

# R / Bioconductor: Curso Intensivo

Leonardo Collado Torres

Licenciatura en Ciencias Genómicas, UNAM

[www.lcg.unam.mx/~lcollado/index.php](http://www.lcg.unam.mx/~lcollado/index.php)

Cuernavaca, México

Oct-Nov, 2008

# Datos Bi/Multivariados y Lattice

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- 1 Bivariados
- 2 Gráficas Bivar.
- 3 Correlaciones
- 4 Regresiones lineales
- 5 Multivariados
- 6 Gráficas Multivar.
- 7 Iris
- 8 Un paréntesis
- 9 Lattice
- 10 Histogramas
- 11 Comp. Distr.
- 12 Visualizar Datos
- 13 Scatterplots
- 14 Terminando

# La clase de hoy...

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- En la anterior clase aprendimos a explorar datos univariados. Nos fijábamos en medidas de tendencia central y de dispersión.
- Con datos bi y multivariados vamos a querer encontrar relaciones entre los datos de dos o más variables.
- Vamos a ver varios tipos de gráficas que sirven para explorar estos tipos de datos y el paquete `lattice`.

# Pares de datos categóricos

- Si vienen resumidos, podemos usar tablas de contingencia para comparar los datos.

```
> x <- matrix(c(56, 2, 8, 16), nrow = 2)
> dimnames(x) <- list(padre = c("con",
+      "sin"), hijo = c("con", "sin"))
> library(xtable)
> table(x)
```

	con	sin
con	56.00	8.00
sin	2.00	16.00

Table 1: Uso del cinturón en California

- En este ejemplo claramente vemos la relación entre los datos bivariados.

## Matrices

Algunas veces tendrán los datos en vectores diferentes. Estos los pueden juntar para hacer una matriz con funciones como **rbind** y **cbind**. A diferencia de los vectores, ahora los nombres no se ponen con **names** pero con **rownames** y **colnames**. Alternativamente pueden usar **dimnames** donde incluso pueden usar una lista para pasar los nombres.

- Cuando estás comparando dos variables, a veces también quieres ver la distribución marginal. Esta la puedes recuperar de varias formas siendo las principales `margin.table` y `addmargins`.
- En fin, la forma más sencilla es usar `table(x,y)` aunque también `prop.table` puede ser bastante útil:

	con	sin
con	0.8750	0.1250
sin	0.1111	0.8889

Table 2: Uso del cinturón en California

## Example

Usen el set de datos `grades` del paquete `UsingR`. Queremos ver si la calificación del examen anterior<sup>1</sup> tiene alguna relación con la calificación del examen actual<sup>2</sup>. Encuentren las distribuciones marginales y hagan una tabla de proporciones. Hay que usar proporciones por columnas o renglones? Qué concluyen?

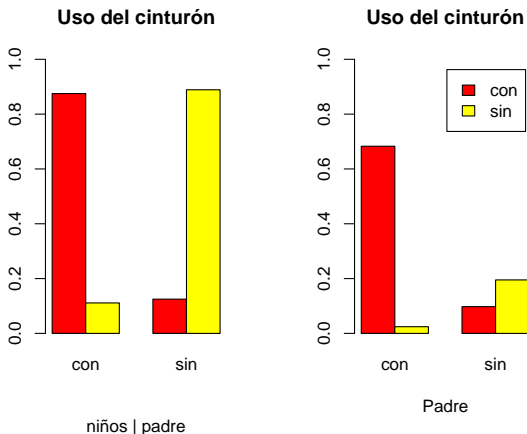
---

<sup>1</sup>Variable `prev`

<sup>2</sup>Variable `grade`

# Con gráficas de barras

- Otra forma de resumir tablas de contingencia es con gráficas de barras. Simplemente usen **barplot**





# Comparando distribuciones gráficamente

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- A veces tenemos datos de dos variables pero no sabemos si son de la misma población. Por ahora podemos compararlas gráficamente usando funciones ya conocidas como:
  - ▶ **boxplot**
  - ▶ **plot** junto con **density** ya que no es fácil juntar 2 histogramas.
  - ▶ **stripchart**
- Aquí les muestro un par de ejemplos:

# Boxplot y density

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

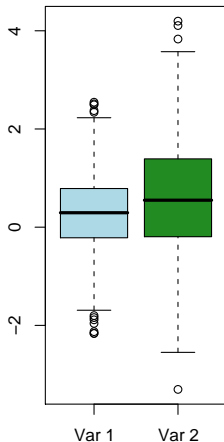
Lattice

Histogramas

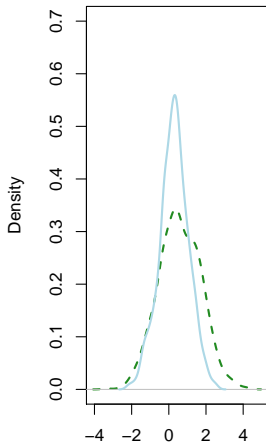
Comp. Distr.

Visualizar  
Datos

**Boxplots**



**Densidades de vars 1 y 2**



N = 1000 Bandwidth = 0.2556

# Stripchart

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

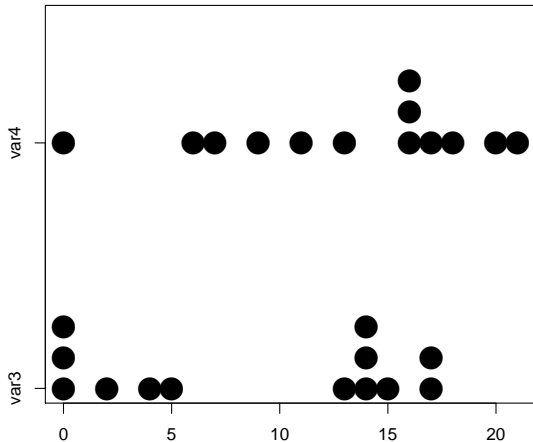
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos



# qqplot

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- Si se acuerdan, `boxplot` usa los cuartiles para hacer la gráfica. Una gráfica más poderosa es **qqplot**.
- `qqplot` te grafica los cuantiles de una distribución contra los de otra. Si obtienes una línea recta, significa que tus distribuciones tienen formas similares.
- Si no es el caso, la línea será curva de tal forma que podrás interpretar los resultados.
- Puedes usar la función **qqnorm** para comparar tu distribución con una normal. **qqline** también te puede ser útil.
- Aquí pueden más ejemplos de `qqplot` y en [wiki](#) pueden leer más al respecto.

