

Programación y Análisis de Datos

Estadística descriptiva en 

Daniel Fraiman

Maestría en Ciencia de Datos, Universidad de San Andrés

Estadística Descriptiva en

OBJETIVOS:

- 1) ¿Cómo resumir la información?
- 2) ¿Cómo organizar y presentar la información?

Medidas de Resumen: $\left\{ \begin{array}{l} \text{centralidad o posición} \\ \text{dispersión o variabilidad} \end{array} \right.$

MEDIDAS DE RESUMEN

Medidas de posición

Medidas de posición: {
promedio (media) muestral
mediana muestral
media α podada

Medidas de posición

Promedio muestral: $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$

- > mean(datos)
- > mean(datos, na.rm=T) # no tiene en cuenta los NA

Mediana muestral: \tilde{x}

- > median(datos) # si hay NA's median(datos, na.rm=T)

Media α podada: \bar{x}_α

\bar{x}_α = es el promedio de los datos una vez que eliminamos el $\alpha 100\%$ de los datos más chicos y el $\alpha 100\%$ de los datos más grandes.

- > mean(x, trim=0.1) # trim es una proporción $\in (0,1)$

Medidas de dispersión:

Medidas de dispersión: $\left\{ \begin{array}{l} \text{rango muestral} \\ \text{desvío estándar muestral} \\ \text{distancia intercuartil} \\ \text{MAD} \end{array} \right.$

- Todas las medidas de dispersión son ≥ 0 .

Medidas de dispersión:

Rango Muestral: $RM = \text{valor máximo} - \text{valor mínimo} = x_{(n)} - x_{(1)}$

> `diff(range(x))` # `diff` hace la resta de lo que devuelve `range`.

Desvió Estándar Muestral: $S = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}}$

> `sd(x)` # o bien `sqrt(var(x))` donde *var* es la varianza muestral

Distancia Intercuartil: $IQR = \text{tercer cuartil muestral} - \text{primer cuartil muestral} = q_3 - q_1$.

> `IQR(x)`

> `IQR(x)/1.349` # IQR estandarizado

Desvío absoluto mediano: $MAD = \text{mediana } |x_i - \tilde{x}|$

> `mad(x, constant = 1)` # este es el mad

> `mad(x)` # mad estandarizado (`mad(x)/0.6745`)

GRAFICOS


Gráficos

“The greatest value of a picture is when it forces us to notice what we never expected to see.” - John W. Tukey

Descubrir:

- La distribución para entender fenómenos subyacentes.
- Sesgos o errores sistemáticos.
- Variabilidad inesperada en los datos.
- Dependencias o patrones en los datos.

Gráficos

Plots con  básico y con la librería *ggplot2*.

ggplot2

Para que todo sea fácil toda la información que queremos mostrar debería estar en un `data.frame`.

Algunos recursos ggplot2:

- <https://ggplot2.tidyverse.org/>
- <https://www.r-graph-gallery.com/>
- <https://r4ds.had.co.nz/data-visualisation.html#the-layered-grammar-of-graphics>
- “R Graphics Cookbook: Practical Recipes for Visualizing Data”
W. Chang

