorreu em  $G_3$ , a terceira em  $G_4$  e a última em  $G_6$ . O processo inteiro evidencia a **genealogia dos alelos** sorteados. A coalescência de Kingman descreve os tempos de espera médios até ocorrer cada uma das coalescências marcadas em amarelo.

Esses tempos podem ser visualizados na figura abaixo:

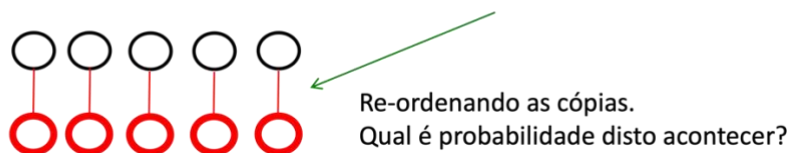
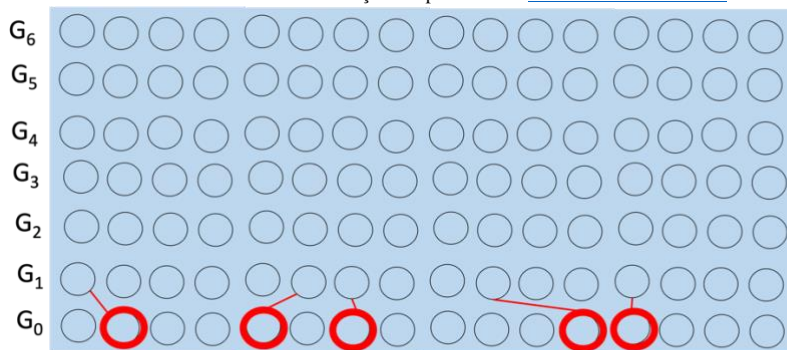


Desta forma,  $t_1$  é o tempo até a primeira coalescência,  $t_2$  é o tempo **entre** as coalescências 1-2,  $t_3$  é o tempo **entre** as coalescências 2-3 e  $t_4$  é o tempo **entre** as coalescências 3-4.

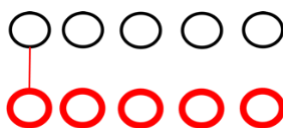
A estratégia de Kingman foi descrever o processo matemático que **reduz progressivamente o número de cópias alélicas que devem coalescer**. Inicialmente, temos  $k = 5$  cópias. Após a primeira coalescência, esse número reduz para  $k = 4$ , depois da segunda coalescência, o número reduz para  $k = 3$ , depois  $k = 2$  até finalmente chegar na molécula mãe das 5 cópias de  $G_0$ . Podemos visualizar esse processo nesta figura:



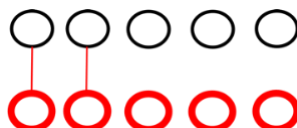
Assim como vários outros casos em Genética de Populações, ele modelou o processo acima usando o oposto: a probabilidade das cópias **não coalescerem**. Ou seja, em  $G_0$ , quando existem  $k = 5$  cópias amostradas, qual seria a probabilidade de, em 1 geração para o passado, o processo abaixo acontecer:



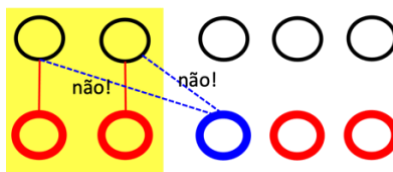
Para entendermos a lógica do cálculo desta probabilidade, devemos iniciar o processo escolhendo ao acaso qualquer uma das  $k = 5$  cópias. A linha vermelha mostra a sua molécula ancestral:



Já estudamos anteriormente que a probabilidade da segunda cópia escolhida coalescer no mesmo ancestral da primeira é  $1/2N$  (i.e., chance de dois alelos amostrados coalescerem em 1 geração). Portanto, a probabilidade de não coalescência (figura abaixo) será  $1 - 1/2N$ :



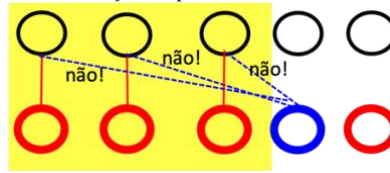
Agora devemos pegar a terceira cópia. Para não ocorrer coalescência em 1 geração, esta terceira cópia **não poderá escolher nenhum dos dois ancestrais já conectados**:



A cópia azul tem chance de escolher no *pool* gamético o mesmo ancestral da primeira cópia com chance  $1/2N$ . Ela também tem chance de escolher do *pool* o ancestral da segunda cópia com chance  $1/2N$ . Assim, a chance da cópia azul encontrar no *pool* um ancestral já conectado é  $2(1/2N)$ . Portanto, a chance não encontrar será  $1 - 2(1/2N)$ .

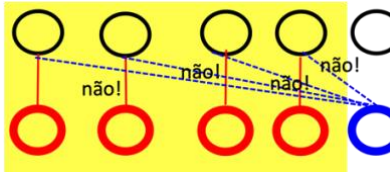
Chegamos na quarta cópia. Ela não pode coalescer com nenhuma das três já conectadas:



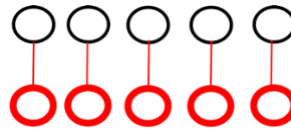


Usando a mesma lógica anterior, a chance **não sortear** do *pool* (= não coalescer com) um alelo ancestral já conectado com as três anteriores será  $1 - 3(1/2N)$ .

Por fim, chegamos na quinta cópia. A chance de não coalescer com nenhuma das quatro já conectadas será  $1 - 4(1/2N)$ :



Agora podemos efetivamente calcular qual é a chance das cinco cópias **não coalescerem** em 1 geração:



Este fenômeno somente ocorrerá se cada um dos processos descritos acima ocorrerem. Ou seja,  $1 - 1/2N$  e  $1 - 2(1/2N)$  e  $1 - 3(1/2N)$  e  $1 - 4(1/2N)$ . Isso é o produto de todos os termos. Podemos afirmar que a chance de não acontecer uma coalescência de  $k = 5$  cópias em uma geração é:

$$\Pr(k = 5) = 1 - \frac{1}{2N} \cdot \left(1 - \frac{2}{2N}\right) \cdot \left(1 - \frac{3}{2N}\right) \cdot \left(1 - \frac{4}{2N}\right)$$

A fórmula acima já deixa evidente que podemos generalizar a derivação para **qualquer valor  $k$  de cópias**:

$$\Pr(k = k) = 1 - \frac{1}{2N} \cdot \left(1 - \frac{2}{2N}\right) \cdot \left(1 - \frac{3}{2N}\right) \cdots \left(1 - \frac{k-1}{2N}\right)$$

Ora, se derivamos a fórmula para **não coalescência** em 1 geração de  $k$  cópias, a chance de uma das  $k$  cópias coalescerem em 1 geração e tornarem-se  $(k - 1)$  cópias será um menos o valor acima:

$$\Pr(k \rightarrow k - 1) = 1 - \Pr(k = k)$$

Esse valor é aproximadamente:

$$\Pr(k \rightarrow k - 1) \approx \frac{k(k - 1)}{4N}$$

A leitura da equação acima é “a chance de uma das  $k$  cópias alélicas amostradas coalescerem em 1 geração para o passado e, portanto, serem reduzidas em  $(k - 1)$  cópias é..”

**Mas o tempo de espera?**

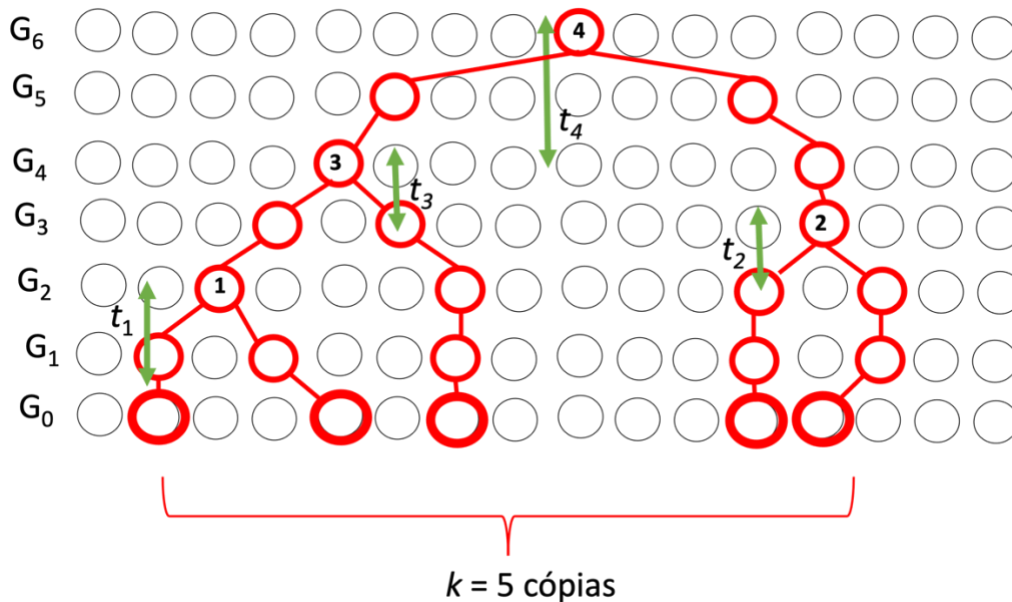
Lembremos do caso mais simples analisado no estudo anterior. Sabemos que um par de alelos coalescerá em 1 geração com chance  $1/2N$  e que o tempo médio disto ocorrer é exponencialmente distribuído com parâmetro  $\lambda = 1/2N$ . Isso significa um tempo médio  $\bar{t} = \lambda^{-1} = 2N$ . Essa foi a dedução de Wright.

