Introducción al análisis estadístico

Teoría de probabilidad y Modelamiento estadístico

Daniel Jiménez M.

Universidad Nacional de Colombia

12 -10 -2020

Distribución Binomial

Es una distribución de probabilidad que estudia entre el número de repeticiones `n` de un evento hasta llegar a comprender el número de exitos obtenidos. Los valores de este tipo de distriución estan entre 0 y 1.

Ya que ustedes son Muchachos de poca fe, acá les muestro la formula

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

Distribución Binomial

Analicemos el siguiente problema: Suponga que esta haciendo un examen con 20 preguntas, cada pregunta tiene cuatro posibles respuestas y solo una de ellas es la correcta. Encuentre la probabilidad de que al menos ocho respuestas sean las correctas.

Tenga presente que : $1/4 = 0.25$ es la probabilidad que una respuesta sea la correcta.

```
pbinom(8,size = 20,prob = 1/4)
```

```
## [1] 0.9590748
```

La probabilidad de que responda al menos ocho preguntas de manera correcta es del 95%.

Distribución Binomial

Las propiedades de esta distribución son las siguientes:

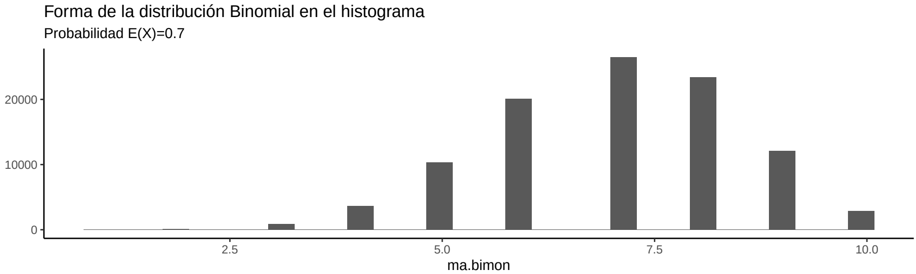
$E(X) = np$

$Var(X)=np(1-p$

$m_x(t)=(pe^t+1-p)^n$

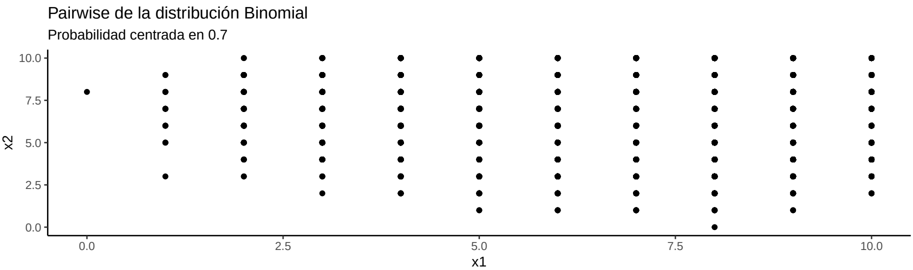
Distribución Binomial

El comportamiento de una variable binomial es el siguiente



Distribución Binomial

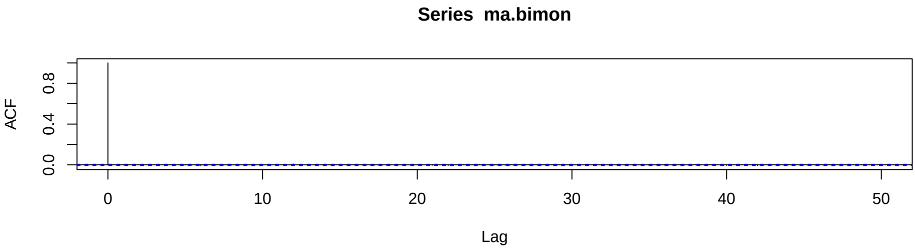
El comportamiento de los datos pareados es el siguiente



Distribución Binomial

La función de Autocorrelación (Esto es super útil cuando quiere hacer forecasting) tiene el siguiente comportamiento

```
acf(ma.bimon)
```



Distribución Poisson

Estudia el número de ocurrencias de algún evento durante un intervalo de tiempo. Este tipo de aplicaciones es super útil cuando quiere calcular la probabilidad de ocurrencia de un evento donde mide la cantidad de individuos o eventos que ocurriran en un momento especifico, como por ejemplo: El número de personas que atenderá un cajero al medio día.