GRAFICOS 1 VARIABLE

Distribución de una variable categórica

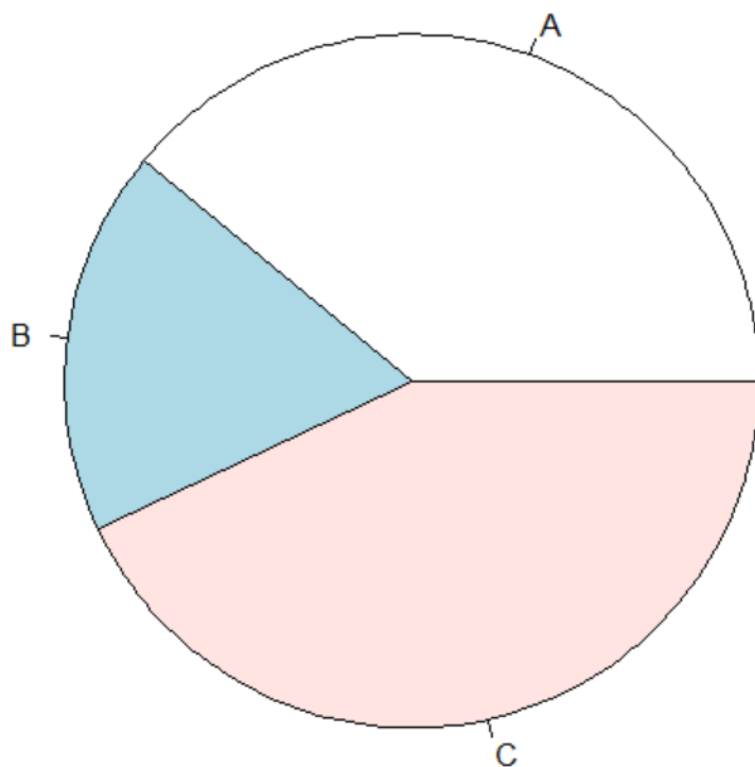
Diagrama de torta

```
> pie(x)
```

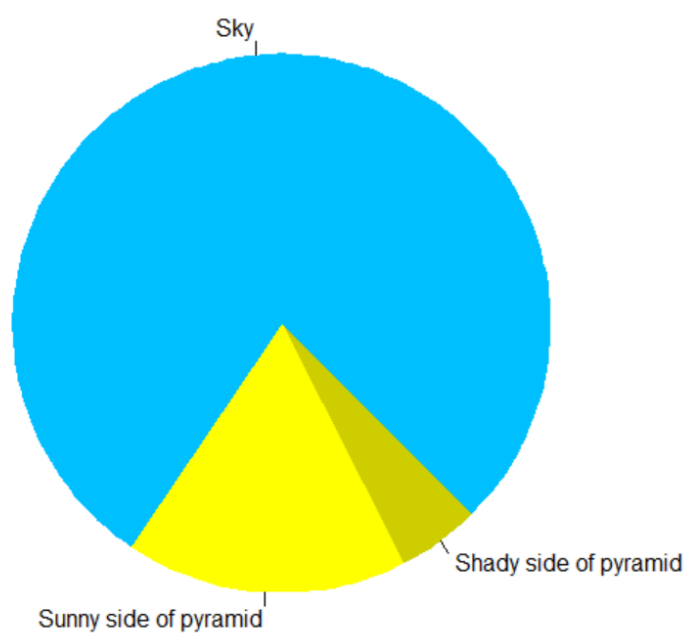
Diagrama de barra

```
> barplot(x,ylab="variable",main="Título",col=color)
```

Distribución de una variable categórica

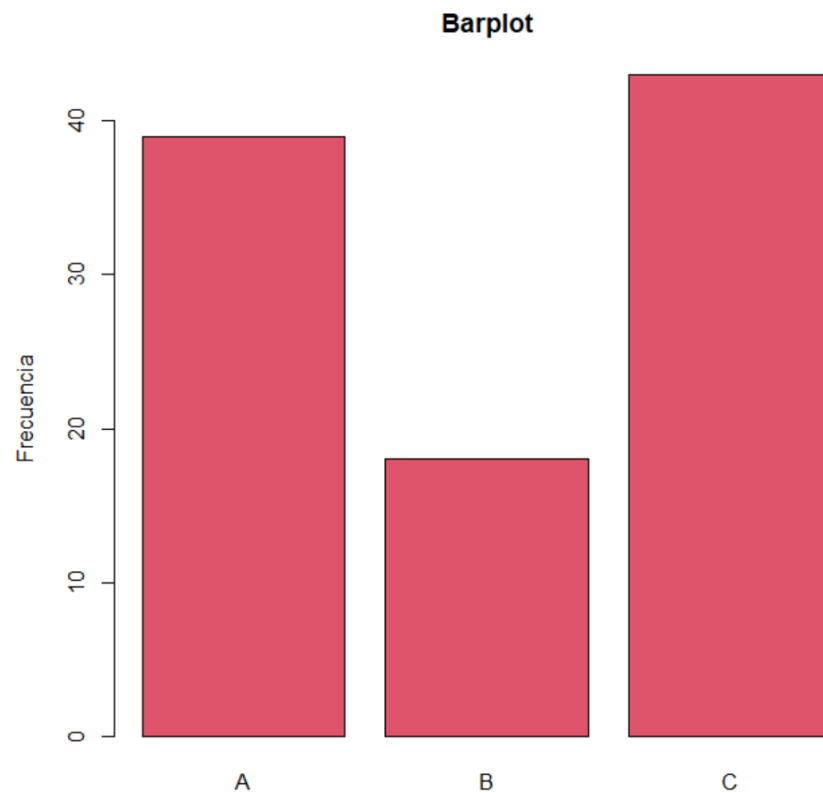


Distribución de una variable categórica



Jajaja, malísimo

Distribución de una variable categórica



Distribución de una variable numérica

Histograma

> hist(x)

