

6648 Bioestadística 1º Cuatrimestre

Se recomienda que los alumnos tengan un conocimiento básico de estadística y manejo del programa estadístico R. Se recomienda la realización de alguno de los siguientes cursos disponibles en internet:

[Curso de R básico](#): Curso de introducción al lenguaje estadístico R (OCW Universidad de Cádiz)

[Introduction to R](#): Curso de introducción al lenguaje estadístico R (Babraham Institute)

[A little book of R for bioinformatics](#): Guía de realización de algunas tareas bioinformáticas en R

TEMA 1. Fundamentos de análisis estadístico de datos experimentales (Juana Mari)

TEMA 2. Análisis estadístico de datos multivariantes (Juana Mari)

TEMA 3. Análisis de modelos estadísticos de comparación y de predicción (Manuel Franco)

TEMA 4. Análisis bayesiano de datos experimentales (Diego Salmerón)

Sistema de Evaluación

EVALUACIÓN CONTINUA

- Observación del trabajo del estudiante: evaluación de la actividad realizada en las horas de clase por el estudiante, así como en las tutorías (10%).
- Resolución de prácticas: evaluación de la calidad de los trabajos prácticos resueltos por el estudiante, con el fin de medir la adquisición de competencias relacionadas con la actividad (90%).
- **3 DE DICIEMBRE**
- **17 DE DICIEMBRE**

EVALUACIÓN EN CONVOCATORIA OFICIAL (100%)

- **Pruebas escritas o en ordenador:** examen escrito o en ordenador para medir las competencias adquiridas por el estudiante.
- **Presentación oral y defensa de trabajos:** evaluación de la presentación oral de los trabajos asignados, así como la respuesta a las preguntas planteadas, con el fin de medir la adquisición de competencias relacionadas con la actividad.

TEMA 1. Fundamentos de análisis estadístico de datos experimentales

1. Revisión del programa estadístico R

R es un lenguaje de programación con funciones orientadas a objetos que permite implementar técnicas estadísticas para el análisis estadístico y gráfico.

Grandes ventajas de utilizar R ...

(1) software libre en el proyecto General Public Licence (GNU) de Free Software Foundation (<http://www.gnu.org/>) en forma de código fuente, y en una amplia variedad de plataformas tales como UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS. **R versión 4.3.2 (octubre-2023)** en <https://www.r-project.org/>.

(2) entorno flexible que integra una multitud de comandos y funciones básicas en la librería denominada base, la cual constituye el núcleo de R, cuya funcionalidad se puede ampliar con librerías específicas.

(3) Enorme calidad del apoyo y soporte disponible en: <https://cran.r-project.org/manuals.html>

An Introduction to R is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics.

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

R Data Import/Export describes the import and export facilities available either in R itself or via packages which are available from CRAN.

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

R Installation and Administration

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

Writing R Extensions covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces.

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

A draft of **The R language definition** documents the language *per se*. That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions.

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

R Internals: a guide to the internal structures of R and coding standards for the core team working on R itself.

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

[HTML](#) | [PDF](#) | [EPUB](#)

The R Reference Index: contains all help files of the R standard and recommended packages in printable form. (9MB, approx. 3500 pages)

[PDF](#)

[PDF](#)

[PDF](#)

Más manuales en: <https://cran.r-project.org/other-docs.html>

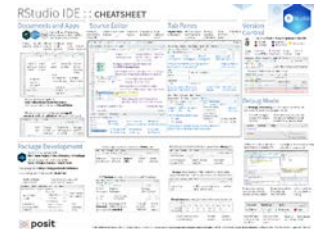
The R Journal es una revista científica de libre acceso indexada en JCR, revisada por expertos y publicada por la Fundación R: <https://journal.r-project.org/index.html>

useR! es el principal encuentro de la comunidad mundial de usuarios y desarrolladores de R, bajo [@R Foundation](#).



Cheatsheet con funciones útiles y lista de métodos abreviados de teclado implementados en RStudio.

<https://github.com/rstudio/cheatsheets/raw/master/rstudio-ide.pdf>



Cursos y libros online, <https://datos.gob.es/es/noticia/cursos-para-aprender-mas-sobre-r>

Trabajando con RStudio...

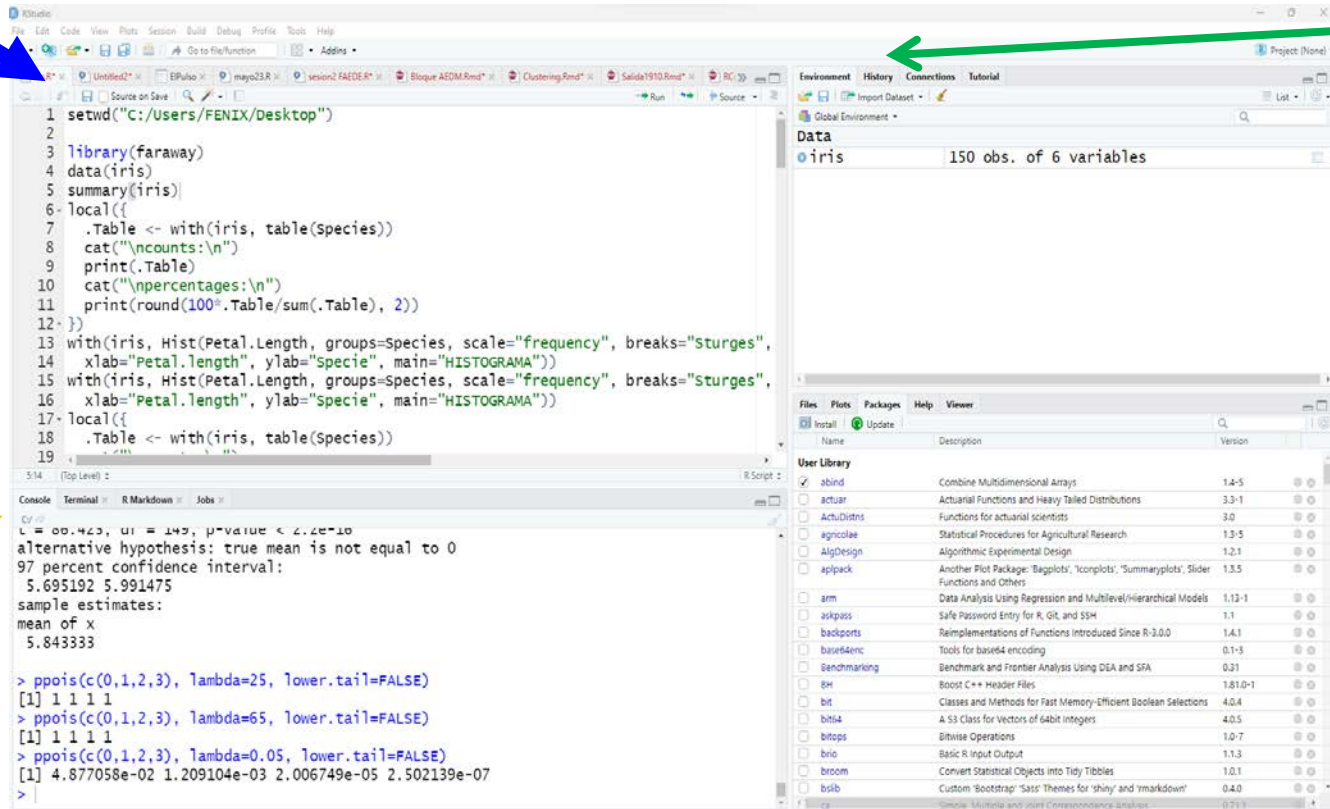
Editor de código

Crear, abrir y editar scripts*, y visualizar datos.

*Comentario precedido de celdilla #.
scripts (*.R)
datos (*.RData o *.rda).

R-consola: Console

espacio de trabajo interactivo, reconocible por el prompt ">"



Environment lista de los objetos creados en memoria

(*.RData o *.rda)

History

histórico de las líneas de código ejecutadas en R,
(*.Rhistory)

Files, árbol de directorios.

Plots, pantalla de gráficos.

Packages, administración de R-paquetes.

Help, páginas de ayuda.

Viewer, contenidos de una web local.

R-Packages

#Instalación de un paquete

`install.packages("nombrepaquete")`

`library("nombrepaquete")` #Para cargar una librería instalada

`citation("nombrepaquete")` #Para citar R o una librería

Buenas prácticas...

```
#Identificación del directorio de trabajo al iniciar R
getwd()
[1] "C:/Users/Usuario/Documents"
#Personalización del directorio de trabajo
setwd("C:/Master/Bioinf/PracticaR")
```

Calculando con R...

```
> 1+2 # Escribir 1+2 en la R-Consola y pulsar Enter
> 3*7
> 7/4
> 7%%4

> abs(-4)
> log(0)
> 2/0
> sqrt(-4)
```

...más que una calculadora científica: Estructuras de datos en R

❖ Vectores

```
> c(1,2,3,4,5) # vector numérico
> c(F,T,T,F,F) # vector lógico
> c("Juan","Pepe","Pedro","Antonio") # vector de caracteres
> 1:5
> 5:1
> v=c(1:5,10:5,12)
> v=c(1:5,10:5,12);v
> v[4]

> x<-c(1,2,3)
> y<-c(T,F,T)
> c(x,y)

> seq(from=0, to=10)
> seq(0, 10, 0.5)
> seq(from=5, by=-0.5, length.out=7)
```


❖ Factores

```
> factor(letters[1:20], labels="letter")
```

❖ Listas

```
> milista=list(hombre="Pedro",mujer="María",casados=T,n.hijos=3,  
+ edad.hijos=c(1,2,4))
```

❖ Matrices

```
> matrix(1:12)  
> matrix(1:12,nrow=3)  
> m <- matrix(1:12, nrow=3, byrow=T)  
> colnames(m) <- c("Dato 1", "Dato 2", "Dato 3","Dato 4")  
> rownames(m) <- c("Primero", "Segundo", "Tercero")
```

Identifica la utilidad de las funciones sobre matrices, ejecutando las líneas:

```
> dim(m)  
> length(m)  
> m[1,]  
> m[,1:3]  
> m[-1,]  
> rbind(m,c(1,1,1,1))  
> cbind(m,c(1,1,1))
```

❖ Data frames

```
L3 <- LETTERS[1:3]
fac <- sample(L3, 10, replace = TRUE)
d <- data.frame(x = 1, y = 1:10, fac = fac)
is.data.frame(d)
```

*Familia apply

Las funciones de esta familia son: [apply\(\)](#), [lapply\(\)](#), [sapply\(\)](#), [vapply\(\)](#), [mapply\(\)](#), [rapply\(\)](#), [tapply\(\)](#).

```
apply(d[,1:2], 2, summary)
apply(d[,c(1,2)], 2, sd)
lapply(d[, -3], skewness)
sapply(d[,1:2], kurtosis)
tapply(d$y, d$fac, summary)
```

*Funciones (objetos de R)

```
CVar<-function(x)
{
  resultado<-sqrt(var(x))/abs(mean(x))
  return(resultado)
}
```

```
tapply(d$y,d$fac,CVar)
```

Gráficos de R

Las funciones gráficas de R se clasifican en:

- funciones gráficas de alto nivel que permiten crear los gráficos básicos (plot, hist, boxplot, pairs,...).
- funciones gráficas de bajo nivel que permiten modificar los gráficos creados (points, lines, text, axis, abline...).

```
data(iris) #dataframe de base
```

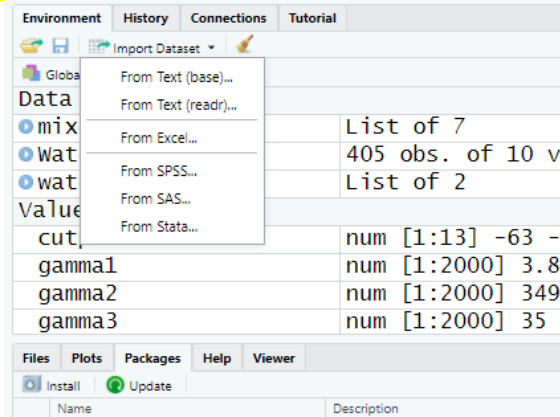
```
o <- par(mfrow=c(1,2)) #cambio de configuración
```

```
hist(iris$Sepal.Length)
```

```
boxplot(iris$Sepal.Length~iris$Species, col=c(2,3,4))
```

```
par(o)
```

Lectura e introducción de datos en R



Ficheros de datos...