Matemáticamente se describe como :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{!x}$$

Distribución Poisson

Suponga que usted tiene **Tinder** y en promedio en un día hace 12 Match. Calcule la probabilidad de que haga 15 Match en un día.

Se calculará la probabilidad como 14 ó mas Matchs

```r
ppois(q = 14,lambda = 12,lower.tail = FALSE)
```

```
## [1] 0.2279755
```

Distribución Poisson

Las propiedades de la distribución de probabilidad es la siguiente

$$E(X)=\lambda$$

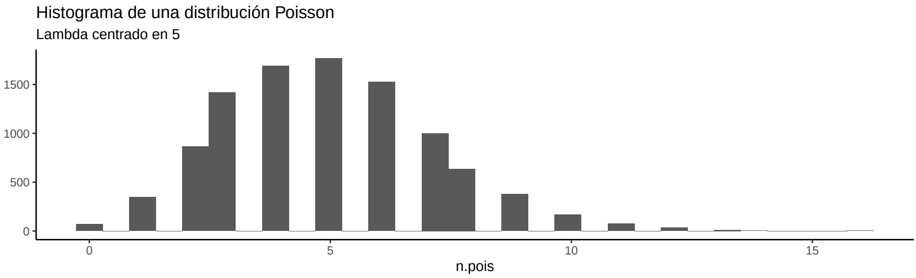
$$Var(X)=\Lambda$$

$$m_x(t)=exp\lambda(e^t-1)$$

Distribución Poisson

El comportamiento del histograma es el siguiente

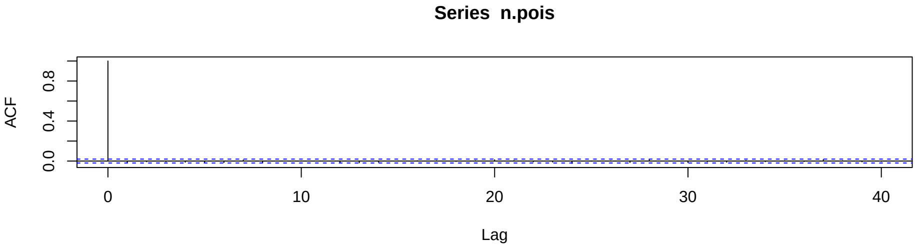
```
n<-10000
n.pois<-rpois(n,lambda = 5)
qplot(n.pois,geom = "histogram")+
  labs(title = "Histograma de una distribución Poisson",
       subtitle = "Lambda centrado en 5")
```



Distribución Poisson

La función de autocorrelación es la siguiente

```r
acf(n.pois)
```



Distribución Exponencial

Calcula la probabilidad de ocurrencia de dos eventos en intervalos de tiempo, como por ejemplo, el tiempo que transcurre hasta recibir una llamada.

Matemáticamente se ve de la siguiente manera:

$$f(x) = \frac{1}{\theta} e^{-x/\theta} I_{(0,\infty)}$$

Distribución Exponencial

Suponga que el tiempo promedio en que un cajero de Juan Valdez vende un producto es de dos minutos. Calcule la probabilidad de que ejecute pagos en al menos 1 minuto.

```r
pexp(1,rate = 1/2)
```

```
## [1] 0.3934693
```

Distribución Exponencial

Las propiedades de esta distribución son:

$$E(X) = \theta$$

$$Var(X) = \theta^2$$

$$m_x(t) = \frac{1}{1-\theta t}$$

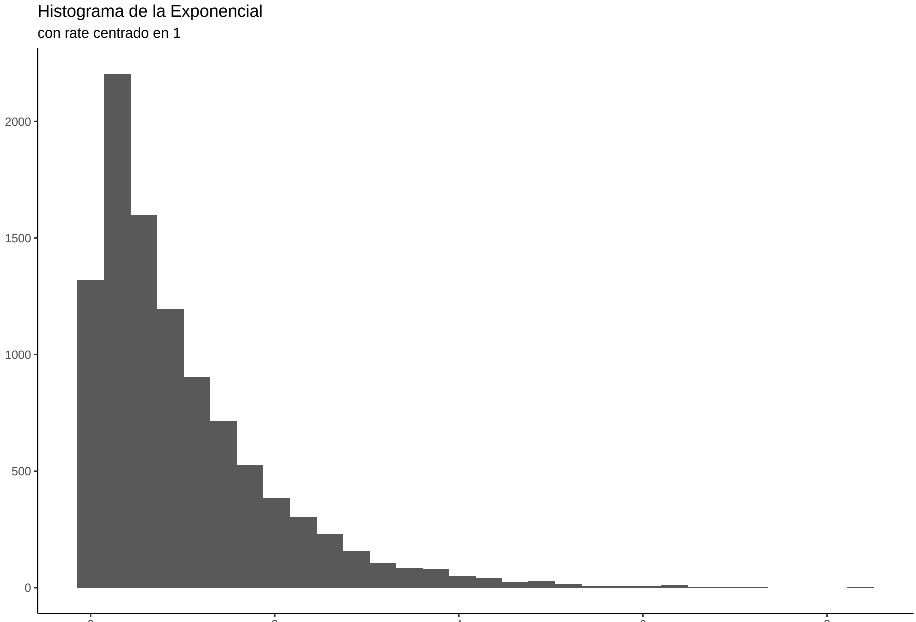
Dato curioso: Notesé que la varianza teórica de la distribución es el cuadrado de la esperanza, por lo tanto a un conjunto de datos continuos mayores a cero, cuando la varianza tiende a parecerse a la esperanza de los valores, podremos decir que esta es una exponencial.

Distribución Exponencial

El histograma de la exponencial tiene el siguiente comportamiento

```
n<-10000
ma.exp<-rexp(n,rate = 1)
qplot(ma.exp,geom="histogram")+
  labs(title = "Histograma de la Exponencial",
       subtitle = "con rate centrado en 1")
```

Distribución Exponencial

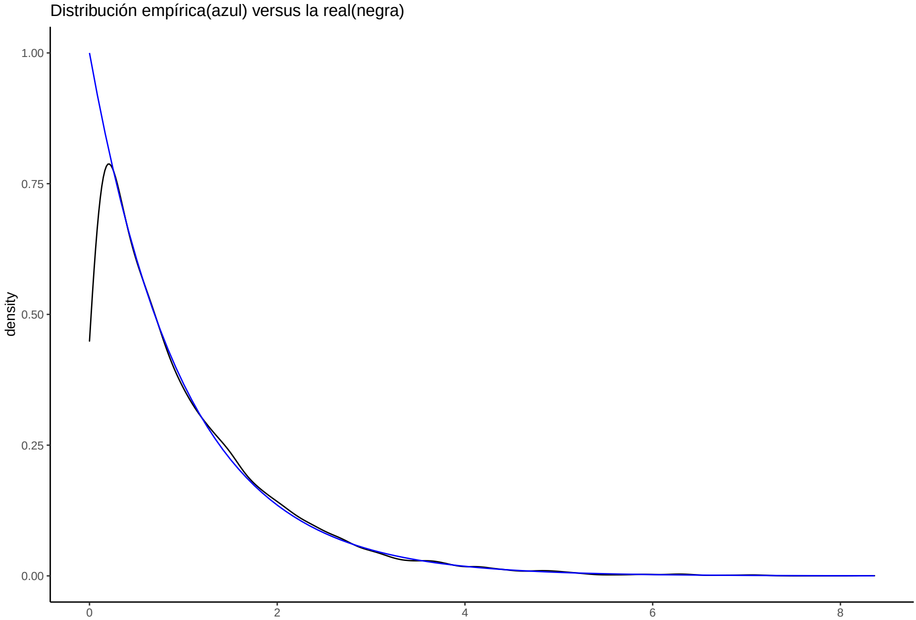


Distribución Exponencial

Una particularidad importante es densidad empírica que consiste en la forma de la distribución como se da a nivel matemático versus su realidad.