Pois bem, se a chance de coalescência em 1 geração de  $k$  cópias é  $\frac{k(k-1)}{4N}$ , então similarmente o tempo médio para ocorrer esse evento será o recíproco deste valor:

$$\bar{t} = \left( \frac{k(k-1)}{4N} \right)^{-1} = \frac{4N}{k(k-1)}$$

Perceba que a derivação acima funciona inclusive se  $k = 2$  cópias, resultando no tempo de  $2N$ .

Vamos então finalmente calcular os valores dos quatro tempos representados na figura abaixo:



A primeira coalescência marca a redução dos alelos amostrados de  $k = 5$  para 4 cópias. O tempo médio para isso acontecer será ( $k = 5$ ):

$$\bar{t}_1 = \frac{4N}{k(k-1)} = \frac{4N}{5(5-1)} = \frac{4N}{20}$$

O segundo tempo médio ( $t_2$ ) marca a redução de  $k = 4$  para 3 cópias:

$$\bar{t}_2 = \frac{4N}{k(k-1)} = \frac{4N}{4(4-1)} = \frac{4N}{12}$$

O terceiro tempo médio ( $t_3$ ) marca a redução de  $k = 3$  para 2 cópias:

$$\bar{t}_3 = \frac{4N}{k(k-1)} = \frac{4N}{3(3-1)} = \frac{4N}{6}$$



Por fim, o quarto tempo ( $t_4$ ) marca a redução de  $k = 2$  para 1 cópias (a molécula mãe das 5 cópias):

$$\bar{t}_4 = \frac{4N}{k(k-1)} = \frac{4N}{2(2-1)} = \frac{4N}{2} = 2N$$

Por exemplo, se  $N = 100$ , a primeira coalescência ocorrerá, em média, em  $\frac{4(100)}{20} = 20$  gerações, a segunda em  $\frac{4(100)}{12} = 33.3$  gerações **após a primeira**. A terceira em  $\frac{4(100)}{6} = 66.67$  gerações **após a segunda** e a última demorará, em média,  $2(100) = 200$  gerações para acontecer.

Fica evidente que, **conforme o número de cópias alélicas vai reduzindo, o tempo de espera para a próxima coalescência vai aumentando**. Além disso, **a última coalescência é a que mais demora**.

Segue um código R implementando a fórmula de Kingman para os tempos de espera:

```
N = 100
k = 5
cores = hcl.colors(k)

# cálculo do parâmetro lambda da exponencial
lambda = (k * (k - 1)) / (4*N)

# gráfico da última coalescência (k = 5)
curve(dexp(x, rate = lambda), col=cores[k], xlim=c(0,400), lwd=2, xlab="Tempo de espera para coalescer
(gerações)", ylab="Densidade")
segments(lambda^(-1), -1, lambda^(-1), 1, col=cores[k], lty=2)

# gráficos das coalescências (k = 4) até (k = 2)
for(i in c( (k - 1):2)){

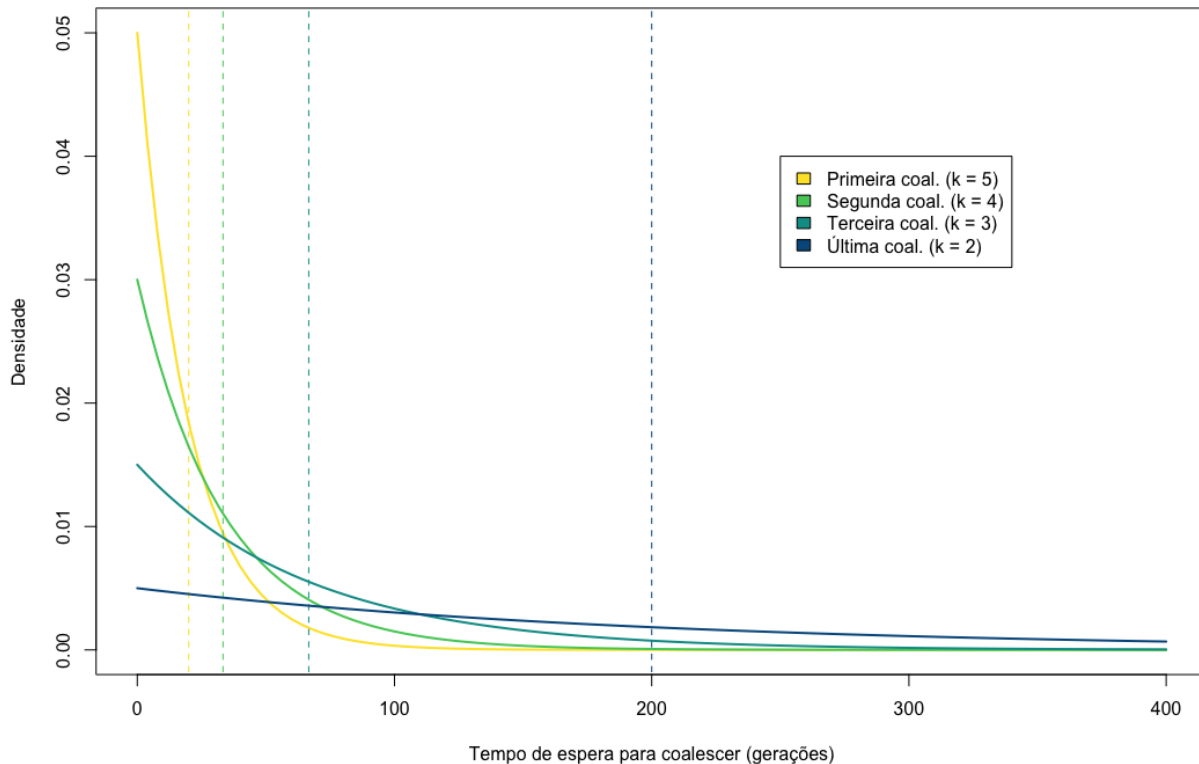
  lambda = (i * (i - 1)) / (4*N)

  curve(dexp(x, rate = lambda), col=cores[i], xlim=c(0,400), lwd=2, add=T)

  segments(lambda^(-1), -1, lambda^(-1), 1, col=cores[i], lty=2)

}

legend(250,0.04,legend=c("Primeira coal. (k = 5)", "Segunda coal. (k = 4)", "Terceira coal. (k = 3)",
"Última coal. (k = 2)"), fill=cores[5:2] )
```



Cada

tempo de espera até a coalescência é exponencialmente distribuído. As linhas pontilhadas indicam as médias ( $\lambda^{-1}$ ) em cada um dos tempos de espera. É notável que a primeira coalescência (curva amarela) é a que possui a variância menor. A última coalescência (curva azul) é aquela com maior variância. Desta forma, é mais **o erro associado à estimativa do tempo da primeira coalescência é o menor de todos.**

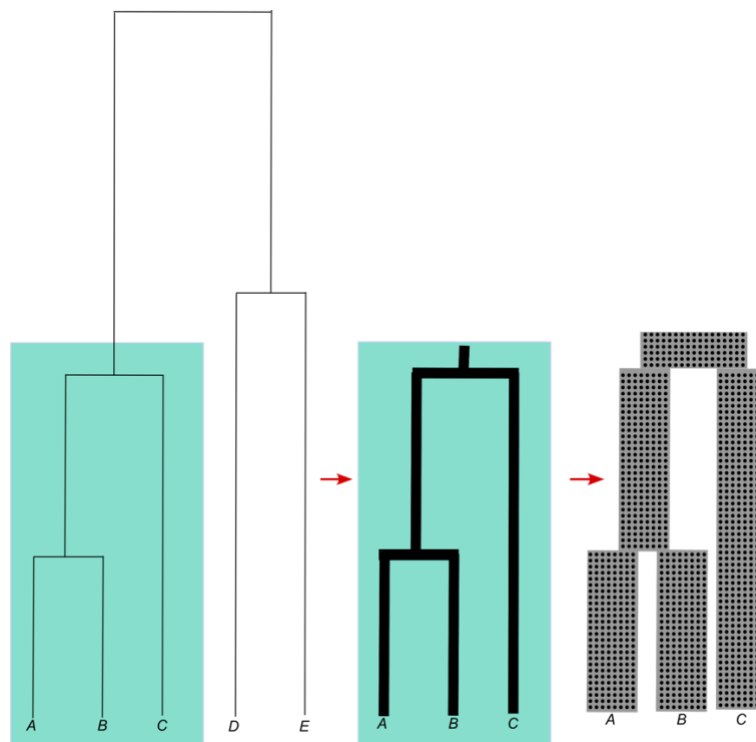
## Referências

Kingman, J.F.C., 1982. The coalescent. *Stochastic Processes and their Applications* 13, 235–248.  
[https://doi.org/10.1016/0304-4149\(82\)90011-4](https://doi.org/10.1016/0304-4149(82)90011-4)

