¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Introducción a R.

José Víctor Moreno Mayar.
mayar@ibt.unam.mx
jmoreno@lcg.unam.mx

IBT - UNAM

18 de enero del 2010

Visión General de la Clase.

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

1. ¿Qué es R?
2. Usando R.
3. Objetos.
4. Expansiones.
5. Leyendo Datos.
6. Datos Aleatorios y Simulaciones.

Visión General de la Clase.

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

7. Correlaciones, Regresiones y Pruebas estadśticas.
8. Visualización.
9. R en paralelo.

Un poco de historia.

- R es una implementación del lenguaje S.
- S fue creado por John Chambers en los laboratorios Bell.
- R es un proyecto GNU creado por Ross Ihaka y Robert Gentleman en 1993.
- Actualmente lo mantiene el *R Development Core Team*.

¿Por qué R?

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

- R funciona en varias plataformas UNIX, MacOS, Windows.
- R es altamente expandible por medio de paquetes.
- R está relacionado con Bioconductor que es un proyecto enfocado al manejo y análisis de datos genómicos.
- R es un lenguaje de programación/ambiente para realizar cómputo estadístico y gráficas.
  - R cuenta con un amplio conjunto de operadores matemáticos para vectores y matrices.
  - Pruebas estadísticas, modelos y gráficas incluidos.

Instalando R.

- El instalador y/o código para R se descargan de la página del CRAN[1].
- Las noticias más recientes acerca del desarrollo de R se publican aquí.
- Así mismo, se puede descargar una gran diversidad de paquetes.

[1] Comprehensive R Archive Network

Habiendo Instalado R.

- Dependiendo de la plataforma en uso, es posible abrir R GUI o desde una terminal[2].
- Si se usa la terminal se puede accesar a R ingresando el comando R en la misma.
- El comando más importante:

```
> quit()
```

- Si tienes algun objeto interesante o que no quieras perder, será recomendable que guardes el espacio de trabajo.

[2] También existen editores como emacs.

Pidiendo Ayuda.

- Existen diferentes formas para buscar ayuda en R.
- Todo depende de lo que se esté buscando y cómo.
- Si se conoce el nombre exacto de la función sobre la cual se desea saber más, el comando es el siguiente:

```
> `?`(sample)
```

- Si no se conoce exactamente la función de la cual se requiere ayuda, pero se sabe alguna parte del nombre, se puede usar apropos(); así el siguiente paso es usar ?.

```
> apropos("norm")
```

Pidiendo Ayuda.


```
 [1] "dlnorm"          "dnorm"
 [3] "normalizePath"   "plnorm"
 [5] "pnorm"           "qlnorm"
 [7] "qnorm"           "qqnorm"
 [9] "qqnorm.default"  "rlnorm"
[11] "rnorm"
```

Más Ayuda.

- Otra opción para buscar ayuda no tan específica y más amigable:

```
> help.start()
```

- El mejor lugar para obtener tips que pueden ser funcionales para muchos usuarios es la mailing list.
- Otra función muy útil es args(), la cual regresa los argumentos de una función determinada.

Objetos.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

- R puede ser usado como una simple calculadora.
- R trabaja con vectores variables.
- Para asignar variables se usa el operador $<-$ [3].
- Uno de los operadores más usados en R es c(), el cual sirve para asignar varios elementos en un vector.
- También se puede operar con objetos de R.
- Siempre hay que tomar en cuenta la regla del reciclaje.

```
> 3 + 5
[1] 8
> 6 * 3/9
[1] 2
```

Objetos.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
> V <- 33
> V <- c(33, 55)
> M <- c(33, 77, 53, 27, 21)
> M * V

[1] 1089 4235 1749 1485  693
```

[3] Similar a $=$, sólo que este se usa dentro de las funciones.

Mode.

- Los objetos pueden ser caracteres, numéricos y lógicos.
- Algunas funciones requieren que sus argumentos sean de un tipo determinado.
- Siempre se puede hacer una forma de casting.

```
> text <- c("Esto", "es", "character")
> mode(V)

[1] "numeric"

> mode(text)

[1] "character"

> v <- as.character(V)
> mode(v)
```

Mode.


```
[1] "character"
```

```
> logical <- c(rep(1, 5), rep(0,
+     5))
> as.logical(logical)
```

```
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
 [7] FALSE FALSE FALSE FALSE
```

Más Tipos de Datos.

- Además de los arreglos/vectores, existen matrices, data frames, y listas.
- Los data frames pueden tener variables de diferente tipo, al igual que las listas y a diferencia de las matrices.
- Existen diferentes métodos para accesarlos. [4]

```
> mat <- matrix(1:25, 5, 5)
> mat
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    6   11   16   21
[2,]    2    7   12   17   22
[3,]    3    8   13   18   23
[4,]    4    9   14   19   24
[5,]    5   10   15   20   25
```

Más Tipos de Datos.


```
> mat[, 1]

[1] 1 2 3 4 5

> mat[1, ]

[1]  1  6 11 16 21

> mat[4, 4]

[1] 19

> type <- factor(c(rep("a", 5), rep("b",
+     5)))
> vals <- 1:10
> df <- data.frame(vals, type)
> df
```

Más Tipos de Datos.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

```
   vals type
1     1    a
2     2    a
3     3    a
4     4    a
5     5    a
6     6    b
7     7    b
8     8    b
9     9    b
10   10    b
```

```
> df$vals
 [1]  1  2  3  4  5  6  7  8  9 10
```

Más Tipos de Datos.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
> lista <- list(mat = mat, df = df)
> lista[[1]]
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    6   11   16   21
[2,]    2    7   12   17   22
[3,]    3    8   13   18   23
[4,]    4    9   14   19   24
[5,]    5   10   15   20   25
```

```
> lista$mat
```

Más Tipos de Datos.

Métodos Estadísticos y Analíticos de Datos Genómicos.
Víctor Moreno
¿Qué es R?
Usando R.
Objetos.
Expansiones.
Leyendo Datos.
Datos Aleatorios y Simulaciones.
Correlaciones, Regresiones y Pruebas estadśticas.
Visualización.
R en paralelo.

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    6   11   16   21
[2,]    2    7   12   17   22
[3,]    3    8   13   18   23
[4,]    4    9   14   19   24
[5,]    5   10   15   20   25
```

```
> lista[[1]][1, 1]
[1] 1
```

-Se pueden hacer estructuras tan complejas como se desee.

```
> lista2 <- list(lista1 = lista,
+     lista1_1 = lista)
> lista2[[2]]$mat[1, 2]
[1] 6
```

-Los factores sirven para datos agrupados.

[4] Se recomienda utilizar siempre la misma forma para accesar a los objetos como buena práctica de programación.

Instalando paquetes.

- Como ya se mencionó R tiene múltiples posibilidades de expansión.
- Esto se logra a través de paquetes.
- Los paquetes deben ser instalados y agregados a la sesión.

```
> install.packages("lattice")
> library(lattice)
```

- Algo similar ocurre con los paquetes de Bioconductor..
- Sin embargo estos se descargan con otro comando y hay que instalarlo primero.
- Más adelante se hablará más al respecto.