# qqplot: simétricas pero una es más larga

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

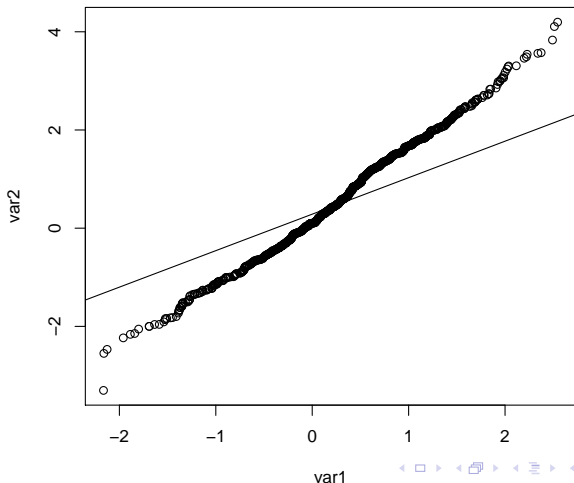
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# Scatterplot con plot

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- Muchas veces tienes datos numéricos pareados y quieres simplemente ver si tienen alguna relación.
- Para esto simplemente podemos hacer un *scatterplot* con la función **plot**.
- A continuación les muestro tres ejemplos. Un scatterplot:
  - ▶ de los valores de 6841 casas en 1970 que luego en el 2000 cuando fueron re-evaluadas.
  - ▶ entre la temperatura máxima en el día y el cambio en el Dow Jones Industrial Average en mayo del 2003.
  - ▶ de la altura vs el peso de unos niños de 0 a 12 años
- Claramente se puede ver en que datos hay una relación y en cuales no.

# Scatterplot: casas

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

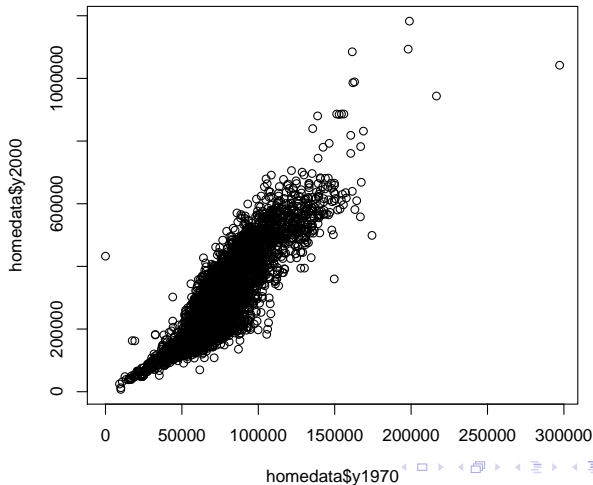
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# Scatterplot: clima y bolsa

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

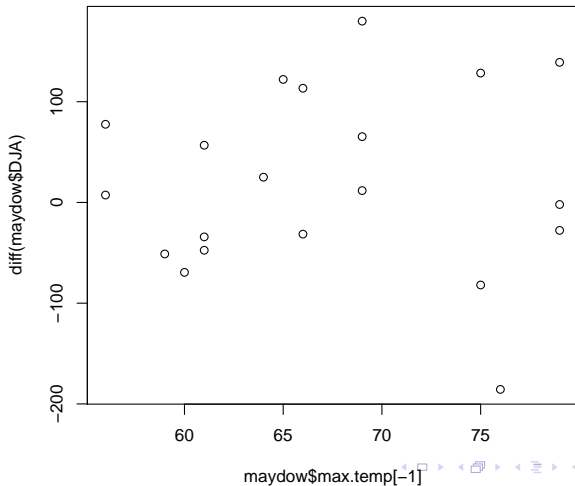
Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

Max. temp vs cambio diario en DJA





# Scatterplot: pesos de niños

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

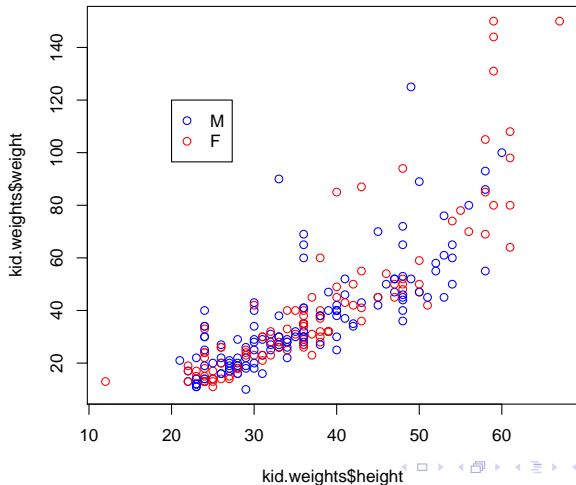
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



- Muchas veces cuando tengas dos variables van a querer saber si están correlacionadas. La correlación, tal como lo dice [wiki](#):
  - ▶ *indica la fuerza y la dirección de una relación lineal entre dos variables aleatorias. Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si tenemos dos variables (A y B) existe correlación si al aumentar los valores de A lo hacen también los de B y viceversa. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad*
- Con R podemos encontrar sin tanto esfuerzo el coeficiente de correlación de Pearson y la correlación de Spearman.

# Correlación de Pearson

- Nos dice que tan correlacionadas están dos variables<sup>3</sup>.
- En fin, les recomiendo que chequen [esta página](#) por si quieren revisar más al respecto.
- En R podemos encontrar la correlación de Pearson muestral con la función **cor**

```
> cor(homedata$y1970, homedata$y2000)
```

```
[1] 0.8962155
```

```
> cor(maydow$max.temp[-1], diff(maydow$DJA))
```

```
[1] 0.01028846
```

```
> cor(kid.weights$height, kid.weights$weight)
```

```
[1] 0.8237564
```

---

<sup>3</sup>Es independiente de la escala de medidas de las variables

# Correlación de Spearman

- Esta nos sirve principalmente si la relación entre las dos variables no es lineal, pero crece.
  - Aquí no usamos los datos como tal, si no que los ordenamos de menor a mayor con la función **rank**.
  - En R podemos calcular esta correlación con `cor(rank(x), rank(y))` o simplemente usando el argumento `method` de la función `cor`.
  - Para más información, lean [esto](#).
- ```
> cor(rank(homedata$y1970), rank(homedata$y2000))  
[1] 0.8878185  
  
> cor(maydow$max.temp[-1], diff(maydow$DJA),  
+      method = "spearman")
```

# Correlación de Spearman

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

```
[1] 0.1315711
```

```
> cor(kid.weights$height, kid.weights$weight,  
+      method = "s")
```

```
[1] 0.8822136
```

# Recordando

- Vamos a ver rapidamente las regresiones lineales. La siguiente ecuación debe serles familiar:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (1)$$

- Aquí,  $\epsilon_i$  es error,  $\beta_0$  y  $\beta_1$  son los coeficientes de la regresión,  $x$  es la variable independiente &  $y$  la dependiente.
- En estadística, generalmente conocemos la  $x$  pero necesitamos estimar el resto. Por ejemplo, por el método de mínimos cuadrados.