## Unificando genética de populações e filogenética: o modelo da coalescência multi-espécies

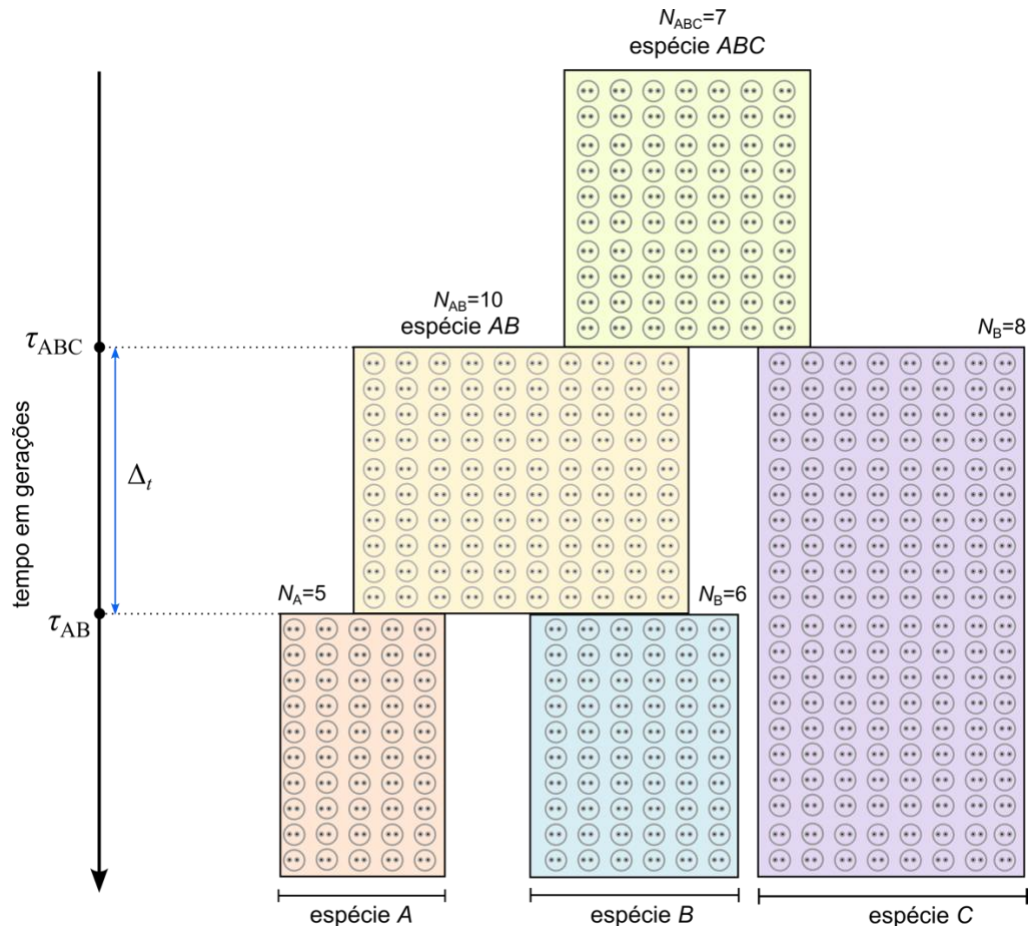
Anteriormente, verificamos que o processo de coalescência de cópias alélicas numa população pode ser usado para estudar a parcela do total de mudanças genéticas entre duas espécies, ou seja, a distância genética (medida em número de substituições por sítio de nucleotídeo), ocorreu antes do completo isolamento reprodutivo de duas espécies. Esse valor,  $4N\mu$ , portanto depende do tamanho efetivo da espécie ancestral além da taxa de mutação por geração.

Podemos estender essa análise para um caso simples de três espécies que divergiram conforme a figura abaixo. Ao final da década de 1980, este problema foi estudado em detalhes por Pamilo e Nei (1988), que estenderam as análises iniciais de Tajima (1983) e Hudson (1983). Assim como fizemos no início do curso, usaremos uma abordagem conceitual que compreende que os ramos de uma filogenia são fundamentalmente uma sequência cronológica (ancestral-descendente) de populações. Como o entendimento contemporâneo da espécie consiste em uma unidade populacional, podemos também interpretar essa sequência cronológica como a mudança **anagenética** que ocorre na espécie no decorrer das gerações.

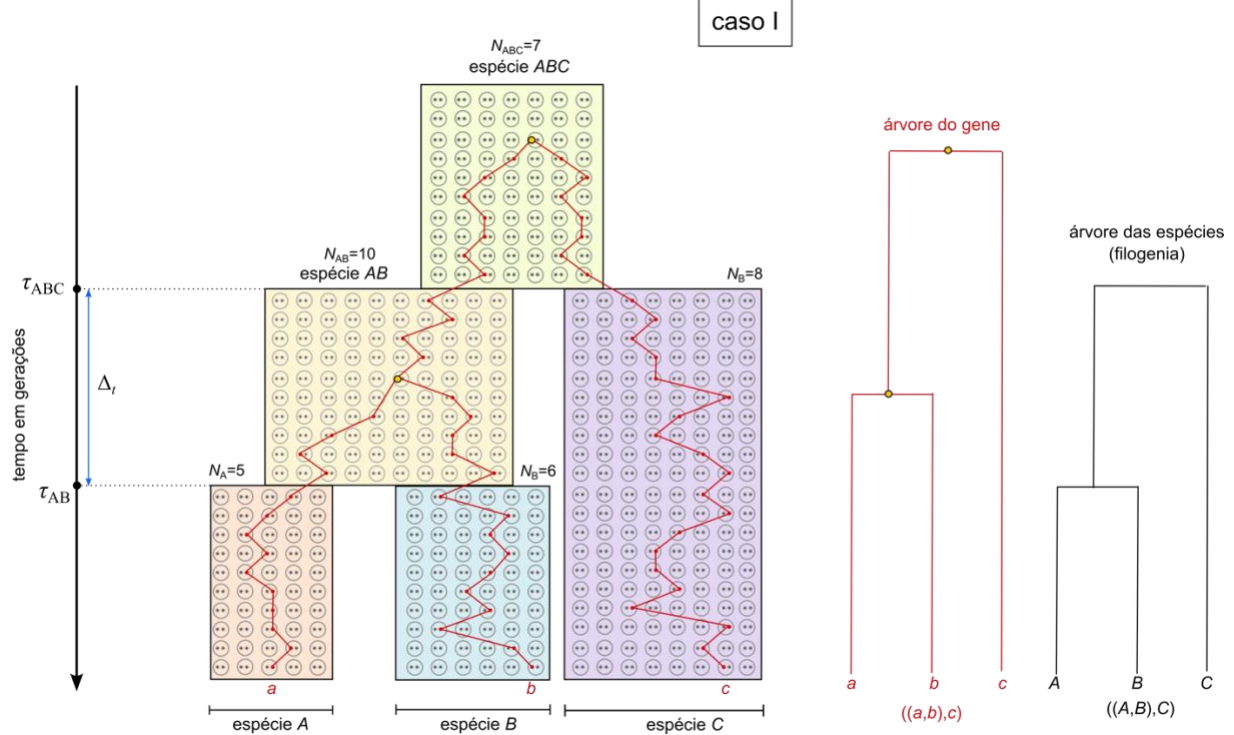


Ao aproximar o zoom nas três espécies investigadas, observamos essencialmente um mosaico de cinco processos Wright-Fisher "colados". Cada um desses processos representa a sequência de transformações da composição genética de cada uma das espécies:  $A$ ,  $B$ ,  $C$ , a espécie ancestral de  $AB$  e a espécie ancestral de  $ABC$ . Nesta figura também vemos representados os tamanhos efetivos de cada uma das populações W-F ( $N_A$ ,  $N_B$ ,  $N_C$ ,  $N_{AB}$ ,  $N_{ABC}$ ) assim como os tempos de **isolamento reprodutivo completo** das espécies  $A$  e  $B$  ( $\tau_{AB}$ ) e  $A$ ,  $B$  e  $C$  ( $\tau_{ABC}$ ). Como estamos usando o arcabouço conceitual da genética de populações, os tempos de isolamento são medidos em **gerações**. Verifiquemos na figura a seguir, onde cada círculo demarca um indivíduo com duas cópias alélicas

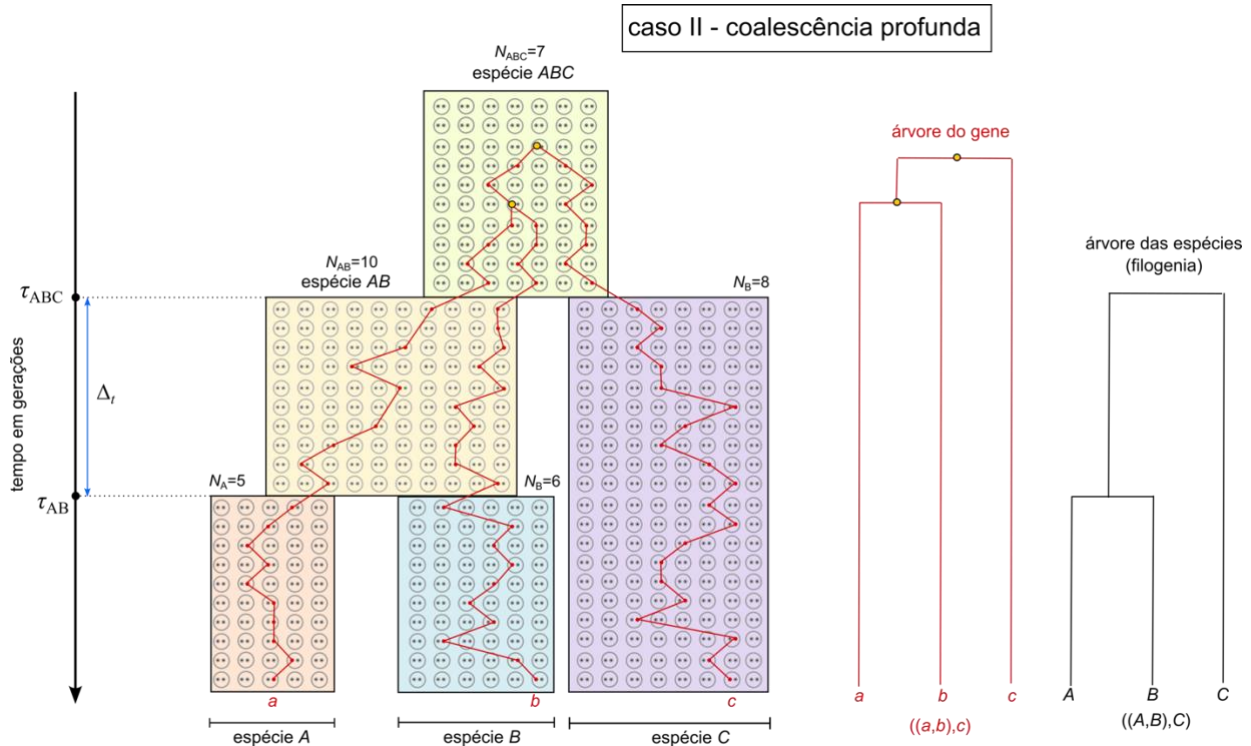
(diplóide). Assim como os estudos prévios, cada linha demarca uma geração e o processo de reprodução entre as gerações ocorre por amostragem do pool gamético (Wright-Fisher):