```
> source("http://bioconductor.org/biocLite.R")
```

Instalando paquetes.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
> biocLite()
```

-Los paquetes de Bioconductor se instalan y se usan de igual forma.

```
> biocLite("affy")
```

```
> library(affy)
```

read.table

- En muchas ocasiones, los datos que serán analizados o manipulados no se generan en R.
- Pueden ser la salida de algún otro programa.
- R cuenta con funciones para leer distintos tipos de archivos.
- read.table() es la función más importante para leer archivos.
- Siempre regresa un data.frame.
- Otras funciones interesantes son read.delim() y read.csv()

```
> freqs <- read.table(file = "freqmatrix",
+     header = T, sep = "\t")
> class(freqs)
```

read.table


```
[1] "data.frame"
```

```
> head(freqs)
```

```
  freqs71 freqs72 freqs81 freqs82
1     243     249     836     793
2      17      22     239     236
3     113     127     256     243
4     339     363     521     480
5     252     265     414     409
6     133     154     365     352
          gene
1 RHE_CH00001
2 RHE_CH00002
3 RHE_CH00007
4 RHE_CH00008
```

read.table

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
5 RHE_CH00009
6 RHE_CH00010
```

Secuencias y Muestreo.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

- Como ya se mencionó R es muy potente cuando se trata de cómputo estadístico.
- En esta rama, el muestreo aleatorio es muy importante.
- Cuando este muestreo se relaciona con datos genómicos, la generación de secuencias aleatorias es indispensable.[5]

```
> sample(1:10, 30, replace = T, prob = seq(from = 0.0
+     to = 0.95, by = 0.1))

 [1]  6  8  6  9  8  7 10  7  2 10  4  4
[13]  8  9 10 10  8  4  9  8 10  4  9 10
[25] 10  8  9  4  8  5

> sample(c(rep("A", 10), rep("T",
+     10), rep("G", 15), rep("C",
+     15)), 50, replace = F)
```

Secuencias y Muestreo.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
 [1] "C" "C" "T" "G" "C" "A" "C" "A" "G"
[10] "A" "C" "T" "A" "T" "T" "C" "A" "G"
[19] "G" "G" "G" "A" "C" "G" "T" "T" "C"
[28] "T" "G" "T" "T" "G" "G" "T" "G" "A"
[37] "C" "G" "C" "A" "C" "C" "A" "C" "C"
[46] "A" "C" "G" "G" "G"
```

[5] Bioconductor cuenta con paquetes especializados en esto.

Distribuciones.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

- Igualmente, se pueden pedir diferentes distribuciones con determinados parámetros.
- Esto será de utilidad cuando se quieran comparar datos obtenidos con distribuciones teóricas.
- Aquí algunos ejemplos[6]:

```
> rnorm(100, 0, 1)
> runif(100, 0, 1)
> rgamma(100, 2, 4)
```

[6] El uso de args() sería apropiado en este caso.

Correlaciones.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

- Cuando se tienen diferentes variables, es importante saber cómo se relacionan.
- El coeficiente de correlación es una buena medida para este propósito.[7]
- cor tiene distintos métodos para calcular correlación como Pearson y Spearman.

```
> x <- seq(0, 50, 0.5)
> y <- rnorm(length(x), 0, 1)
> z <- rnorm(length(x), 0, 1)
> w <- rnorm(length(x), 0, 1)
> cor(x, y)
```

```
[1] 0.0918007
```

Correlaciones.


```
> cor(x, y, method = "spearman")

[1] 0.07641235
```

[7] Siempre es importante recordar que las corrlaciones no lineales no serán detectadas.

Regresiones Lineales.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

- Al igual que las correlaciones, una regresión lineal va a servir para observar cómo se comportan(relacionan) dos variables.
- Una regresión lineal, puede ser útil para interpolar o extrapolar.
- lm utiliza la notación de fórmula que es muy socorrida en R
- lm incluso tiene métodos como predict para extrapolar.[8]

```
> xy <- data.frame(x = x, y = y)
> head(xy)
```

Regresiones Lineales.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

```
    x           y
1 0.0  2.36628808
2 0.5 -1.61989514
3 1.0 -0.72976982
4 1.5  0.85819146
5 2.0  0.82799512
6 2.5  0.02509898
```

```
> reg <- lm(y ~ x, data = xy)
> reg
```

Regresiones Lineales.


```
Call:
lm(formula = y ~ x, data = xy)

Coefficients:
(Intercept)            x
  -0.174961     0.006085
```

[8] Más adelante se visualizará esto.

Pruebas Estadísticas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

- Una de las ventajas de R es la inclusión de diversas pruebas estadísticas paramétricas y no paramétricas.
- Student's t, Kolmogorov-Smirnov, Shapiro-Wilk, χ^2 son algunas de las numerosas pruebas que R es capaz de realizar.

```
> dist <- rnorm(100, 0, 1)
> distmov <- rnorm(100, 10, 1)
> dist2 <- rnorm(100, 0.01, 1)
> distunif <- runif(100, 0, 10)
> shapiro.test(distunif)
```

Pruebas Estadísticas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

```
Shapiro-Wilk normality test

data:  distunif
W = 0.9614, p-value = 0.005093

> shapiro.test(dist)

Shapiro-Wilk normality test

data:  dist
W = 0.971, p-value = 0.02649

> t.test(dist, mu = 50)
```

Pruebas Estadísticas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

```
One Sample t-test

data:  dist
t = -503.3379, df = 99, p-value <
2.2e-16
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 -0.2596753  0.1350278
sample estimates:
  mean of x
-0.06232371
```

```
> t.test(dist, mu = 1)
```

Pruebas Estadísticas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

```
One Sample t-test

data:  dist
t = -10.6808, df = 99, p-value <
2.2e-16
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 -0.2596753  0.1350278
sample estimates:
  mean of x
-0.06232371
```

```
> t.test(dist, distmov)
```

Pruebas Estadísticas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

```
Welch Two Sample t-test

data:  dist and distmov
t = -70.5241, df = 197.764, p-value
< 2.2e-16
alternative hypothesis: true difference in means is n
95 percent confidence interval:
 -10.378009  -9.813406
sample estimates:
  mean of x   mean of y
-0.06232371 10.03338379
```

```
> t.test(dist, dist2)
```

Pruebas Estadísticas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

```
Welch Two Sample t-test

data:  dist and dist2
t = 0.5188, df = 197.893, p-value =
0.6045
alternative hypothesis: true difference in means is n
95 percent confidence interval:
 -0.2020753  0.3463458
sample estimates:
  mean of x   mean of y
-0.06232371 -0.13445894
```

```
> ks.test(dist, dist2)
```

Pruebas Estadísticas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
Two-sample Kolmogorov-Smirnov test

data:  dist and dist2
D = 0.17, p-value = 0.1111
alternative hypothesis: two-sided
```

```
> ks.test(dist, distmov)

Two-sample Kolmogorov-Smirnov test

data:  dist and distmov
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Gráficas.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

- Otra de las ventajas de R es su capacidad para hacer gráficas útiles en la exploración de los datos.
- Algunas gráficas básicas son plot, hist, barplot, boxplot, qqplot, además de un largo etc.

Plot.

Sin duda una de las más usadas.

```
> leg <- c("y~x", "z~x", "w~x", "lm(y~x)")
> cols <- c("red", "purple", "orange",
+     "blue")
> plot(y ~ x, type = "l", col = "red",
+     main = "X vs. Y")
> lines(z ~ x, col = "purple")
> points(w ~ x, col = "orange")
> abline(reg, col = "blue")
> legend("bottomright", leg, col = cols,
+     lty = 1)
```