# Funciones

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- Uno de sus principales es hacer predicciones de la variable independiente.
- En R podemos hacer regresiones simples con **lm** (**formula.modelo**) donde usamos la notación y TILDE x para la fórmula. Por ejemplo:

```
> res <- lm(homedata$y2000 ~ homedata$y1970)
> res
```

Call:

```
lm(formula = homedata$y2000 ~ homedata$y1970)
```

Coefficients:

|             |                 |
|-------------|-----------------|
| (Intercept) | homedata\$y1970 |
| -1.040e+05  | 5.258e+00       |

- Sin embargo, es mejor almacenar el resultado en una variable. Así luego podemos graficarla u obtener más información de esta regresión con funciones como:
  - ▶ `coef` nos da los coeficientes
  - ▶ `residuals` nos regresa los residuos
  - ▶ `predict` nos predice un valor dado una  $x$
- Les aviso que no usen regresiones lineales para predecir valores en un rango no válido para la variable independiente.
- A veces pueden transformar sus datos para que el modelo lineal sea más apropiado.
- Chequen a las funciones `lqs` y `rlm` ya que estas son más resistentes a *outliers*.



# Graficando

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

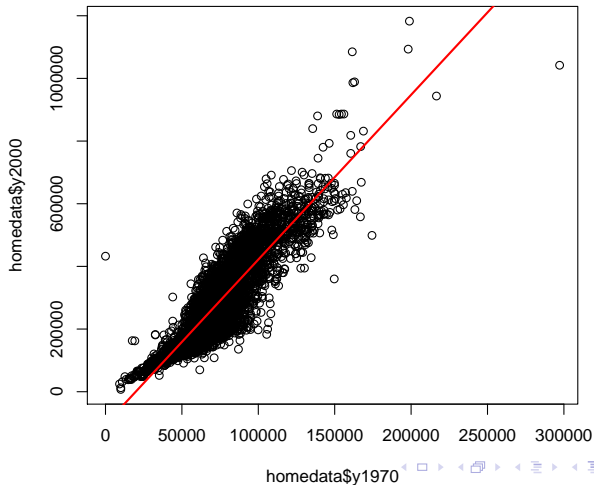
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# Graficando: con una transformación

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

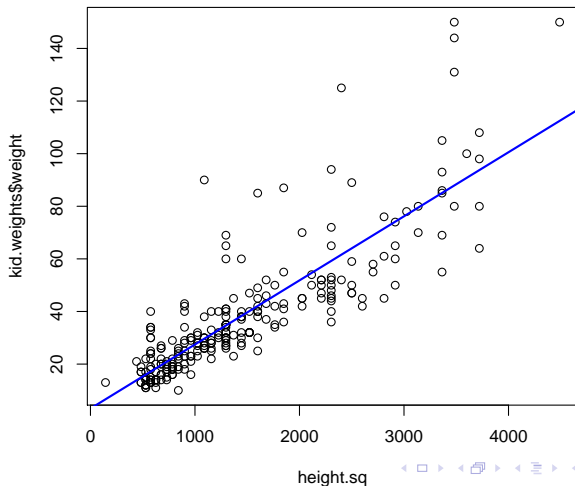
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# Multivariados

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- Visualizar y comparar datos multivariados no es sencillo. Lo que se puede hacer es mantener constantes a algunas de las variables mientras observamos al resto.
- En general en R, se pueden usar las mismas herramientas para datos uni y bivariados con los multivariados. Aunque hay algunas diferencias.
- Vamos a entrar un poco en el paquete de **lattice** ya que este nos permite explorar datos multivariados más a fondo.

# Datos categóricos

- Podemos simplemente ver los datos como una matriz. Sin embargo, se vuelve difícil encontrar tendencias en los datos. Por ejemplo, chequen el set de datos `student.expenses`
- La otra opción<sup>4</sup> es hacer tablas de contingencia con 3 variables con `table`. Por cada valor de la variable 1 vamos a ver una tabla de contingencia para las variables 2 y 3.
- De igual forma, podemos obtener tablas de 2 variables cuando  $n$  están fijas.  

```
> attach(student.expenses)  
> table(cell.phone, car, cable.modem)
```

# Datos categóricos

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

```
, , cable.modem = N
```

```
      car  
cell.phone N Y  
      N 1 2  
      Y 2 3
```

```
, , cable.modem = Y
```

```
      car  
cell.phone N Y  
      N 0 0  
      Y 1 1
```

---

<sup>4</sup>Más recomendable

- Para resumir las tablas anteriores podemos usar **ftable**. Aquí se comprimen las tablas y nos ahorran espacio :)

```
> ftable(table(cell.phone, car, cable.modem),
+        col.vars = c("cable.modem",
+                      "car"))
```

|            | cable.modem |  |   |   |
|------------|-------------|--|---|---|
|            | N           |  | Y |   |
|            | car         |  | N | Y |
| cell.phone |             |  |   |   |
| N          |             |  | 1 | 2 |
| Y          |             |  | 2 | 3 |

```
> detach(student.expenses)
```

# Boxplot

- La función **boxplot** te puede hacer este tipo de gráficas para varias variables de un solo golpe.
- Sin embargo, tengan cuidado ya que es fácil cometer errores en estos casos:
  - ▶ puedes repetir datos sin darte cuenta.
  - ▶ hay que añadir los nombres manualmente.
  - ▶ puedes olvidar de alguna variable que agrupe a tus datos en diferentes categorías.
- Vamos a usar el set de datos `ewr` que tiene info sobre los tiempos de espera para llegar a la terminal o para despegar en un aeropuerto.

```
> attach(ewr)
> boxplot(AA, CO, DL, HP, NW, TW,
+         US, US)
```

# Encuentren los errores

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

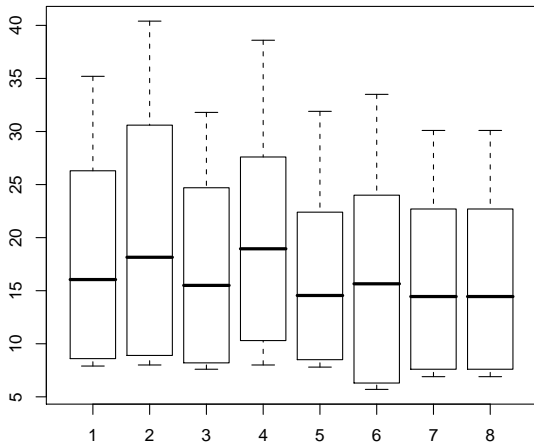
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos





# Versión correcta

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

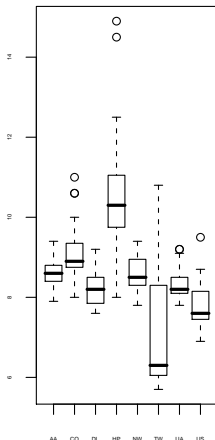
Lattice

Histogramas

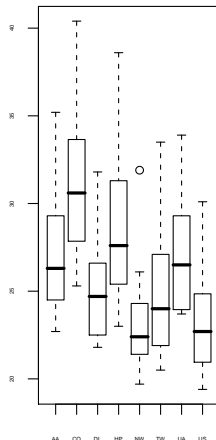
Comp. Distr.