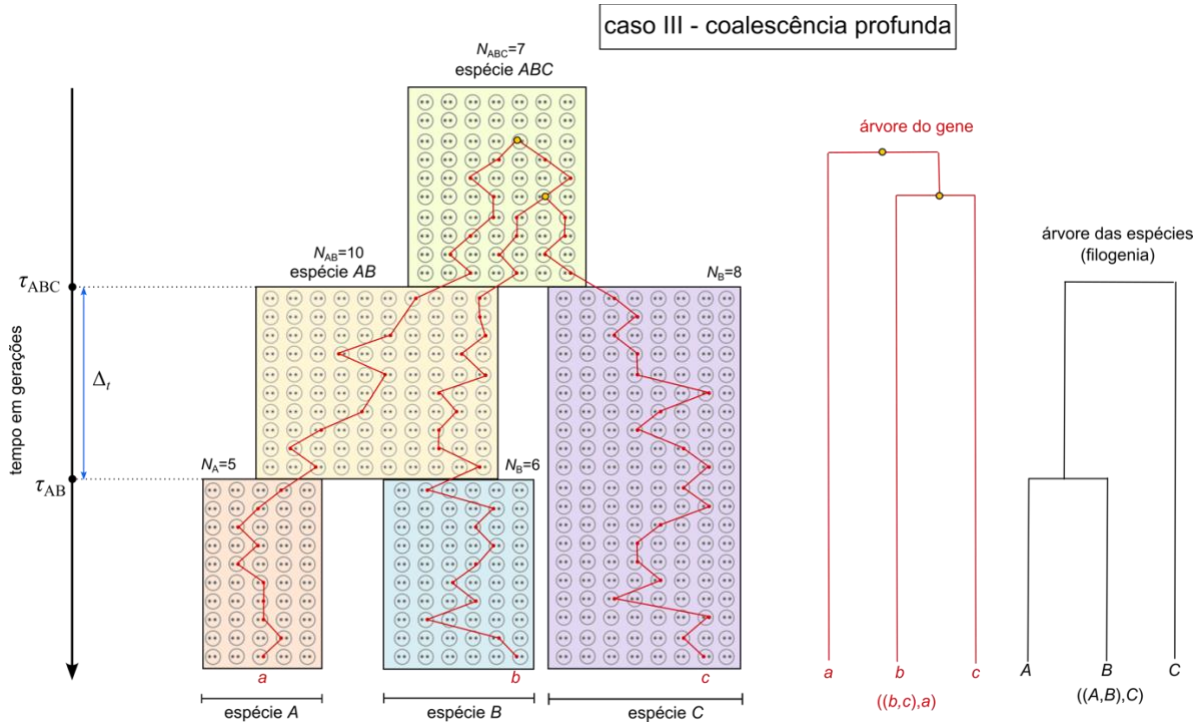
Após definir essas variáveis básicas, podemos tentar acompanhar o processo de coalescência de alelos amostrados em cada uma das espécies até suas moléculas de DNA ancestrais. Nesta próxima figura, mostramos as quatro possibilidades de genealogias dos genes (alelos) coletados em cada uma das três espécies. No caso I abaixo, observamos que a coalescência das cópias oriundas das espécies *A* e *B* ocorre na espécie ancestral *AB*, ou seja, no intervalo de tempo  $\Delta_t = \tau_{ABC} - \tau_{AB}$ . Neste caso, temos que **topologia da árvore** da genealogia dos genes é a mesma da **filogenia das espécies** *A*, *B* e *C*. Podemos afirmar, portanto, que a topologia da **árvore do gene**, representada pelo **formato Newick**  $((a, b), c)$ , é a mesma da árvore das espécies:  $((A, B), C)$ :



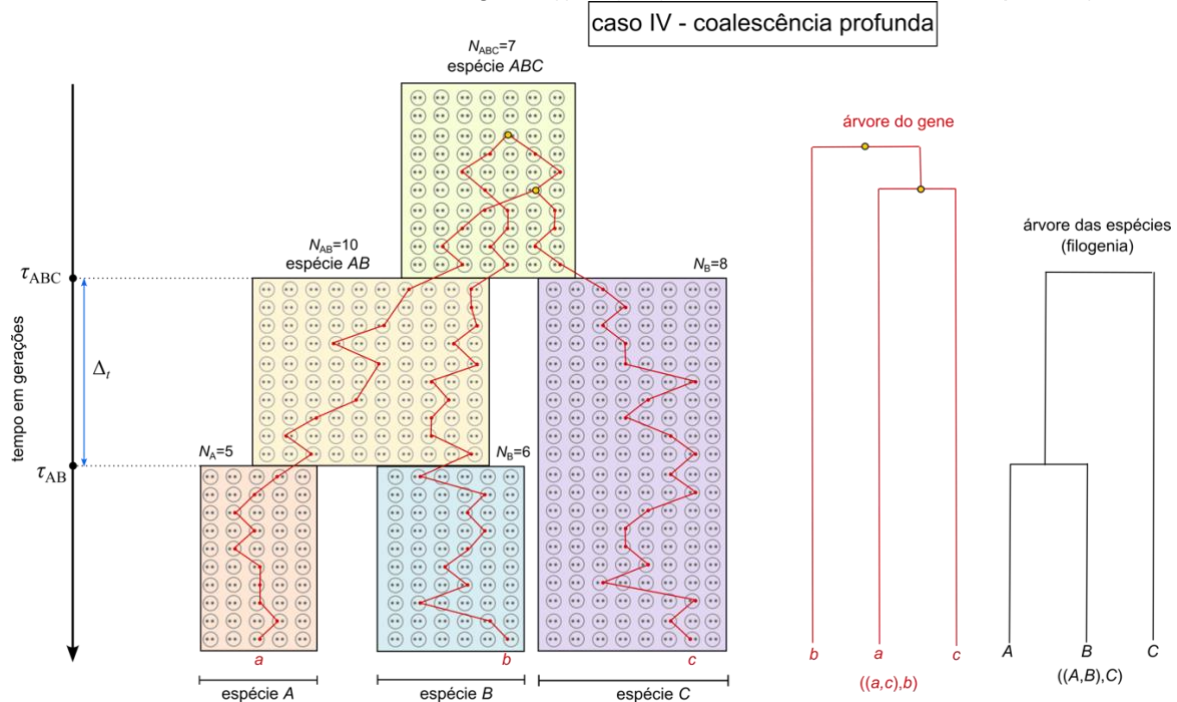
No caso II a seguir, embora a coalescência das cópias alélicas  $a$  e  $b$  ocorra na espécie ancestral  $ABC$ , a topologia da árvore do gene -  $((a, b), c)$  - também é igual à filogenia (=árvore das espécies)  $((A, B), C)$ . Entretanto, como a coalescência das cópias alélicas  $a$  e  $b$  não ocorre dentro do intervalo  $\Delta_t$ , este é um caso de **coalescência profunda**:



Os casos III e IV mostram duas possibilidades interessantes. Ambos representam casos de coalescência profunda. Entretanto, como as linhagens alélicas *a* e *b* chegam na espécie ancestral *ABC* sem coalescer, é possível que ocorra, ao acaso, coalescência entre *b* e *c* na população *ABC* (caso III). Neste caso, a topologia da árvore de gene resultante será  $((b, c), a)$ . **Esta árvore de gene é diferente da filogenia das espécies:**



Similarmente, podemos conceber um cenário onde as cópias alélicas *a* e *c* coalescem no ancestral *ABC*, também resultando numa árvore de gene  $((a, c), b)$  diferente da árvore de espécie (caso IV):







Desta forma, unindo as possibilidades descritas nos quatro casos acima, temos dois casos que resultam em árvore de genes idênticas à árvore das espécies e dois casos que resultam em diferenças entre árvore de genes e a filogenia. Os casos de coalescência profunda potencialmente resultam num fenômeno chamado de **incomplete lineage sorting (ILS)**. Devido a existência de polimorfismo alélico na população ancestral, é comum afirmar que as coalescências profundas estão relacionadas à **retenção de polimorfismo ancestral**. Consequentemente, podemos obter topologias de árvores de genes (a genealogia dos alelos) diferentes da filogenia das espécies. **Como quantificar a probabilidade de cada uma dessas possibilidades ocorrerem?**

Podemos usar os princípios da teoria da coalescência para calcular as probabilidades dos casos I, II, III e IV. Começemos com o caso I.