>

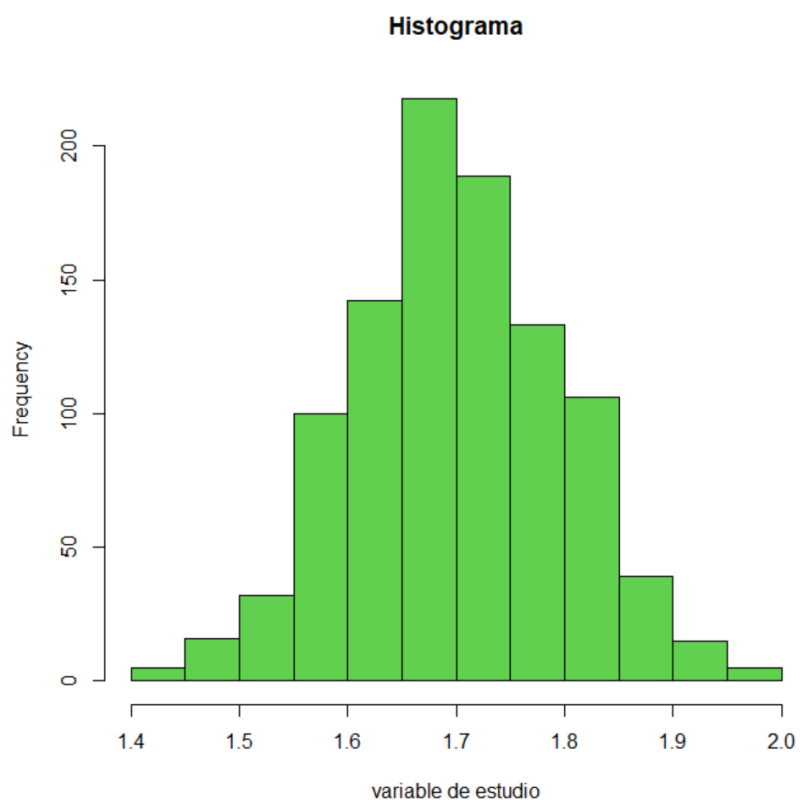
hist(x,xlab="variable",ylab="Frecuencia",main="Título",col=color,
breaks=cortes)

Boxplot

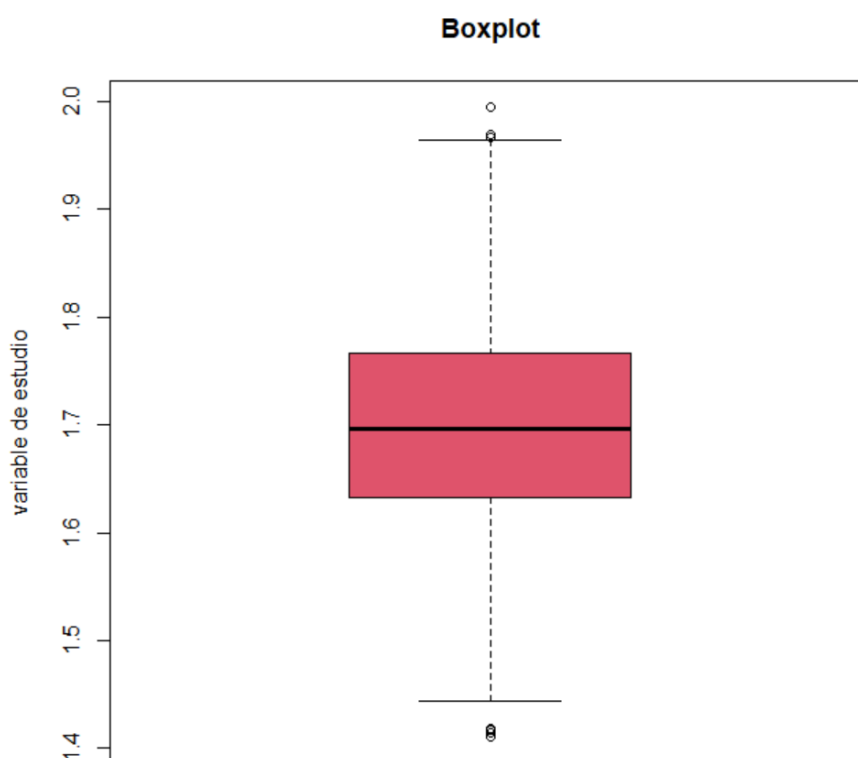
> boxplot(x)

> boxplot(x,ylab="variable",main="Título",col=color)