- en formato texto (From Text...),
- en hojas de cálculo Excel (From Excel...),
- de paquetes estadístico como
 - a. SPSS (From SPSS...),
 - b. SAS (From SAS...) y
 - c. Stata (From Stata...)

```
library(readr) #Importar fichero ElPulso.txt (desde Environment)
```

```
ElPulso <- read_table2("ElPulso.txt")
```

```
#estructura de un dataframe
```

```
names(ElPulso)
```

```
ElPulso[,1]
```

```
ElPulso$Activity
```

```
ElPulso[, "Smokes"] # también ElPulso[["Smokes"]]
```

```
summary(ElPulso)
```

Pulse1	Pulse2	Ran	Smokes	Sex	Height	Weight	Activity
Min. : 48.00	Min. : 50	Min. :1.00	Min. :1.000	Min. :1.00	Min. :61.00	Min. : 95.0	Min. :1.000
1st Qu.: 64.00	1st Qu.: 68	1st Qu.:1.00	1st Qu.:1.000	1st Qu.:1.00	1st Qu.:66.00	1st Qu.:125.0	1st Qu.:2.000
Median : 71.00	Median : 76	Median :2.00	Median :2.000	Median :1.00	Median :69.00	Median :145.0	Median :2.000
Mean : 72.87	Mean : 80	Mean :1.62	Mean :1.696	Mean :1.38	Mean :68.72	Mean :145.2	Mean :2.132
3rd Qu.: 80.00	3rd Qu.: 85	3rd Qu.:2.00	3rd Qu.:2.000	3rd Qu.:2.00	3rd Qu.:72.00	3rd Qu.:155.5	3rd Qu.:2.000
Max. :100.00	Max. :140	Max. :2.00	Max. :2.000	Max. :2.00	Max. :75.00	Max. :215.0	Max. :3.000
						NA's :1	

#Etiquetar los codigos de las variables cualitativas

```
ElPulso$Actividad=factor(ElPulso$Activity, levels=1:3,  
+ labels=c("Suave","Moderada","Alta"))  
ElPulso$Fumar <- factor(ElPulso$Smokes, levels=1:2, labels=c("Fuma","No Fuma"))  
ElPulso$Sexo <- factor(ElPulso$Sex, levels=1:2, labels=c("Hombre","Mujer"))  
ElPulso$Correr <- factor(ElPulso$Ran,levels=c(1,2), labels=c("No","Si"))
```

#Calcular nuevas variables

```
ElPulso$Peso.kg <- with(ElPulso, round(ElPulso$Weight*0.454,1))  
ElPulso$Altura.cm <- with(ElPulso, round(ElPulso$Height*2.54,1))
```

#Transformar una variable "cuanti" en "cuali", agrupando

library("car") #cargar libreria car para usar la funcion recode

Loading required package: carData

library(carData)

summary(ElPulso\$Peso.kg)

```
ElPulso$Peso.int<-recode(ElPulso$Peso.kg, '40:60="N"; 60.1:80="S";80.1:100="M" '  
+, as.factor=TRUE)
```

#Renombrar variables

```
names(ElPulso)[c(1,2)] <- c("Pulso1","Pulso2")
```

attach(ElPulso) #carga en la memoria las variables del dataframe

dettach(ElPulso)

2. Revisión de Modelos de Probabilidad Usuales

Los modelos son herramientas de la estadística matemática muy útiles y poderosas para encontrar “estructuras” en la forma en la que se comportan los sistemas biológicos. En general:

- Los modelos infraparametrizados conducen a un pobre ajuste a los datos observados,
- El incremento de parámetros en un modelo (mediante extensiones) generalmente incrementa la flexibilidad del mismo y por tanto mejora su ajuste a los datos observados
- **Ojo!** los modelos sobreparametrizados pueden conducir a una predicción pésima de sucesos futuros

Si el modelo se ajusta a la realidad...

- Estructura interna de sistemas biológicos.
- Comparación de sistemas biológicos.
- Explicación de comportamiento

Clasificación de variables aleatorias

Variable Aleatoria	Los Posibles Resultados	Clasificación
$X = \text{Resultado de lanzar un dado}$	$\{1, 2, 3, 4, 5, 6\}$	V.A. DISCRETA
$X = \text{N}^\circ \text{ de entrecruzamientos entre dos loci durante una meiosis}$	$\{0, 1, \dots, N\}$, N n° máximo de entrecruzamientos	V.A. DISCRETA
$X = \text{N}^\circ \text{ de bases coincidentes al comparar dos strands de ADN, de longitud } N$	$\{0, 1, \dots, N\}$.	V.A. DISCRETA
$X = \text{Porcentaje de recombinación entre dos loci}$	$[0, 0.5]$	V.A. CONTINUA
$X = \text{Tiempo de degradación de una molécula celular}$	$(0, \infty)$	V.A. CONTINUA
$X = \text{Peso molecular de una molécula de ARN elegida al azar}$	$(0, M)$, M límite superior	V.A. CONTINUA
$X = \text{Log del nivel de expresión génica relativa de un gen}$	$(-\infty, +\infty)$	V.A. CONTINUA

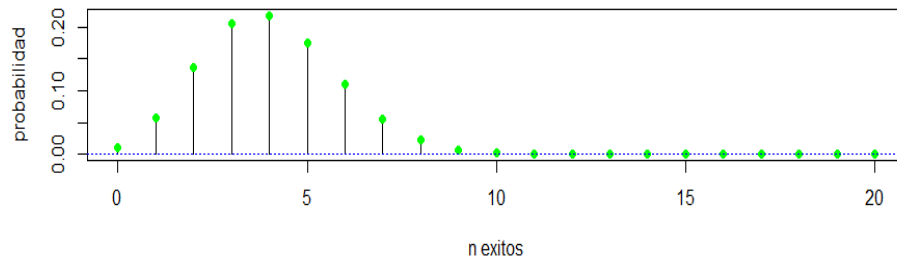
Función de probabilidad puntual

$$P(X = x_i) = p(x_i) = p_i$$

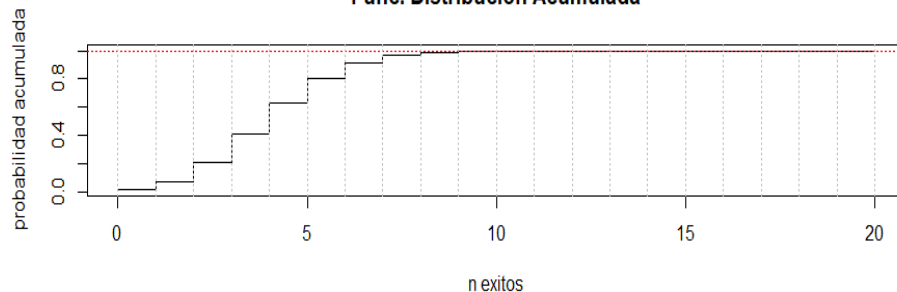
Función de distribución (acumulada)

$$F(x_i) = P(X \leq x_i) = p_1 + p_2 + \dots + p_i$$

Func. Probabilidad Puntual



Func. Distribución Acumulada



- $p_i \geq 0$ (línea azul f. probabilidad puntual)
- $\sum_x p(x) = 1$ (línea roja f. distribución)

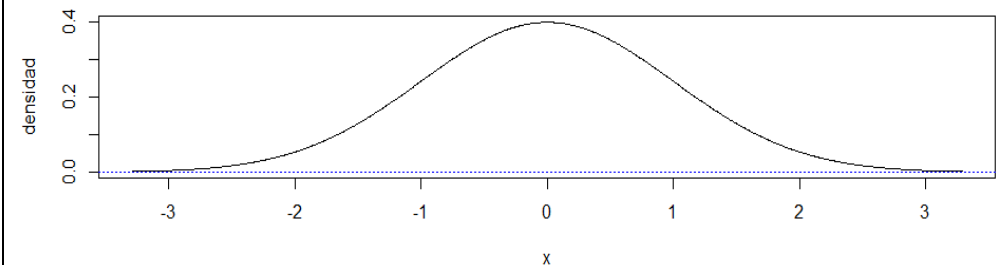
Función de densidad de probabilidad

$$f(x)$$

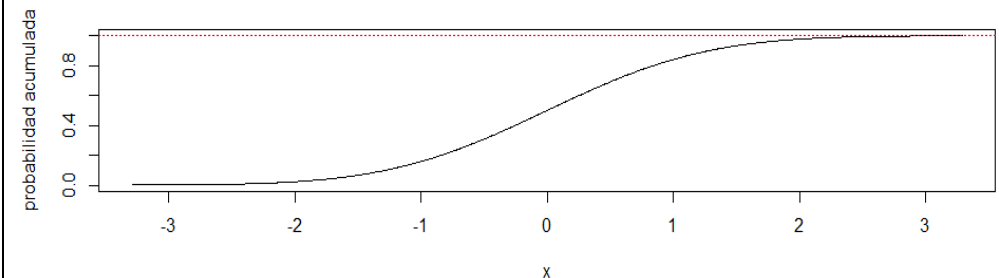
Función de distribución (acumulada)

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Func. Densidad



Func. Distribucion



- $f(x) \geq 0$ (línea azul f. densidad)
- $0 \leq F(x) \leq 1$ (línea roja f. distribución)

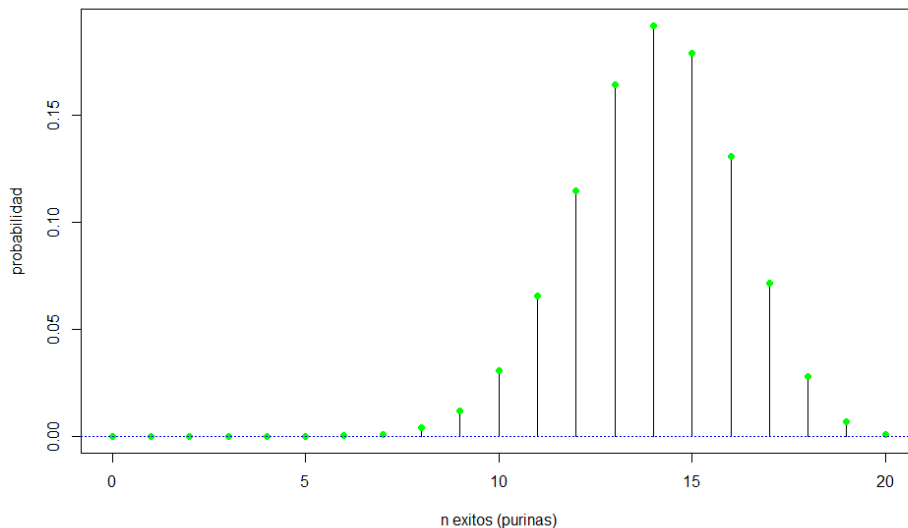
Variables Aleatorias DISCRETAS: Distribución Binomial: $X \sim \mathcal{B}(n, p)$

Función de probabilidad puntual

$$P(X = x_i) = p_i = \binom{n}{x_i} p^{x_i} (1 - p)^{n - x_i}, x = 0, 1, \dots, n$$

`dbinom(x, size=n, prob=p)`

Func. Probabilidad Puntual

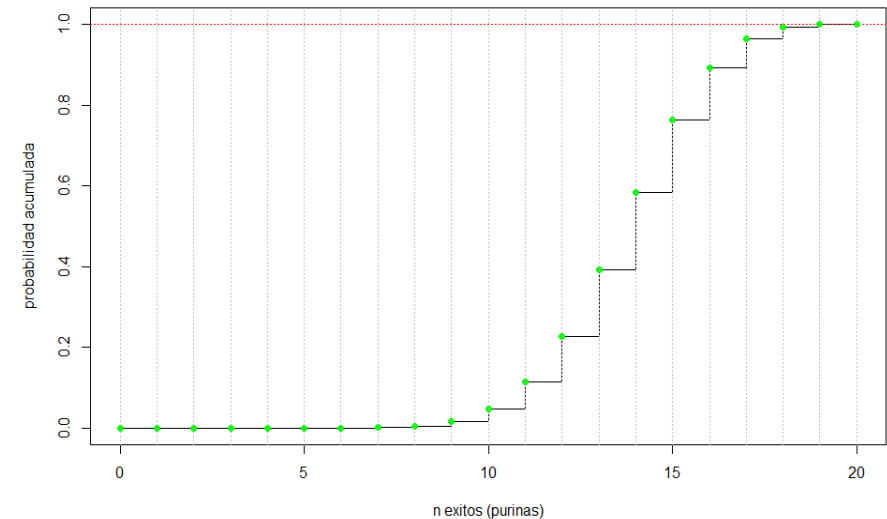


Función de distribución acumulada

$$F(x_i) = P(X \leq x_i) = p_1 + p_2 + \dots + p_i$$

`pbinom(x, size=n, prob=p)`

Func. Distribucion Acumulada



En secuencias de ARN de **tamaño 20**, la **probabilidad de purina sigue una binomial con probabilidad 0.7**.

1. Representa la función de probabilidad y la función de distribución.

2. ¿Cuál es la probabilidad de que haya 10 purinas? $P(X = 10)$

3. ¿Cuál es la probabilidad de que haya menos de 10 purinas? $P(X \leq 10)$

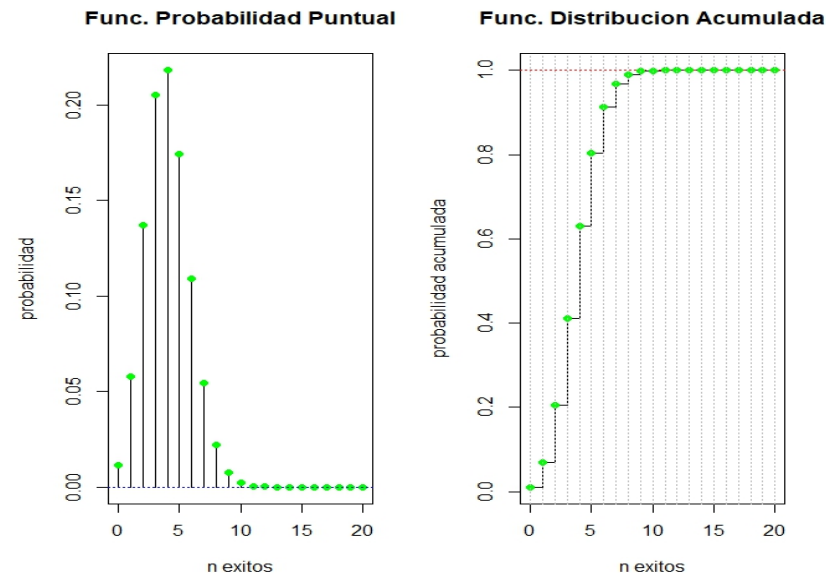
4. ¿Cuál es la probabilidad de que haya más de 10 purinas? $P(X > 10) = 1 - P(X \leq 10)$

En secuencias de ARN de **tamaño 20**, la **probabilidad de purina** sigue una **binomial con probabilidad 0.7**.

1. Representa la función de probabilidad y la función de distribución.

```
o=par(mfrow=c(1,2))
x<-0:20
plot(x,dbinom(x,size=20,prob=0.2),xlab="n exitos", ylab="probabilidad", xlim=c(0,20),main="Func. Probabilidad Puntual",type="h")
points(x,dbinom(x,size=20,prob=0.2),pch=16,col="green")
abline(h=0,col="blue",lty = 3)

plot(x,pbinom(x,size=20,prob=0.2),xlab="n exitos", ylab="probabilidad acumulada",main="Func. Distribucion Acumulada",type="s")
abline(h=1,col="red",lty = 3)
points(x,pbinom(x,size=20,prob=0.2),pch=16,col="green")
for(i in 0:20){abline(v=i,col="grey",lty=9)}
```



```
par(o)
```

2. ¿Cuál es la probabilidad de que haya 10 purinas? $P(X = 10)$

```
> dbinom(10, 20, 0.7)
```

```
[1] 0.03081708
```



¿Cómo obtener todas las probabilidades puntuales?

