MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

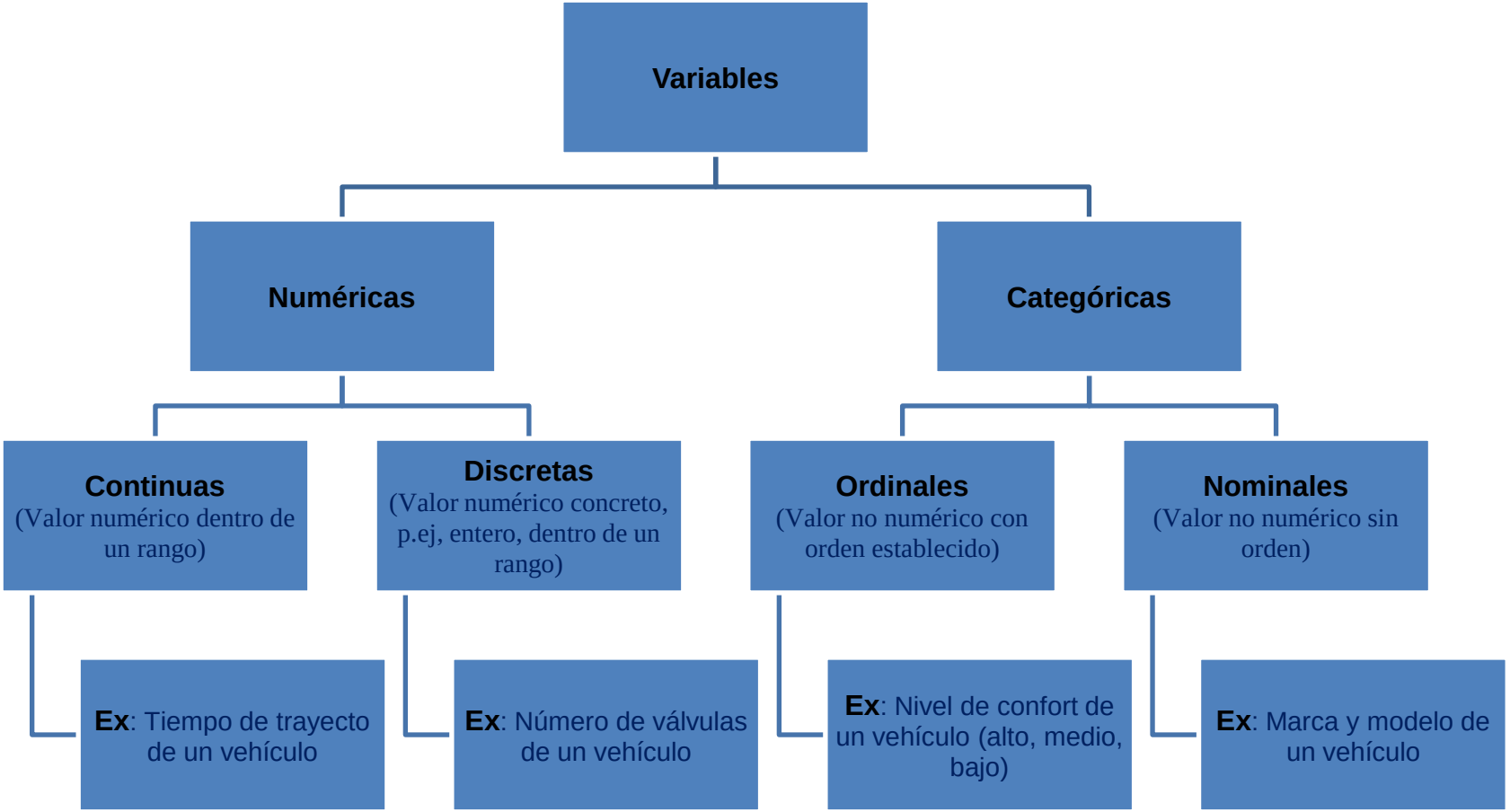
laSalle ENG
Universitat Ramon Llull

Estadística Descriptiva

Noción

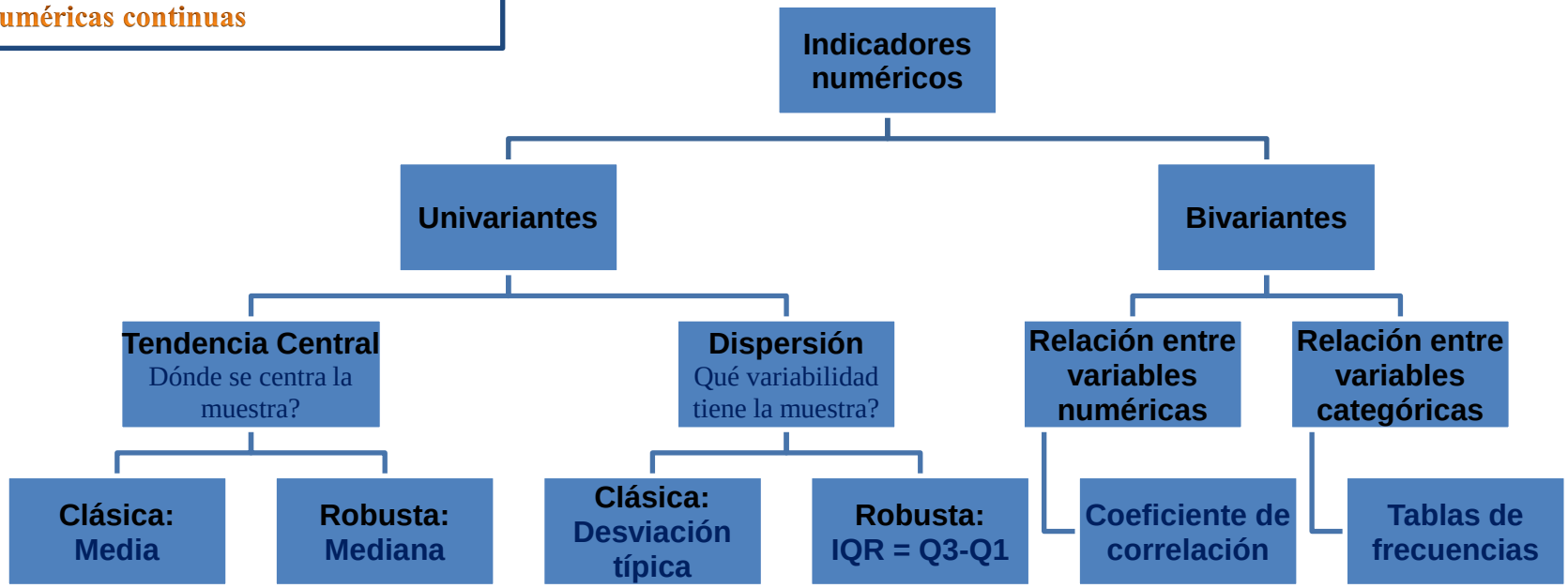
- Objetivo:
  - Univariante: Describir (estadísticamente) las variables de una muestra una a una.
  - Bivariante: Describir (estadísticamente) las relaciones existentes entre dos variables en una muestra
- Herramientas:
  - Indicadores numéricos: media, mediana, desviación típica…
  - Gráficos: histograma, boxplot, diagrama de barras…
- Independientemente del análisis a realizar, siempre se debe realizar una descriptiva de los mismos para conocerlos.