Plot.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

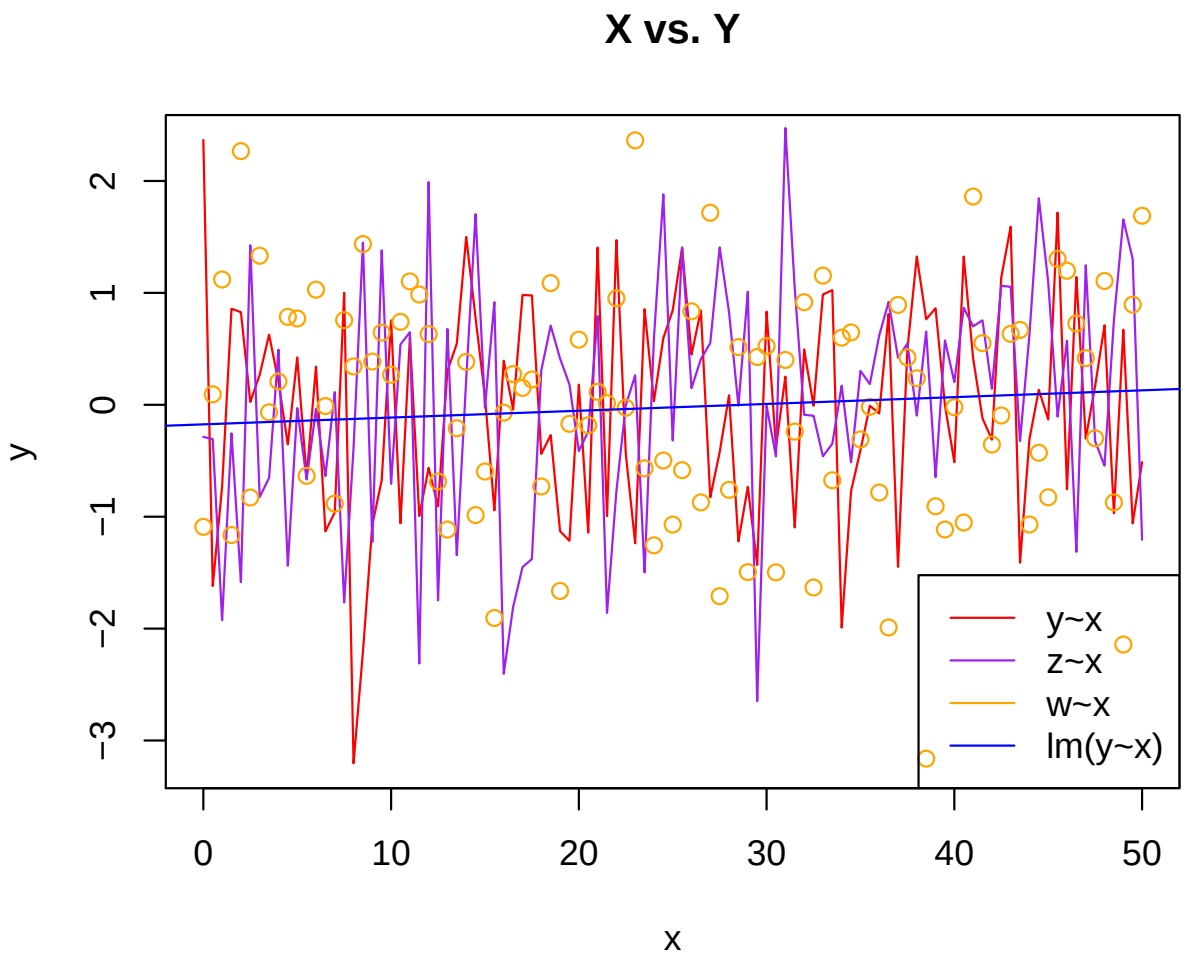
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.



Hist.

Una de las primeras gráficas que hay que hacer al momento de explorar las mediciones.

```
> hist(dist, col = "red", main = "Distribucion de Fre
+      breaks = 20)
```

Hist.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

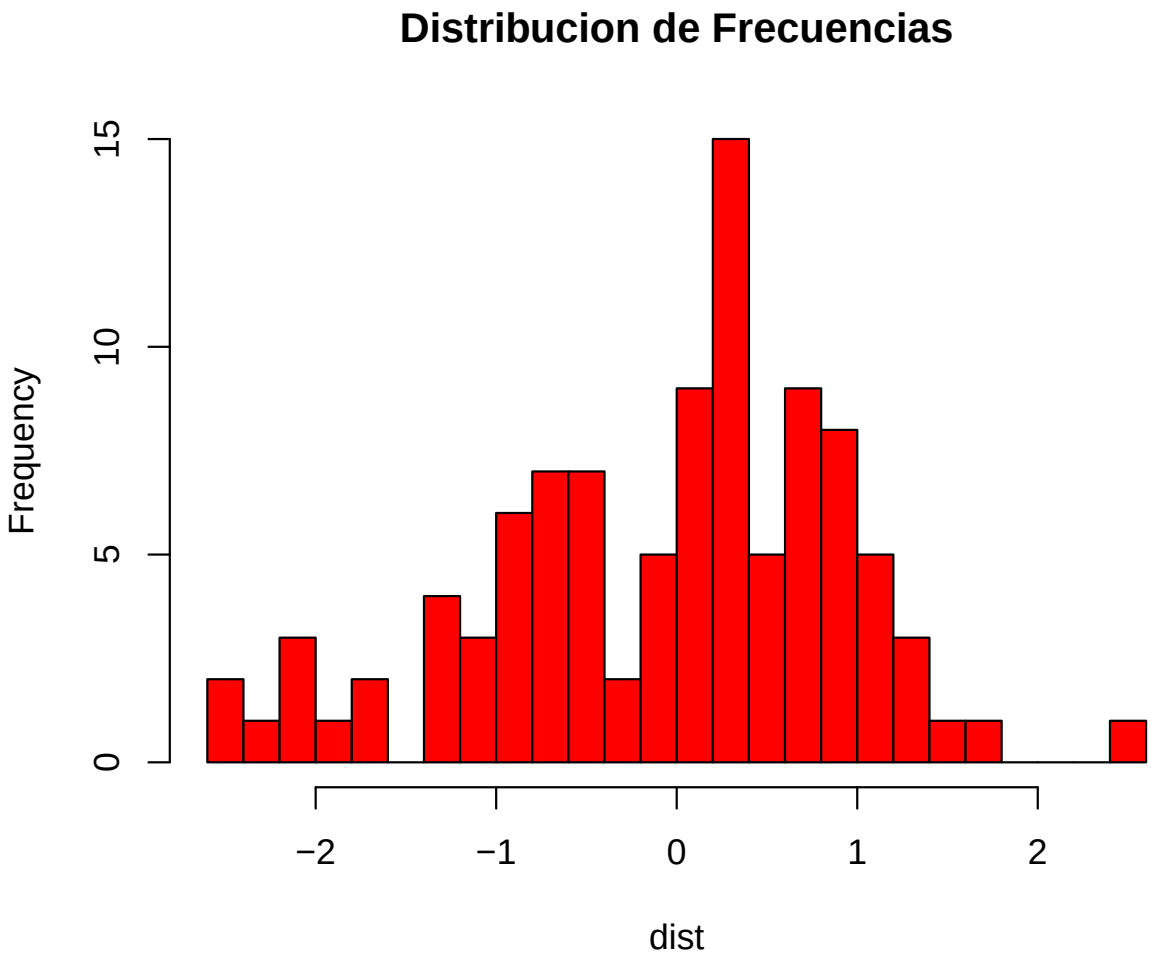
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.



Hist.

- En muchas de las ocasiones, sólo se desea tener líneas de densidad para la variable.

```
> plot(density(dist), col = "red",
+      main = "Distribucion de Frecuencias")
```