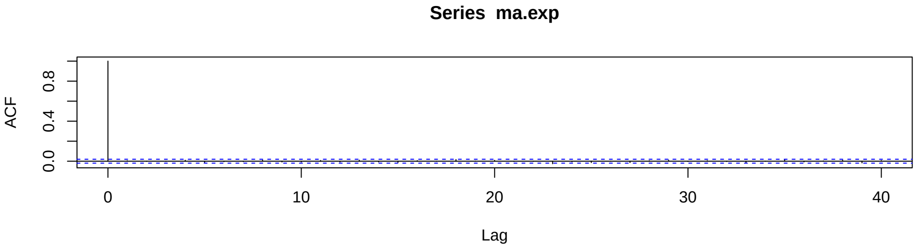
```
h<-data.frame(x=ma.exp)
h%>%
  ggplot(aes(x))+
  geom_density()+
  stat_function(fun = dexp,geom = "line",col="blue")+
  labs(title = "Distribución empírica(azul) versus la real(neg
```

Distribución Exponencial



Distribución Exponencial

```r
acf(ma.exp)
```



Prueba Kolmogorov - Smirnov

Imaginesé que quiere comprobar si una distribución proviene de una familia estadística especifica, par comprobarlo debe usar una prueba de hipótesis

H_0 : Los datos provienen de una distribución específica
H_1 : Los datos no provienen de dicha distribución

Ahora viene el mejor amigo de todos, el p-value: si este es menor que α, cuando $\alpha = 0.05$, entonces se rechaza la hipótesis nula.

Prueba Kolmogorov - Smirnov

```r
ks.test(ma.exp,"pexp",rate=1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  ma.exp
## D = 0.0071654, p-value = 0.6836
## alternative hypothesis: two-sided
```

Se acepta la distribución exponencial.

Distribución Normal

También se le conoce como distribución gaussiana, es la más utilizada en teoría estadística gracias a su amplitud de aplicaciones en temas sociales, naturales y en psicología. El poder de esta distribución radica que asume eventos incontrolables como independientes en cada una de sus observaciones.

Matemáticamente se describe de la siguiente manera:

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Distribución Normal

Suponga que el ranking para adquirir un prestamos de vivienda debe ser de 80 (promedio) puntos, la desviación del mismo esta en 30. ¿Cúal es la probabilidad de que personas que accedan al crédito este por encima de 90?

```r
pnorm(90,mean = 80,sd = 30,lower.tail = FALSE)
```

```
## [1] 0.3694413
```

Distribución Normal

Las propiedades de la normal son :

$$E(X){=}\mu$$

$$Var(X){=}\sigma^2$$

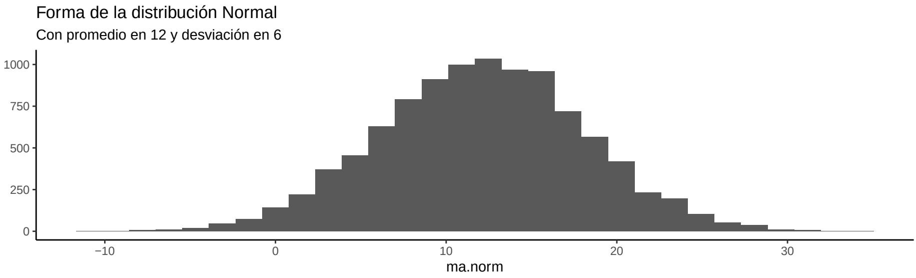
$$m_x(t){=}exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$$

Distribución Normal

El histograma de la normal es el siguiente :

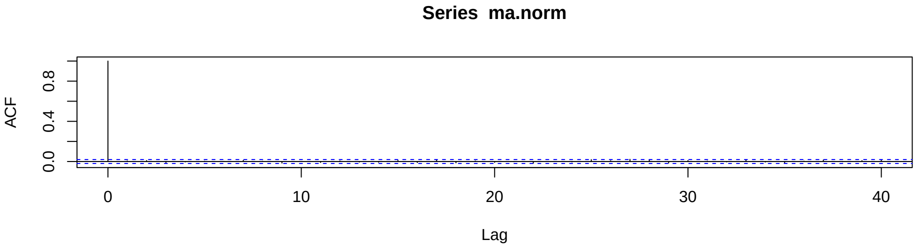
```
ma.norm<-rnorm(10000,mean = 12,sd = 6)
qplot(ma.norm,geom = "histogram")+
  labs(title = "Forma de la distribución Normal",
       subtitle = "Con promedio en 12 y desviación en 6")
```

Distribución Normal



La función de autocorrelación es la siguiente :

```r
acf(ma.norm)
```



Distribución Weibull

Es conocida también como análisis de sobrevivencia, ya que estudia el tiempo transcurrido hasta que llegue un evento de fallecimiento o falla de un fenómeno estudiado.