Tipos de variables

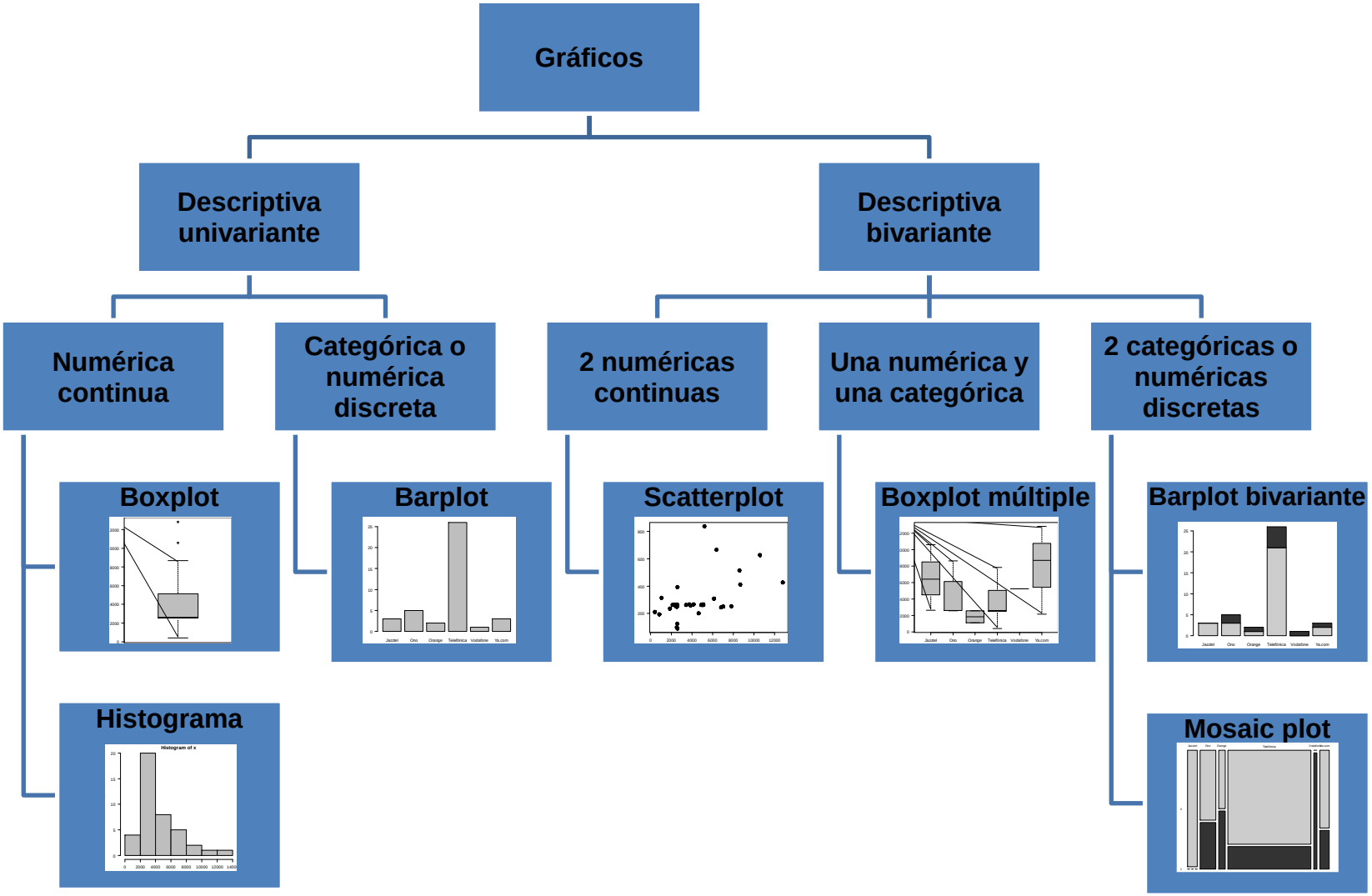


Indicadores numéricos

Los indicadores univariantes, sobretodo aplican a variables numéricas continuas



Gráficos descriptivos



Datos-Ejemplo

Descripción

- Con un conjunto de datos concretos, se analizará como llevar a cabo la descriptiva en distintas situaciones:
- Univariante:
  - Variable numérica
  - Variable categórica
- Bivariante
  - Numérica vs Numérica
  - Numérica vs Categórica
  - Categórica vs Categórica