Sugerencia: Posibles resultados $\{0,1,2,\dots,n\}$

```
> x=0:20; dbinom(x, 20, 0.7)
```

```
[1] 3.486784e-11 1.627166e-09 3.606885e-08 5.049639e-07 5.007558e-06 3.738977e-05 2.181070e-04  
[8] 1.017833e-03 3.859282e-03 1.200665e-02 3.081708e-02 6.536957e-02 1.143967e-01 1.642620e-01  
[15] 1.916390e-01 1.788631e-01 1.304210e-01 7.160367e-02 2.784587e-02 6.839337e-03 7.979227e-04
```

3. ¿Cuál es la probabilidad de que haya menos de 10 purinas? $P(X \leq 10)$

```
> pbinom(10, 20, 0.7)
[1] 0.0479619
```



¿ $P(X \leq 10) = P(X < 10)$?
NO

```
> pbinom(9, 20, 0.7)
[1] 0.01714482
```



¿Cómo obtener todas las probabilidades acumuladas?
> x = 0: 20; pbinom(x, 20, 0.7)

```
[1] 3.486784e-11 1.662034e-09 3.773088e-08 5.426947e-07 5.550253e-06 4.294002e-05 2.610470e-04
[8] 1.278880e-03 5.138162e-03 1.714482e-02 4.796190e-02 1.133315e-01 2.277282e-01 3.919902e-01
[15] 5.836292e-01 7.624922e-01 8.929132e-01 9.645169e-01 9.923627e-01 9.992021e-01 1.000000e+00
```

4. ¿Cuál es la probabilidad de que haya más de 10 purinas? $P(X > 10) = 1 - P(X \leq 10)$

```
> pbinom(10, 20, 0.7, lower.tail=F)
```

```
[1] 0.9520381
```



¿ $P(X > 10) = P(X \geq 10)$?

NO

```
> pbinom(9, 20, 0.7, lower.tail=F)
```

```
[1] 0.9828552
```

Ejemplo. Si una persona que ha sufrido cáncer de colon tiene una probabilidad de mutación en el gen p53 de 70%, para una muestra aleatoria de 10 pacientes con este tipo de cáncer. Calcular:

1. La probabilidad de que 2 pacientes tengan el gen mutado, $P(X=2)$.

```
dbinom(2,10,0.7)
```

```
[1] 0.001446701
```

2. La probabilidad de que como mucho 2 pacientes tengan el gen mutado, $P(X \leq 2)$.

```
pbinom(2,10,0.7)
```

```
[1] 0.001590386
```

3. La probabilidad de que al menos 2 pacientes tengan el gen mutado, $P(X \geq 2)$.

```
pbinom(1,10,.7,lower.tail = F)
```

```
[1] 0.9998563
```

4. Los cuartiles de esta variable aleatoria, i.e., qué valor de X es tal que el 25%, el 50% y el 75% (respectivamente) de todos los valores están por debajo.

```
qbinom(c(.25,.5,.75),10,0.7)
```

```
[1] 6 7 8
```

5. Generar una muestra aleatoria de tamaño 50 de una distribución binomial de parámetros $n=10$ y $p=0.5$. Importante: ¡semilla de aleatorización!

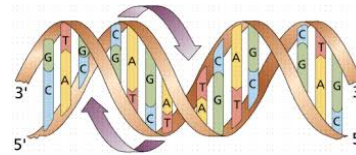
```
set.seed(12345);rbinom(50,10,0.5)
```

```
[1] 6 7 6 7 5 3 4 5 6 9 2 3 6 1 5 5 5 5 4 8 5 4 8 6 6 5 6 5 4 5 6 1 4 6 4 4 7 7 5 3 6 5 7
```

```
[44] 6 4 4 3 2 3 6
```

Extensión del modelo binomial: **Distribución multinomial**

Distribución probabilística básica para modelizar la forma en la que se distribuyen las secuencias.



- Nucleótidos

$N_{\text{ADN}} = \{A, C, G, T\}$ $N_{\text{ARN}} = \{A, C, G, U\}$ (Tamaño 4)

Codones

- $C = \{AAA, AAC, AAG, AAT, ACA, ACC, \dots, TTT\}$ (Tamaño $4^3 = 64$)

Se dice entonces que la v.a. k -dimensional $X = (X_1, \dots, X_k)$ sigue un modelo de distribución multinomial con parámetros n y $p = (p_1, \dots, p_k)$, $X \sim \mathcal{M}(n, p)$ y función de probabilidad

$$f(x; n, p) = \binom{n}{n_1, \dots, n_k} p_1^{n_1} \dots p_k^{n_k}$$

En secuencias de ADN de tamaño N se asumiría que los nucleótidos son independientes, y $X_i \sim B(N, p_i)$, $i = A, C, G, T \Rightarrow (X_A, X_C, X_G, X_T) \sim \mathcal{M}(N, p)$, donde $p = (p_A, p_C, p_G, p_T)$, con $p_A + p_C + p_G + p_T = 1$

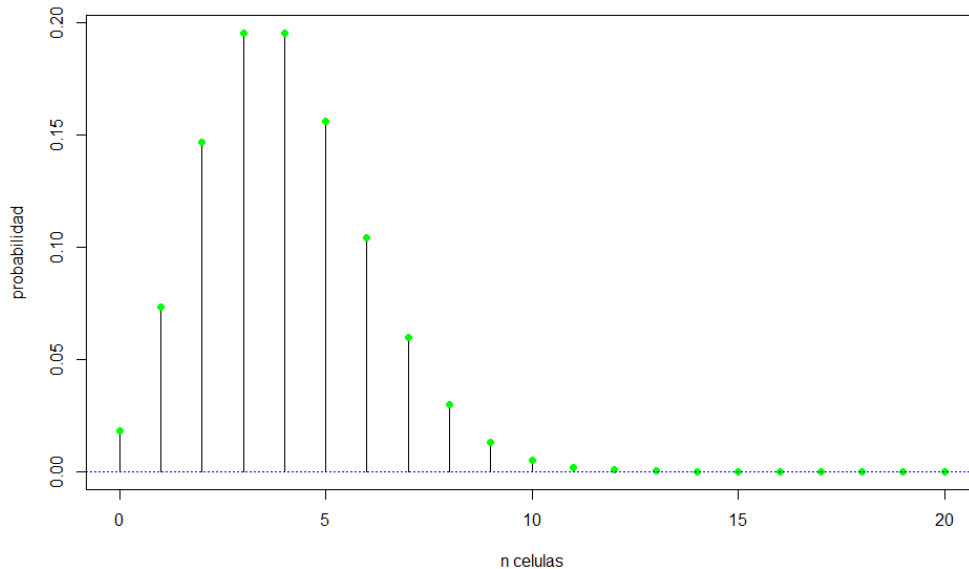
Variables Aleatorias DISCRETAS: Distribución de Poisson $X \sim \mathcal{P}(\lambda)$

Función de probabilidad puntual

$$P(X = x_i) = p_i = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2 \dots$$

dpois(x, lambda= λ)

Func. Probabilidad Puntual

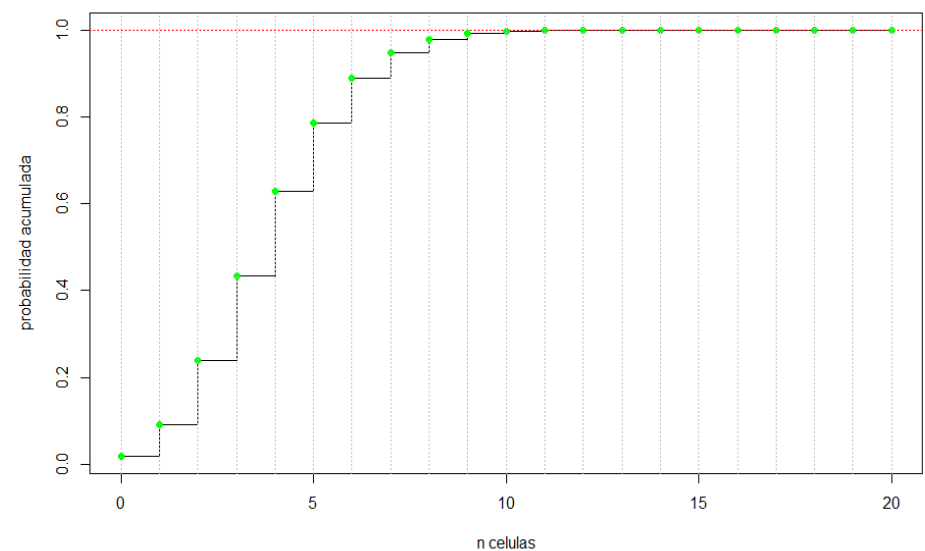


Función de distribución acumulada

$$F(x_i) = P(X \leq x_i) = p_1 + p_2 + \dots + p_i$$

ppois(x, lambda= λ)

Func. Distribucion Acumulada



Si el número medio de células en un cultivo de 20 μm^2 es 5, y se distribuyen de forma estable.

¿Cuántas células podríamos esperar en 16 μm^2 ?

Calcular la probabilidad de que no haya ninguna célula en un cultivo de 16 μm^2 , y representar las correspondientes funciones de probabilidad y de distribución.

Si el número medio de células en un cultivo de $20 \mu\text{m}^2$ es 5, y se distribuyen de forma estable.

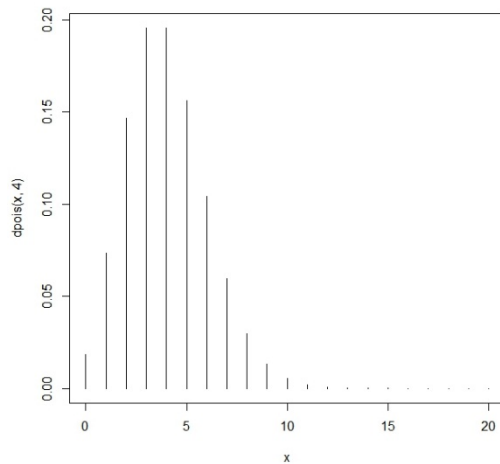
¿Cuántas células podríamos esperar en $16 \mu\text{m}^2$? $\lambda=4$

Calcular la probabilidad de que no haya ninguna célula en un cultivo de $16 \mu\text{m}^2$, $P(X=0)=\text{dpois}(0,4)$

[1] 0.01831564

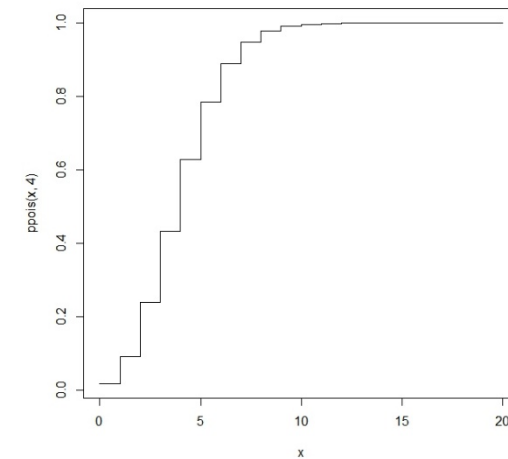
Representar las correspondientes funciones de probabilidad y de distribución.

Función de Probabilidad Puntual



```
x=0:20  
plot(x,dpois(x,4),type="h")
```

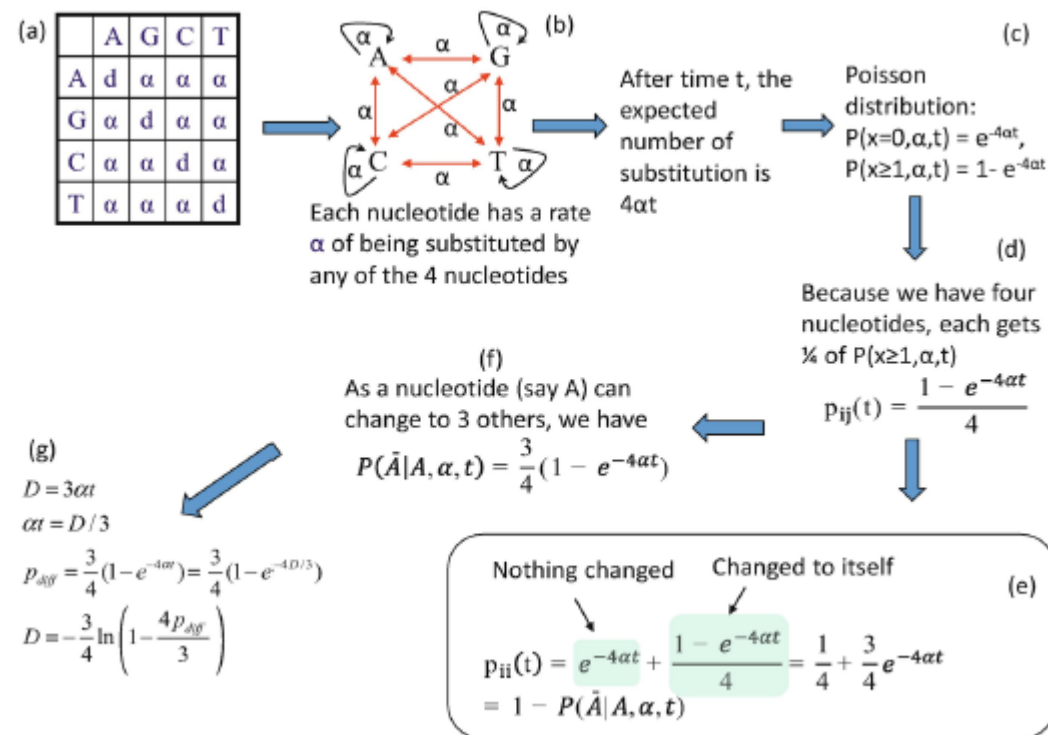
Función de Probabilidad Acumulada



```
x=0:20  
plot(x,ppois(x,4),type="s")
```

Aproximación usual... El número de entrecruzamientos por meiosis y cromosoma (o región cromosómica) se suele aproximar a una distribución de Poisson.