Distribución de una variable numérica



Distribución de una variable numérica



Distribución de una variable numérica con ggplot2

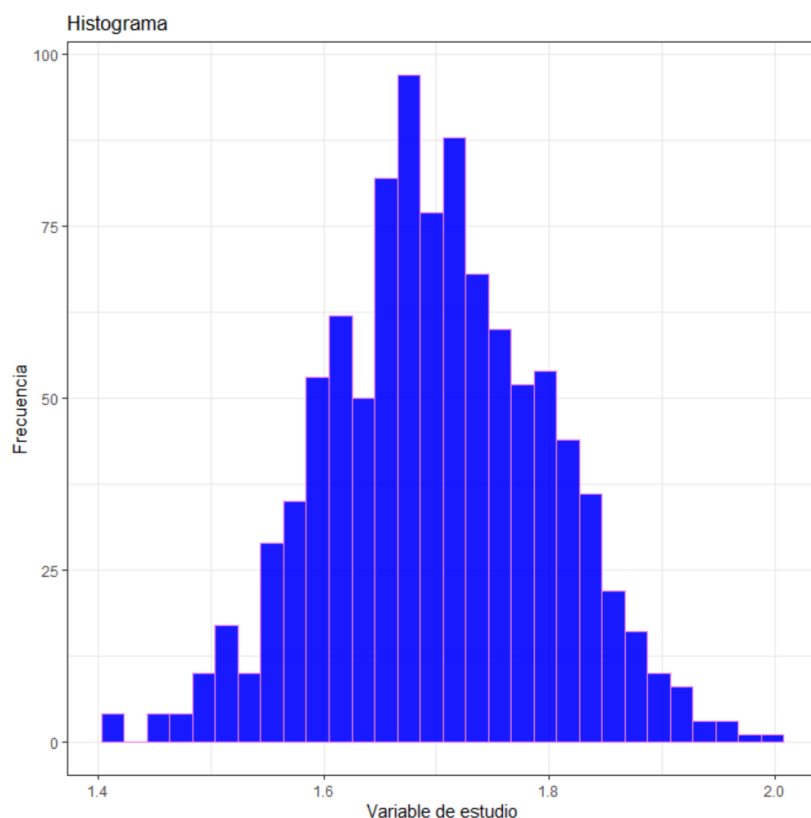
Histograma ggplot2

```
> ggplot() +  
  geom_histogram(aes(x = variable_numerica), color =  
    "violet", fill="blue", alpha=0.9) +  
  labs(y = "Frecuencia", title = "Título") +  
  theme_bw() # opcional "fondo de pantalla"  
              # theme_classic(), theme_minimal()
```

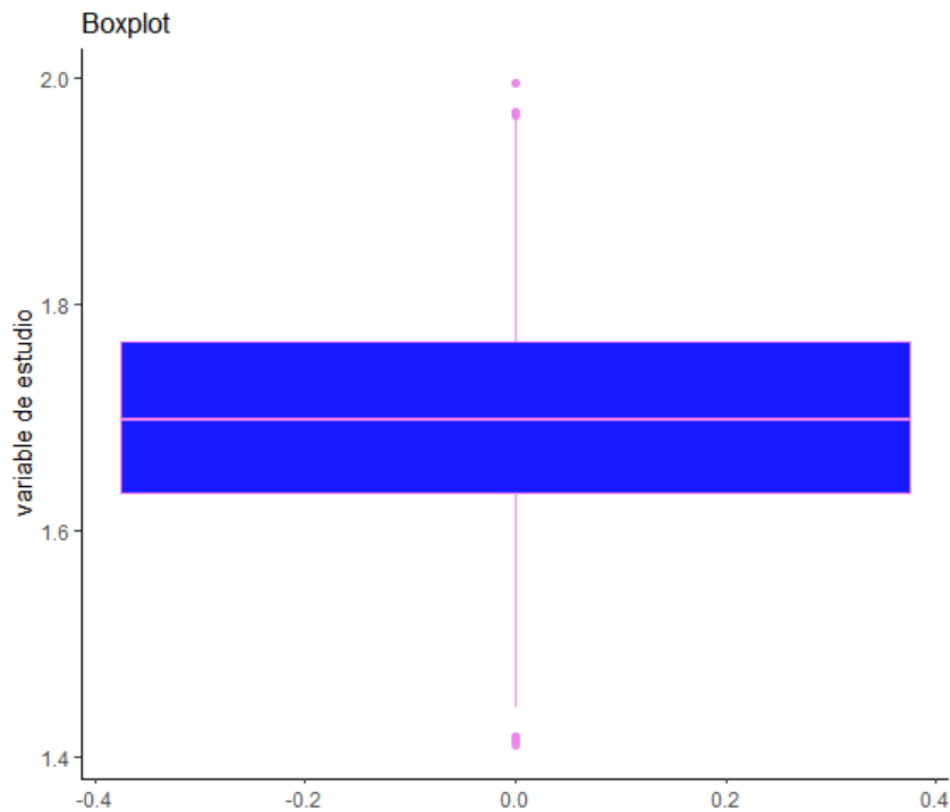
Boxplot ggplot2

```
> ggplot() +  
  geom_boxplot(aes(x = variable_numerica), color="violet",  
    fill="blue", alpha=0.9) +  
  labs(y = "Variable numerica", title="Título")
```