### Probabilidade de coalescência na espécie ancestral AB (caso I):

Sabemos que o **tempo de espera para a coalescência** de um par de alelos numa população é dado por uma distribuição exponencial onde  $\lambda = 1/2N$ . Assim, como a distribuição exponencial é dada por  $f(t; \lambda) = \lambda e^{-\lambda t}$ , temos:

$$f(t; \lambda) = \frac{1}{2N} e^{-\frac{1}{2N}t}$$

Para fins didáticos, vamos plotar no R esse gráfico para  $N = 10$ , que é o tamanho da população da espécie ancestral AB:

```
N = 10 # tamanho da população ancestral AB

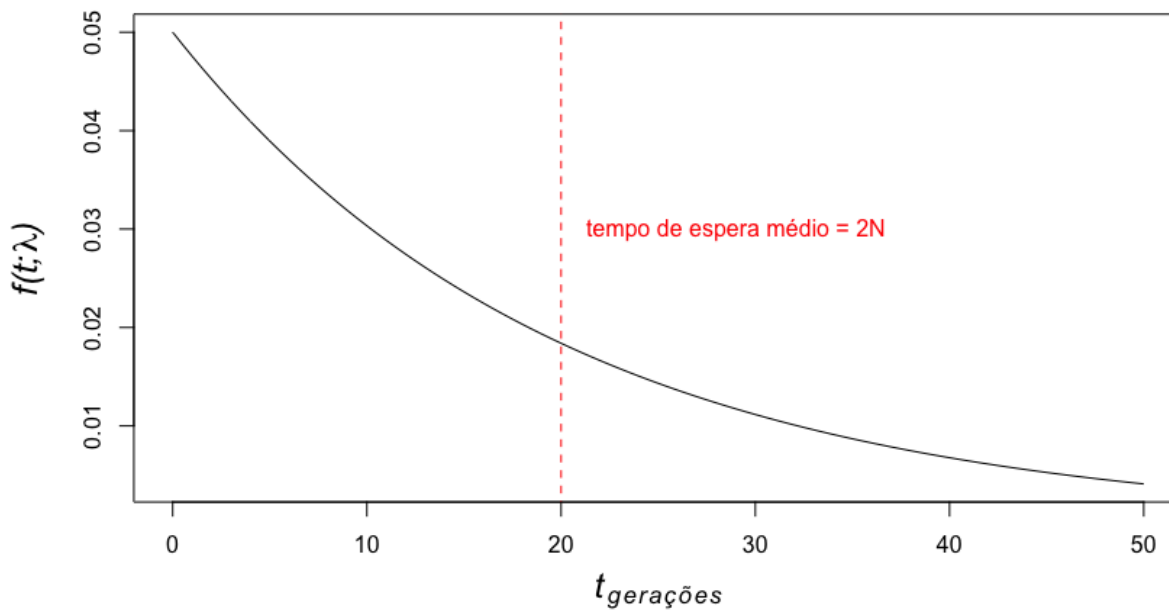
# Ajustando a posição as margens, labels e títulos dos eixos
par(mar = c(5, 5, 4, 2) + 0.1)
par(mgp = c(3, 1, 0))

curve(dexp(x, rate = 1/(2*N)), xlim=c(0,50),ylab=expression(italic( paste("f(t;", lambda, ")"))),
xlab=expression(italic(t[gerações])), cex.lab=1.5, main="Distribuição do tempo de espera para um par de
alelos (N = 10)")

segments(20,0,20,1, col="red", lty=2)

text(29,0.03, "tempo de espera médio = 2N", col="red")
```

### Distribuição do tempo de espera para um par de alelos (N = 10)



Sabemos que o tempo de espera médio é dado por  $\bar{t} = \lambda^{-1} = 2N$ . Entretanto, queremos calcular qual é valor da probabilidade da coalescência dos alelos *a* e *b* ocorrer num tempo igual ou menor que  $\Delta_t$ . Pois, se isso acontecer, teremos o caso I.

No exemplo mostrado  $\Delta_t = 10$  gerações. Podemos visualizar no R essa probabilidade usando o gráfico acima:

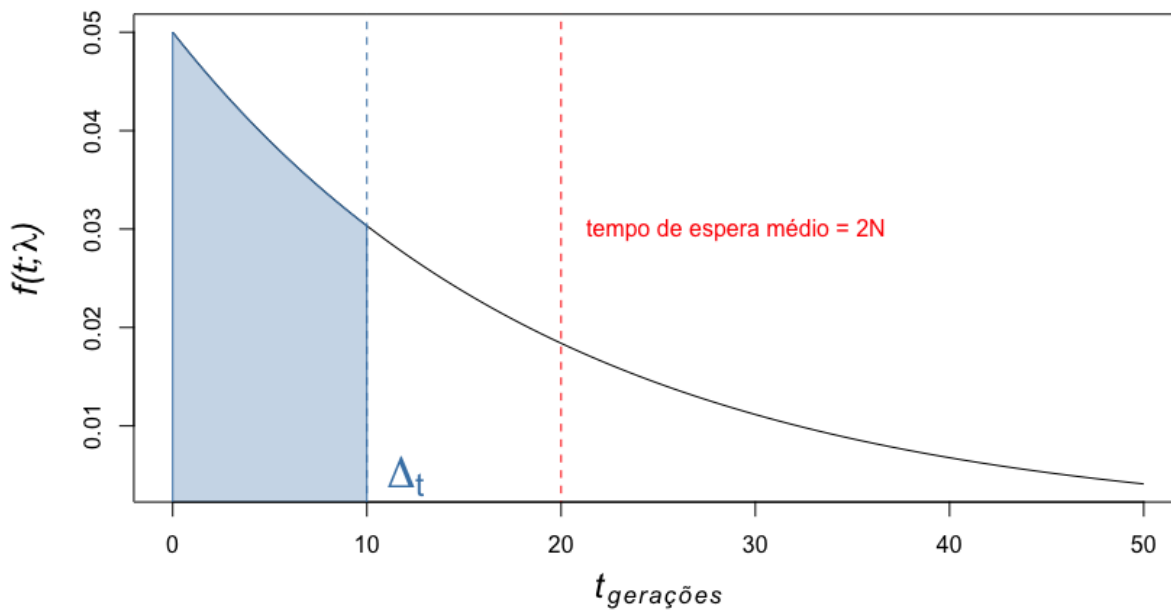
```
segments(10,0,10,1, col="steelblue", lty=2)

t_fill <- seq(0, 10, length.out = 100)
y_fill <- dexp(t_fill, rate = 1/(2*N))

# plotar um polígono definido pelos pontos acima sob a curva:
polygon(c(t_fill, rev(t_fill)), c(rep(0, length(t_fill)), rev(y_fill)), col = rgb(70/255, 130/255, 180/255, 0.3), border = rgb(70/255, 130/255, 180/255) )

text(12, 0.005, expression(Delta[t]), col="steelblue", cex=2)
```

### Distribuição do tempo de espera para um par de alelos (N = 10)



Como a função  $f(t; \lambda)$  é uma **função de densidade de probabilidade**, a área sob a curva retorna justamente a probabilidade associada. O valor da área total sob a curva é 1 (ou seja, 100%). Mas estamos interessados apenas na área entre 0 e  $\Delta_t$ . Esse valor pode ser obtido pela seguinte integral:

$$\int_0^{\Delta_t} f(t; \lambda) dt = \int_0^{\Delta_t} \frac{1}{2N} e^{-\frac{1}{2N}t} dt$$

Podemos resolver essa integral no chatGPT ( $d = \Delta_t$ ):



**You**

Suponha que uma variável  $t$  seja exponencialmente distribuída com valor de taxa de decaimento  $\lambda = 1/(2*N)$ . Calcule a integral definida de 0 a  $d$ . Apresente cada passo da solução da integral.



**ChatGPT**

Para calcular a integral definida da variável  $t$ , que segue uma distribuição exponencial com taxa de decaimento  $\lambda = \frac{1}{2N}$ , de 0 a  $d$ , vamos seguir os seguintes passos:

1. **Definir a Função:** A função da distribuição exponencial é dada por  $f(t) = \lambda e^{-\lambda t}$ . No nosso caso,  $\lambda = \frac{1}{2N}$ , então  $f(t) = \frac{1}{2N} e^{-\frac{t}{2N}}$ .
2. **Escrever a Integral Definida:** A integral que queremos calcular é  $\int_0^d \frac{1}{2N} e^{-\frac{t}{2N}} dt$
3. **Calcular a Integral:** Para calcular esta integral, usamos a regra de integração para funções exponenciais. A integral de uma função exponencial  $e^{at}$  é  $\frac{1}{a} e^{at} + C$ , onde  $C$  é a constante de integração. Então, aplicando essa regra, temos:

$$\int \frac{1}{2N} e^{-\frac{t}{2N}} dt = -e^{-\frac{t}{2N}} + C$$

4. **Aplicar os Limites da Integral:** Agora, aplicamos os limites de 0 a  $d$  na integral calculada.

Substituindo  $t$  por  $d$  e 0, obtemos:

$$\left[ -e^{-\frac{d}{2N}} + C \right] - \left[ -e^{-\frac{0}{2N}} + C \right]$$

Simplificando, considerando que  $e^0 = 1$ , temos:

$$\begin{aligned} & -e^{-\frac{d}{2N}} - (-1) \\ & = 1 - e^{-\frac{d}{2N}} \end{aligned}$$

Portanto, a integral definida de 0 a  $d$  da variável  $t$  distribuída exponencialmente com taxa de decaimento  $\lambda = \frac{1}{2N}$  é  $1 - e^{-\frac{d}{2N}}$ .



Portanto,

$$\int_0^{\Delta_t} f(t; \lambda) dt = \int_0^{\Delta_t} \frac{1}{2N} e^{-\frac{1}{2N}t} dt = 1 - e^{-\frac{\Delta_t}{2N}}$$

Assim, a probabilidade do caso I acontecer é  $1 - e^{-\frac{\Delta_t}{2N}}$ .