Modelos de sustitución



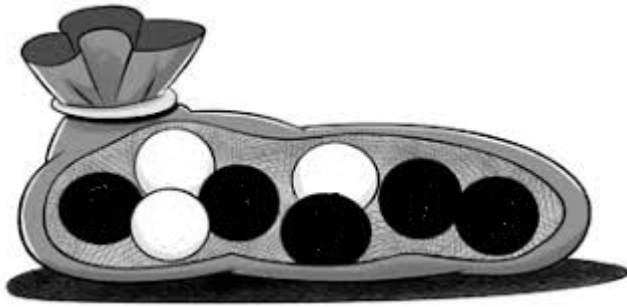
Xia, X. (2018). Nucleotide substitution models and evolutionary distances. In *Bioinformatics and the Cell* (pp. 269-314). Springer, Cham

Extracción con reemplazamiento

1ª bola blanca: $P(1^a B) = 3/8$

2ª bola blanca cuando 1ª bola blanca: $P(2^a B | 1^a B) = 3/8$

3ª bola blanca cuando 1ª y 2ª bola blanca: $P(3^a B | 1^a B \cap 2^a B) = 3/8$



Extracción sin reemplazamiento

1ª bola blanca: $P(1^a B) = 3/8$

2ª bola blanca cuando 1ª bola blanca: $P(2^a B | 1^a B) = 2/7$

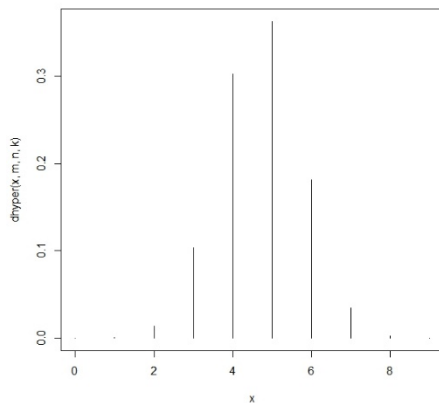
3ª bola blanca cuando 1ª y 2ª bola blanca: $P(3^a B | 1^a B \cap 2^a B) = 1/6$

Variables Aleatorias DISCRETAS: Distribución Hipergeométrica $X \sim H(m, n, k)$

Función de probabilidad puntual

$$P(X = x) = \frac{\binom{m}{x} \binom{n}{k-x}}{\binom{m+n}{k}}, x = 0, 1, \dots, \min\{m, k\}$$

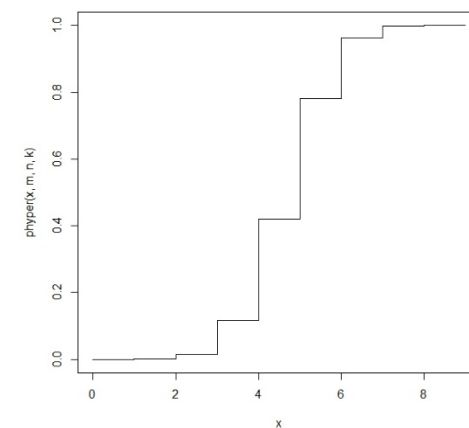
dhyper (x, m,n,k)



Función de distribución acumulada

$$F(x_i) = P(X \leq x_i) = p_1 + p_2 + \dots + p_i$$

phyper (x,m,nk)



La distribución hipergeométrica viene a cubrir esta necesidad de modelizar procesos de Bernoulli con probabilidades no constantes (sin reemplazamiento).

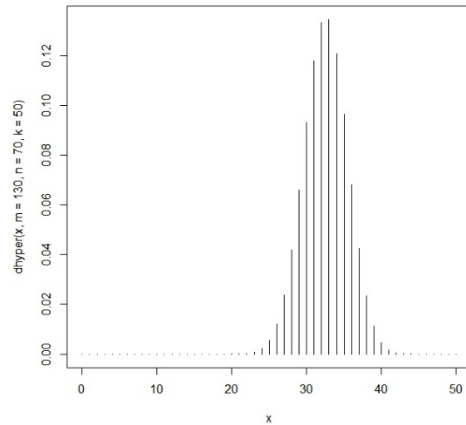
Ejemplo. El análisis de sobrerrepresentación, a través de la distribución hipergeométrica, evalúa el conjunto de genes diferencialmente expresados para los que podrían formar parte de una vía biológica, considerando cuatro atributos para llegar a una decisión

1. Número total de genes en el ensayo considerado ($m+n$)
2. Los genes expresados diferencialmente (m)
3. Genes en la vía objetivo del número total de genes (k)
4. Genes expresados diferencialmente en la vía objetivo (x)

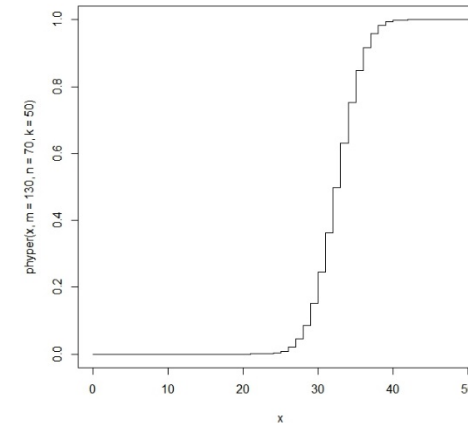
Si trabajamos con una distribución $H(m = 130, n = 70, k = 50)$, entonces

- Representar la función de probabilidad puntual y acumulada.
- Calcular la probabilidad de que exactamente 17 genes se expresen diferencialmente en la vía de objetivo.
- Calcular la probabilidad de que más de 17 genes se expresen diferencialmente en la vía de objetivo.

Función de Probabilidad Puntual



Función de Probabilidad Acumulada



```
x=0:50
plot(x,dhyper(x,m=130,n=70,k=50),type="h")
```

```
x=0:50
plot(x,phyper(x,m=130,n=70,k=50),type="s")
```

```
curve(dhyper(x,130,70,50),0,50,51,type="h")
```

```
curve(dhyper(x,130,70,50),0,50,51,type="s")
```

```
dhyper(x = 17, m = 130, n = 200 - 130, k = 50)
```

```
[1] 1.799491e-07
```

```
phyper(q = 17, m = 130, n = 200 - 130, k = 50,lower.tail=F)
```

```
[1] 0.9999998
```

```
qhyper(p = 0.75, m = 130, n = 200 - 130, k = 50)
```

```
[1] 34
```

Analysing the Protein-DNA Binding Sites in *Arabidopsis thaliana* from ChIP-seq Experiments

Almagro-Hernández, G.; Vivo, J.-M.; Franco, M.; Fernández-Breis, J.T. *Analysing the Protein-DNA Binding Sites in Arabidopsis thaliana from ChIP-seq Experiments*. *Mathematics* **2021**, *9*, 3239. <https://doi.org/10.3390/math9243239>

Our basic assumptions are that the results of a ChIP-seq experiment (set of peaks) are a random vector following a multivariate hypergeometric distribution, and we can model the genome according to the expected characteristics of this type of random experiment.

Let S be a finite population formed by m elements which are classified into k mutually exclusive classes, i.e., each element belongs to one and only one of the k classes. Let S_i

be the subpopulation of all the elements of the i -th class, being m_i its subpopulation size ($i = 1, 2, \dots, k$) and $m = \sum_{i=1}^k m_i$. Then, the random experiment consisting in drawing without replacement n elements of S is represented by the random vector $\mathbf{X} = (X_1, \dots, X_k)$, where each X_i denotes the number of elements of the S_i class in the sample. The random vector \mathbf{X} follows a multivariate hypergeometric distribution with parameters n , (m_1, \dots, m_k) and $m = \sum_{i=1}^k m_i$, $\mathbf{X} \sim MH(n, (m_1, \dots, m_k), m)$, whose joint probability mass function is given by $P_{\mathbf{X}}(x_1, \dots, x_k) = \frac{1}{\binom{m}{n}} \prod_{i=1}^k \binom{m_i}{x_i}$ where $0 \leq x_i \leq m_i$ and

$$\sum_{i=1}^k x_i = n.$$

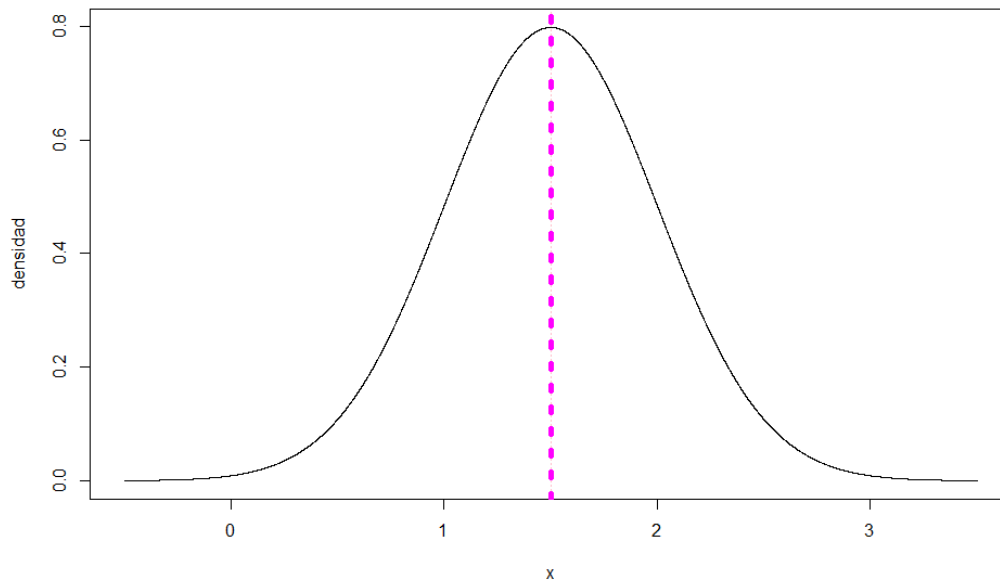
Variables Aleatorias CONTINUAS: Distribución Normal $X \sim \mathcal{N}(\mu, \sigma)$

Función de densidad de probabilidad

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

`dnormal(x, mu=μ, sigma=σ)`

Func. Densidad

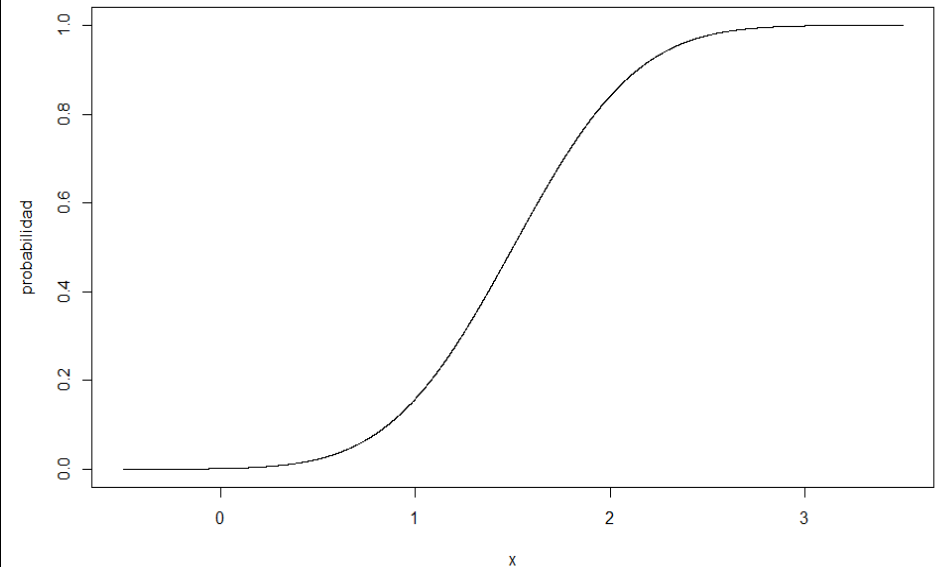


Función de distribución

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

`pnormal(x, mu=μ, sigma=σ)`

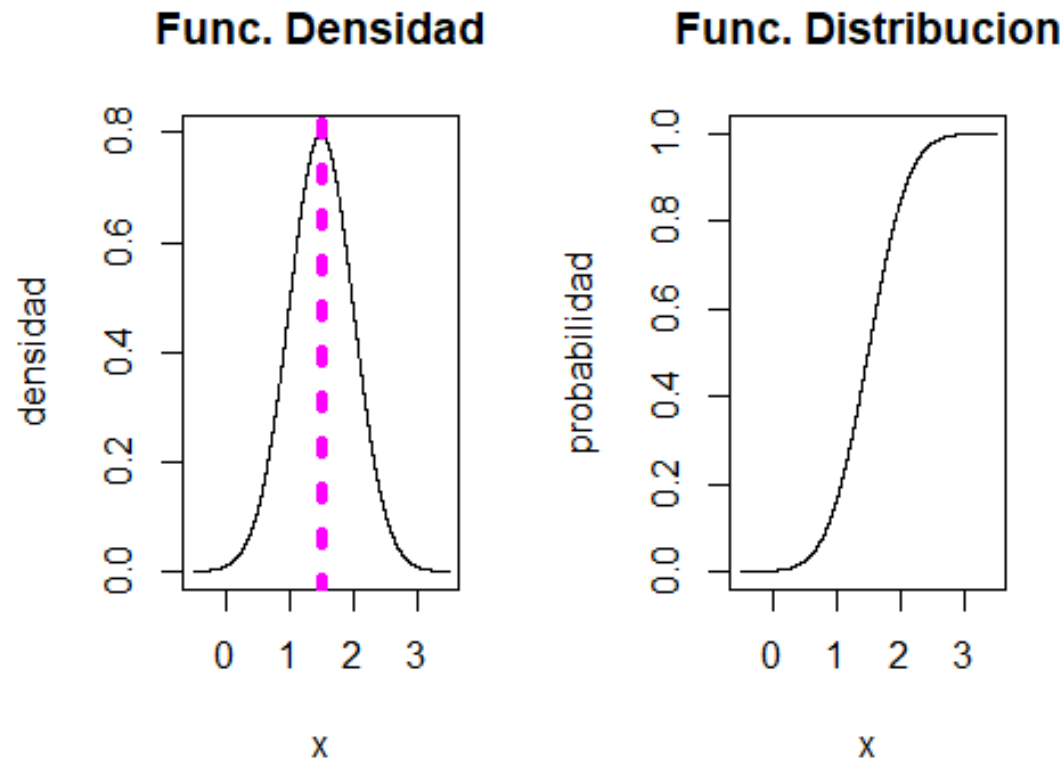
Func. Distribucion



La distribución de los valores de expresión del gen Zyxin en paciente ALL sigue una $N(\mu=1.5, \sigma=0.5)$.