Hist.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

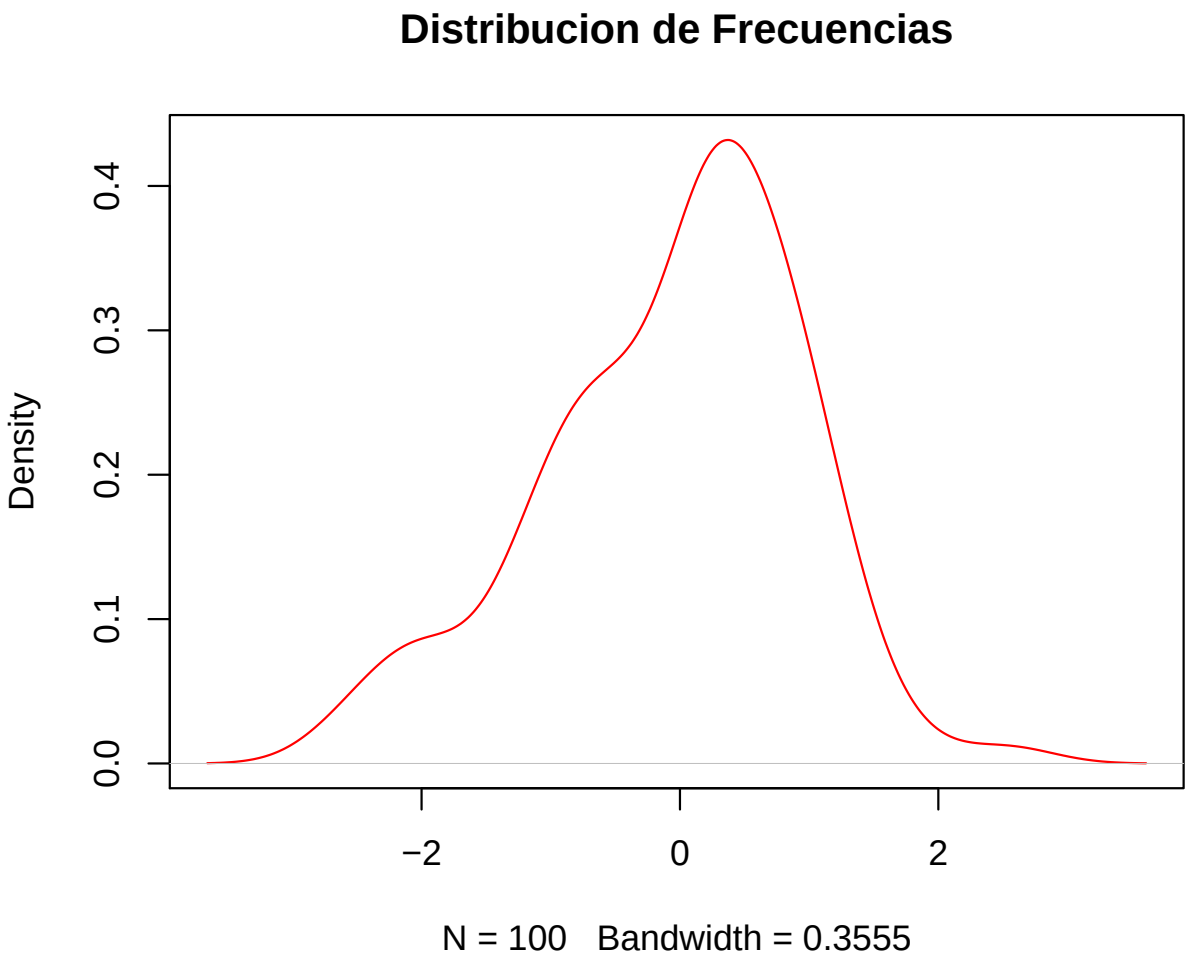
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.



Barplot.

Una de las más famosas cuando se tienen diversas frecuencias.

```
> barplot(freqs[1:20, 2], col = rainbow(20),
+     names.arg = freqs[1:20, 5],
+     las = 2, cex.names = 0.5, ylim = c(0,
+         1500), main = "Frecuencias")
```

Barplot.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

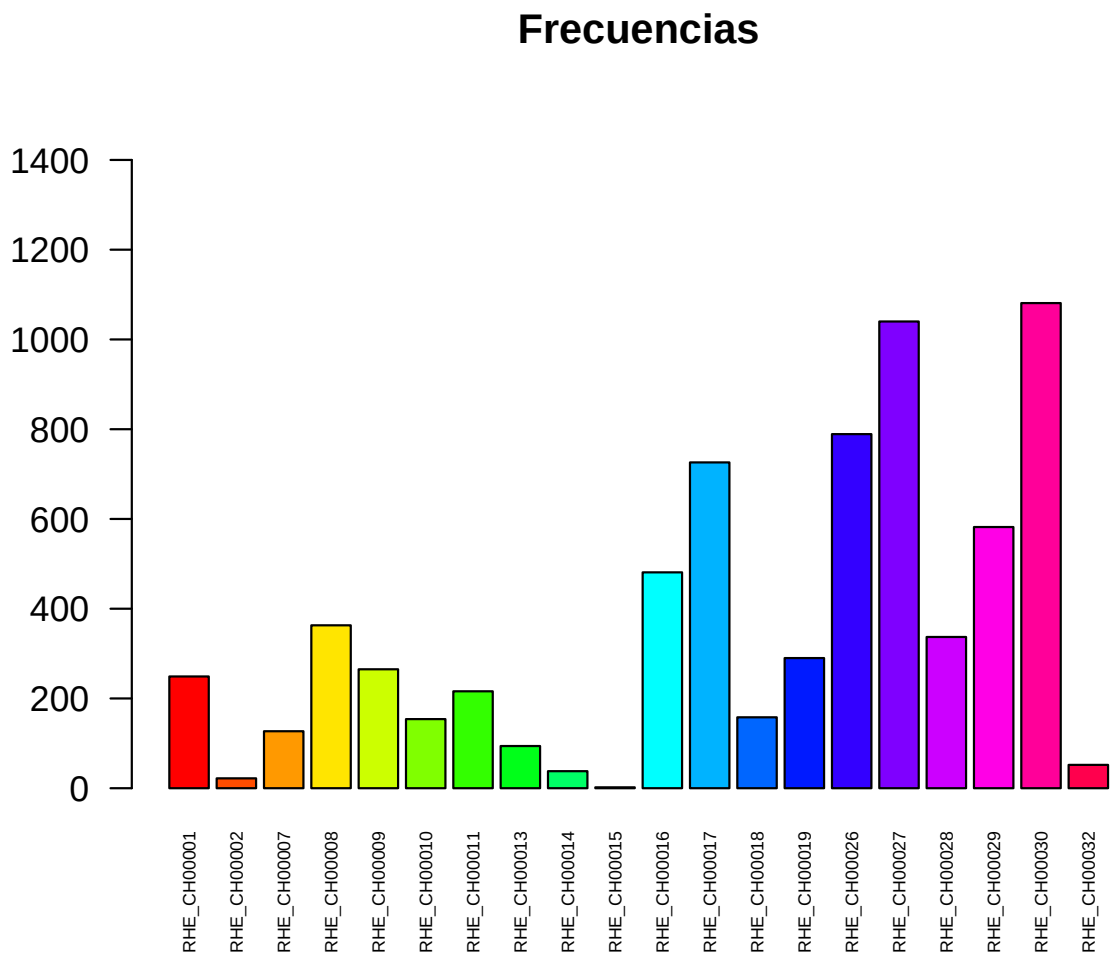
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.



Boxplot.

Una forma efectiva de comparar distintas distribuciones.

```
> boxplot(dist, dist2, distmov, distunif,
+      col = rainbow(4), names = c("dist",
+          "dist2", "distmov", "distunif"),
+      ylim = c(-5, 20))
```