Visualizar  
Datos

**Llegar a terminal**



**Para despegar**



# Scatterplots

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

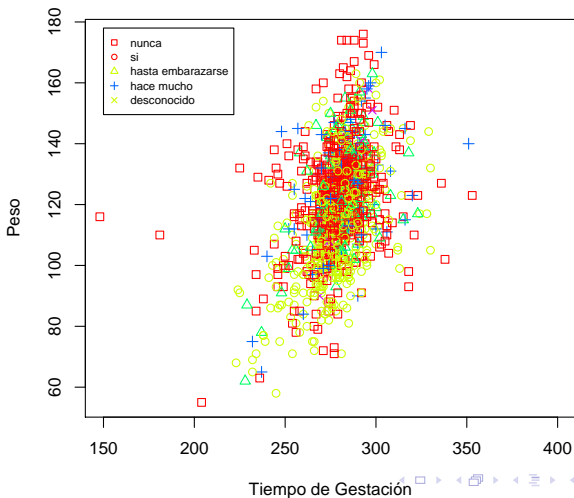
- Para investigar relaciones entre variables, ya habíamos visto a los *scatterplots*. Puedes graficar todos los datos en una sola imagen diferenciando a los puntos con colores y/o formas.
- Alternativamente, puedes hacer gráficas para cada par de variables como las que ya habíamos visto antes. Les muestro un ejemplo y luego vamos a jugar con los datos de *Iris*.

## Funciones para gráficas

Rápidamente les menciono varias funciones de bajo y alto nivel. Las de alto nivel hacen una gráfica nueva mientras que las de bajo nivel escriben encima de alguna gráfica existente. **plot**, **points**, **lines**, **abline**, **curve**, **rug**, **arrows**, **text**, **title** y **legend**.

# Un scatterplot sobre bebés

**Peso vs t gestación vs mamás fumadoras**



R /  
Bioconductor:  
Curso  
Intensivo

Bivariados  
Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

# Conociendo los datos

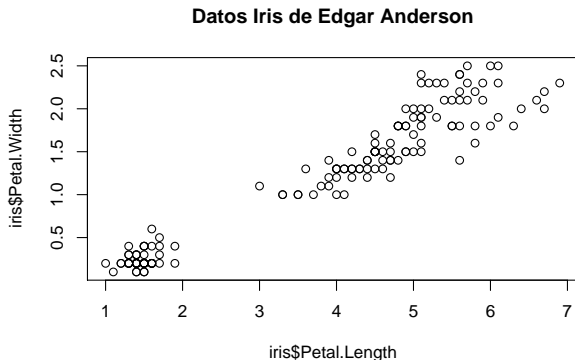
- Para empezar, hay que conocer nuestros datos ^^ . Utilizen los siguientes comandos para darse una idea:

```
> class(iris)
> head(iris)
> tail(iris)
> colnames(iris)
> iris$Petal.Length
```

# Un scatterplot sencillo

- Hagan un scatterplot sencillo con dos variables:  

```
> plot(iris$Petal.Length, iris$Petal.Width,  
+      main = "Datos Iris de Edgar Anderson")
```



# Un truco :D

- Ahora chequen el siguiente truco :)

```
> unclass(iris$Species)[c(1:3, 75:77,  
+ 112:114)]
```

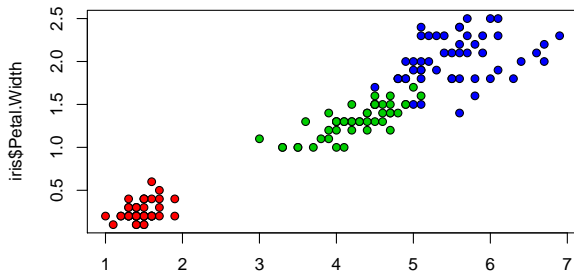
```
[1] 1 1 1 2 2 2 3 3 3
```

- **unclass** nos sirve para romper datos con tipo factor en números del 1 hasta  $n$  categorías.
- Con este truco ahora podemos hacer una gráfica con símbolos y/o colores diferentes por especie.

# Con colores

```
> plot(iris$Petal.Length, iris$Petal.Width,  
+      pch = 21, bg = c("red", "green3",  
+                        "blue")[unclass(iris$Species)],  
+      main = "Datos Iris de Edgar Anderson")
```

Datos Iris de Edgar Anderson



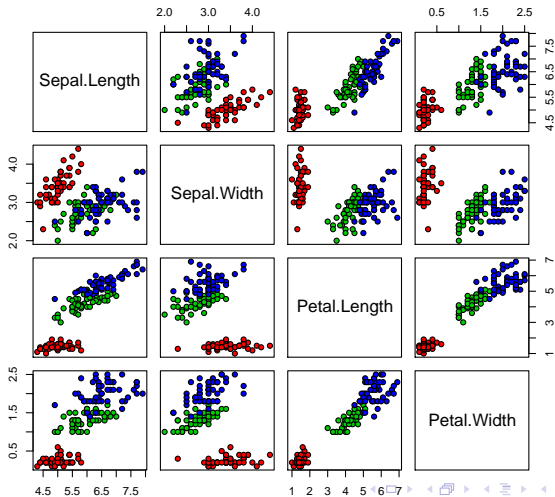
- Ahora queremos ver todos los scatterplots. Usamos la función **pairs** así:

```
> pairs(iris[1:4], pch = 21, bg = c("red",  
+   "green3", "blue")[unclass(iris$Species)],  
+   main = "Datos Iris de Edgar Anderson")
```
- Esta función nos va a generar los scatterplots por pares de variables. Si fuera el caso de datos tridimensionales, podríamos usar esta función para visualizarlos.
- En nuestro ejemplo, casi cualquier par de variables nos podría servir para separar las especies.



# Pares

Datos Iris de Edgar Anderson



R /  
Bioconductor:  
Curso  
Intensivo

Bivariados  
Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

# Con más info

- A veces van a querer tener otro tipo de información en vez de todos los pares. Si se fijaron, cada gráfica sale dos veces.
- Vamos a hacer una función para obtener los coeficientes de correlación.

```
> panel.pearson <- function(x, y,  
+   ...) {  
+   horizontal <- (par("usr")[1] +  
+     par("usr")[2])/2  
+   vertical <- (par("usr")[3] +  
+     par("usr")[4])/2  
+   text(horizontal, vertical,  
+     format(abs(cor(x, y)),
```

# Con más info

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

```
+             digits = 2))  
+ }  
  
> pairs(iris[1:4], pch = 21, bg = c("red",  
+   "green3", "blue")[unclass(iris$Species)],  
+   upper.panel = panel.pearson,  
+   main = "Datos Iris de Edgar Anderson")
```

# Con correlaciones de Pearson

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados  
Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

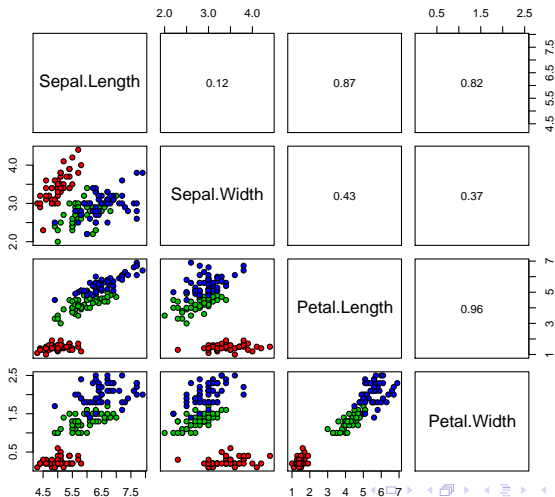
Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