Probabilidade de coalescência na espécie ancestral ABC (casos II, III e IV):

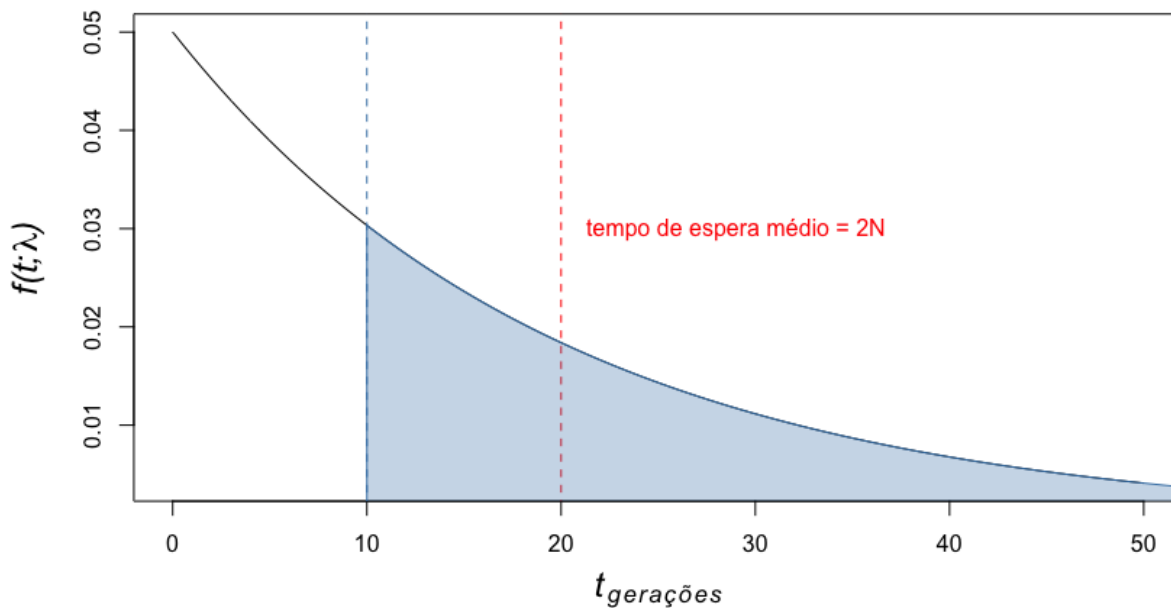
Após resolver o caso I, é relativamente simples calcular os casos II, III e IV, ou seja, os casos de coalescência profunda. A probabilidade total de cada um desses três casos pode ser visualizada no seguinte gráfico:

```
segments(10,0,10,1, col="steelblue", lty=2)

t_fill <- seq(10, 100, length.out = 100)
y_fill <- dexp(t_fill, rate = 1/(2*N))

polygon(c(t_fill, rev(t_fill)), c(rep(0, length(t_fill)), rev(y_fill)), col = rgb(70/255, 130/255, 180/255, 0.3), border = rgb(70/255, 130/255, 180/255) )
```

Distribuição do tempo de espera para um par de alelos (N = 10)



Essa área sob a curva representa a chance dos alelos *a* e *b* não coalescerem no intervalo  $\Delta_t$ , ou seja, na espécie ancestral *AB*. Obviamente, como o valor da área total sob a curva é 1 e sabemos que a área de 0 até  $\Delta_t$  é  $1 - e^{-\frac{\Delta_t}{2N}}$ , a área da cauda restante mostrada na figura acima será

$$1 - \left(1 - e^{-\frac{\Delta_t}{2N}}\right) = e^{-\frac{\Delta_t}{2N}}$$

Notemos que o valor  $e^{-\frac{\Delta_t}{2N}}$  se refere à **probabilidade total** dos casos II, III e IV. Como essas são coalescências profundas que ocorrem na espécie ancestral *ABC* e, portanto, antes mesmo destas espécies virem a existir, **não há razão para que algum desses casos seja mais provável que outro**. Podemos, portanto, dividir essa probabilidade total entre os três casos para calcular o valor de II, III e IV. **Cada um dos três casos**, ocorrerá com chance:

$$\frac{e^{-\frac{\Delta_t}{2N}}}{3}$$

Podemos sumarizar nossos cálculos na seguinte tabela:

| Caso | Topologia da árvore de gene | Probabilidade de ocorrência          |
|------|-----------------------------|--------------------------------------|
| I    | $((a,b),c)$                 | $1 - e^{-\frac{\Delta_t}{2N}}$       |
| II   | $((a,b),c)$                 | $\frac{e^{-\frac{\Delta_t}{2N}}}{3}$ |
| III  | $((b,c),a)$                 | $\frac{e^{-\frac{\Delta_t}{2N}}}{3}$ |
| IV   | $((a,c),b)$                 | $\frac{e^{-\frac{\Delta_t}{2N}}}{3}$ |

Com esses valores, podemos calcular a **probabilidade de equivalência topológica entre as árvores de genes e a árvore de espécies**. Essa equivalência ocorre nos casos I e II. Desta forma, a probabilidade de a árvore do gene (a *gene tree*, *gt*) ser topologicamente igual à árvore das espécies (a *species tree*, *st*) será

$$\Pr(gt = st) = \left(1 - e^{-\frac{\Delta_t}{2N}}\right) + \left(\frac{e^{-\frac{\Delta_t}{2N}}}{3}\right) = 1 - \frac{2e^{-\frac{\Delta_t}{2N}}}{3}$$

O que exatamente esta equação descreve? Uma interpretação desta equação é: **a probabilidade de uma reconstrução filogenética (com algum método filogenético) usando uma região genômica, homóloga entre as três espécies, resultar numa topologia idêntica à filogenia correta entre as espécies é justamente  $1 - (2e^{-\Delta_t/2N})/3$** . Alternativamente, podemos interpretar que se um número muito grande de marcadores moleculares (regiões genômicas) forem amostrados, uma proporção igual a este mesmo valor terá topologia equivalente à filogenia correta. Entretanto, **essas regiões devem estar em equilíbrio de ligação**. Regiões ligadas apresentarão interferência espacial, pois não segregam independentemente. Desta forma, as histórias de suas coalescências não serão independentes. Portanto, não teremos a independência estatística necessária para aplicar a fórmula acima.

Além disto, esta equação assume que as regiões genômicas que evoluem sem a interferência da seleção natural. **A seleção afetará a taxa de coalescência dos genes e, portanto, esse parâmetro não será compartilhado entre as regiões genômicas**, violando um dos pressupostos do modelo de coalescência multi-espécies.

Em 2001, Chen e Li (2001) usaram esta equação para calcular o tamanho efetivo da espécie ancestral de humanos e chimpanzés. Eles amostraram 53 loci autossômicos e reconstruíram a árvore filogenética de cada um deles (árvores de genes). Verificaram que apenas 69% dessas árvores eram topologicamente idênticas à árvore de espécies (*(Homo, Pan), Gorilla*). Assumindo que o tamanho do ramo interno que separa as divergências de *(Homo-Pan-Gorilla)* e *(Homo-Pan)*, ou seja,  $\Delta_t$ , tem aproximadamente 2 milhões de anos (1 geração = de 15 a 20 anos), usando a equação acima, eles forneceram uma estimativa para o tamanho efetivo da espécie ancestral (1 geração = 17.5 anos):



$$\Pr(gt = st) = 1 - \frac{2e^{-\frac{\Delta t}{2N}}}{3}$$
$$0.69 = 1 - \frac{2e^{-\frac{\frac{2 \times 10^6}{17.5}}{2N}}}{3}$$
$$N \approx 74625$$

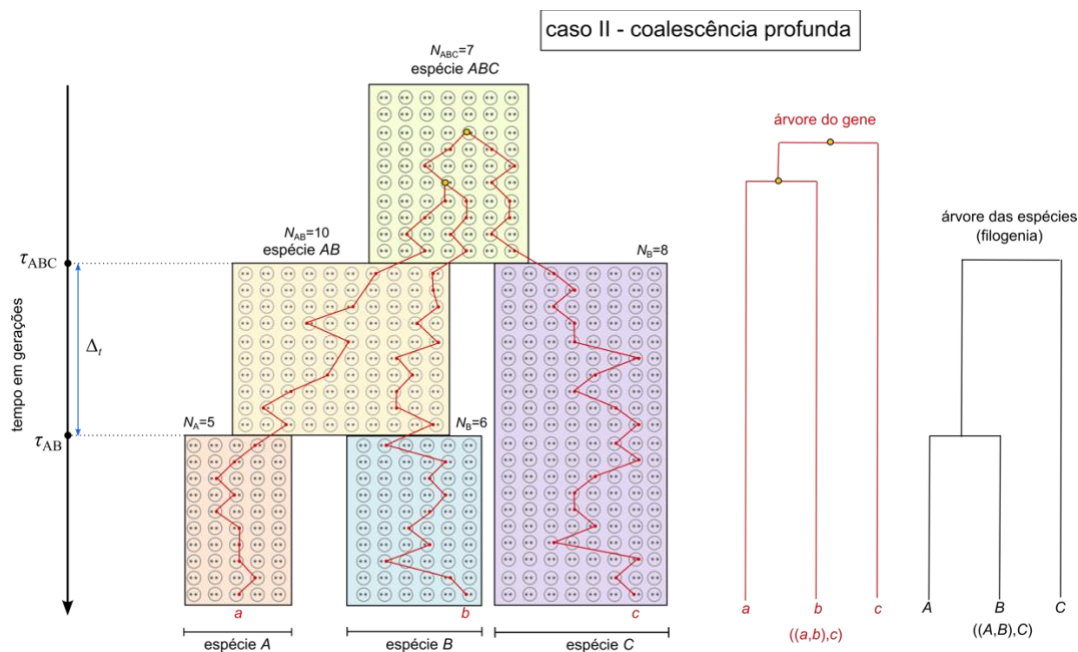
Evidentemente, esse valor está sujeito a um erro associado muito grande, pois ele depende de outras estimativas que já possuem erros em seus cálculos.