Boxplot.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

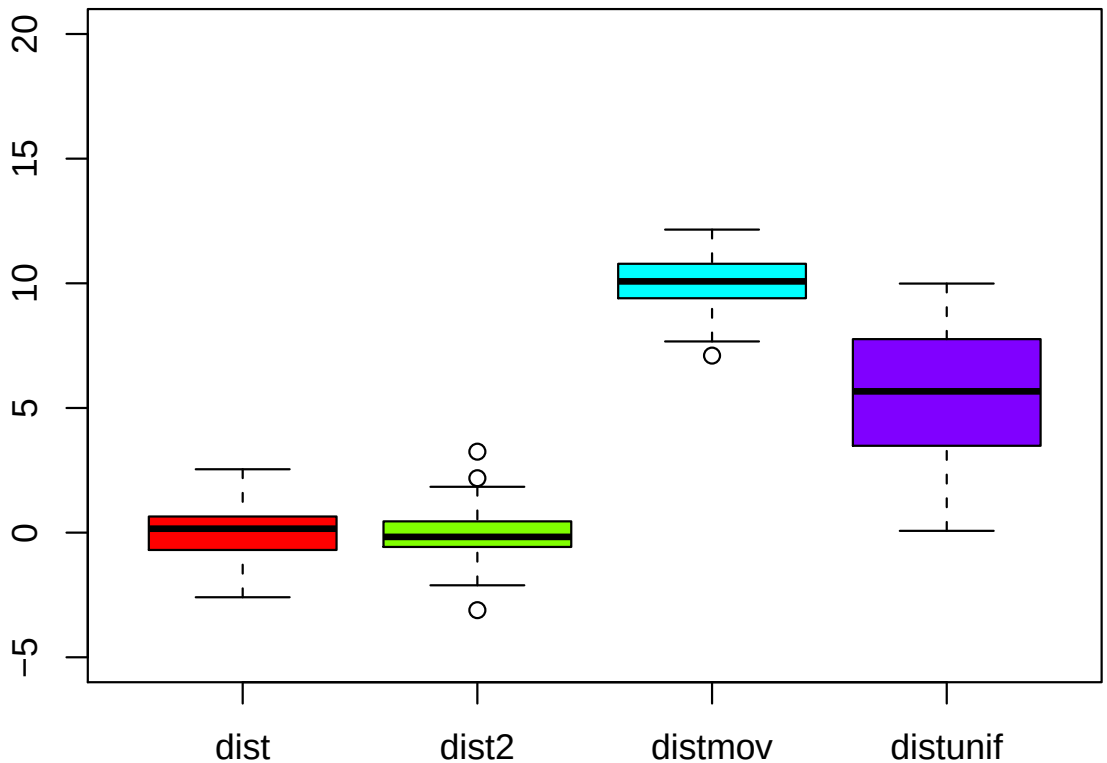
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.



Qqplot.

Útil para comparar una distribución empírica con una teórica.

```
> qqplot(dist, dist2, col = "red",
+      main = "dist vs dist2")
```

Qqplot.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

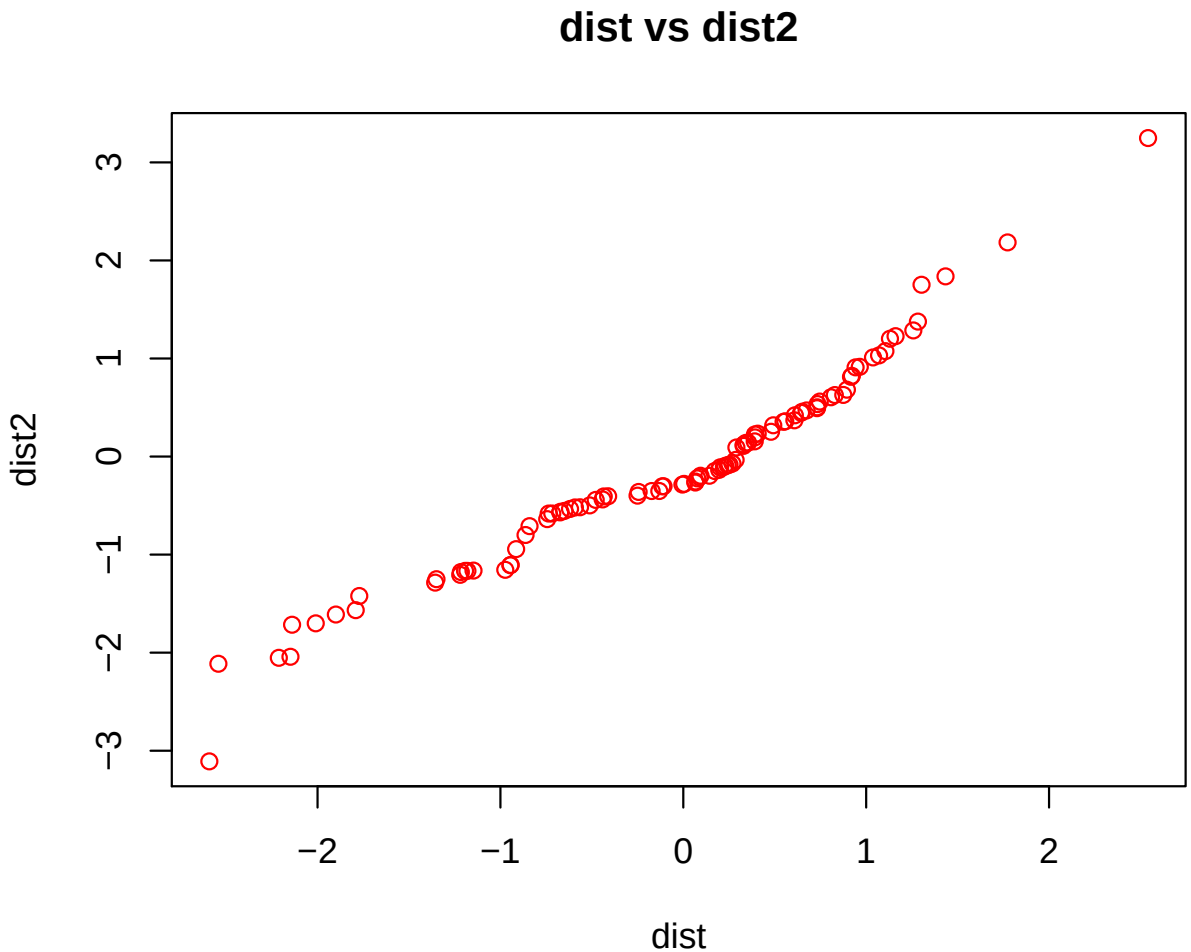
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.