Datos-Ejemplo

Descripción

- Datos correspondientes a los teléfonos móviles de 179 estudiantes de una universidad australiana (*Mobiles* del paquete *MindOnStats*)
- Variables:
  - Gender
  - Age
  - Faculty
  - Brand
  - Colour
  - Provider
  - PlanType
  - Bill
  - PrimaryUse
  - No.Phones

Datos-Ejemplo

Inspección

```
## Instalar y cargar paquete MindOnStats
install.packages("MindOnStats")
library(MindOnStats)

## Cargar datos
data(Mobiles)

## Inspeccionar datos
dim(Mobiles)        # Número de filas y columnas
names(Mobiles)      # Nombre de las variables
summary(Mobiles)    # Descriptiva univariante de las variables
```

Numérica - indicadores clásicos

Media, varianza y desviación típica

Media

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 22.22$$

n: longitud de la muestra

x_i: observación i-ésima

Varianza

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 40.53$$

Desviación típica o estándar

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 6.37$$

```
> ## Calcular media de la edad
> mean(Mobiles$Age)
[1] 22.21788

> ## Calcular varianza de la edad
> var(Mobiles$Age)
[1] 40.53091

> ## Calcular desviación típica de la
edad
> sd(Mobiles$Age)
[1] 6.366389
```

La varianza (V) y la desviación típica (DT) son medidas de la variabilidad de la muestra.: La DT indica cuánto se alejan en promedio mis observaciones de la media. La V es el cuadrado de la DT.

Numérica - indicadores robustos

Mediana, Cuartiles y IQR

Q1: Cuartil 1

Posición Q1 = (n+1)/4 = (179+1)/4 = 45

Q1 = $X_{(45)}$ = 18

Q2: Mediana o Cuartil 2

Posición Q2 = (n+1)/2 = (179+1)/2 = 90

Q2 = $X_{(90)}$ = 20

Q3: Cuartil 3

Posición Q3 = 3·(n+1)/4 = 135

Q3 = $X_{(135)}$ = 24

Rango Intercuartílico

IQR = Q3 - Q1 = 24 – 18 = 6

Mediana (Q2) : valor que deja el 50% de las observaciones por debajo (50% por encima)

Cuartil 1 (Q1): valor que deja el 25% de las observaciones por debajo (75% por encima)

Cuartil 3 (Q3): valor que deja el 75% de les observaciones por debajo (25% por encima)

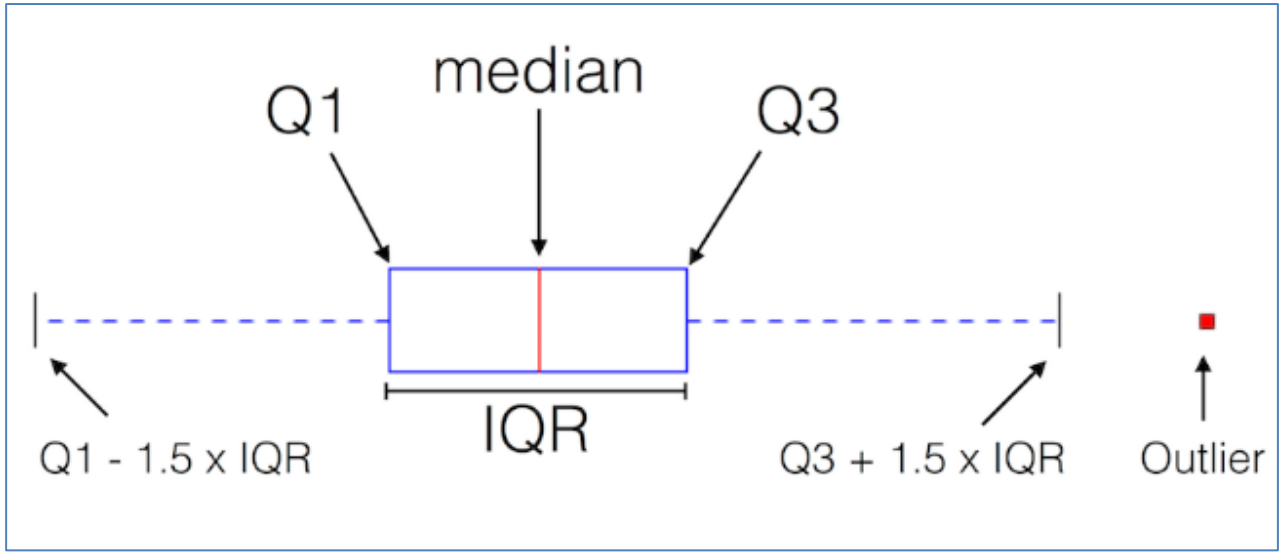
Si la posición es un entero, el cuartil correspondiente es el valor que ocupe esa posición en la lista de valores ordenados crecientemente. Si no es un entero, se ponderan los valores anterior y posterior según la parte decimal de la posición.

```
> ## Calcular Q1,Q2 y Q3 entre otros
> summary(Mobiles$Age)
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
17.00   18.00  20.00 22.22   24.00 55.00
> ## Calcular IQR > IQR(Mobiles$Age)
[1] 6
```

Numérica - Gráficos

Diagrama de caja (*Boxplot*)

- Representa los indicadores robustos y los *outliers* (datos extremos). Partes:
  - **Caja**. Delimitada por el Q1 y por el Q3. La línea interior representa la mediana.
  - **Bigotes**. Desde la caja con una longitud máxima de 1.5 veces el IQR (=longitud de la caja). Si el mínimo o máximo de la muestra no llegan a esta distancia, el bigote llega hasta estos valores.
  - **Outliers**. Puntos más allá de los bigotes. Se consideran valores anómalos.