### Referências

- Chen F-C, Li W-H. 2001. Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees. *The American Journal of Human Genetics* 68:444–456.
- Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* 23:183–201.
- Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5:568–583.
- Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

## Interpretando filogenias num contexto populacional

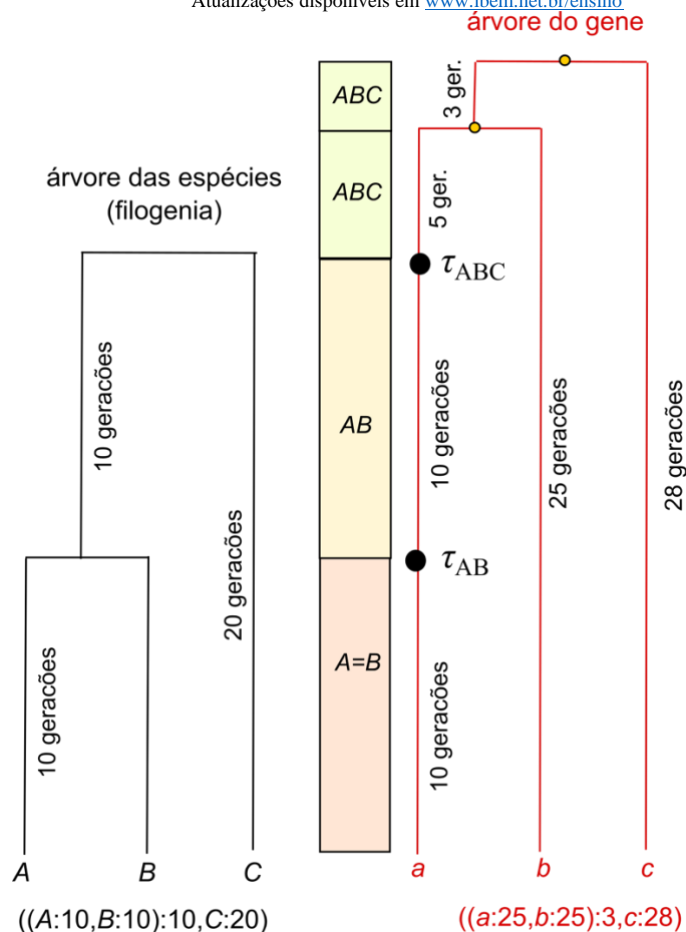
O arcabouço teórico estabelecido pelo modelo da coalescência multi-espécies pode ser usado para análise de vários dos resultados comumente obtidos em filogenética molecular. Neste estudo faremos uma avaliação mais detalhada desses casos. Inicialmente precisamos, entretanto, esclarecer melhor como transformar os resultados estudados no estudo anterior (as expectativas teóricas) com o tipo de dado ou resultado que observamos na prática da reconstrução filogenética. Uma primeira questão que precisamos resolver está relacionada às unidades dos tamanhos dos ramos das árvores de genes e da filogenia. Voltemos ao caso II estudado anteriormente, que apresenta uma coalescência profunda com topologia do gene idêntica à filogenia:



Observemos que a escala de tempo das árvores apresentadas está em número de gerações. Nas populações correspondentes às espécies A e B, verificamos que se passaram 10 gerações desde o isolamento reprodutivo completo ( $\tau_{AB} = 10$ ). Afinal, existem 10 linhas de indivíduos Wright-Fisher nestas populações. A população da espécie ancestral AB também tem duração de 10 gerações, portanto,  $\Delta_t = 10$  e  $\tau_{ABC} = \tau_{AB} + \Delta_t = 20$  gerações. Por fim, as duas coalescências destacadas com os pontos amarelos ocorrem na espécie ancestral ABC. A primeira delas, a coalescência (a,b) ocorre após 5 gerações e a coalescência final ((a,b),c) ocorre 3 gerações depois.

Vamos então inserir esses valores na árvore do gene e compará-los com a árvore de espécies (=filogenia):



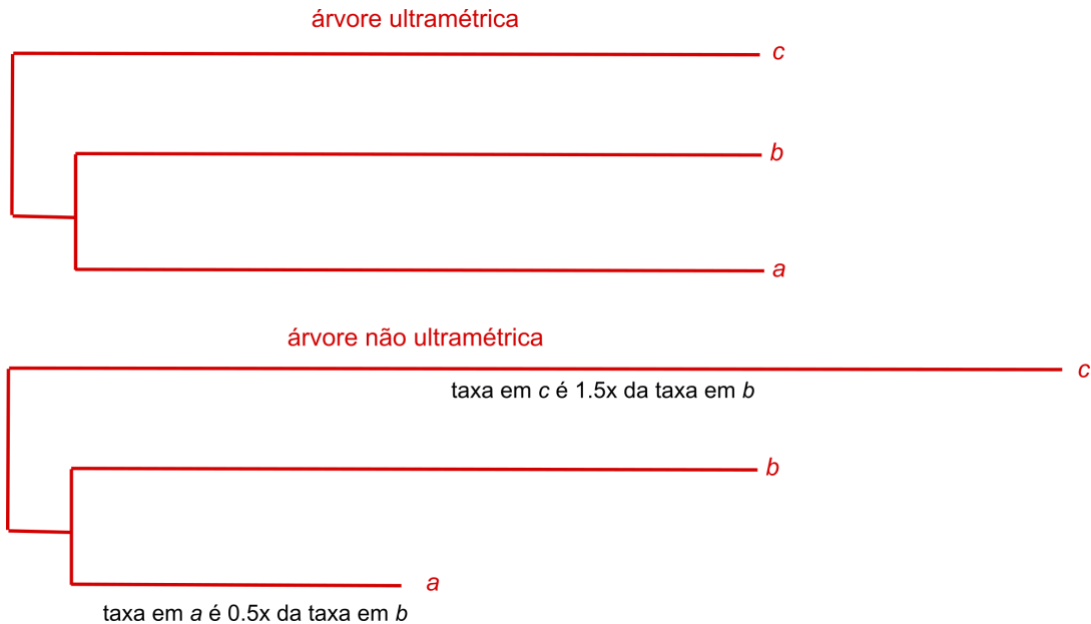


É notável que, embora exista correspondência topológica entre a árvore de gene e a árvore de espécies, os tamanhos dos ramos (medidos em número de gerações) é bastante diferente entre elas. Isso pode ser sumarizado usando a **notação Newick** de ambas as árvores – ((A:10, B:10):10, C:20) e ((a:25, b:25):3, c:28). Portanto, fica evidente que **as idades das especiações  $\tau_{AB}$  e  $\tau_{ABC}$ , ou seja, os isolamentos reprodutivos completos de A-B e AB-C, não são equivalentes às idades dos nós na árvore dos genes.**

A mensuração dos tamanhos dos ramos em número de gerações é, entretanto, pouco útil. Na prática, quando reconstruímos filogenias moleculares, não é possível medir a diferença entre as sequências em número de gerações. Efetivamente, a unidade mais utilizada é o **número de substituições por sítio** de nucleotídeo ou aminoácido. Para transformarmos as unidades acima neste valor, é necessário conhecer a taxa de substituição por geração em cada uma das espécies (viventes e ancestrais) da árvore de espécies. Novamente, não possuímos uma medida desse valor. As melhores medidas que possuímos são oriundas de estudos de sequenciamento de genomas de trios familiares. Bergeron et al. (Bergeron et al., 2023) fez um estudo comparativo do valor da taxa de substituição geracional, ou seja, da linhagem germinativa, entre vários vertebrados. Conforme reportado pelos autores, existe uma variação considerável desta taxa entre as espécies. Para humanos, o valor é aproximadamente  $1.5 \times 10^{-8}$  substituições por sítio por geração. Usemos esse valor para transformar as árvores acima, basta multiplicar o número de gerações pela taxa por geração:

| Árvore    | Gerações                | Subst./sítio/geração                              |
|-----------|-------------------------|---|
| Gene      | ((a:25, b:25):3, c:28)  | ((a: 3.75e-07, b: 3.75e-07): 4.5e-08, c: 4.2e-07) |
| Filogenia | ((A:10, B:10):10, C:20) | ((A: 1.5e-07, B: 1.5e-07): 1.5e-07, C: 3.0e-07)   |

As árvores medidas em substituição/sítio/geração são **ultramétricas**, ou seja, os tamanhos dos ramos são proporcionais às coalescências dos genes ou das especiações (filogenia). Isso acontece, pois as unidades geracionais foram multiplicadas pela mesma taxa ( $1.5 \times 10^{-8}$ ). A constância de taxas evolutivas em todas as espécies é o pressuposto do relógio molecular. É mais comum, entretanto, que as taxas apresentem variação entre as espécies, resultando em árvores **não ultramétricas**. Por exemplo:

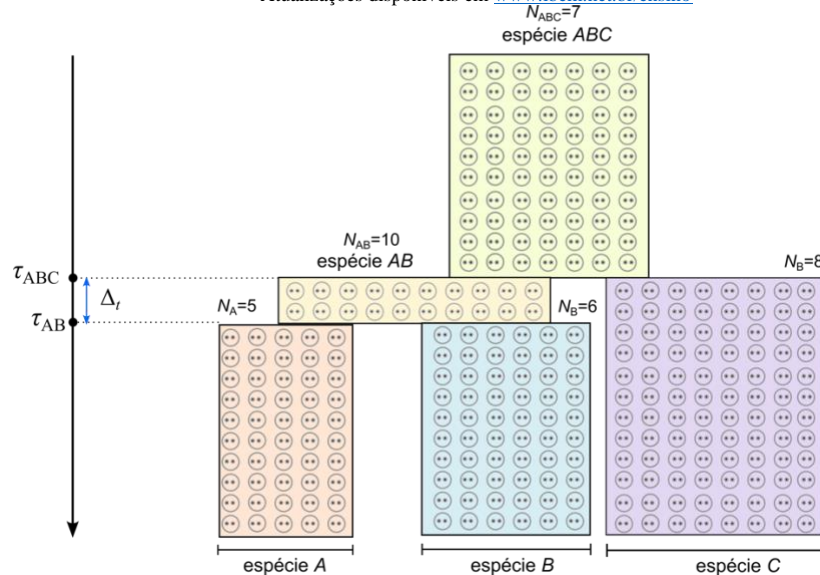


Unidades de coalescência e probabilidade de estimar a filogenia correta das espécies