Qqplot.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
> qqplot(dist, distunif, col = "purple",
+     main = "dist vs distunif")
```

Qqplot.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

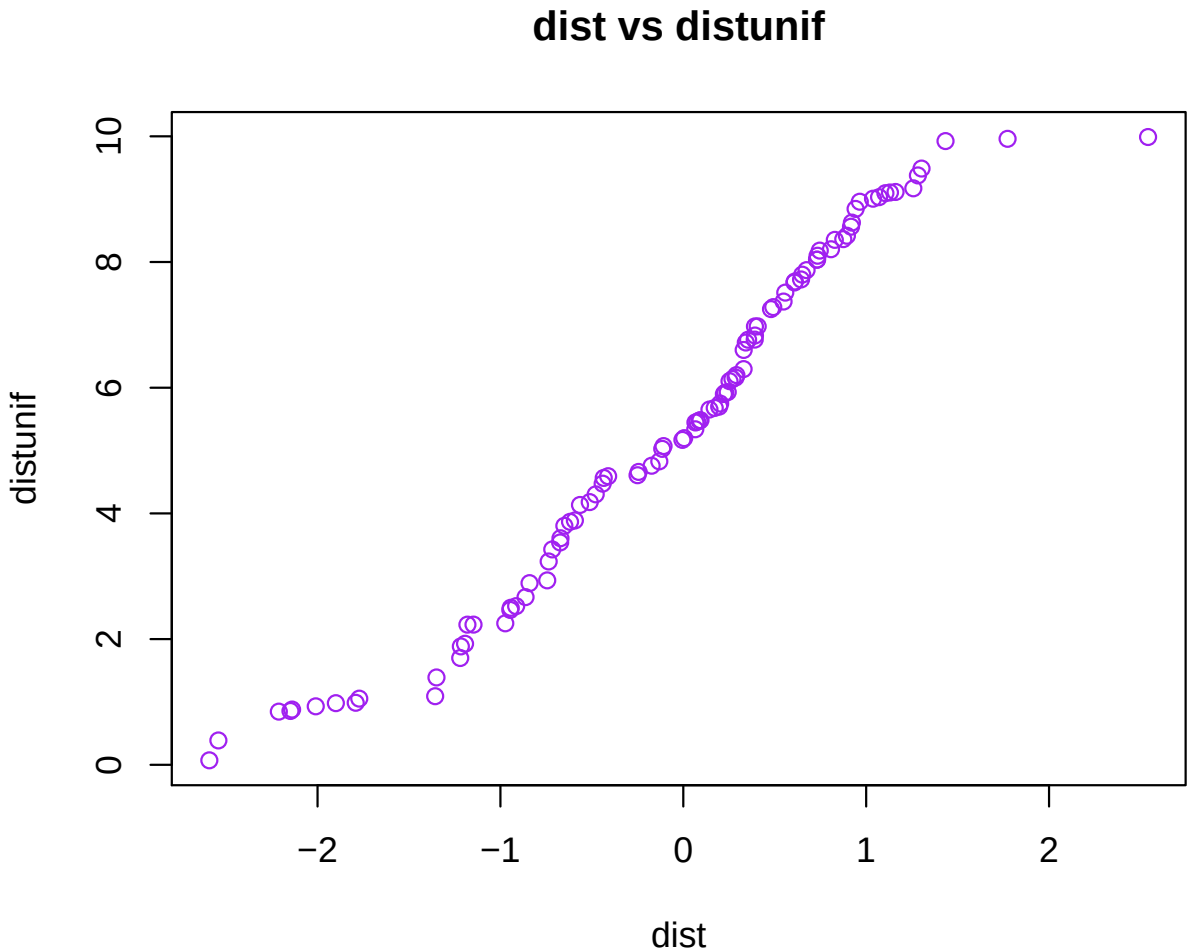
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.

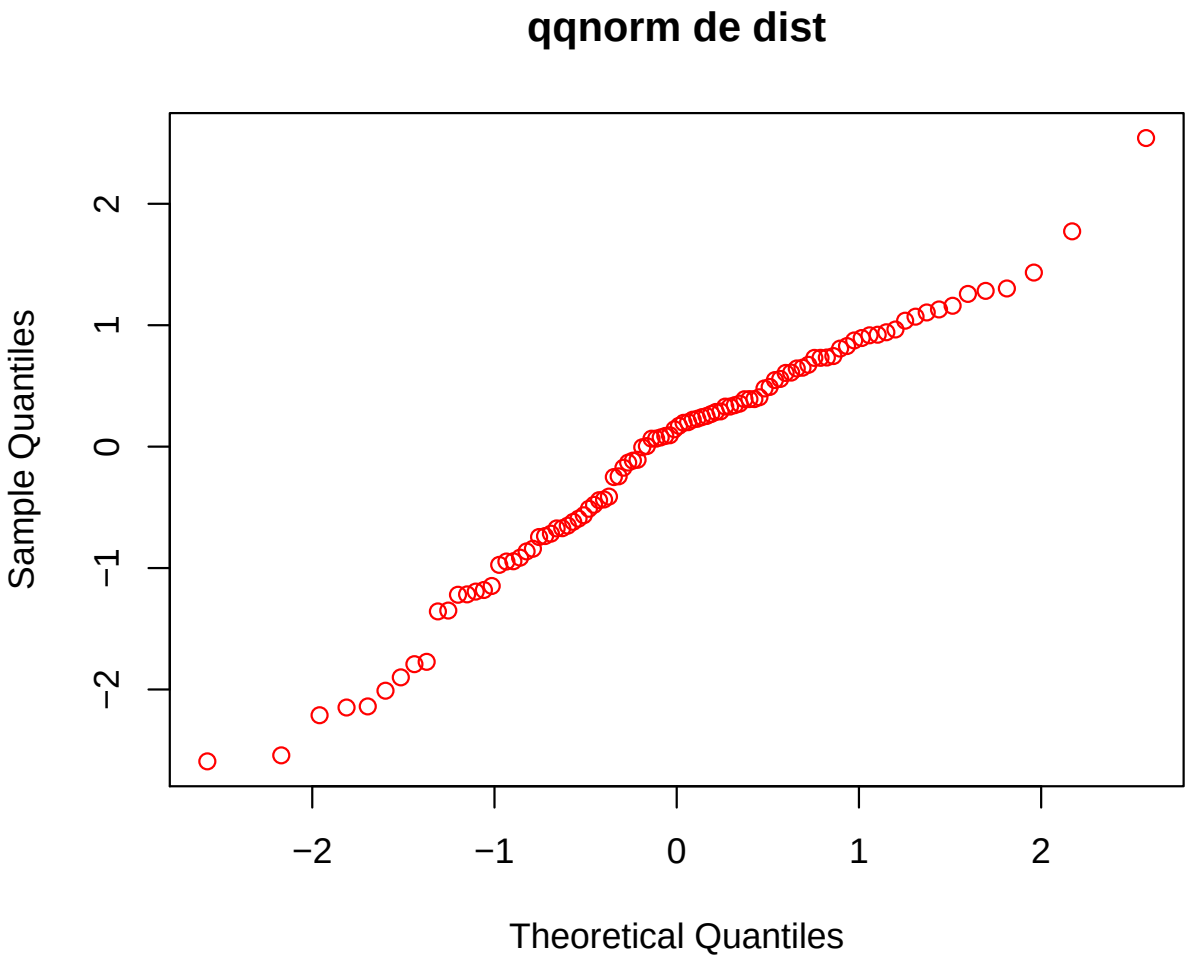


Qqnorm.

Compara una distribución empírica con una normal.

```
> qqnorm(dist, col = "red", main = "qqnorm de dist")
```

Qqnorm.


qqnorm de dist
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

Qqnorm.


```
> qqnorm(distunif, col = "purple",
+     main = "qqnorm de distunif")
```

Qqnorm.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

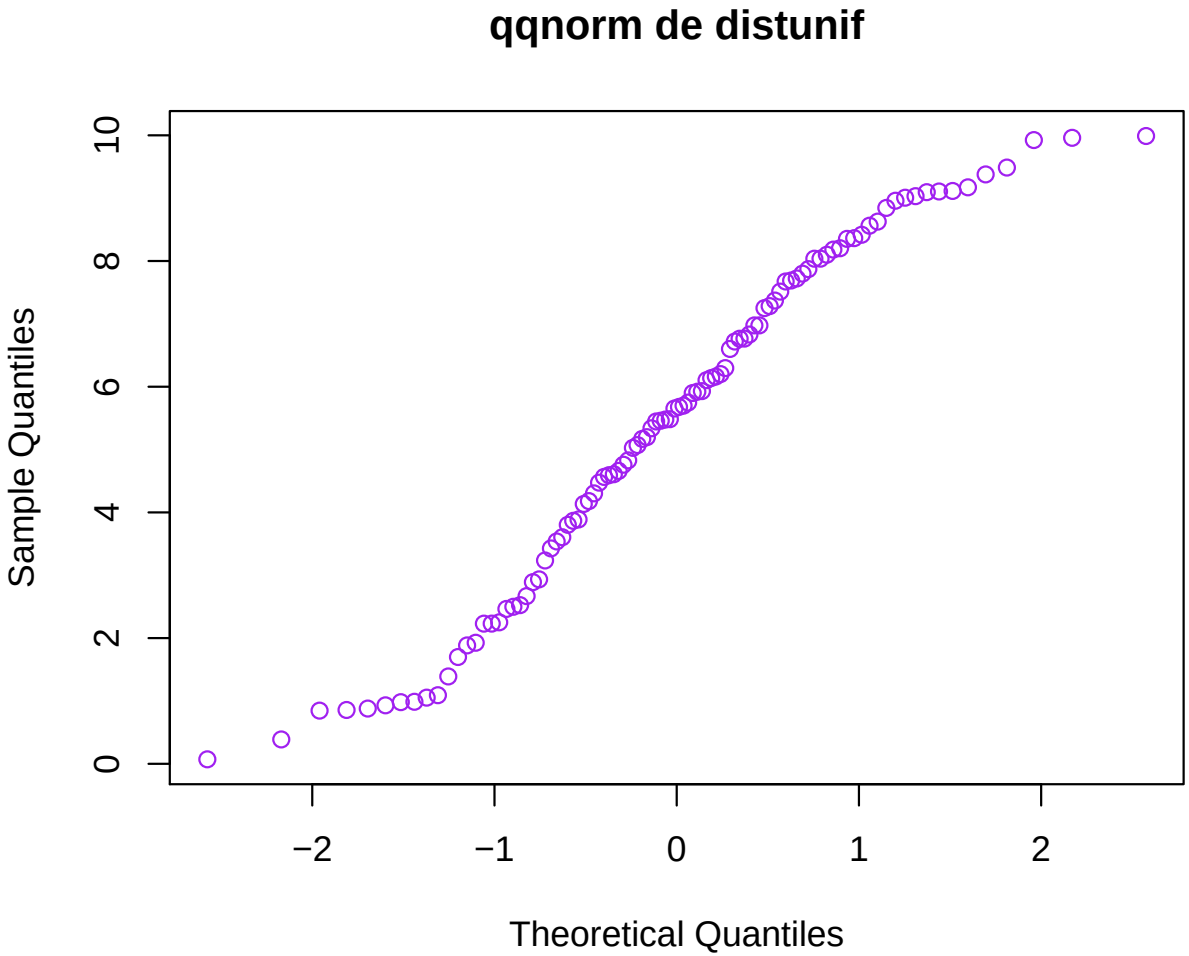
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.



Curve.

Si se quiere evitar tabular y graficar una función que se tiene, se puede emplear esta función.