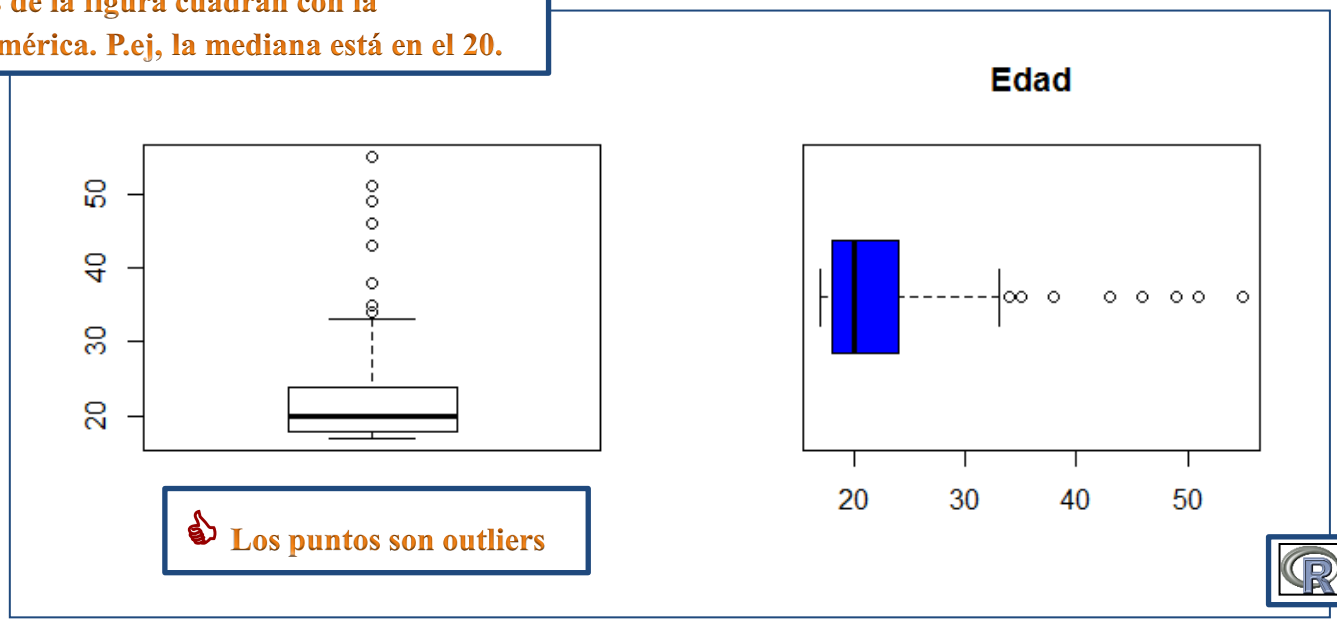
Numérica - Gráficos

Diagrama de caja (*Boxplot*)

La instrucción *par* sirve para fijar parámetros gráficos

```
> ## Boxplot simple y más elaborado
> par(mfrow=c(1,2))                    # Dos ventanas gráficas
> boxplot(Mobiles$Age)                 # Boxplot simple
> boxplot(Mobiles$Age,col="blue",horizontal=TRUE,main="Edad",las=1)  # Complejo
```

Los valores de la figura cuadran con la descriptiva numérica. P.ej, la mediana está en el 20.



Los puntos son outliers

?boxplot para ayuda

Numérica - Gráficos

Outliers

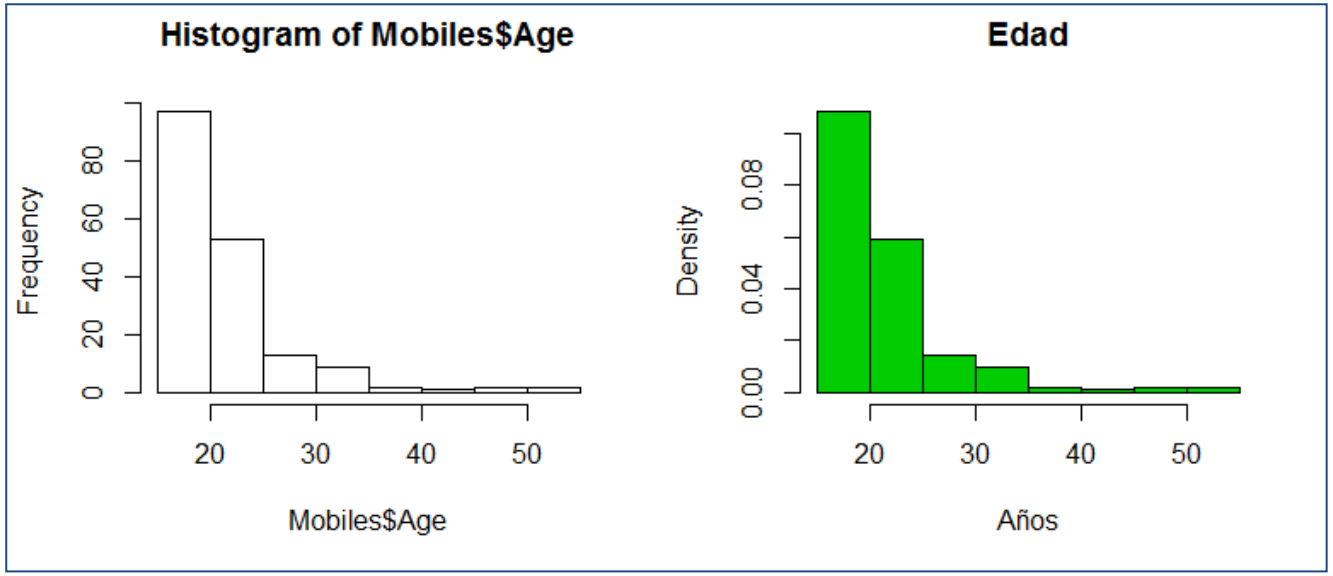
- Los *outliers* son valores anómalos dentro de nuestra muestra
- El primer paso es verificar si se tratan de errores de recogida de los datos
- En caso de no tratarse de errores, no es aconsejable eliminarlos para hacer el análisis más confortable.
- En algunos casos, existen técnicas para modificar su valor de forma que no distorsionen los análisis