Distribución de una variable numérica



Distribución de una variable numérica

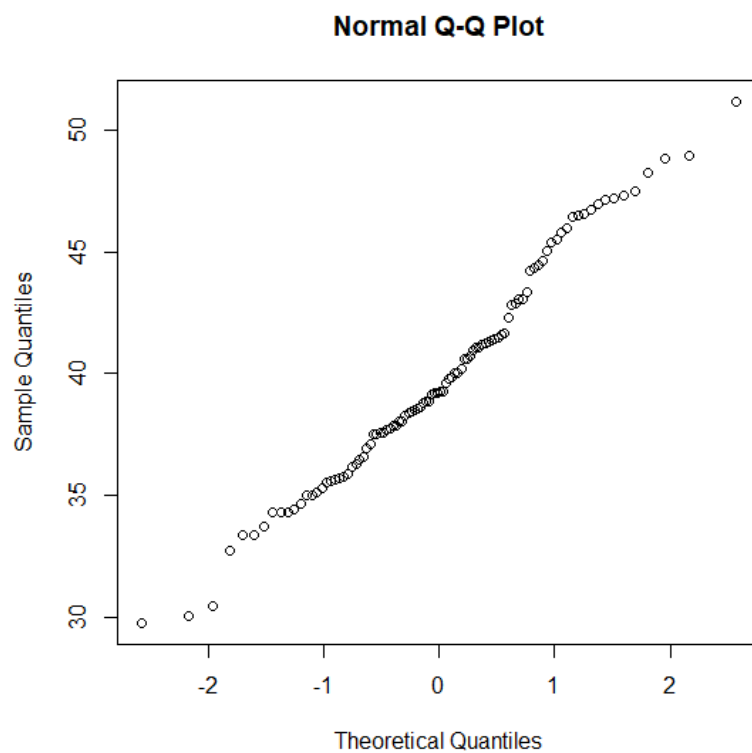


Distribución de una variable numérica

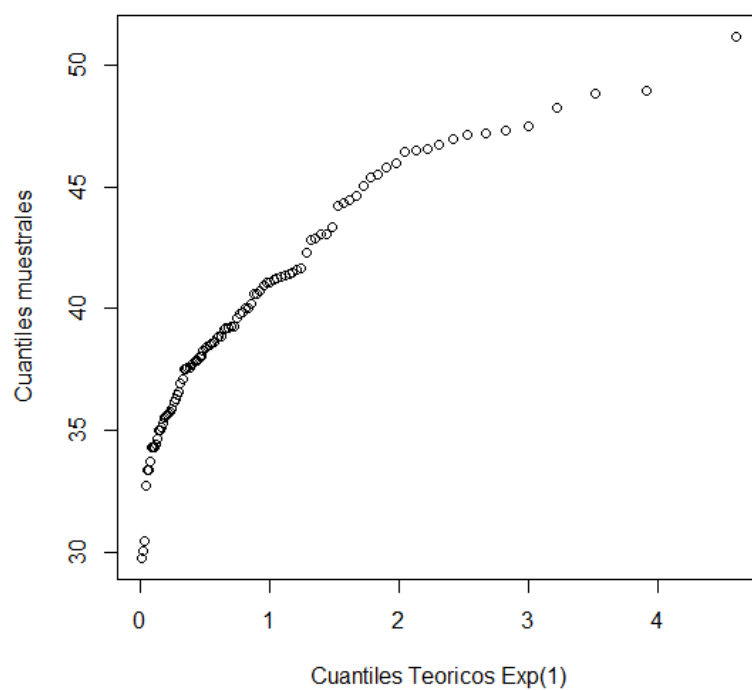
QQ-plot

- > `x= rnorm(100,40,5)`
- > `qqnorm(x)` # para comparar con una normal
- > `cuantiles_teo=qexp(c(1:length(x))/(length(x)+1),1)`
- > `qqplot(cuantiles_teo,x,xlab="Cuantiles Teoricos Exp(1)",ylab="Cuantiles muestrales")`
- > `y= rexp(100,0.1)`
- > `qqplot(cuantiles_teo,y,xlab="Cuantiles Teoricos Exp(1)",ylab="Cuantiles muestrales")`