```
> curve(log(x^2), from = -100, to = 100,
+       col = "red")
```

Curve.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

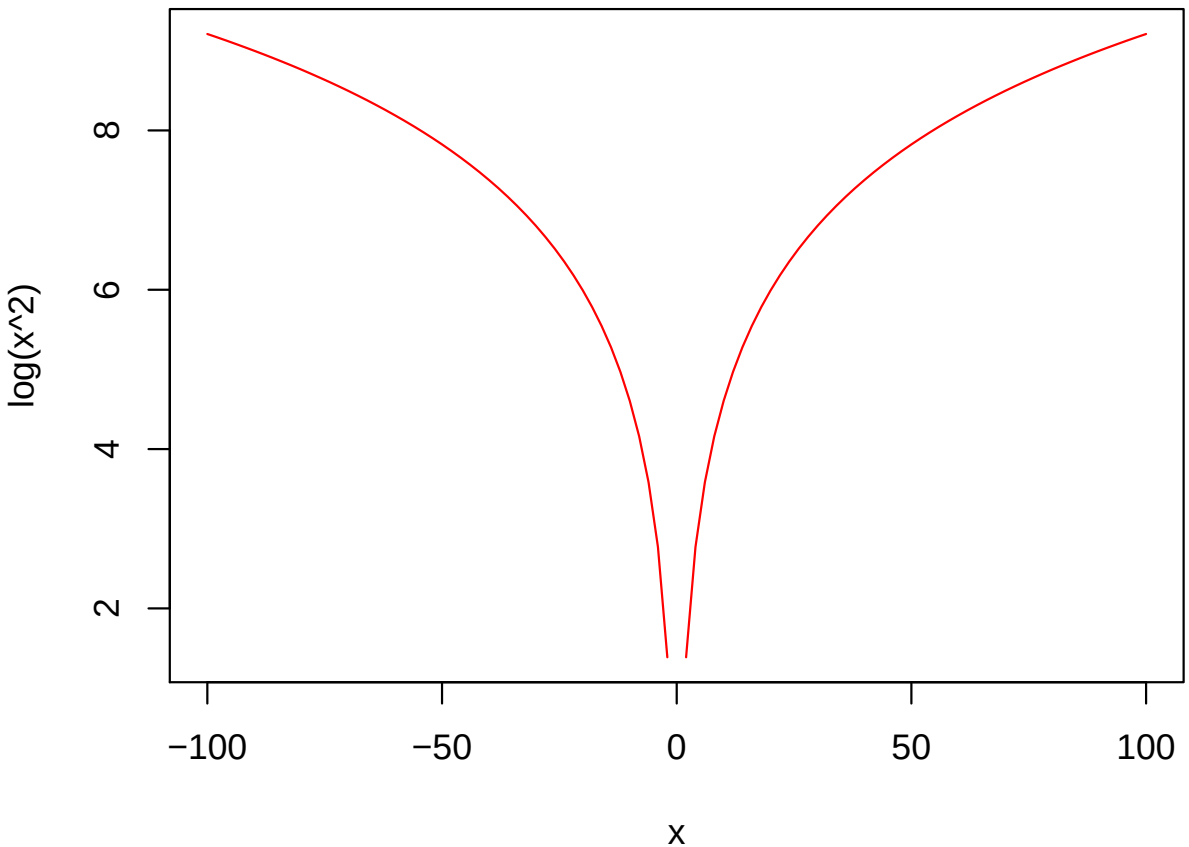
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.



Lattice.

- Lattice es un paquete creado por Deepayan Sarkar, el cual hace uso de layouts para visualizar varias variables agrupadas de un solo golpe.
- Gráficas como hist, xyplot y qqplot están incluidas en lattice.

```
> dataset <- data.frame(x = x, y = y,
+     nat = sample(c("MX", "DK",
+         "UK", "AR"), length(x),
+         replace = T), gen = sample(c("M",
+         "F"), length(x), replace = T))
> head(dataset)
```

Lattice.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.

```
    x           y nat gen
1 0.0  2.36628808  MX   F
2 0.5 -1.61989514  UK   F
3 1.0 -0.72976982  UK   M
4 1.5  0.85819146  MX   M
5 2.0  0.82799512  MX   M
6 2.5  0.02509898  DK   M
```

```
> library(lattice)
> print(xyplot(y ~ x | nat, groups = gen,
+     data = dataset, auto.key = T))
```

Lattice.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

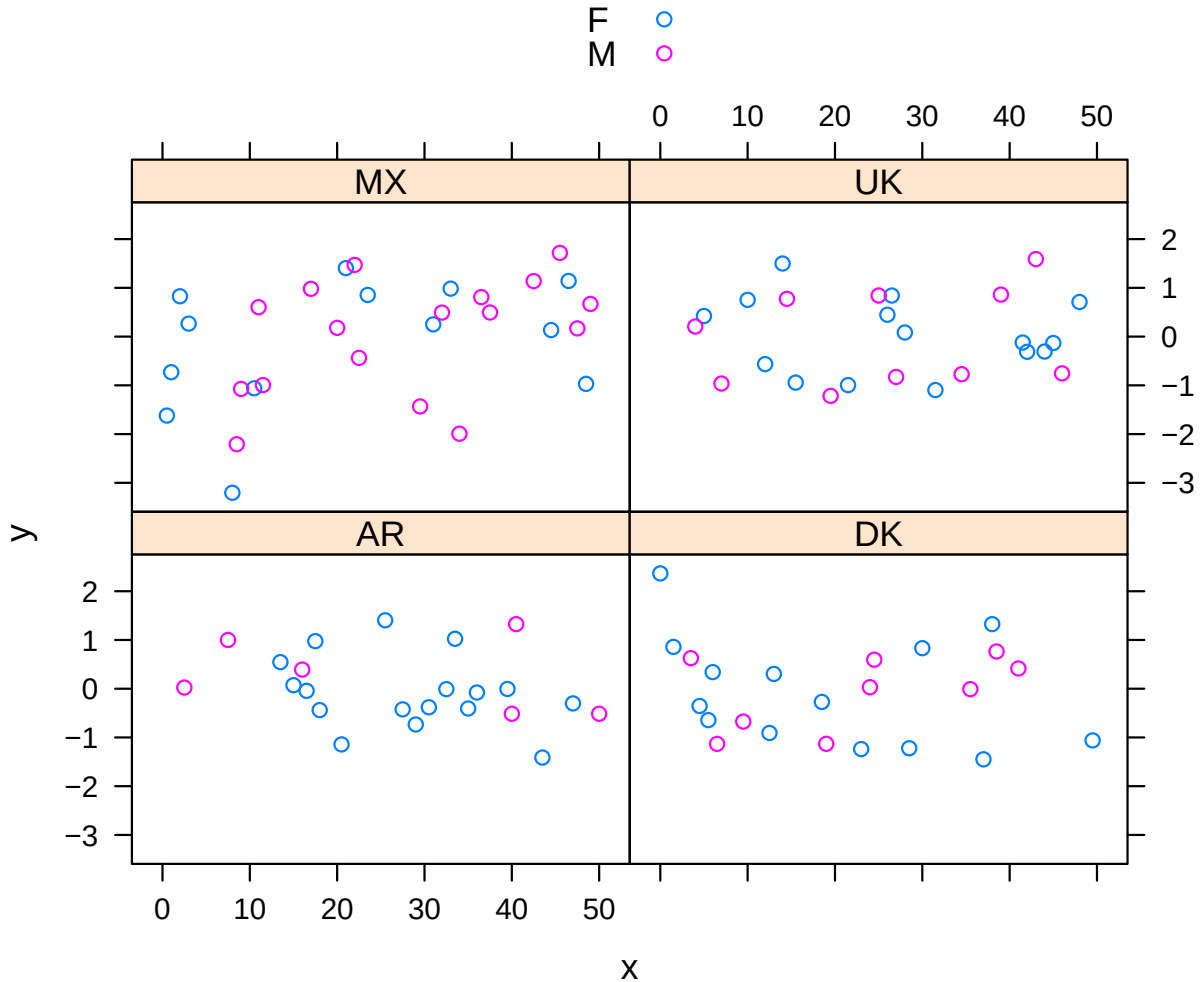
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadsticas.

Visualización.

R en paralelo.



Hierobarplot.

- R cuenta con gráficas tan poderosas que se pueden visualizar 4 variables a la vez.
- Tal es el caso de esta función que forma parte del paquete plotrix.
- Se usará el ejemplo de default de esta gráfica.