Datos Iris de Edgar Anderson



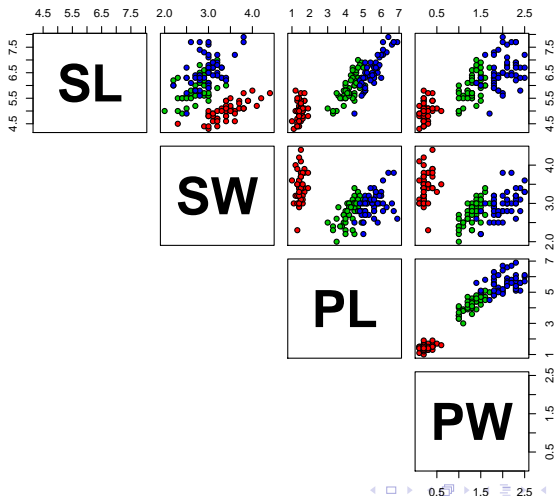
# Una última

- Ahora sin los paneles inferiores :)

```
> pairs(iris[1:4], pch = 21, bg = c("red",  
+   "green3", "blue")[unclass(iris$Species)],  
+   lower.panel = NULL, labels = c("SL",  
+   "SW", "PL", "PW"), font.labels = 2,  
+   cex.labels = 4.5, main = "Datos Iris de And")
```

# Sin inferiores

Datos Iris de Anderson — 3 especies



# Más funciones: usen ?

## Data frames

Rápidamente les menciono unas funciones que luego les pueden ayudar a manipular data frames:

- **subset** que es equivalente a usar [...]
- **split** sirve para separar una variable por los niveles de un factor,
- **stack** es inverso a split,
- **xtabs** nos genera tablas de contingencia para un data frame

# Algo del manual

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

## Fórmulas

Sobre la notación de fórmula, les recomiendo que chequen la sección *11.1 Defining statistical models; formulae* del manual de introducción a R disponible en esta [esta página](#). Tengan cuidado al usar esta notación porque símbolos como  $+$ ,  $-$ , etc hacen otras cosas.



# Intro al paquete

- **lattice** es un paquete muy importante para hacer gráficas con datos multivariados que no podrías hacer con R básico.
- Utilizen los siguientes comandos para ver una pequeña descripción y varias funciones que ofrece el paquete:  

```
> library(lattice)
```

```
> `?`(Lattice)
```
- **lattice** usa la sintaxis de fórmula para sus funciones generalmente así: `var.dep TILDE var.indep | condición` aunque a veces no hay una `var.dep`
- Con eso ya terminé y son expertos en **lattice**! :P

# Set de datos

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

- Gran parte de lo que veremos lo tomé del material de apoyo sobre lattice que les puse en la página.
- Primero vamos a usar los datos de Chem97 del paquete `mlmRev`. Usaremos las variables:
  - ▶ `score`: califs de un examen con valores (0, 2, 4, 6, 8)
  - ▶ `gcsescore`: promedio de los exámenes GCSE. Es continua y puede servir para predecir valores de `score`
  - ▶ `gender`: bueno... sus valores son M ó F
- Cargen los datos:

```
> data(Chem97, package = "mlmRev")  
> head(Chem97)
```

# Histogramas poderosos

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

**Histogramas**

Comp. Distr.

Visualizar  
Datos

- Con `lattice` podemos hacer los histogramas comunes.
- Sin embargo, lo poderoso es hacer un histograma con la función `histogram` de una variable para cada intervalo dado de otra variable.
- En la primera gráfica van a ver una distribución simétrica unimodal. Con la segunda podemos aprender más.  

```
> histogram(~gcsescore, data = Chem97)  
> histogram(~gcsescore | factor(score),  
+           data = Chem97)
```

# Histograma simple

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

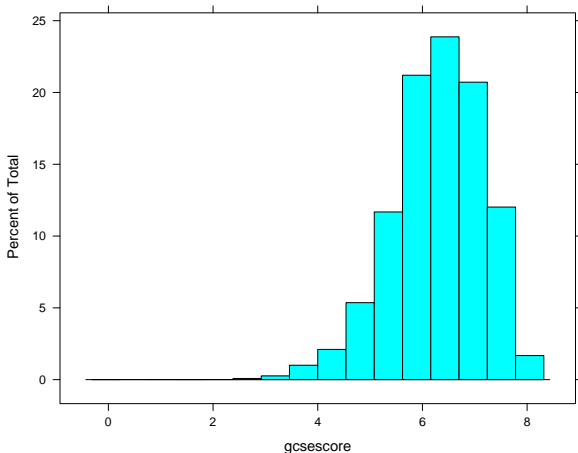
Un paréntesis

Lattice

**Histogramas**

Comp. Distr.

Visualizar  
Datos



# Histograma poderoso

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados  
Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

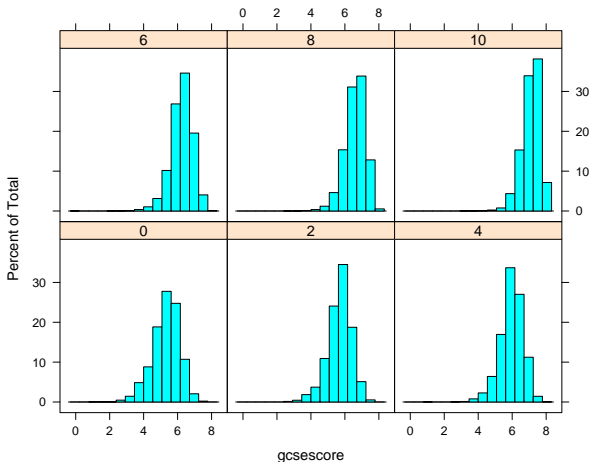
Un paréntesis

Lattice

**Histogramas**

Comp. Distr.

Visualizar  
Datos



- Ahora queremos separar a los hombres y mujeres. Como no es fácil sobreponer histogramas, podemos usar densidades.

- Esto lo podemos hacer con la función **densityplot**:

```
> densityplot(~gcsescore | factor(score),  
+           Chem97, groups = gender, plot.points = FALSE,  
+           auto.key = TRUE)
```

# Densidades

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

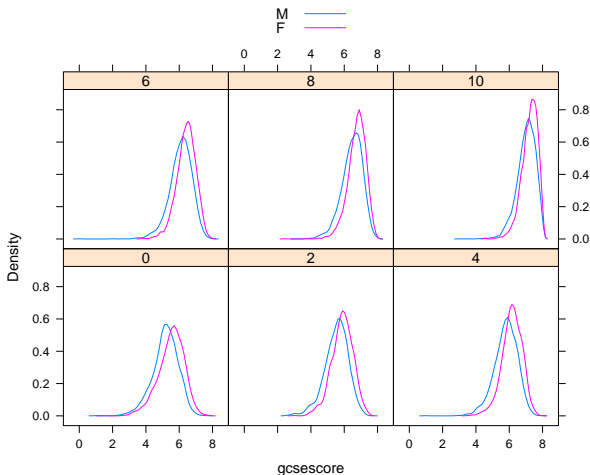
Lattice

Histogramas

Comp. Distr.