1. Representar la función de densidad y la función de distribución.

```
x=seq(-0.5,3.5,length=10000)
par(mfrow=c(1,2))
plot(x,dnorm(x,mean=1.5,sd=0.5), xlab="x", ylab="densidad", main="Func. Densidad", type="l")
abline(v=1.5,col="magenta",lty = 3,lwd=5)
plot(x,pnorm(x,mean=1.5,sd=0.5), xlab="x", ylab="probabilidad",main="Func. Distribucion",type="l")
```



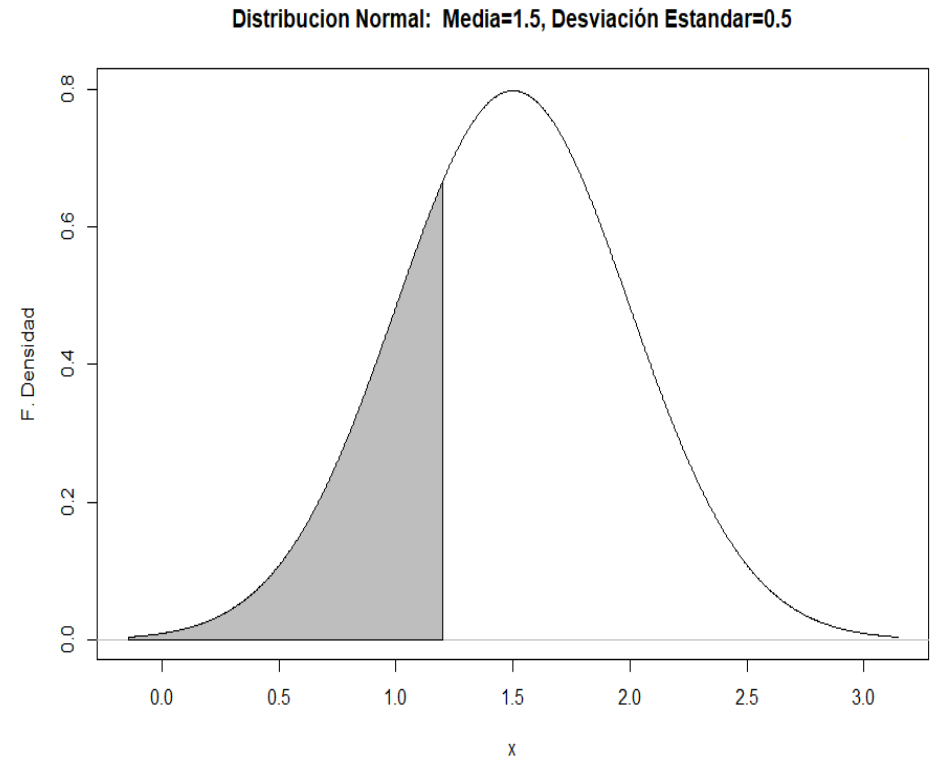
```
par(par(mfrow=c(1,2)))
```

2. ¿Cuál es la probabilidad de que los valores de expresión sean **menores que 1.2**?

$$P(X \leq 1.2) = P(X < 1.2) = \int_{-\infty}^{1.2} f(t) dt$$

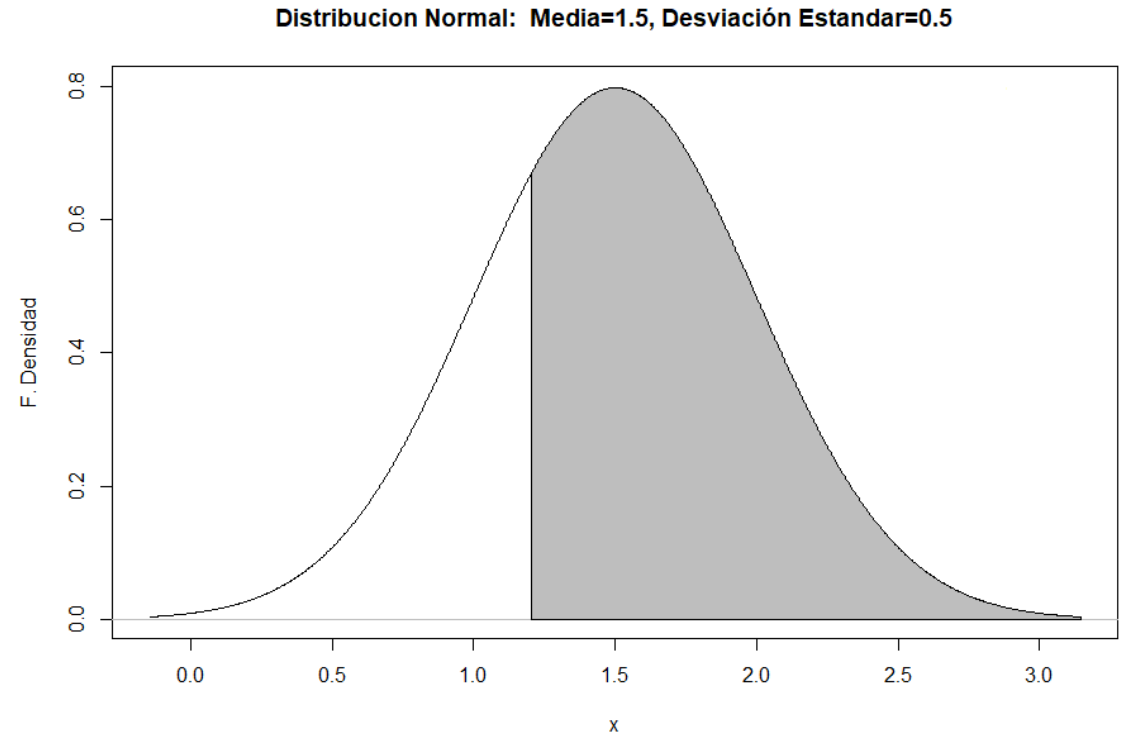
```
pnorm(1.2, mean=1.5, sd=0.5)
```

```
[1] 0.2742531
```



3. ¿Cuál es la probabilidad de que los valores de expresión sean mayores que 1.2?

$$\begin{aligned} P(X \geq 1.2) &= P(X > 1.2) = \\ &= 1 - P(X < 1.2) = \\ &= 1 - P(X \leq 1.2) = \\ &= \int_{1.2}^{\infty} f(t) dt \end{aligned}$$

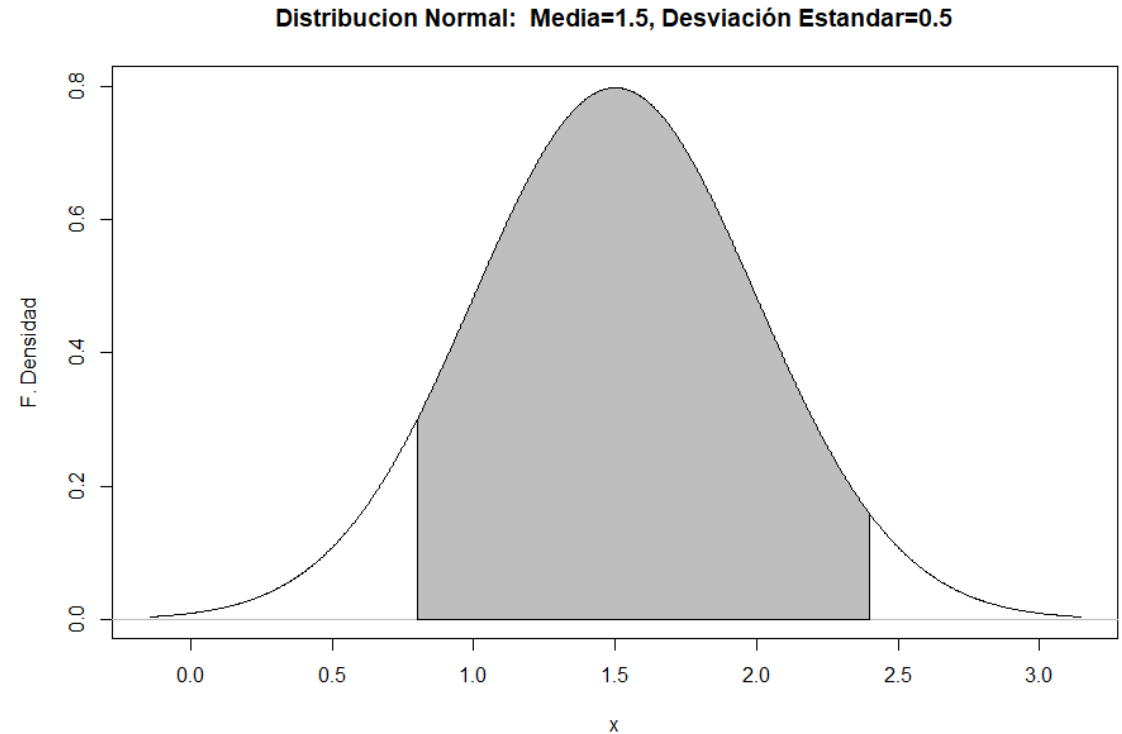


```
pnorm(1.2, mean=1.5, sd=0.5, lower.tail=F)
```

```
[1] 0.7257469
```

4. ¿Cuál es la probabilidad de que los valores de expresión estén entre 0.8 y 2.4?

$$\begin{aligned}
 P(0.8 \leq X \leq 2.4) &= P(0.8 \leq X < 2.4) \\
 &= P(0.8 < X \leq 2.4) \\
 &= P(0.8 < X < 2.4) \\
 &= \int_{0.8}^{2.4} f(t) dt = F(2.4) - F(0.8)
 \end{aligned}$$



`pnorm(2.4, mean=1.5, sd=0.5) - pnorm(0.8, 1.5, 0.5)`

`[1] 0.883313`

5. Genera una muestra aleatoria de tamaño 1000 de la población dada, esto es, que se distribuya según una normal de parámetros con $\mu = 1.5$ y $\sigma = 0.5$.