```
> library(plotrix)
> test.df <- data.frame(Age = rnorm(100,
+     25, 10), Sex = sample(c("M",
+     "F"), 100, TRUE), Marital = sample(c("D",
+     "M", "S", "W"), 100, TRUE),
+     Employ = sample(c("Full Time",
+         "Part Time", "Unemployed"),
+         100, TRUE))
```

Hierobarplot.


```
> test.col <- list(Overall = "green",
+     Employ = c("purple", "orange",
+         "brown"), Marital = c("#1affd8",
+         "#caeecc", "#f7b3cc", "#94ebff"),
+     Sex = c(2, 4))
> hierobarp(formula = Age ~ Sex +
+     Marital + Employ, data = test.df,
+     ylab = "Mean age (years)",
+     main = "Show only the final breakdown",
+     errbars = F, col = test.col$Sex)
```

Hierobarplot.

Métodos Estadísticos y Analíticos de Datos Genómicos.

Víctor Moreno

¿Qué es R?

Usando R.

Objetos.

Expansiones.

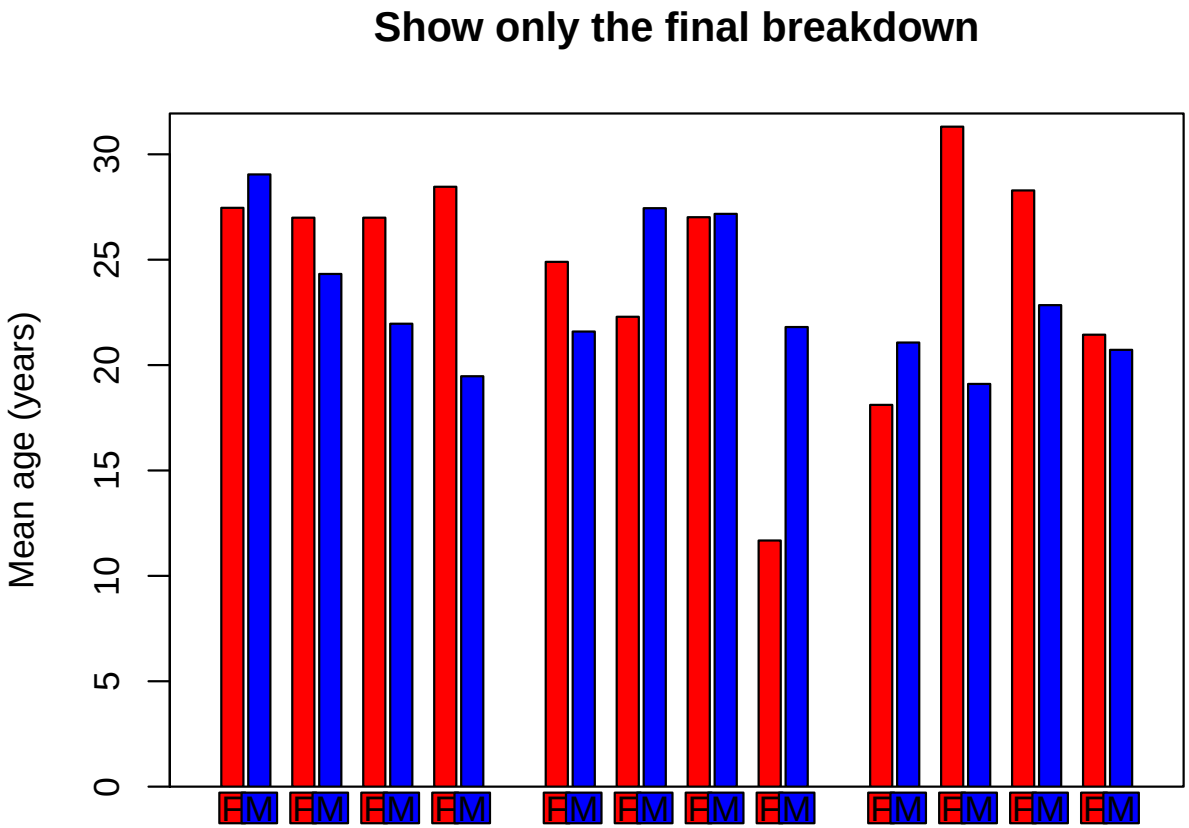
Leyendo Datos.

Datos Aleatorios y Simulaciones.

Correlaciones, Regresiones y Pruebas estadśticas.

Visualización.

R en paralelo.



Rmpi.

- Algunas funciones en R se pueden beneficiar del cómputo en paralelo.
  - Tareas de remuestreo MTP.
  - Clustering.
- El paquete Rmpi tiene funciones para paralelizar procesos en R.
- Un tutorial básico se puede obtener aquí.
- A continuación se muestra un ejemplo de programa en paralelo escrito en R.

Rmpi.


```
> if (!is.loaded("mpi_initialize")) {
+     library("Rmpi")
+ }
> mpi.spawn.Rslaves()
> .Last <- function() {
+     if (is.loaded("mpi_initialize")) {
+         if (mpi.comm.size(1) >
+             0) {
+             print("Please use mpi.close.Rslaves() t
+             mpi.close.Rslaves()
+         }
+         print("Please use mpi.quit() to quit R")
+         .Call("mpi_finalize")
+     }
+ }
```

Rmpi.


```
> mpi.remote.exec(paste("I am", mpi.comm.rank(),
+     "of", mpi.comm.size()))
> mpi.quit()
```

Session Info.


```
> sessionInfo()

R version 2.10.0 (2009-10-26)
x86_64-apple-darwin9.8.0

locale:
[1] C

attached base packages:
[1] stats     graphics  grDevices
[4] utils     datasets  methods
[7] base

other attached packages:
[1] plotrix_2.7-2   affy_1.24.0
```

Session Info.


```
[3] Biobase_2.6.0   lattice_0.17-26

loaded via a namespace (and not attached):
[1] affyio_1.13.5
[2] grid_2.10.0
[3] preprocessCore_1.7.9
```