Visualizar

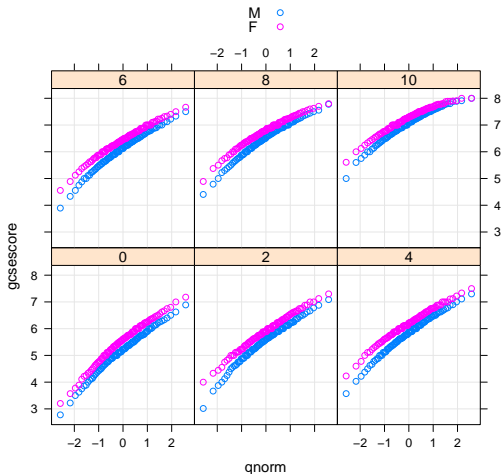
Datos



- Ahora podemos hacer lo equivalente a muchas `qqnorm` con la función **qqmath**.

```
> qqmath(~gcsescore | factor(score),  
+       Chem97, groups = gender, f.value = ppoints(  
+       auto.key = TRUE, type = c("p",  
+       "g")), aspect = "xy")
```
- De aquí podemos ver que las distribuciones están sesgadas a la izquierda. Además, el cambio en la pendiente sugiere que la varianza cambia.
- Si quieren hacer las `qqplots` pareadas usen simplemente la función **qq**.





- Ahora hagamos boxplots condicionales con la función **bwplot**  

```
> bwplot(factor(score) ~ gcsescore /  
+       gender, Chem97)
```
- En esta gráfica podemos notar que la varianza decrece mientras aumenta el score.
- Además, podemos notar los sesgos hacia la izquierda.

# bwplot

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

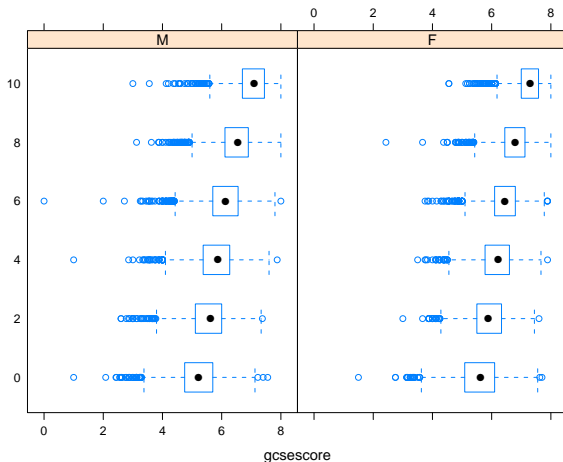
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos



# bwplot cambiada

- Con el argumento `layout` podemos cambiar cuantos paneles tenemos.  

```
> bwplot(gcsescore ~ gender | factor(score),  
+       Chem97, layout = c(6, 1))
```
- Junto con el cambio de orden en las variables, podemos comparar directamente los pares de M y F.

# bwplot 2

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

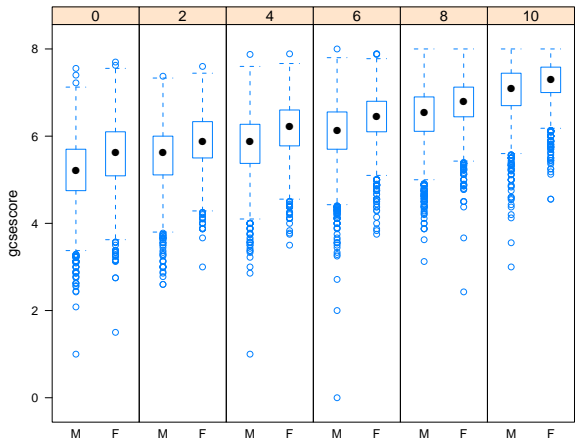
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# stripplot

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

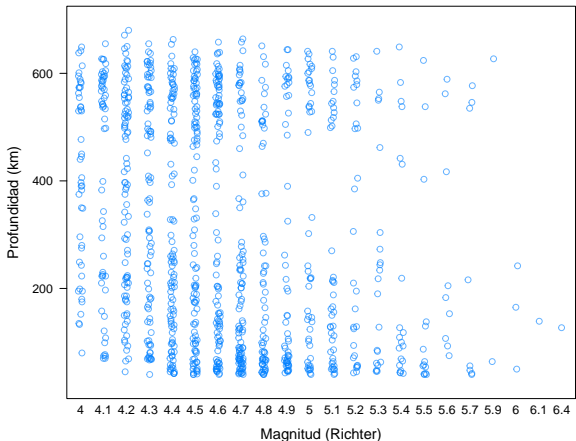
Visualizar  
Datos

- A veces, simplemente ver todos los datos nos puede dar información. Claro, preferencialmente cuando el set de datos no es monstruoso.
- Usemos la función **stripplot** para visuar unos datos sobre temblores.  

```
> stripplot(depth ~ factor(mag),  
+           data = quakes, jitter.data = TRUE,  
+           alpha = 0.6, main = "Profundidad de los epi  
+           xlab = "Magnitud (Richter)",  
+           ylab = "Profundidad (km)")
```
- Si se fijan, los puntos en la gráfica son parcialmente transparentes por si se sobrelapan mucho.

# stripplot

Profundidad de los epicentros por magnitud



# barchart

- Muchas veces van a tener datos en tablas. Estos los podemos comparar con las funciones **barchart** y **dotplot** de forma similar a lo que veníamos haciendo.

- Al usar **barchart** hay que tener cuidado de que el área de las barras sea proporcional al valor que representa.

```
> VADeathsDF <- as.data.frame.table(VADeaths,  
+   responseName = "Rate")  
  
> barchart(Var1 ~ Rate | Var2, VADeathsDF,  
+   layout = c(4, 1))  
  
> barchart(Var1 ~ Rate | Var2, VADeathsDF,  
+   layout = c(4, 1), origin = 0)
```

- VADeaths tiene datos sobre muertes en Virginia en 1941.