Matemáticamente se describe como:

$$f(x) = \frac{k}{\theta^k} x^{(k-1)} exp\{-\frac{x^k}{\theta^k}\} I_{(0,\infty)}$$

Distribución Weibull

Las propiedades de la Weibull son :

$$E(X)=\theta T(1+\frac{1}{k})$$

$$Var(X)=\theta^2[T(1+\frac{2}{k})-(T(1+\frac{1}{k}))^2]$$

Distribución Weibull

Las partes de un Computador *Compaq* tiene una duración de vida $\alpha = 4$ y $\beta = 3$, ¿Calcule la fiabilidad que no fallen a las .70 horas?

```
pweibull(.70,4,scale = 3^(1/4),lower.tail = FALSE)
```

```
## [1] 0.9230856
```

Modelamiento Estadístico

El modelamiento estadístico sirve para:

- Identificar patrones en los datos;
- Clasificar Datos;
- Detectar multiples influencias en los datos;
- Evaluar fuerza de la evidencia de los datos

Modelamiento Estadístico

Algunas definciones necesarias:

Modelo: Representación de la realidad;
Modelo matemático: Construcción matemática de objetos
Modelo estadístico: Entrenamiento de datos para construir objetos.

Modelamiento Estadístico

Bases de datos: conjunto de matriz que se caracteriza por tener nombres de variables y valores.

Suponga que quiere averiguar el promedio del sepalo (largo) por especie

```
library(mosaic)
data("iris")# Cargar data
mean(Sepal.Length~Species,data=iris) # Esta es una forma eleg
```

```
##     setosa versicolor  virginica
##      5.006      5.936      6.588
```

Construyendo un modelo: Suponga que quiere construir un modelo que describa el largo del sepalo, de tal manera que cada variable nueva en el set de datos se pueda calcular. Para ello trabajaremos con la función `lm`

```
modelo<-lm(Sepal.Length~Species,data=iris)
modelo
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Species, data = iris)
##
## Coefficients:
##       (Intercept)  Speciesversicolor   Speciesvirginica
##             5.006              0.930              1.582
```

Modelamiento Estadístico

```r
modelo%>%summary()
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Species, data = iris)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6880 -0.3285 -0.0060  0.3120  1.3120
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.0060     0.0728  68.762  < 2e-16 ***
## Speciesversicolor   0.9300     0.1030   9.033 8.77e-16 ***
## Speciesvirginica    1.5820     0.1030  15.366  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Lo anterior quiere decir

Largo del sepalo $=$ intercepto$(5.006)+(\beta_1$ versicolor$*(0.930)+(\beta_1$ virginica

Modelamiento Estadístico

Una forma practica aunque con falta de rigor para hallar los datos que sirvan para pronosticar el largo del sepalo es con la función `step` y el criterio de Akaike

```
modelo1<-lm(Sepal.Length~.,data=iris)
step(modelo1,direction = 'both',trace = 1)
```

```
## Start:  AIC=-348.57
## Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width + S
##
##                Df Sum of Sq    RSS     AIC
## <none>                      13.556 -348.57
## - Petal.Width   1    0.4090 13.966 -346.11
## - Species       2    0.8889 14.445 -343.04
## - Sepal.Width   1    3.1250 16.681 -319.45
## - Petal.Length  1   13.7853 27.342 -245.33

##
```

Para conocer los intervalos de confianza

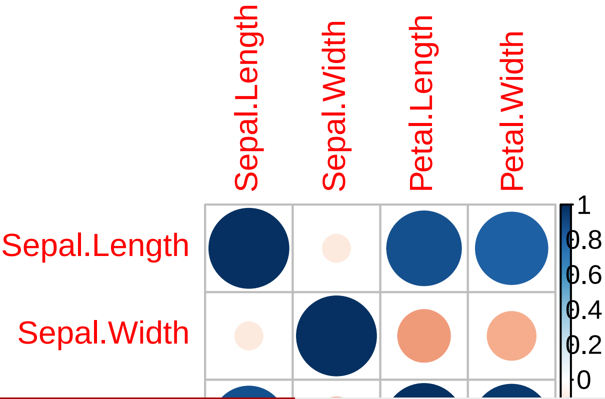
```r
confint(modelo1)
```

```
##                        2.5 %      97.5 %
## (Intercept)        1.6182321  2.72430044
## Sepal.Width        0.3257653  0.66601260
## Petal.Length       0.6937939  0.96469395
## Petal.Width       -0.6140049 -0.01630542
## Speciesversicolor -1.1982739 -0.24885002
## Speciesvirginica  -1.6831329 -0.36386273
```

Modelamiento Estadístico

Una buena practica para desarrollar modelos estadísticos es validar su nivel de correlación