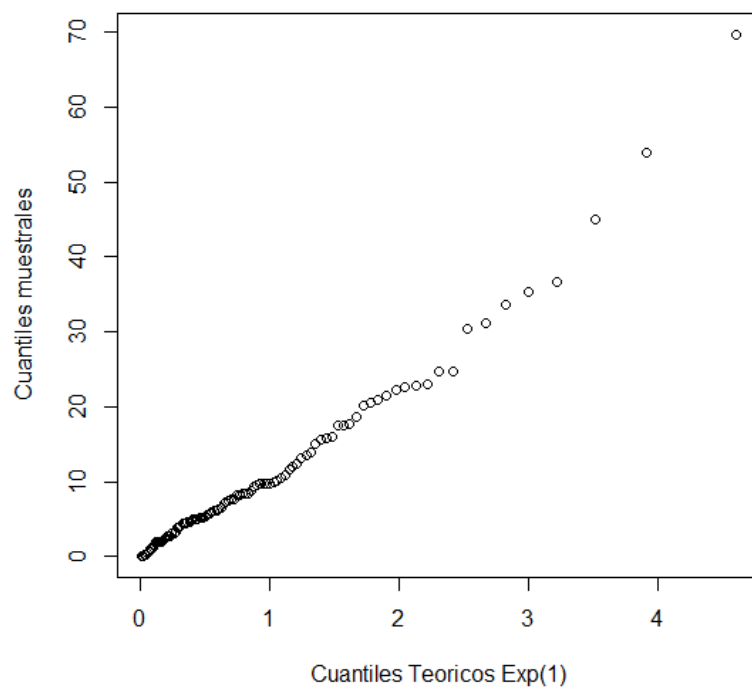
Distribución de una variable numérica



Distribución de una variable numérica



Distribución de una variable numérica



GRAFICOS 2 VARIABLES

Dependencia entre 2 variables

Alternativas entre 2 variables

- Numérica vs Numérica (dep_NN)
- Numérica vs Categórica (dep_NC)
- Categórica vs Categórica (dep_CC)

Gráficos para estudiar dep_NN

Scatter plot

```
> plot(x,y)  
> plot(x,y,xlab="var X",ylab="var Y",pch=2,col=3,  
type="b",main="titulo")
```

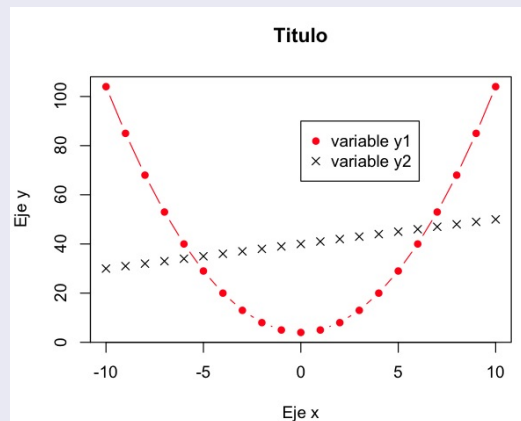
Scatter plot ggplot2

```
> ggplot(data=datos,aes(x=x,y=x))+  
+ geom_point()
```

Scatter y Line plots (dep_NN)

R básico

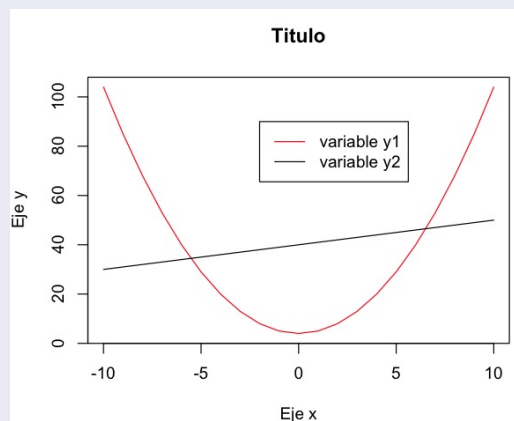
- > `x=seq(-10,10,1)`
- > `y1=x2+4`
- > `y2=x+40`
- > `plot(x,y1,xlab="Eje x",ylab="Eje y",pch=16,col=2,type="b",main="Titulo")`
- > `points(x,y2,pch=4,type="p",col=1)`
- > `legend(0,90,col=c(2,1),pch=c(16,4),c("variable y1","variable y2"))`



Scatter y Lines plots (dep_NN)