`rnorm(1000, 1.5, 0.5)`

```
[ 1] 0.93896813 0.89697629 1.44260205 1.30792540 1.36157064 1.41973011 1.64400245 2.74674565 1.32575667 1.68267054 1.55195269 1.97693198 2.34799083
[14] 1.30324616 1.36685609 1.69224908 1.54412218 1.92124311 0.86626366 2.34454134 1.30556311 1.38752615 1.18847506 1.46178929 1.10674758 1.75363223
[27] 1.26883263 1.57795289 0.45992934 1.42570347 0.53041234 0.80179968 0.28100900 1.29138994 0.90444274 1.36666684 1.37183272 0.91242815 1.35042095
[40] 1.41555271 1.83292702 2.23934298 1.90370123 1.42364027 2.09094876 2.17747840 1.81919223 1.22163409 0.22187200 1.01487789 1.25021935 2.06701549
[53] 0.86773585 0.41077302 1.17390500 1.19247215 0.79998348 1.29393451 0.78570805 1.67381812 1.57166573 1.17132848 1.15165343 1.37286831 1.62270146
[66] 1.28115039 2.49708184 1.75785046 2.24131925 0.79729023 1.60093943 1.76001688 1.62282683 1.45442012 1.08396558 2.09908610 1.93017604 0.20184165
[79] 1.10398629 2.12919469 0.90110189 0.99381559 1.93466406 1.70712444 1.13520497 0.94857011 1.55001430 1.46180261 1.43252515 1.76632904 1.93497984
[92] 1.76246359 1.06711970 1.38977309 2.15127974 1.01915790 0.89473757 0.67499195 1.31912237 1.31557282 1.00970850 0.49815514 2.47785981 0.67540452
[105] 1.68895718 1.99071534 1.63595016 1.70882408 1.44359996 2.37825064 1.07337911 1.02146857 0.29746341 1.71235498 0.43308729 1.09707986 1.35785080
[118] 1.81019063 1.74509606 1.91389528 1.63727761 2.11648822 2.23441452 1.92906742 2.09087685 1.45514367 2.02811412 1.61909346 1.97340067 1.42100060
[131] 1.19956744 2.31112844 1.75425213 1.26464264 1.37123495 1.32947901 0.60635507 1.18778483 1.58634606 1.81560484 1.62792231 1.82747893 1.88355696
[144] 1.31430404 0.61275471 1.05091564 0.88012598 1.36865562 1.26256428 1.59828195 2.09944303 2.01966800 1.20712849 1.68104760 1.06531857 1.98504994
[157] 1.14889904 2.01427563 2.19635902 2.34526951 0.93040825 1.01839042 1.56945418 1.77310417 1.38169346 2.14954704 1.55255902 1.75580872 2.63191770
[170] 1.64622267 1.53771982 1.91393593 1.24760541 2.40733307 2.35111696 1.93610580 0.97424776 1.30687794 2.36447651 2.01724794 2.06002352 1.29147870
[183] 1.95368473 1.72029822 1.01622644 1.61646794 1.10891905 0.87612410 2.14454557 1.18162997 1.12230113 2.22330502 1.41768003 1.97651009 2.30237647
[196] 1.21586435 0.78010522 1.61764953 0.84858946 1.76627173 0.87397609 1.79206325 0.85418848 2.06116296 1.18756645 1.48966325 1.79369008 1.07051236
[209] 1.30693428 1.77570490 1.31640762 1.81924493 1.33398374 1.42678096 1.23176775 0.97744925 0.93202973 1.14608907 1.63989872 1.74162470 1.62106291
[222] 1.67985613 2.11320117 1.5525743 1.75234807 2.22654932 1.23803268 1.44570889 1.30745711 1.81797537 1.07576347 0.53594344 2.24213462 1.64344633
[235] 2.08835179 -0.02154912 1.08800390 1.37340442 1.20201326 0.65894365 1.57192935 0.93174229 1.97125382 0.90285872 1.59684539 1.91883469 0.35049480
[248] 0.48937448 1.32655093 1.54518803 1.47770779 1.96214718 2.00433595 1.84752092 1.13667314 1.21724984 2.15213271 2.30032852 0.86310274 0.78744753
[261] 1.13007073 1.92934468 0.22785453 1.01975818 2.00970832 1.73415088 0.48629023 1.68608493 0.92229813 1.79668748 1.73763422 1.93621836 1.23817191
[274] 1.94148432 1.64953582 1.57934900 2.45522135 1.72762528 1.94525034 1.56223779 0.86699425 0.94200794 1.64834415 1.30559810 1.50054711 1.78247390
[287] 1.01399198 1.88537543 1.25780557 0.89780599 1.49904641 1.94764265 1.20203590 1.70325196 1.88385166 2.39792093 1.33633349 1.57540716 2.15362506
[300] 1.37415103 0.43700905 1.13050379 1.23379578 1.59438547 2.00155364 1.32437004 1.68608493 0.92229813 1.52418672 2.21450132 0.92998995 0.81455529
[313] 1.76136755 1.86213496 1.84690513 2.10284113 1.98919602 1.84548986 0.92418864 1.85612406 1.00963097 1.46776446 1.02506338 0.62613093 1.58373831
[326] 1.88839373 1.64850753 1.12917920 1.71201437 0.97498641 1.29580018 0.44447726 1.29697868 1.66307103 1.26672535 1.26084499 1.62769923 1.11094810
[339] 1.23513287 2.32242258 2.40595843 1.32913605 2.16810695 0.41495239 1.59773304 0.84012372 2.29695572 0.86310383 1.46427715 1.12775055 1.97261560
[352] 1.87445292 0.51257623 1.08040689 1.12579500 1.86953305 1.01679047 1.88698969 1.55092706 1.19746097 1.03674493 2.07605833 1.00401638 1.85113128
[365] 1.12416601 2.02188030 1.54456472 1.25161766 1.65177136 1.05503293 1.59299288 2.07880527 1.99704612 2.19407120 2.26227453 1.14802438 1.80591535
[378] 2.04309627 1.15451509 1.32529973 1.26677096 1.81078570 1.87281705 1.96281976 1.60456796 1.60972770 1.25925936 1.40071425 1.38575347 1.36285917
[391] 1.73864385 1.12971554 1.95422333 2.12383452 1.48329888 1.34521931 1.93654961 1.58759428 1.24613257 1.34922879 1.52282639 1.90145066 1.62319411
[404] 1.38238696 0.98553458 1.65054736 1.72889410 1.59977559 1.79945944 1.10555275 1.76698371 2.18919237 0.76370150 1.60233896 1.66997342 1.26283549
[417] 2.17757640 1.51084561 1.04309272 1.99590591 0.56906088 2.04605325 1.39086977 1.32952604 1.29269065 1.28215731 2.11641774 2.40699155 1.32174092
[430] 0.39873874 2.15184386 2.68388797 1.15216967 1.75493383 1.18772117 2.30525222 1.48096342 1.74456270 0.72681892 0.80539213 1.18298575 0.56402517
[443] 1.47276791 1.05475645 0.54204524 1.39519338 0.77219784 0.28788990 1.65033970 2.06642234 1.27410721 1.28592623 1.83775376 1.89247575 2.22486594
[456] 0.67558446 1.43334653 2.11232249 1.76503933 2.10635304 1.65914437 1.06215755 1.09431560 2.24171390 1.49143688 1.68542354 2.05699357 2.07565706
[469] 0.53465342 1.57850484 2.21610290 1.30267855 -0.06755553 1.26343126 1.37052020 0.40535525 1.25179446 1.63350444 1.26909998 0.29381784 1.19048542
[482] 1.42841320 1.62436694 0.97331986 1.58238109 1.83297334 1.67594161 1.46458916 1.90659617 1.78025515 1.87276893 0.87159263 1.97628291 1.58674484
[495] 1.98867800 1.88847990 2.09604497 1.47371737 1.79308812 1.34329160 0.96069863 1.63991409 0.04882101 1.99887454 1.40848988 2.21767070 0.70166915
[508] 2.71860895 1.24845204 1.99873327 2.03781452 1.86725646 1.62000936 0.87672205 1.99419646 1.24812438 2.14374659 1.55054287 1.66121206 1.11542742
[521] 1.32182173 0.89426006 0.87742234 1.43495690 1.67279577 1.06280841 1.35157731 1.53029089 1.87993744 1.34872845 0.81806238 1.49084468 1.34803658
[534] 1.23811627 0.79559103 1.03323672 1.76020148 1.56814232 1.56677042 0.87811721 1.04534745 0.98552787 1.39317888 1.96634154 0.79226858 1.96653279
[547] 2.23691014 1.39077621 1.45249967 0.98634265 1.84228217 1.30234988 1.21524199 1.88826821 0.56933499 1.95965386 1.02434718 1.74594383 1.17194482
[560] 1.08747798 1.10054744 1.08446159 1.41559265 1.60875504 1.96536226 1.57009547 0.62658356 2.25627544 1.27197130 1.80271575 2.99128175 1.53064621
[573] 2.34607687 1.52200553 1.22584531 1.03448342 1.86135317 1.43167016 1.57362199 1.67790003 1.14888538 2.04129460 1.83818918 1.69130143 1.82938652
[586] 1.90486147 1.44116866 1.78243671 2.41357463 0.58691458 1.53678248 0.91020742 1.34682411 0.75717134 1.44716602 2.23639488 1.75233513 0.89535922
[599] 1.93392512 1.24312720 1.56088431 1.06864874 0.92539827 1.24429301 1.48613437 1.57911656 1.82098016 0.42180913 2.27609834 1.07518620 1.22730608
```

[612]	1. 60889403	0. 86319270	1. 50801002	0. 93297378	0. 84984117	2. 67577306	1. 73435890	1. 50652189	1. 58867726	0. 98476803	1. 26436440	1. 10942511	0. 41250327
[625]	0. 81864453	1. 36498180	1. 51906954	0. 82267214	0. 63426174	1. 17782419	0. 51285933	2. 38960919	2. 70661688	1. 86054657	1. 43534011	1. 34499980	1. 24068887
[638]	1. 12191124	1. 22078306	1. 02258801	2. 09714432	1. 14059329	2. 07013403	1. 28081778	1. 17439988	1. 39969993	1. 80625018	1. 17900849	1. 29082068	0. 33868034
[651]	0. 95693156	1. 33062945	1. 91687443	2. 03099942	1. 95426377	2. 38971041	1. 99893076	1. 27159786	1. 05129738	1. 73791694	1. 96525215	1. 22283131	1. 22498730
[664]	1. 91134299	1. 30314468	1. 79434264	2. 12075066	1. 75135175	1. 69873094	1. 35982715	2. 01122135	1. 34100679	1. 16427806	1. 95699177	0. 46250513	2. 15299408
[677]	0. 70786489	1. 42591984	2. 18955186	1. 26698811	1. 98971973	1. 69301835	2. 19778533	1. 69410251	1. 84384361	1. 32164969	1. 44945972	0. 95829049	1. 77047919
[690]	1. 92323764	1. 04215713	1. 44818842	2. 04617569	1. 83667145	0. 95964407	1. 04865280	1. 35017884	2. 46445242	1. 04492201	0. 97619706	2. 17984237	2. 44897145
[703]	2. 52396792	2. 30454989	1. 67655603	1. 18086447	1. 85329254	0. 73067515	0. 76303585	1. 96177675	2. 27712924	0. 92145803	1. 67486387	2. 15867837	1. 57608194
[716]	1. 36539593	2. 08054977	1. 88559065	1. 11454358	2. 04490961	1. 73860019	2. 03683002	1. 45323451	1. 31609226	1. 68848321	1. 40358578	0. 47036954	1. 44315181
[729]	2. 07232416	1. 34007342	1. 33984720	0. 46395347	2. 01300924	1. 56262377	1. 74482050	2. 66470780	1. 72377300	1. 26793575	1. 31011789	0. 61530403	1. 13140092
[742]	1. 00315204	0. 62283792	2. 12460296	1. 69875755	1. 73532863	0. 89590859	1. 90264916	1. 31574472	1. 04080333	1. 23237283	2. 09422637	1. 17034405	1. 02668537
[755]	0. 86822079	1. 60567993	0. 46111969	0. 99654149	1. 00834061	1. 57111442	1. 17346677	2. 03276377	1. 47551976	1. 75880414	1. 02998138	1. 31058222	0. 55086665
[768]	1. 59760217	0. 51162872	1. 42571840	1. 35183922	1. 36468033	1. 10137771	0. 98722973	2. 02869576	2. 27957258	2. 50681432	2. 52081955	1. 44832339	0. 70642284
[781]	1. 37561246	1. 25652401	1. 76578658	1. 23632712	1. 65503157	1. 88185816	1. 16054196	2. 16573792	1. 89042748	1. 40691392	0. 58550097	1. 21967721	1. 42611303
[794]	1. 24438633	1. 88060491	1. 01741726	2. 29252726	0. 86985510	1. 27020589	1. 21117152	1. 93481973	1. 70300382	2. 00357931	1. 25748400	1. 19575643	1. 58240725
[807]	1. 25710836	1. 15879547	1. 06605282	2. 13187352	1. 83557497	1. 24662286	2. 29475253	1. 06343398	1. 51140751	1. 73606963	1. 83859403	2. 23353211	1. 22464087
[820]	0. 87864916	1. 90953196	1. 94532928	1. 96536918	1. 69391453	1. 64871219	0. 88826562	1. 90620657	1. 95635478	1. 53582191	2. 43065649	1. 68183689	2. 15318568
[833]	1. 41731806	1. 89446855	1. 44680843	2. 18589715	1. 69240763	2. 34662059	1. 72705003	1. 43921459	1. 47629319	2. 26004783	1. 40824571	1. 90463558	1. 81218143
[846]	1. 50512240	1. 78176340	2. 12258677	1. 23302055	1. 87535118	2. 41423959	0. 99904741	0. 82459206	1. 42545913	1. 27849942	1. 08402153	2. 26577598	1. 28625445
[859]	2. 25197288	1. 68760462	1. 66343156	1. 76106399	1. 81193132	2. 02451450	1. 88195258	1. 43301023	0. 48429210	1. 81303170	1. 01692120	0. 98611518	1. 34428060
[872]	1. 56250767	1. 24455404	1. 53668217	1. 42275990	2. 04788649	1. 19112345	0. 79570590	2. 04550769	1. 64862246	0. 71725097	1. 31950069	0. 73299946	1. 73636894
[885]	1. 31962452	1. 79760683	1. 10743240	0. 98899481	1. 09432444	1. 66808226	0. 85436585	1. 78752183	0. 07968653	1. 04967855	1. 83457012	2. 51018613	1. 59452628
[898]	1. 80579064	0. 76109357	0. 93440059	1. 76479550	1. 28350728	2. 12657486	1. 38691618	2. 02615412	2. 51809156	0. 76312534	0. 89564723	0. 83399761	2. 17223138
[911]	1. 38754210	0. 70006559	1. 90306331	0. 39228744	1. 84895478	1. 61908827	1. 28288806	0. 97155311	3. 00435671	1. 04978993	1. 71861960	1. 27966744	1. 87792657
[924]	1. 10234267	1. 63834111	2. 40340117	1. 47504136	1. 43762804	1. 35353932	1. 64690329	1. 59103965	1. 61219002	1. 79293860	1. 64644019	1. 93273696	1. 95095186
[937]	0. 87509486	0. 83346104	2. 96780674	1. 66531450	0. 20361560	1. 62995653	1. 75053978	0. 80954785	1. 92658309	1. 47159569	1. 48752485	2. 20549995	1. 21150952
[950]	2. 01237175	1. 54947799	0. 74101722	2. 22950152	1. 45917044	2. 43669546	1. 10356313	1. 35124509	1. 49913602	1. 56598601	2. 08617873	2. 36515115	1. 41854400
[963]	2. 01430179	1. 64347550	1. 02960235	1. 72031621	1. 23296613	2. 49557019	1. 26511663	1. 88514896	1. 38247132	1. 81191318	1. 20389346	2. 50372277	1. 24823140
[976]	0. 96351486	1. 89537741	1. 47499145	1. 55176717	0. 36202003	1. 49585425	2. 17260482	1. 47216754	1. 26732898	1. 38461598	2. 19197454	1. 55192831	1. 39041526
[989]	1. 90527214	1. 44209996	1. 71933613	1. 71685433	1. 81909120	0. 77475508	1. 36931512	1. 81643019	1. 35065010	1. 30757098	1. 79284692	1. 74580172	

<i>generador de numeros aleatorios</i>	rdistrib (n, par)
<i>función densidad/probabilidad</i>	ddistrib (x, par)
<i>función distribución</i>	pdistrib (x, par)
<i>función inversa distribución (cuantiles)</i>	qdistrib (x, par)