```
library(corrplot)
cor(iris[,-5])%>%
  corrplot()
```

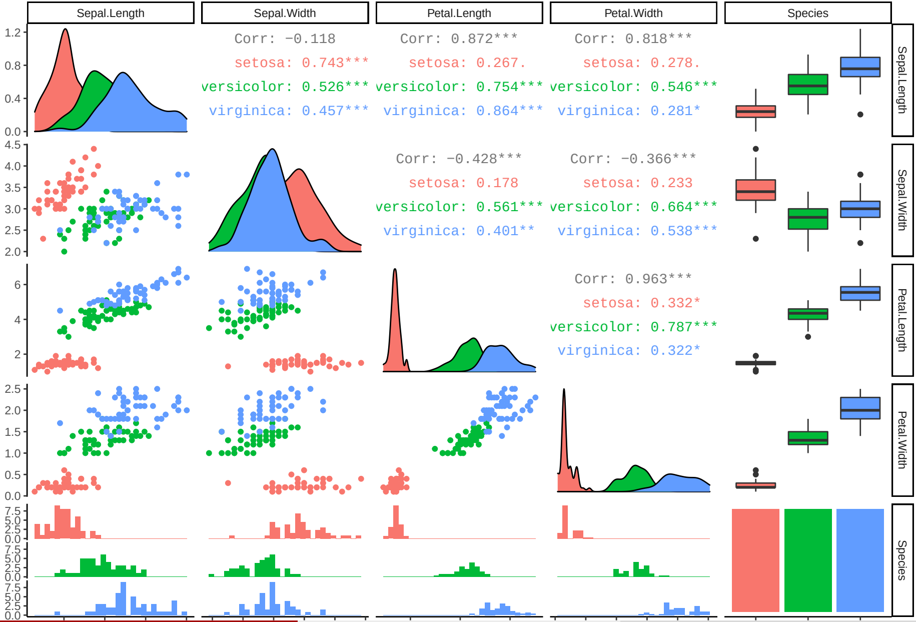


Modelamiento Estadístico

Una manera más elegante y eficiente de trabajar esto es con la función `ggpairs`

```
library(GGally)
iris%>%
  ggpairs(aes(color=Species))
```

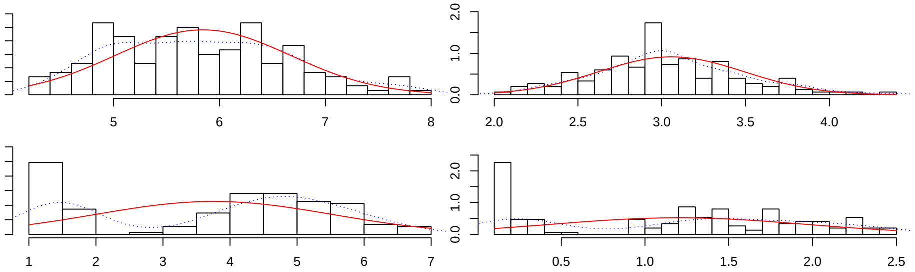
Modelamiento Estadístico



Modelamiento Estadístico

Una última forma de validar estas relaciones es

```
library(psych)
multi.hist(x = iris[,-5], dcol = c("blue", "red"), dlty = c("d
           main = "")
```



Modelamiento Estadístico

Pasos para diseñar un modelo:

- Definir el objetivo;
- Diseñar un modelo con las variables necesarias;
- Entrenar un modelo;
- Evaluar un modelo;
- Testear el modelo;
- Interpretar el modelo.

Modelamiento Estadístico

Una forma excelente de entender la arquitectura de un modelo es con rpart

```
library(rpart)
library(rpart.plot)
modelo_1<-lm(Sepal.Length~.,data=iris)
modelo_2<-rpart(Sepal.Length~.,data=iris)
```

Modelamiento Estadístico