R básico

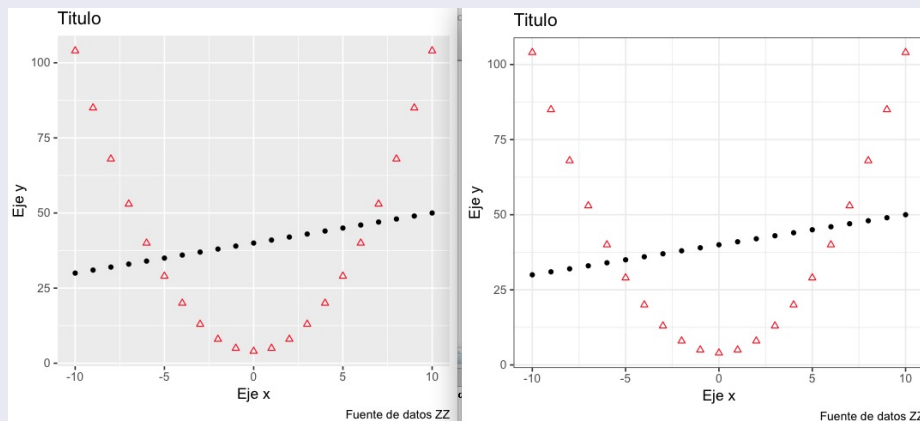
- > `plot(x,y1,xlab="Eje x",ylab="Eje y",col=2,type="l",main="Titulo")`
- > `points(x,y2,type="l",col=1)`
- > `legend(-2,90,col=c(2,1),lty=c(1,1),c("variable y1","variable y2"))`



Scatter y Lines plots (dep_NN)

Con ggplot2: Lógica de capas separadas por “+”

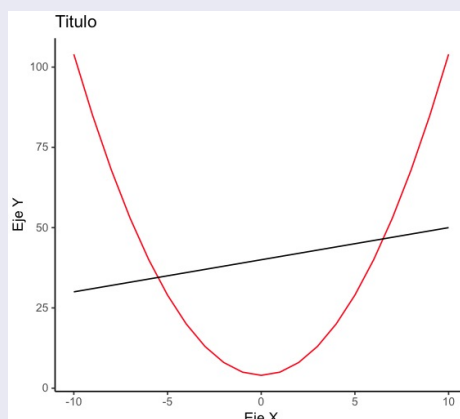
- > `df=data.frame(x,y1,y2) # más cómodo trabajar con data.frames`
- > `require(ggplot2) # o library(ggplot2) para cargar el paquete`
- > `ggplot(df, aes(x=x)) + # objeto gráfico básico`
- > `geom_point(aes(y=y1), colour="red",shape=2) + # primera capa`
- > `geom_point(aes(y=y2), colour="black",shape=16)+ # segunda capa`
- > `labs(x="Eje x",y="Eje y",caption"Fuente de datos ZZ",title="Titulo") + # tercera capa`
- > `theme() # como queremos que sea el fondo, theme_bw()`



Scatter y Lines plots (dep_NN)

Con ggplot2: Lógica de capas separadas por “+”

- > `g <- ggplot(df, aes(x))`
- > `g <- g + geom_line(aes(y=y1), colour="red")`
- > `g <- g + geom_line(aes(y=y2), colour="black")`
- > `g`
- > `g + labs(x="Eje X",y="Eje Y",title="Titulo")+ theme_classic()`



Gráficos para estudiar dep_NC

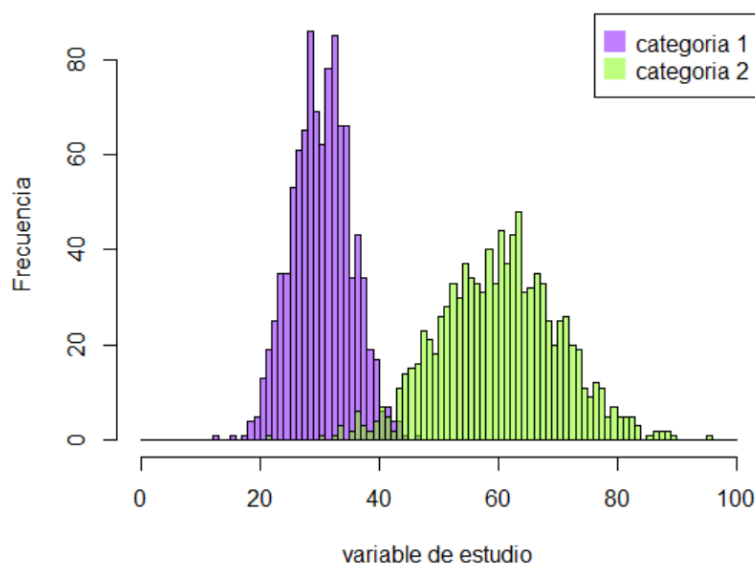
Dos histogramas en figuras separadas

```
> par(mfrow=c(1,2)) # partimos la pantalla grafica en 1 fila y 2 columnas  
> hist(x1, breaks=100, xlim=c(0,100), col=rgb(0,0,1,0.5))  
> hist(x2, breaks=100, xlim=c(0,100), col=rgb(0,0,1,0.5), add=T)  
# hay que usar el mismo intervalo xlim
```