Distribuciones de probabilidad relacionadas con el modelo normal

Distribución *chi-cuadrado de Pearson* con n grados de libertad: χ_n^2

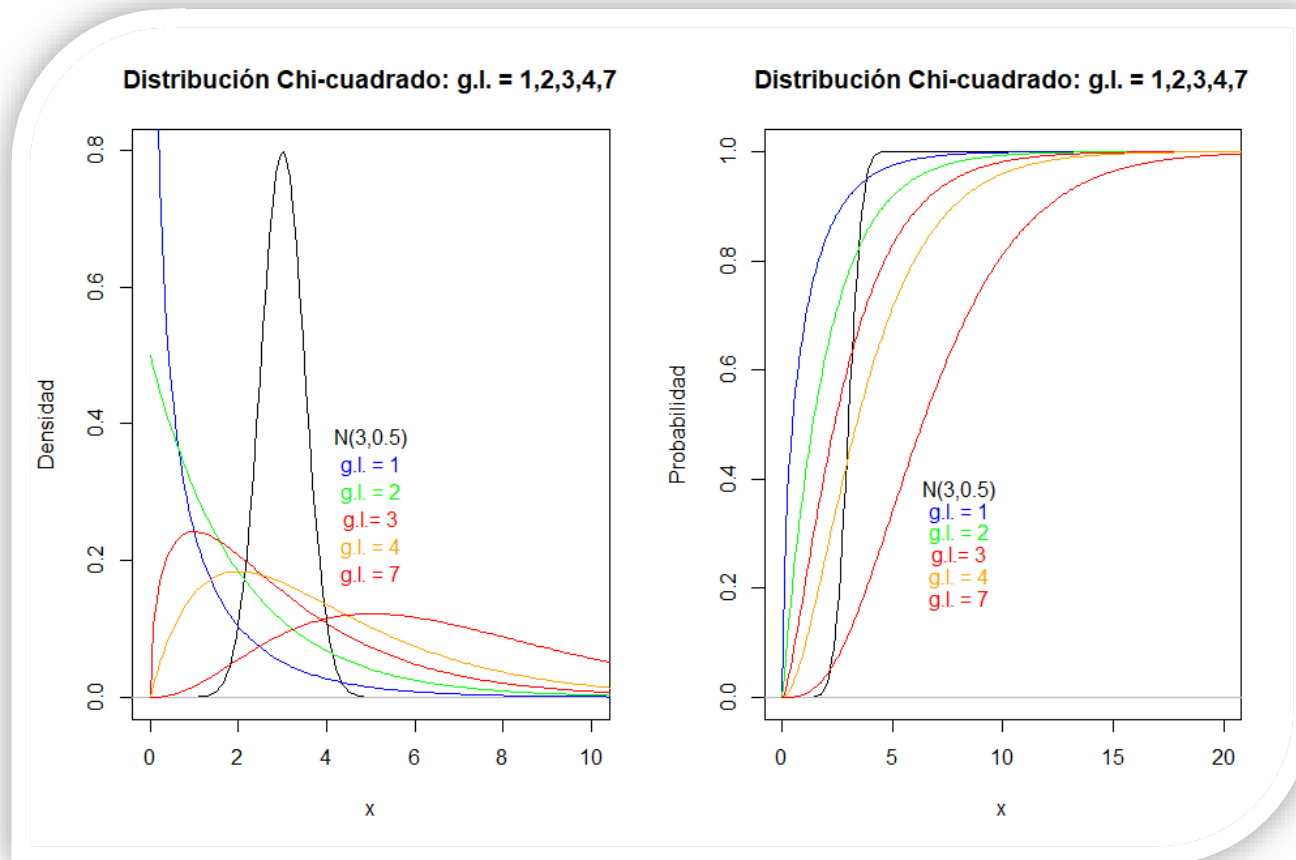
Intuitivamente, esta distribución es de utilidad para obtener información de la varianza poblacional a partir de un conjunto de datos extraídos de una variable normal.

Distribución *t de Student* con n grados de libertad: t_n

Intuitivamente, esta distribución es de utilidad para obtener información o establecer comparaciones entre las medias poblacionales a partir de uno o dos conjuntos de datos extraídos de una variable normal.

Distribución *F de Snedecor* con m y n grados de libertad: $F_{m,n}$

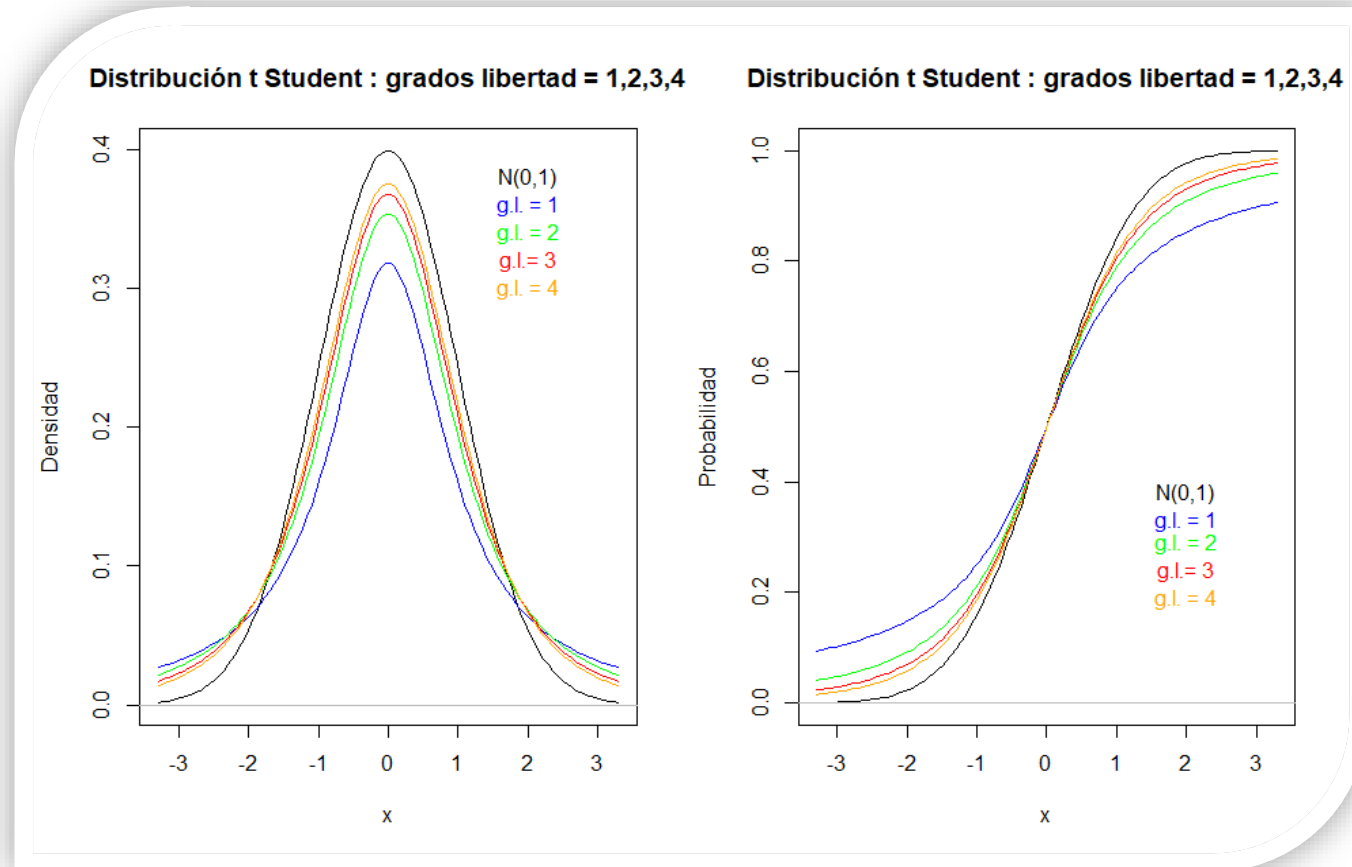
Intuitivamente, esta distribución es de utilidad para establecer comparaciones entre las varianzas poblacionales a partir de dos conjuntos de datos extraídos de una variable normal.



```
x<- seq(0, 30, length=400)
```

```
plot(x, dchisq(x, df=3), xlim=c(0, 10), xlab="x", ylab="Densidad",  
+ main="Distribución Chi-cuadrado: g.l. = 3", type="l")
```

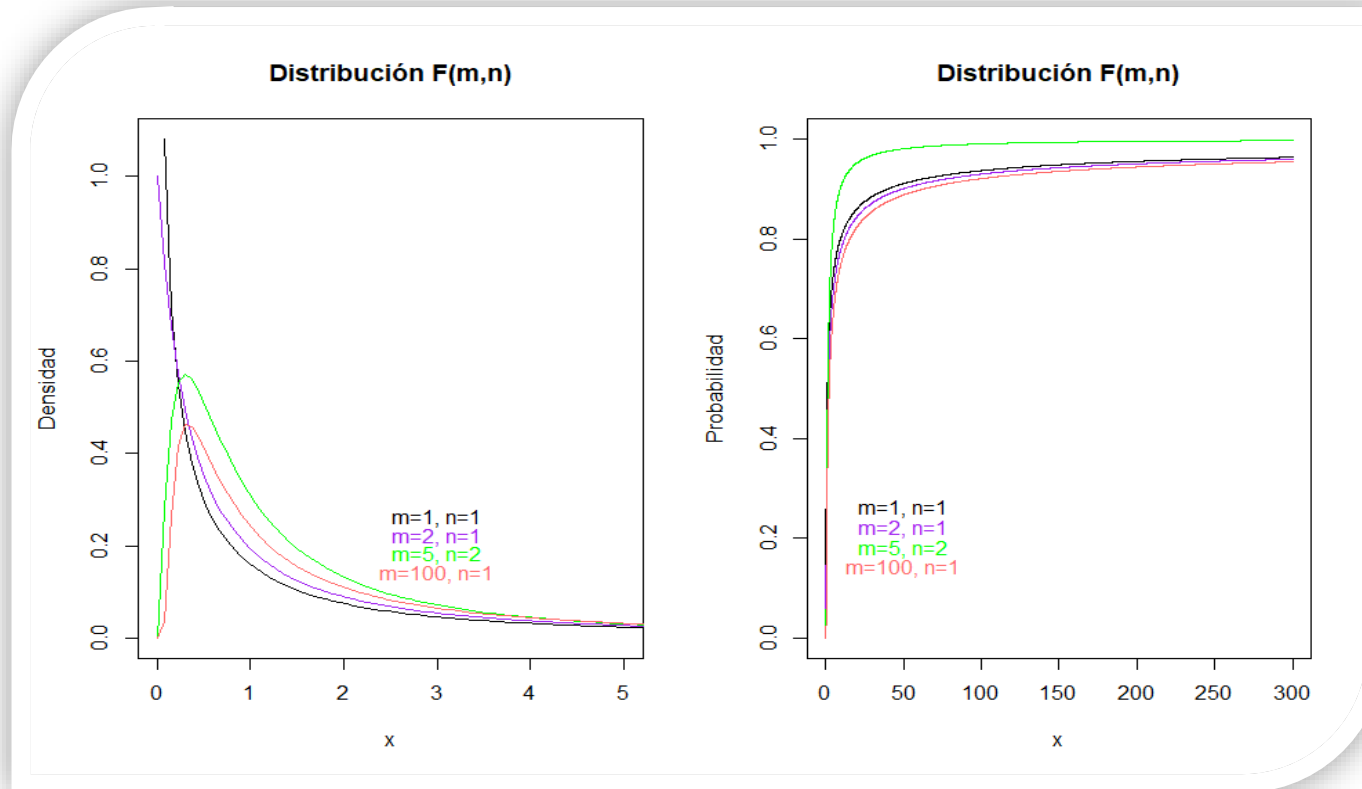
```
plot(x, pchisq(x, df=3), xlim=c(0, 30), xlab="x", ylab="Probabilidad",  
+ main="Distribución Chi-cuadrado: g.l. = 3", type="l")
```

```
x<- seq(- 3. 291, 3. 291, length=100)
```

```
plot(x, dt(x, df=3), col="red", ylab="Densidad",  
+ main="Distribución t Student: g.l. = 3", type="l")
```

```
plot(x, pt(x, df=3), col="red", ylab="Probabilidad",  
+ main="Distribución t Student: g.l. = 3", type="l")
```



```
x<- seq(0, 300, length=4000)
```

```
plot(x, df(x, df1=5, df2=2), xlim=c(0, 5), xlab="x", ylab="Densidad",  
+ main="Distribución F(5, 2)", type="l")
```

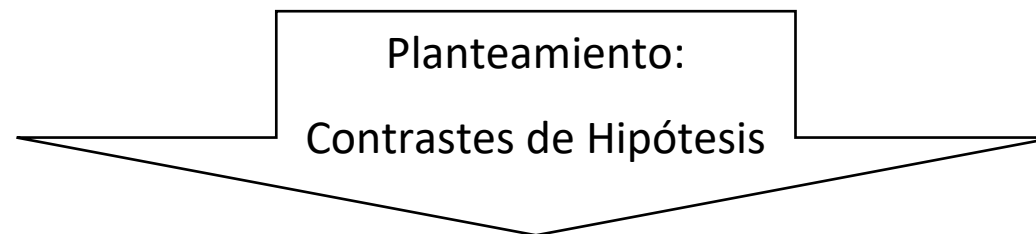
```
plot(x, pf(x, df1=5, df2=2), xlab="x", ylim=c(0, 1), ylab="Probabilidad",  
+ main="Distribución F(5, 2)", type="l")
```

3. Revisión de inferencia estadística básica paramétrica y no paramétrica

La Inferencia Estadística es el conjunto de métodos que permiten trasladar los resultados de una muestra a la población de manera fiable, midiendo la incertidumbre o acierto de los resultados, decisiones y sus conclusiones.

- Inferencia Estadística Paramétrica, si es conocida la forma funcional del modelo de distribución que sigue la variable aleatoria. por lo que sólo tenemos que estudiar los parámetros que la determinan
- Inferencia Estadística No Paramétrica, si el objetivo a estudiar es global, no se centra en sus parámetros.

La bondad de estas deducciones se mide en términos probabilísticos, es decir, toda inferencia se acompaña de su probabilidad de acierto.



Un test o contraste de hipótesis consiste en decidir sobre la veracidad de una hipótesis establecida como supuestamente cierta sobre la población.