```r
modelo_1
```

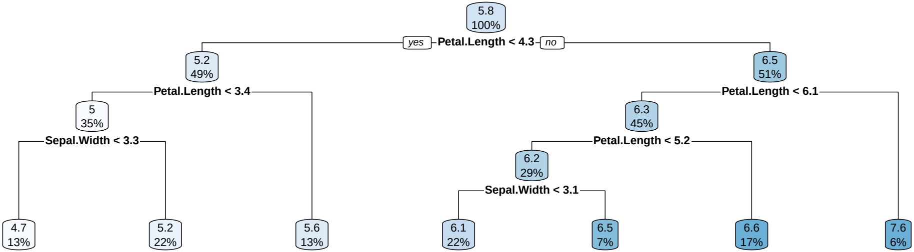
```
##
## Call:
## lm(formula = Sepal.Length ~ ., data = iris)
##
## Coefficients:
##       (Intercept)        Sepal.Width       Petal.Length
##            2.1713             0.4959             0.8292
## Speciesversicolor   Speciesvirginica
##           -0.7236            -1.0235
```

```
modelo_2
```

```
## n= 150
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 150 102.1683000 5.843333
##    2) Petal.Length< 4.25 73  13.1391800 5.179452
##      4) Petal.Length< 3.4 53   6.1083020 5.005660
##        8) Sepal.Width< 3.25 20   1.0855000 4.735000 *
##        9) Sepal.Width>=3.25 33   2.6696970 5.169697 *
##      5) Petal.Length>=3.4 20   1.1880000 5.640000 *
##    3) Petal.Length>=4.25 77  26.3527300 6.472727
##      6) Petal.Length< 6.05 68  13.4923500 6.326471
##       12) Petal.Length< 5.15 43   8.2576740 6.165116
##         24) Sepal.Width< 3.05 33   5.2218180 6.054545 *
```

Modelamiento Estadístico

```r
library(rpart.plot)
rpart.plot(modelo_2)
```



Para calcular nuevos outputs se hace de la siguiente manera

```
new_input<-data.frame("Sepal.Width"=4,"Petal.Length"=1.1,"Peta
predict(modelo_1,newdata = new_input)
```

```
##        1
## 4.940928
```

A través del modelo rpart

```r
predict(modelo_2,newdata = new_input)
```

```
##        1
## 5.169697
```

¿Cúal de los modelos es mejor?

```
output1<-iris$Sepal.Length-predict(modelo_1,newdata = new_inpu
head(output1)
```

```
## [1]  0.15907172 -0.04092828 -0.24092828 -0.34092828  0.059
```

```
output2<-iris$Sepal.Length-predict(modelo_2,newdata = new_inpu
head(output2)
```

```
## [1] -0.06969697 -0.26969697 -0.46969697 -0.56969697 -0.169
```

El que tenga menor error

Modelamiento Estadístico

La forma correcta de seleccionar un modelo es a través de un `Mean Square Error`

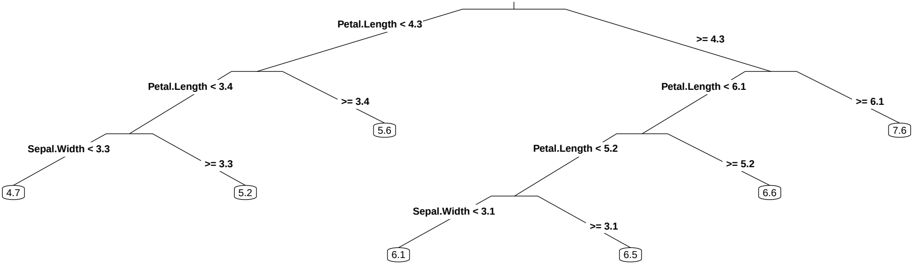
```r
print(mean(output1^2))
```

```
## [1] 1.495457
```

```r
print(mean(output2^2))
```

```
## [1] 1.134908
```

```
prp(modelo_2, type = 3,varlen = 0)
```



Modelamiento Estadístico

Si se quiere detallar mas el modelo

```
modelo_2<-rpart(Sepal.Length~.,data=iris,cp=0.0002)
```

```
prp(modelo_2,type=4,varlen = 0)
```