Numérica - Gráficos

Histograma

- Representa la distribución de la variable
- El eje horizontal contiene los valores de la variable y el eje vertical, las frecuencias (o la densidad)

```
> ## Histograma simple y más elaborado
> par(mfrow=c(1,2))
> hist(Mobiles$Age)
> hist(Mobiles$Age,col=3,xlab="Años",main="Edad",freq=FALSE)
```



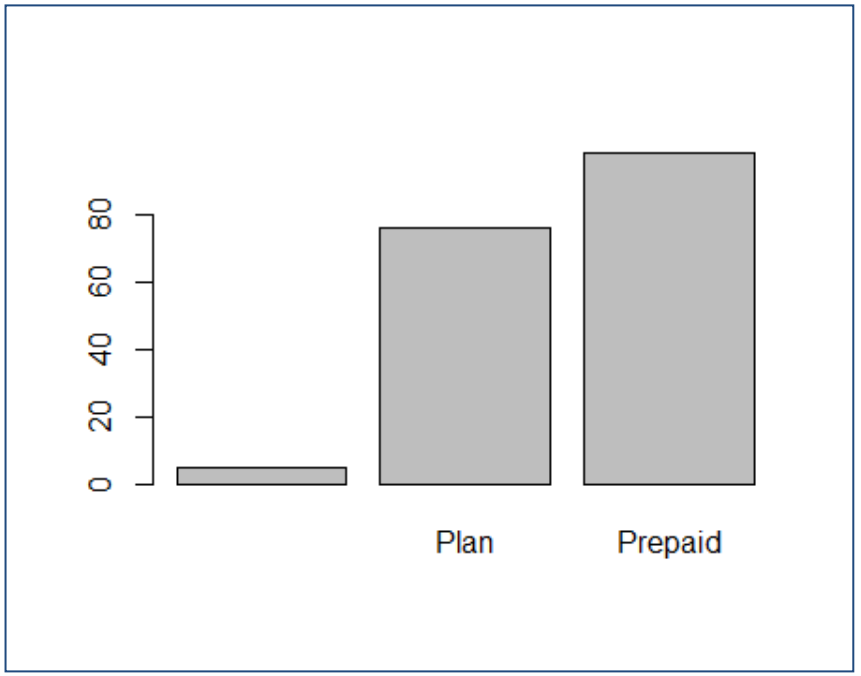
Hay casi 100 usuarios con menos de 20 años

Categórica - indicadores numéricos y gráficos

Tablas y diagramas de barras (*barplots*)

- Representa el número de efectivos (frecuencias) para cada categoría
- El eje horizontal contiene las categorías de la variable y el eje vertical, las frecuencias (o proporciones)

```
> ## Tabla de frecuencias, proporciones y diagrama de barras
> table(Mobiles$PlanType)
   Plan Prepaid
5    76      98
> prop.table(table(Mobiles$PlanType))
                Plan    Prepaid
0.02793296 0.42458101 0.54748603
> barplot(table(Mobiles$PlanType))
```

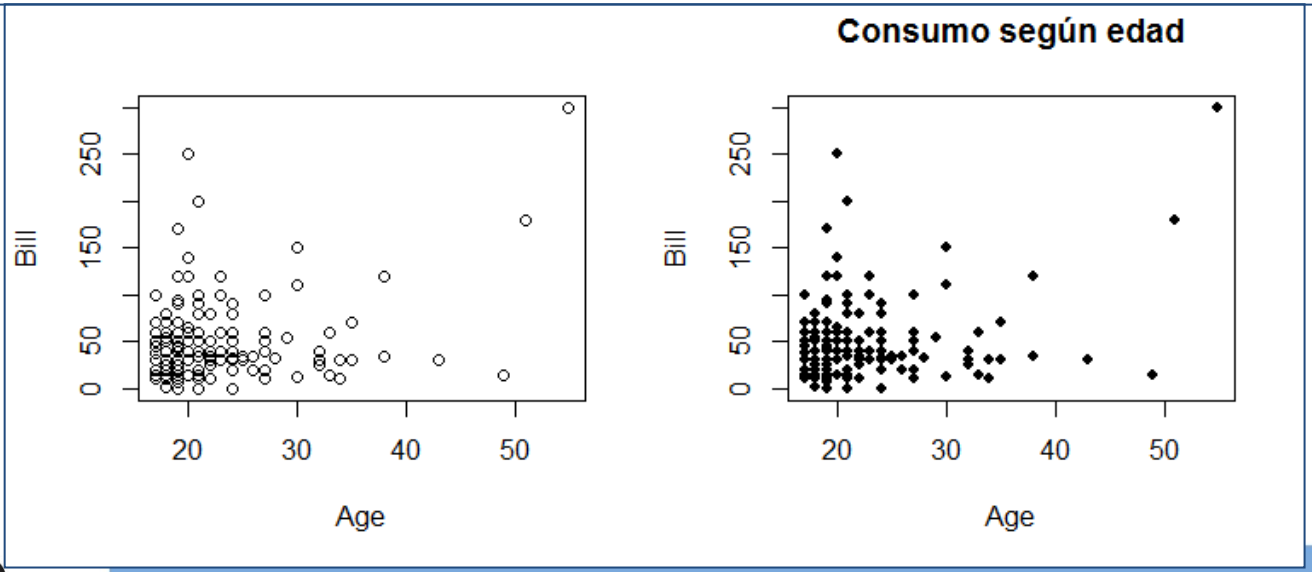


Núm vs Núm - Gráficos

Diagrama de dispersión (*Scatterplot*)

- Representa la distribución bivariante de dos variables numéricas.
- El eje horizontal contiene los valores de la variable explicativa y el eje vertical, de la variable respuesta.