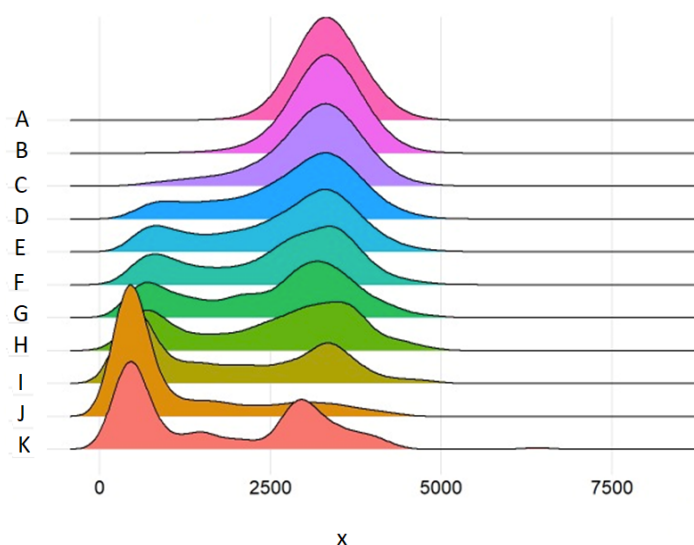
Dos histogramas en una misma figura

```
> hist(x1, breaks=100, xlim=c(0,100), col=rgb(1,0,0,0.5),  
xlab="Variable", ylab="variable", main="Facebook mensual" )  
hist(x2, breaks=100, xlim=c(0,100), col=rgb(0,0,1,0.5), add=T) #  
add=T grafica arriba  
legend("topright", legend=c("categoria 1","categoria 2"),  
col=c(rgb(1,0,0,0.5), rgb(0,0,1,0.5)), pt.cex=2, pch=15 )
```

Gráficos para estudiar dep_NC



Gráficos para estudiar dep_NC



Gráficos para estudiar dep_NC

Boxplots

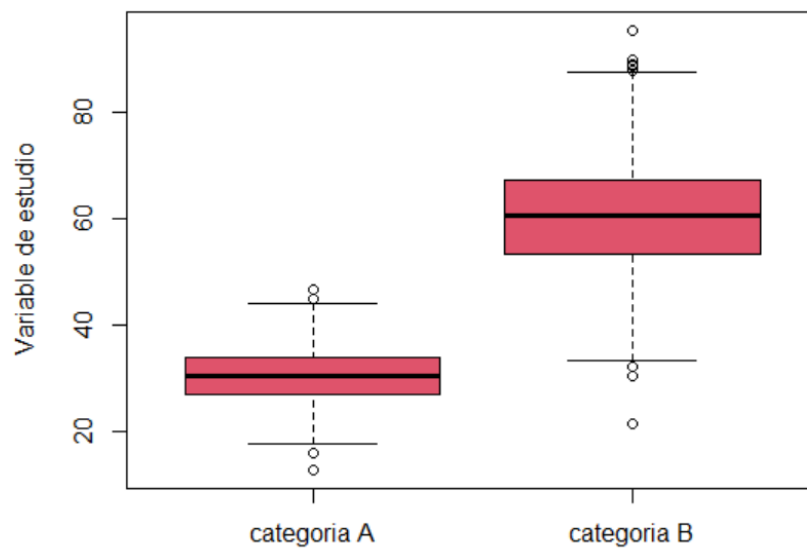
```
> boxplot(x~ y,xlab="Categorías",ylab="Variable",col=2) # x  
numérica y categorica
```

Boxplots en ggplot2

```
> ggplot(data=datos,aes(x=var_categ,y=var_numer))+  
  geom_boxplot()
```

```
> ggplot(data=datos,aes(x=var_categ,y=var_numer))+  
  geom_boxplot( outlier.colour = "blue", outlier.shape = 19,outlier.size  
= 1.5) +  
  theme_bw() +  
  labs(title = "Titulo",caption = "Fuente de datos: XX", x = "Variable  
categorica",y = "Variable numérica")
```

Gráficos para estudiar dep_NC



Gráficos para estudiar dep_CC

2 variables categóricas

- Nivel educativo y voto.
- Sexo y carrera universitaria.
- .

Hacemos una tabla de frecuencias con los resultados, y podemos graficar esta tabla.

Gráficos para estudiar dep_CC