# barchart incorrecto

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

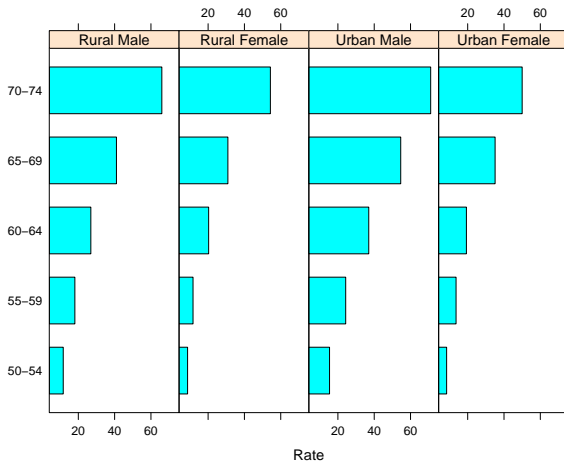
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# barchart correcto

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

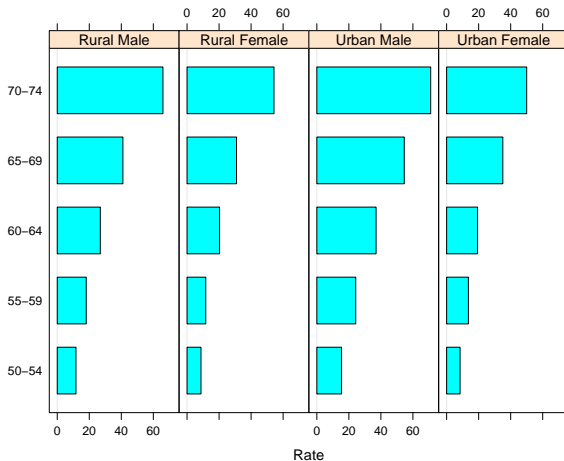
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



- Las gráficas de barras luego te pueden distraer de la comparación que querías hacer. Osea, ver los valores máximos. Para eso usamos **dotplot**

```
> dotplot(Var1 ~ Rate | Var2, VADeathsDF,  
+         layout = c(4, 1))  
> dotplot(Var1 ~ Rate, data = VADeathsDF,  
+         groups = Var2, type = "o",  
+         auto.key = list(space = "right",  
+         points = TRUE, lines = TRUE))
```

- La primera es como la homóloga a las gráficas de barras que hicimos. La segunda nos muestra con mayor claridad relaciones en los datos: fíjense en las curvas de hombres vs las de mujeres.

# dotplot

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

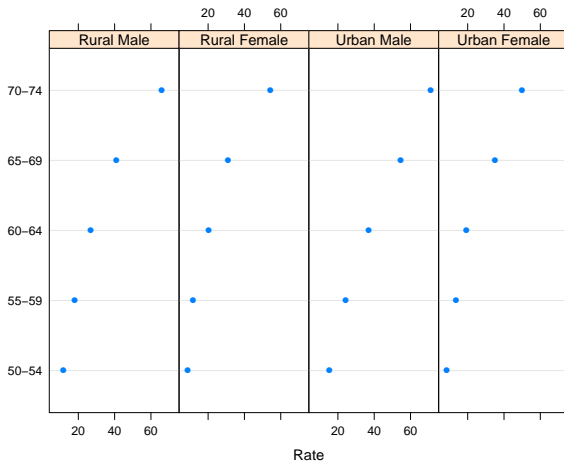
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# dotplot clara

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

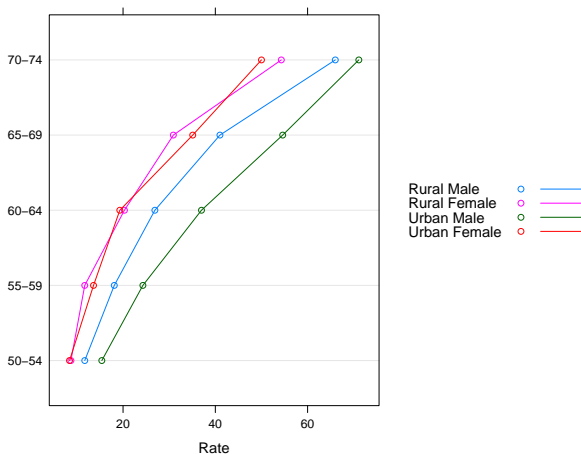
Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos



# Usando dfs

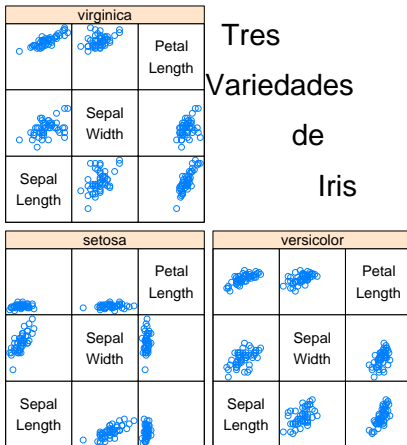
- Las funciones de alto nivel de `lattice` son genéricas con métodos que hacen el trabajo detrás.
- Hemos usado las funciones con la sintaxis de fórmulas, pero también se pueden usar `data frames` en algunas. Ese es el caso de `barchart` y `dotplot`
- Si quieren checar los métodos para alguna función utilicen `methods`. Por ejemplo:  

```
> methods(generic.function = "dotplot")
```

# xyplot, splom, cloud

- Para hacer scatterplots, podemos usar la función **xyplot**.
- Si queremos hacer varias de estas, podemos usar funciones como **splom** y **cloud**.
- Les muestro unos ejemplos con los datos de Iris :).

## Tres Variedades de Iris

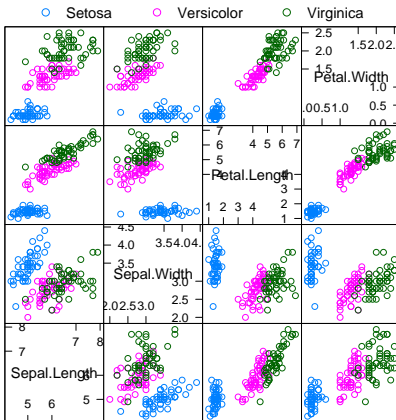


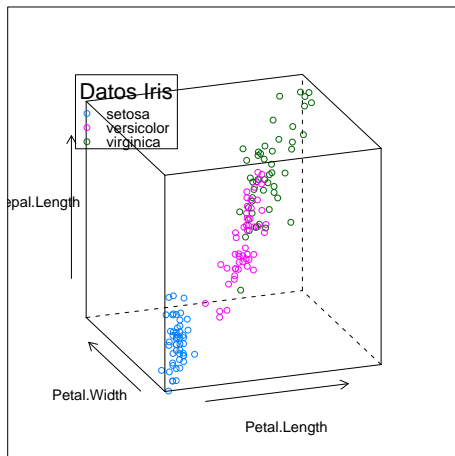
Scatter Plot Matrix



# sploM 2

## Tres Variedades de Iris





- Vamos a ver la relación entre la máxima aceleración horizontal y la distancia de una estación de medición del epicentro de varios terremotos.
- La primera gráfica muestra unos datos sesgados hacia la derecha.
- En la segunda, mejoramos la gráfica al usar ejes logarítmicos. Además, le añadimos una *grid* para poder leer mejor los datos.

# xyplot

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice

Histogramas

Comp. Distr.