Voltemos à equação que descreve a probabilidade de uma árvore de gene ser possuir a mesma topologia da árvore das espécies:

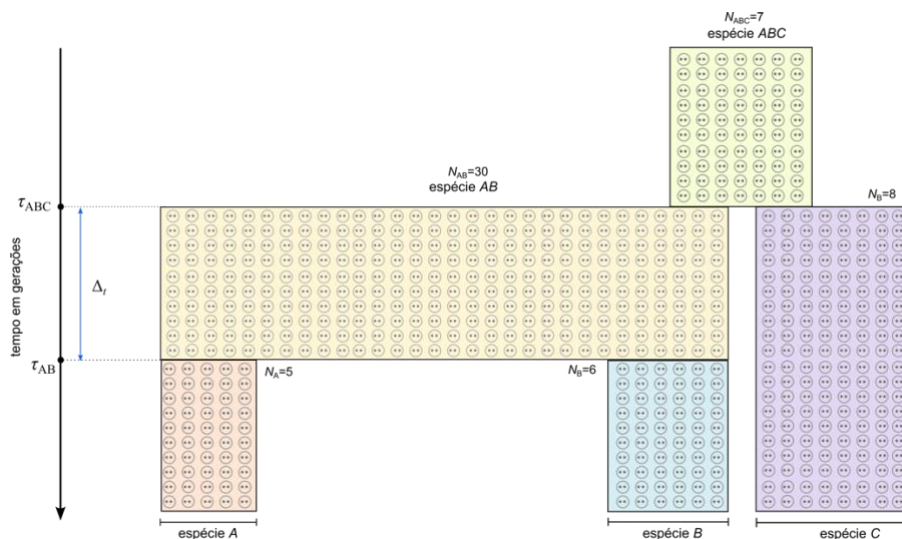
$$\Pr(gt = st) = 1 - \frac{2e^{-\frac{\Delta_t}{2N}}}{3}$$

Nesta equação fica evidente que duas variáveis são relevantes para dizer o quão provável é obter no genoma um gene que possui a mesma topologia da filogenia –  $\Delta_t$ , o tamanho do ramo interno, ou seja, o intervalo entre as especiações e  $N$ , o tamanho populacional efetivo da espécie ancestral que ocupa este ramo interno. Esta equação indica que em situações em que  $\Delta_t$  é muito curto, a probabilidade de equivalência topológica é pequena. Por exemplo:



Neste caso acima, a chance de a árvore de gene ser topologicamente idêntica à filogenia ocorre porque o intervalo entre as especiações  $ABC$  e  $AB$  é tão curto ( $\Delta_t = 2$  gerações apenas) que dificilmente ocorrerá coalescência das linhagens alélicas oriundas da espécie  $A$  e da  $B$  dentro desta população ancestral  $AB$ . A maioria das coalescências ocorrerão na espécie ancestral  $ABC$  e consistirão, portanto, de coalescências profundas. Neste caso, cada uma das topologias possíveis ( $((a,b),c)$ ;  $((b,c),a)$  e  $((a,c),b)$ ) serão igualmente prováveis. Este cenário acima é semelhante a eventos de **radiação adaptativa**, onde os eventos de especiação ficam muito intensos e ocorrem em intervalos muito curtos. Numa situação limite, onde  $\Delta_t = 0$ , a probabilidade de equivalência topológica será apenas  $1/3$ . **Isso indica que, se um pesquisador coletar várias regiões genômicas em equilíbrio de ligação ao longo do genoma, e reconstruir a árvore filogenética com elas, apenas 33.3% (1/3) serão topologicamente iguais à topologia da correta da filogenia.**

Verifiquemos agora outro cenário interessante, quando a o tamanho da população ancestral do ramo interno é muito grande:



Neste novo cenário, a probabilidade de ocorrência de coalescências na espécie ancestral  $AB$  (caso I do estudo anterior) também é menor. Lembremos que as linhagens alélicas oriundas da espécie  $A$  e



da espécie  $B$  terão tempo médio de coalescência igual a  $2N$  gerações. Como o tamanho efetivo de  $AB$  aumentou de 10 para 30, o tempo médio subiu de 20 ( $2 \times 10$ ) para 60 ( $2 \times 30$ ) gerações. Como o tamanho do ramo interno  $\Delta_t$  é de 10 gerações (portanto,  $< 60$  gerações), dificilmente ocorrerá uma coalescência dentro deste intervalo. Novamente, a maioria das coalescências serão profundas.

Desta forma, quando regiões genômicas em equilíbrio de ligação resultam numa frequência pequena de árvores de genes topologicamente equivalentes à filogenia correta, pode ser resultado de um  $\Delta_t$  curto, de um  $N$  ancestral grande ou de ambos. Na prática, é difícil de decompor estes dois valores. Por esta razão, ao trabalhar com o modelo coalescência multi-espécies, é comum medir o tamanho de ramo da filogenia em **unidades de coalescência (u.c.)**, definida justamente por:

$$u. c. = \frac{\Delta_t}{2N}$$

Quando esta razão é igual a  $\Delta_t/2N = 1$ , temos uma unidade de coalescência, ou seja, o tamanho do ramo equivale a  $2N$  gerações. Se  $\Delta_t/2N = 2$ , o tamanho de ramo equivale a  $4N$  gerações ( $2 \times 2N$ ). Fica claro, portanto, que quanto maior for o tamanho de ramo em unidades de coalescência, maior será a chance de ocorrer o caso I, diminuindo a probabilidade de coalescências profundas.

Podemos visualizar no R a probabilidade de equivalência topológica entre árvore de gene e árvore de espécies usando esta nova parametrização

$$\Pr(gt = st) = 1 - \frac{2e^{-u.c.}}{3}$$

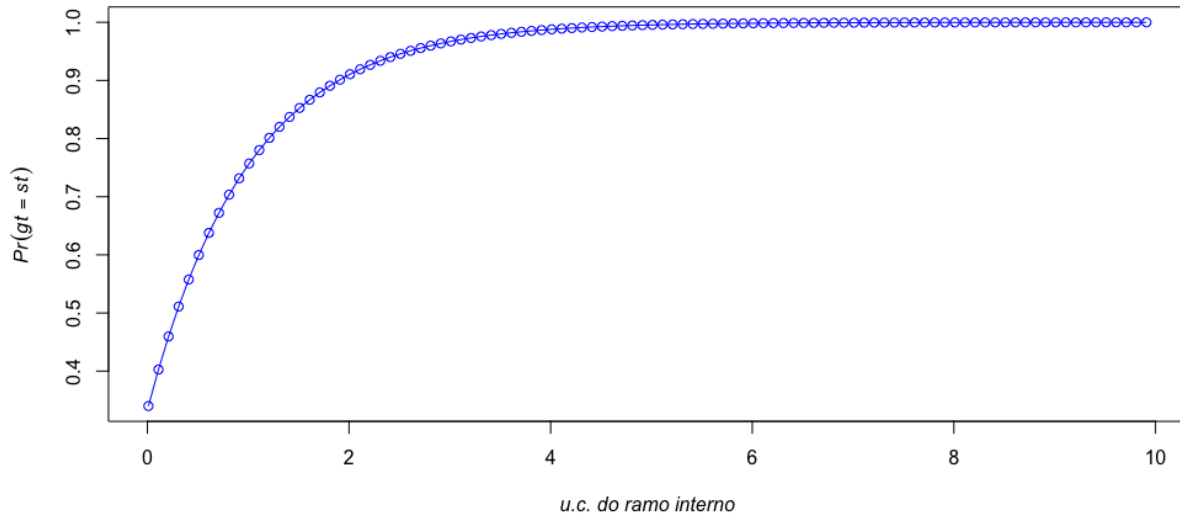
```
uc = seq(from=0.01, to=10, by=0.1) # unidades de coalescência analisadas
pr_gtst = vector()

for(i in uc){
  pr_gtst = c(pr_gtst, 1 - ( (2/3)*exp(-i) ) ) # fórmula da prob. de equivalência topológica
}

plot(pr_gtst ~ uc, ylab=expression(italic(Pr(gt==st)) ), xlab=expression(italic("u.c. do ramo interno")),
main="Probabilidade de equivalência topológica entre\ndárvore de gene e árvore de espécies", col="blue" )

lines(pr_gtst ~ uc, col="blue")
```

### Probabilidade de equivalência topológica entre árvore de gene e árvore de espécies



O gráfico acima mostra que o valor mínimo de probabilidade de equivalência topológica é justamente  $1/3$  (0.333), pois essa é justamente a frequência das topologias  $((a,b),c)$  entre as coalescências profundas. Conforme o tamanho do ramo interno (medido em unidades de coalescência) aumenta, maior será a chance de reconstruir uma árvore de gene e obter a mesma topologia da árvore das espécies, a filogenia.

### Referências

Bergeron, L.A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M.F., Quintard, B., Hoffman, J.I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M.T.P., Schierup, M.H., Zhang, G., 2023. Evolution of the germline mutation rate across vertebrates. *Nature* 1–7. <https://doi.org/10.1038/s41586-023-05752-y>