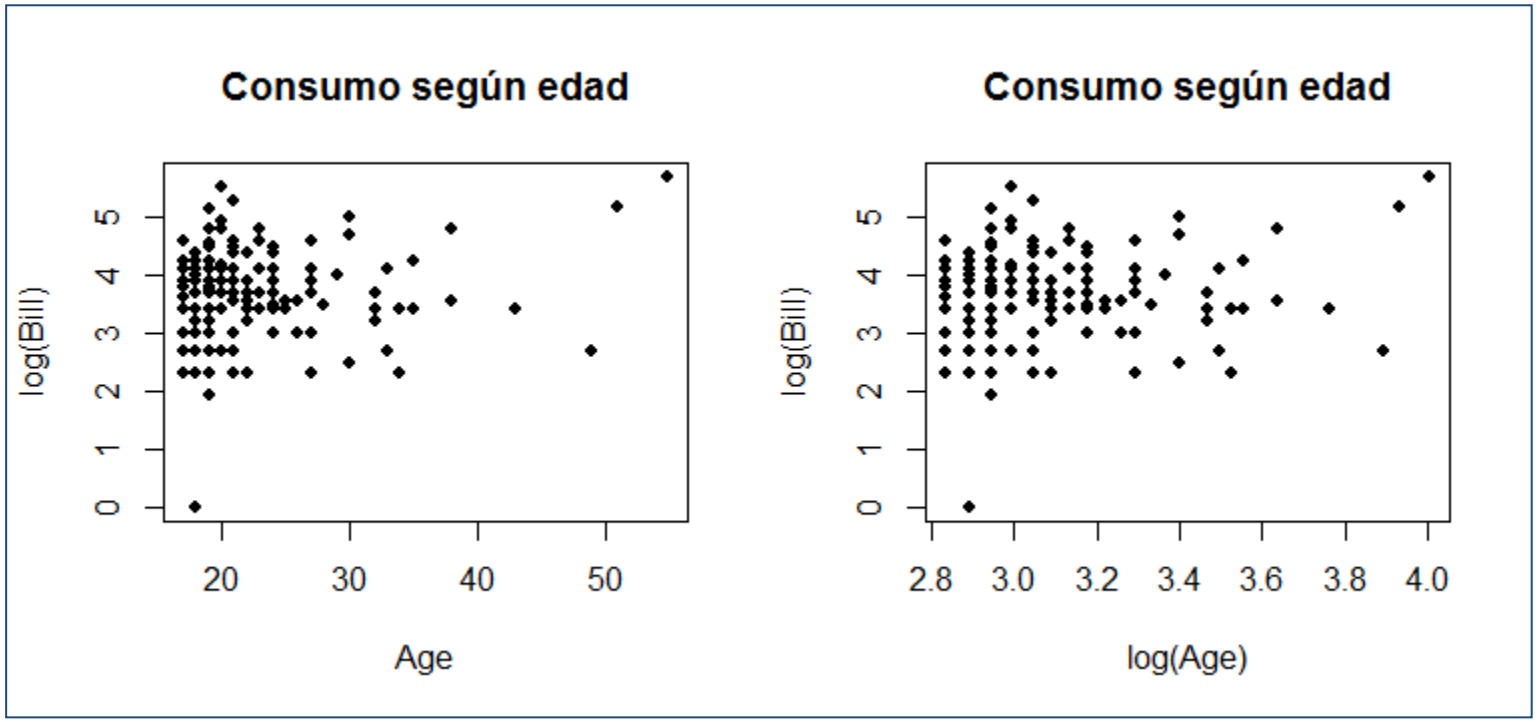
```
##--Numérica vs Numérica
par(mfrow=c(1,2))
plot(Bill~Age,Mobiles)
plot(Bill~Age,Mobiles,pch=19,cex=0.8,main="Consumo según edad")
```