Visualizar  
Datos

```
> data(Earthquake, package = "nlme")
> xyplot(accel ~ distance, data = Earthquake)
> xyplot(accel ~ distance, data = Earthquake,
+       scales = list(log = TRUE),
+       type = c("p", "g", "smooth"),
+       xlab = "Distancia al Epicentro (km)",
+       ylab = "Máxima Aceleración Horizontal")
```

# Scatterplot sencillo

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

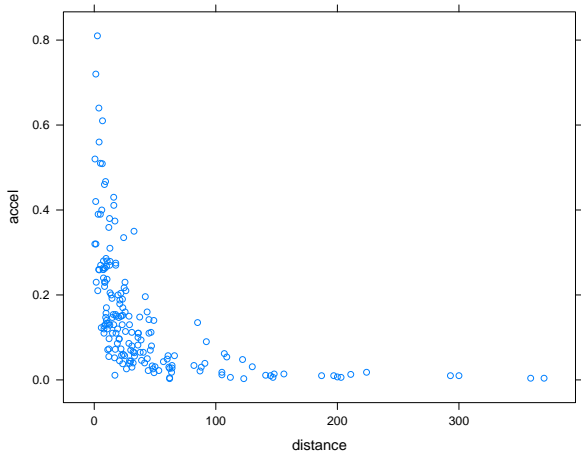
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos



# Con ejes log

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

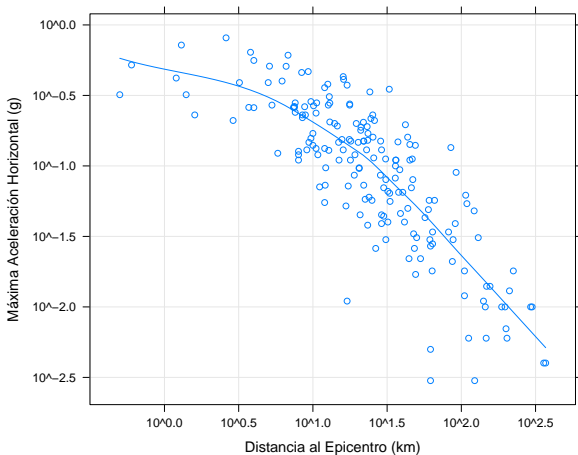
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos



- Con `xyplot` y junto con los llamados *shingles*<sup>5</sup> podemos visualizar unos datos 3D en 2D.

```
> Depth <- equal.count(quakes$depth,  
+      number = 8, overlap = 0.1)  
> xyplot(lat ~ long | Depth, data = quakes)
```

---

<sup>5</sup>Son grupos similares a los que puedes hacer con `cut` junto con `factor` pero se sobrelapan en cierto porcentaje

# shingles

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

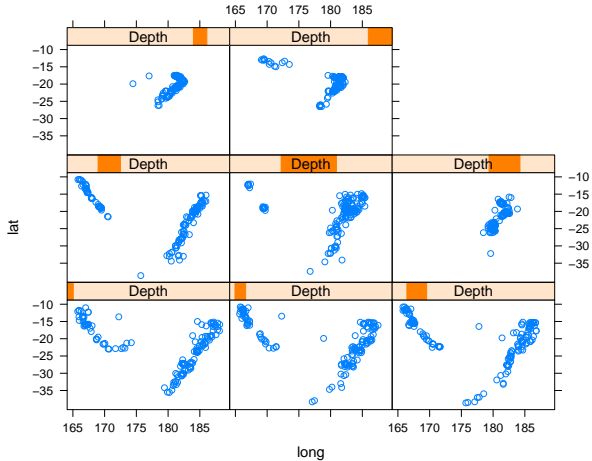
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos





- Claro, que también lo podemos hacer en 3D con la función **cloud**. Por ejemplo:

```
> cloud(depth ~ lat * long, data = quakes,  
+       zlim = rev(range(quakes$depth)),  
+       screen = list(z = 105, x = -70),  
+       panel.aspect = 0.75, xlab = "Longitud",  
+       ylab = "Latitud", zlab = "Profundidad")
```

- El problema es que no podemos interactuar con nuestra gráfica 3D; por ejemplo, para girarla. Hay que hacerlo manualmente con el argumento `screen`.

# 3D

## R / Bioconductor: Curso Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

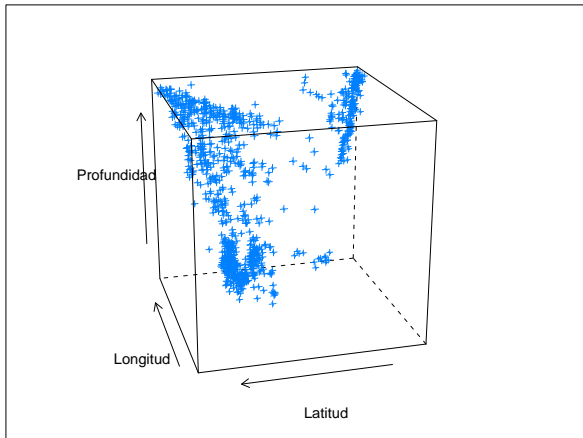
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos



# 3D girado

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

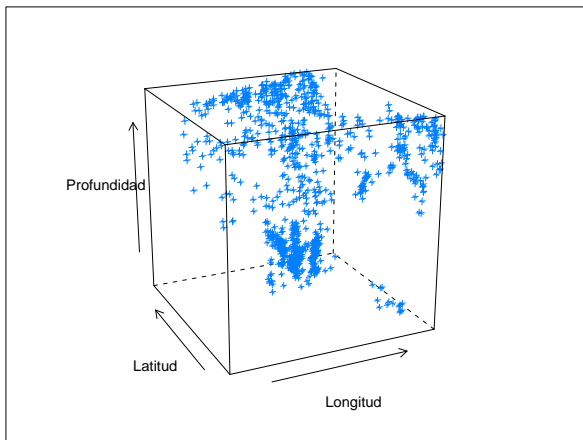
Lattice

Histogramas

Comp. Distr.

Visualizar

Datos



# lattice vs R

R /  
Bioconductor:  
Curso  
Intensivo

Bivariados

Gráficas Bivar.

Correlaciones

Regresiones  
lineales

Multivariados

Gráficas  
Multivar.

Iris

Un paréntesis

Lattice


Histogramas

Comp. Distr.

Visualizar  
Datos

- Una diferencia muy importante entre las gráficas de R y las de `lattice` es que las segundas son objetos de clase "Trellis".
- Esto quiere decir que los podemos almacenar en una variable para luego imprimirlos con `plot` o `print`.
- Para un ejemplo, fíjense en el código<sup>6</sup> con el que hice las gráficas de Iris en `lattice`.
- En fin, hay **muchas** más funciones de `lattice` que les pueden ser interesantes. Si quieren chequen los ejemplos y el archivo `latticeIntro.pdf` ;)

---

<sup>6</sup>A través de la página del curso, porque está invisible en la presentación 

- Sé que fue un bombardeo de info :P
- En fin, hagan el ejercicio 3 en la página del curso.
- **Suerte!**