Núm vs Núm - Gráficos

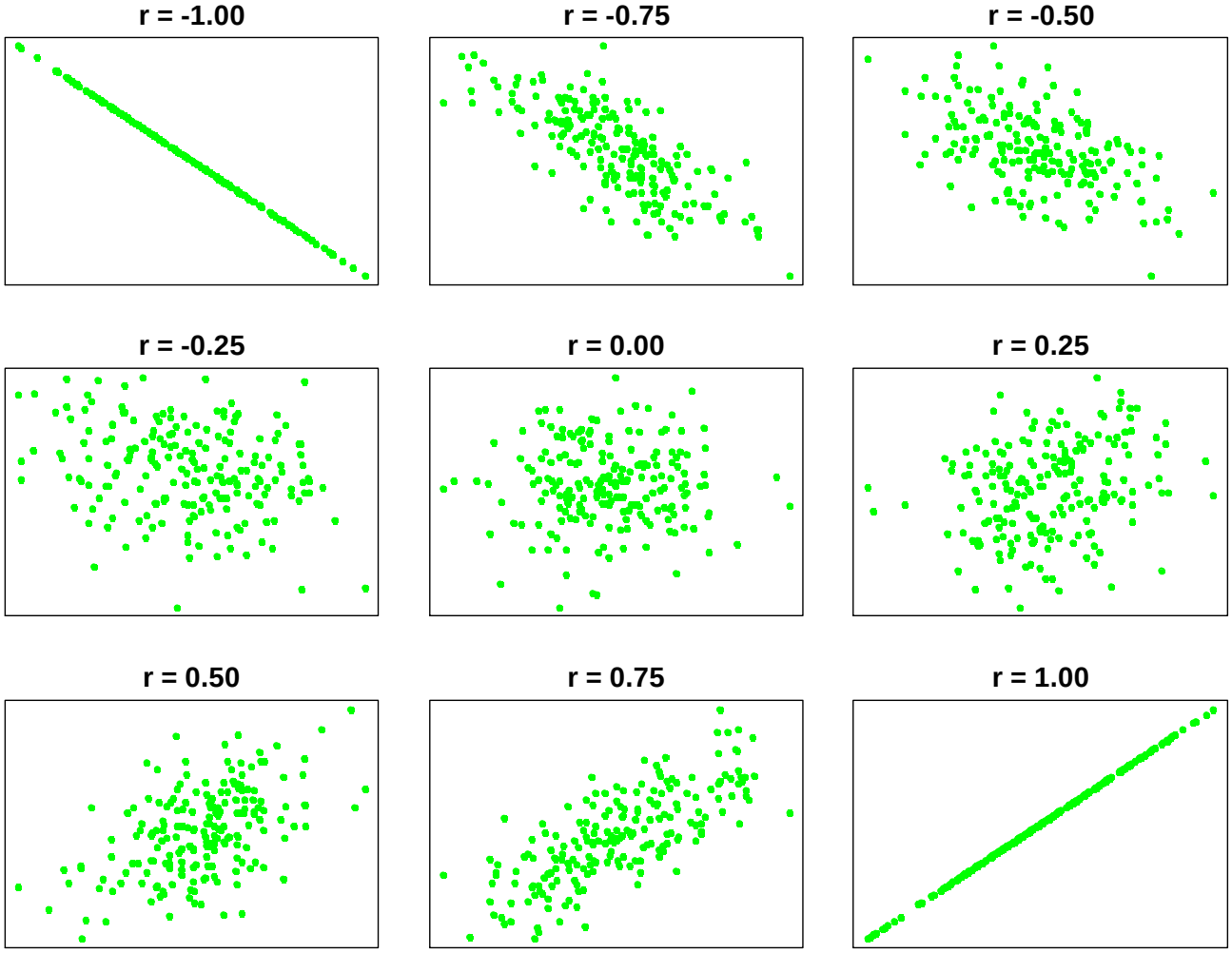
Logaritmos

- En ocasiones, los logaritmos "normalizan las variables" (sólo si toman valores estrictamente positivos) permitiendo ver las relaciones de forma más nítida
- No parece existir una relación entre el consumo de teléfono y la edad



Núm vs Núm – Indicador numérico

Coeficiente de correlación (r)



Núm vs Núm – Indicador numérico

Coeficiente de correlación (r)

- Determina la dirección y la intensidad de una relación lineal entre dos variables numéricas
- Es resultado de dividir la variación conjunta de X e Y (S_{XY}) entre el producto de las desviaciones estándares de X (S_X) e Y (S_Y)

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{S_X S_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

n: longitud de la muestra

x_i, y_i: observación i-ésima de x o y

$\bar{x}$, $\bar{y}$: media de x o y

S_x, S_y: desviación de de x o y

- Es un valor entre -1 i +1
- El signo indica la dirección de la relación: directa (positivo) o inversa (negativo)
- La magnitud en valor absoluto mide la intensidad de la relación
- r_{xy}= 0 indica ausencia de relación lineal
- r_{xy} = 1 ó r_{xy} = -1 indica relación lineal perfecta representable por la recta Y = a + bX

Cat vs Cat – Indicador numérico

Tablas de contingencia

- Representa las frecuencia de una variable categórica dentro de cada categoría de la otra

```
> with(Mobiles,table(PlanType,PrimaryUse))
          PrimaryUse
PlanType          Both Calls SMS
               5     0     0   0
Plan           0    33    22  21
Prepaid        0    35    16  47
```

```
> with(Mobiles,prop.table(table(PlanType,PrimaryUse),1))
          PrimaryUse
PlanType                 Both      Calls        SMS
        1.0000000 0.0000000 0.0000000 0.0000000
Plan    0.0000000 0.4342105 0.2894737 0.2763158
Prepaid 0.0000000 0.3571429 0.1632653 0.4795918
```

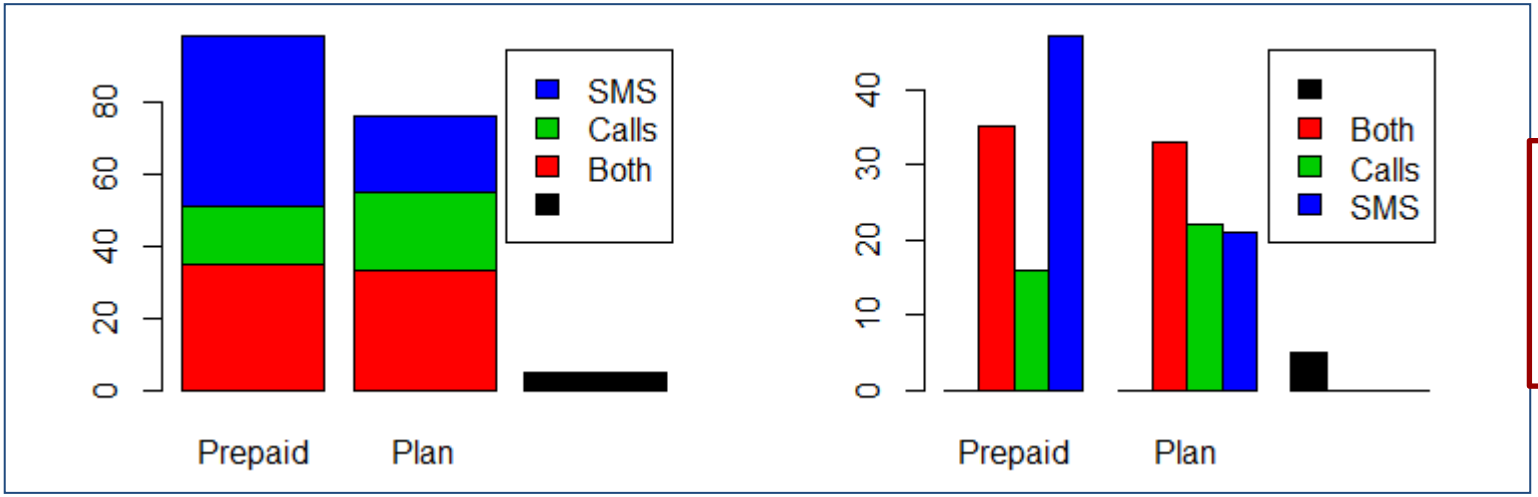
La instrucción prop.table con el parámetro *margin = 1* da la proporción por filas.
Los espacios en blanco representan valores ausentes

Cat vs Cat – Gráficos

Diagrama de barras (barplot)

- Representa las frecuencia de una variable categórica dentro de cada categoría de la otra

```
##-- Cat vs Cat - Barplot
par(mfrow=c(1,2))
with(Mobiles,barplot(table(PrimaryUse,PlanType)[,3:1],col=1:4,legend=TRUE))
with(Mobiles,barplot(table(PrimaryUse,PlanType)[,3:1],col=1:4,legend=TRUE,beside=
TRUE))
```



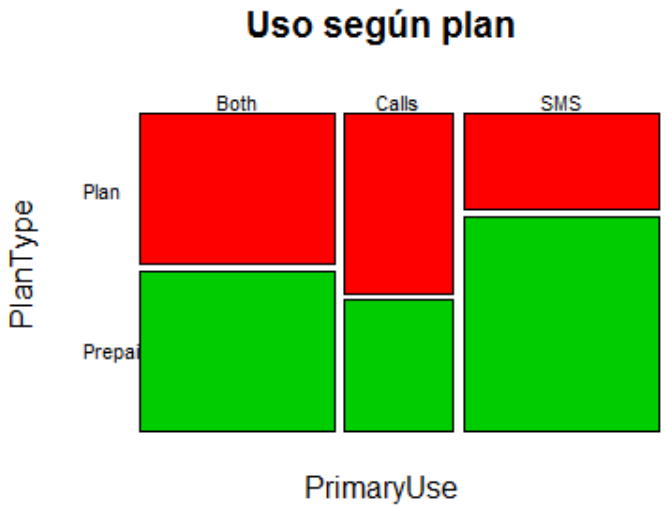
La frecuencia del uso principal del teléfono depende del tipo de plan del cliente

Cat vs Cat – Gráficos

Diagrama de mosaico (*mosaicplot*)

- Es la representación gráfica de una tabla de frecuencias entre dos variables categóricas
- Columna: anchura proporcional a la frecuencia de cada categoría de una les variables
- Files: altura de cada fila en cada columna proporcional a la frecuencia intra-categoría

```
par(mfrow=c(1,1))
Mobiles.Clean <- droplevels(subset(Mobiles,Mobiles$PlanType!="",drop=TRUE))
with(Mobiles.Clean,mosaicplot(PrimaryUse~PlanType,col=2:3,las=1, main="Uso
según plan"))
```



Núm vs Cat – Indicadores numéricos

Indicadores estratificados

- Se hace la descriptiva numérica para cada categoría
- La instrucción *tapply* describe una variable numérica estratificada según otra variable categórica.
- Sintaxi: `tapply (var_num, var_cat, fun)`
  - *var_num*: variable numérica de interés
  - *var_cat*: variable categórica de estratificación
  - *fun*: función a aplicar a la variable numérica

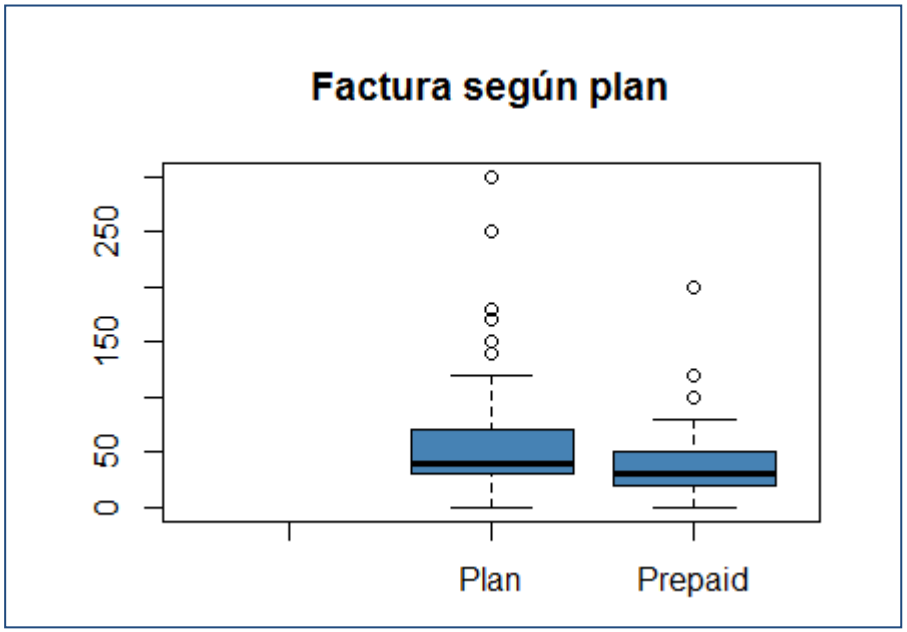
```
> with(Mobiles,tapply(Bill,PlanType,summary))
[[1]]
$Plan
 Min.   1st Qu. Median  Mean 3rd Qu.   Max.
0.00     32.25   40.00 61.17   70.00 300.00
$Prepaid
 Min.   1st Qu. Median  Mean 3rd Qu.   Max.
 0.00    20.00   30.00 35.74 50.00   200.00
```

Num vs Cat – Gráficos

Gráficos estratificados

- Se hace la descriptiva numérica para cada categoria

```
##-- Num vs Cat - Boxplot estratificado
par(mfrow=c(1,1))
boxplot(Bill~PlanType,Mobiles,col="steelblue",main="Factura según plan")
```



MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

laSalle ENG
Universitat Ramon Llull