La **hipótesis nula** es la hipótesis que se desea contrastar, y se denota por H_0

- Se rechaza o no se rechaza
- Rechazar H_0 implica aceptar la **hipótesis alternativa**, denotada como H_1
- H_1 puede aceptarse o no aceptarse

EJEMPLOS DE CONTRASTES DE HIPÓTESIS

Para Una Población

Condición:

La media de expresión del gen Gdf5 en pacientes ALL es 0

$$H_0: \mu_{ALL} = 0$$

$$H_1: \mu_{ALL} \neq 0$$

Para Dos Poblaciones



Condición:

La media de expresión de Gdf5 es igual para pacientes AML y ALL

$$H_0: \mu_{AML} = \mu_{ALL}$$

$$H_1: \mu_{AML} \neq \mu_{ALL}$$

En la decisión de si H_0 es cierta o es falsa, se puede cometer un error:

REALIDAD\DECISIÓN	NO SE RECHAZA H_0	SE RECHAZA H_0
H_0 CIERTA	DECISIÓN CORRECTA 	error de tipo I $\alpha = P(\text{ERROR I})$
H_0 FALSA	error de tipo II $\beta = P(\text{ERROR II})$	DECISIÓN CORRECTA 

α es la probabilidad de error al RECHAZAR H_0 , cuando H_0 es CIERTA.

$$\alpha = P(\text{Rechazar } H_0 | H_0 \text{ CIERTA})$$

β es la probabilidad de error al NO RECHAZAR H_0 , cuando H_0 es FALSA.

$$\beta = P(\text{No Rechazar } H_0 | H_0 \text{ FALSA})$$

Las hipótesis estadísticas se pueden contrastar a partir de la información extraída de la muestra mediante un estadístico (estadístico del contraste), a partir de la probabilidad asociada a una región de rechazo limitada por el valor observado de estadístico asociado el contraste que se conoce como el p-valor, dado por

$$p - \text{valor} = (\text{Rechazar } H_0 \text{ con nuestra muestra particular} | H_0 \text{ CIERTA})$$

Si se ha fijado de antemano el nivel de significación α ,

1. $p - \text{valor} \geq \alpha, \Rightarrow \underline{\text{NO SE RECHAZA } H_0}$

2. $p - \text{valor} < \alpha, \Rightarrow \underline{\text{SE RECHAZA } H_0}$

Ejemplo. Si $\alpha = 0.05$ entonces,

1. $p - \text{valor} \geq 0.05, \Rightarrow \underline{\text{NO SE RECHAZA } H_0}$

2. $p - \text{valor} < 0.05, \Rightarrow \underline{\text{SE RECHAZA } H_0}$

TIPOS DE CONTRASTES DE HIPÓTESIS

CONTRASTE DE HIPÓTESIS PARAMÉTRICOS

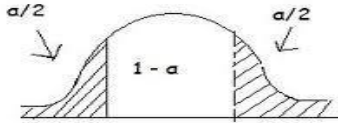
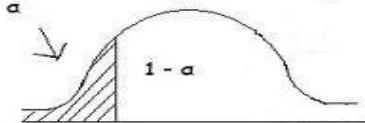
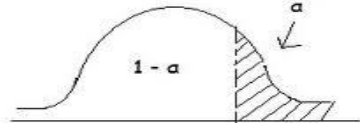
$$H_0: \mu_{ALL} = 0$$

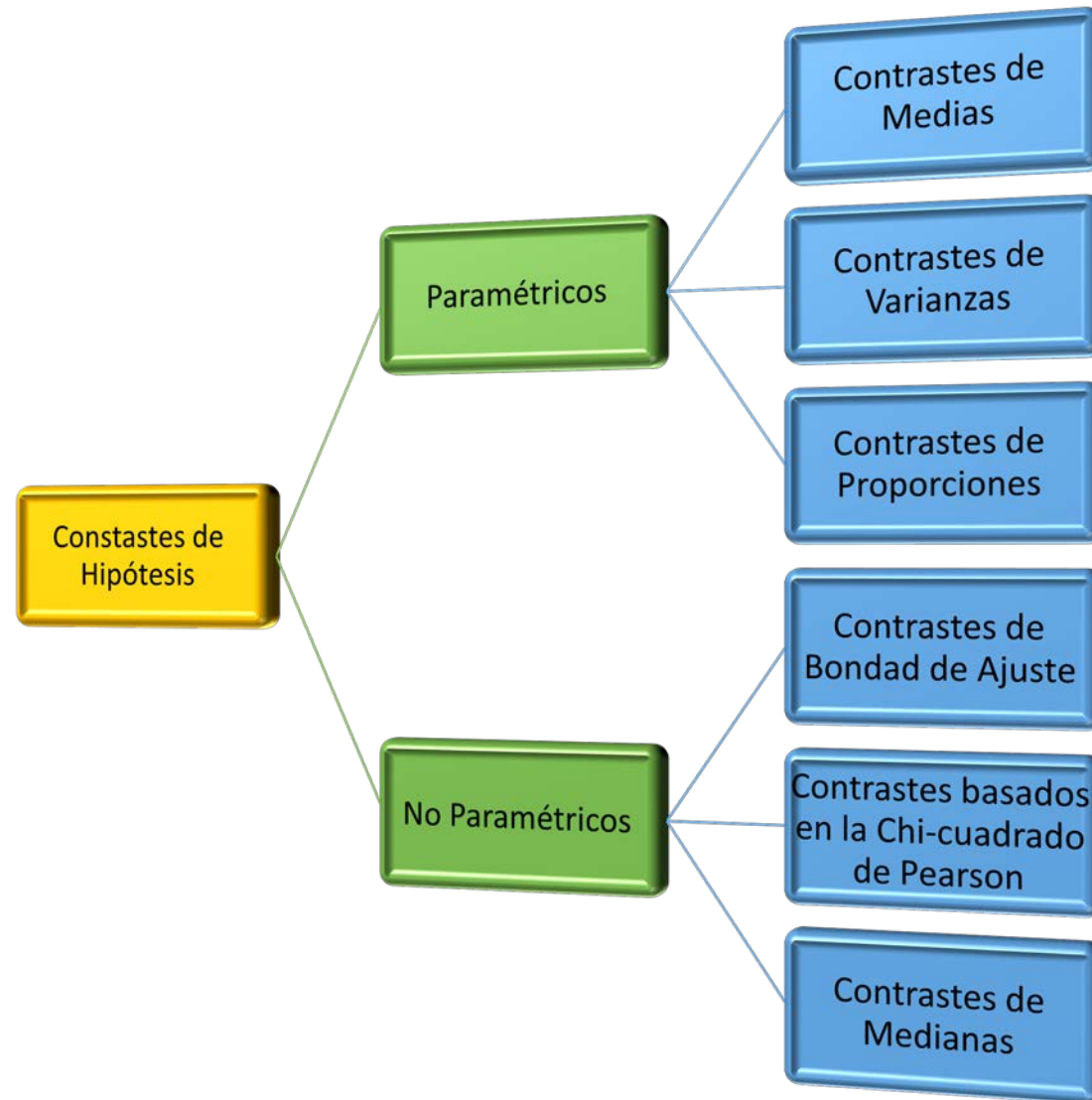
$$H_1: \mu_{ALL} \neq 0$$

CONTRASTE DE HIPÓTESIS NO PARAMÉTRICOS

$$H_0: X_{ALL} \sim N$$

$$H_1: X_{ALL} \not\sim N$$

Contraste de hipótesis para el parámetro θ		
Bilateral ("two.sided")	Unilateral Izquierda ("less")	Unilateral Derecha ("greater")
$H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$	$H_0: \theta \geq \theta_0$ $H_1: \theta < \theta_0$	$H_0: \theta \leq \theta_0$ $H_1: \theta > \theta_0$
Test de conformidad de la media 	Test de conformidad de la media 	Test de conformidad de la media 



Dataset: *leukemia* de Golub et al. (1999) en *spikeslab*

Analizaremos los niveles de expresión de 3571 genes humanos en 72 pacientes con leucemia de los cuales 47 sufrían leucemia linfoblástica aguda (ALL) y los 25 restantes leucemia mieloide aguda (AML). Golub y sus colegas midieron el nivel de expresión génica con el objetivo de detectar un subconjunto de genes que pudiesen ser usado como tests diagnósticos, permitiendo evaluar si un nuevo paciente padece alguno de estos tipos de cáncer.

- Etiquetar los códigos de la variable Y , 0 como ALL y 1 como AML
- Grabar la nueva variable factor en el fichero.

```
library(spikeslab)
```

```
data(leukemia)
```

```
table(leukemia$Y)
```

```
leukemia$factor<- factor(leukemia$Y, levels=0:1, labels=c("ALL", "AML"))
```

```
table(leukemia$factor)
```

Test t de Student para una muestra: Prueba de conformidad de la media

<i>Contraste de Hipótesis</i>	<i>Estadístico del contraste</i>
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \equiv_{H_0} t_{n-1}$

Contrastar si la media de expresión génica del gen Gdf5 de ALL es 0, i.e.,

$H_0: \mu_{ALL} = 0$
$H_1: \mu_{ALL} \neq 0$

`t.test(leukemi a$x. 3395[leukemi a$Y==0])`

One Sample t-test

```
data:  leukemi a$x. 3395[leukemi a$Y == 0]
t = 0.0011339, df = 46, p-value = 0.9991
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1344803  0.1346319
sample estimates:
mean of x
0.00007579787
```



```
t.test(leukemia$x.3395[leukemia$factor=="ALL"])
```

Ejemplo. Asumiendo que los valores de expresión de CCND3 Cyclin D3 para pacientes ALL, contrastaremos si es $\mu_{ALL} \leq 0$

```
t.test(leukemia$x.1040[leukemia$factor=="ALL"], alternative="greater")
```

One Sample t-test

```
data: leukemia$x.1040[leukemia$factor == "ALL"]
t = 41.733, df = 46, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 1.815445      Inf
sample estimates:
mean of x
 1.891529
```

Ej 3.9 Contrastar si el pulso medio antes de la actividad realizada para hombres fumadores es 75.

Test t de Student para dos muestras independientes

Golub et al. (1999) identifica genes expresión diferenciales entre pacientes AML y ALL, siendo uno de ellos el gen CCND3 Cyclin D3. Compruébalo.

$$H_0: \mu_{ALL} = \mu_{AML}$$

$$H_1: \mu_{ALL} \neq \mu_{AML}$$

`t.test(leukemia$x.1040~ leukemia$Y, var.equal=T)`

Two Sample t-test

```
data: leukemia$x.1040 by leukemia$Y
t = 3.118, df = 70, p-value = 0.002642
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0880387 0.4005939
sample estimates:
mean in group 0 mean in group 1
 1.891529      1.647213
```

`t.test(leukemia$x.1040~ leukemia$Y, var.equal=F)`

Welch Two Sample t-test

```
data: leukemia$x.1040 by leukemia$Y
t = 3.068, df = 46.893, p-value = 0.003575
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.08410265 0.40452991
sample estimates:
mean in group 0 mean in group 1
 1.891529      1.647213
```

Contrastes de varianzas: Test F de Snedecor para dos muestras independientes

$$\begin{array}{l} H_0: \sigma_{ALL}^2 = \sigma_{AML}^2 \\ H_1: \sigma_{ALL}^2 \neq \sigma_{AML}^2 \end{array}$$

`var.test(leukemi a$x. 1040~leukemi a$Y)`

F test to compare two variances

data: leukemi a\$x. 1040 by leukemi a\$Y

F = 0.90082, num df = 46, denom df = 24, p-value = 0.7413

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.4248098 1.7633269

sample estimates:

ratio of variances

0.9008223

$$\begin{array}{l} H_0: \sigma_{ALL}^2 = \sigma_{AML}^2 \Leftrightarrow H_0: \sigma_{ALL}^2 / \sigma_{AML}^2 = 1 \\ H_1: \sigma_{ALL}^2 \neq \sigma_{AML}^2 \Leftrightarrow H_1: \sigma_{ALL}^2 / \sigma_{AML}^2 \neq 1 \end{array}$$



```
t.test(leukemia$x.1040~ leukemia$Y, var.equal=T)
```

Two Sample t-test

```
data: leukemia$x.1040 by leukemia$Y
```

```
t = 3.118, df = 70, p-value = 0.002642
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
0.0880387 0.4005939
```

```
sample estimates:
```

```
mean in group 0 mean in group 1
```

```
1.891529
```

```
1.647213
```

$$\begin{array}{l} H_0: \mu_{ALL} = \mu_{AML} \Leftrightarrow H_0: \mu_{ALL} - \mu_{AML} = 0 \\ H_1: \mu_{ALL} \neq \mu_{AML} \Leftrightarrow H_1: \mu_{ALL} - \mu_{AML} \neq 0 \end{array}$$

Ejemplo 3.2 Análogamente, para el gen Gdf5, situado en la columna 3396, la hipótesis nula de igualdad de varianzas y la hipótesis nula de igualdad de medias entre los pacientes ALL y AML puede ser testadas.

Test t de Student para dos muestras dependientes (o apareadas)

$$\begin{array}{l} H_0: \mu_{Pulso1} = \mu_{Pulso2} \Leftrightarrow H_0: \mu_{Pulso1} - \mu_{Pulso2} = 0 \\ H_1: \mu_{Pulso1} \neq \mu_{Pulso2} \Leftrightarrow H_1: \mu_{Pulso1} - \mu_{Pulso2} \neq 0 \end{array}$$

```
t.test( El Pul so$Pul so1, El Pul so$Pul so2, paired=T, conf.level=0.95)
```

Paired t-test

data: El Pul so\$Pul so1 and El Pul so\$Pul so2

t = -5.0769, df = 91, p-value = 2.023e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-9.920273 -4.340597

sample estimates:

mean of the differences

-7.130435

```
t.test( El Pul so$Pul so1, El Pul so$Pul so2, pai red=T, conf.level=0.95)
```

Paired t-test

data: El Pul so\$Pul so1 and El Pul so\$Pul so2

t = -5.0769, df = 91, p-value = 2.023e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-9.920273 -4.340597

sample estimates:

mean of the differences

-7.130435



```
t.test( El Pul so$Pul so1, El Pul so$Pul so2, pai red=T, conf.level=0.99)
```

Paired t-test

data: El Pul so\$Pul so1 and El Pul so\$Pul so2

t = -5.0769, df = 91, p-value = 2.023e-06

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

-10.825552 -3.435317

sample estimates:

mean of the differences

-7.130435

Ejemplo. Contrastar si existen diferencias de medias entre pulso antes y pulso después en fumadores.

Contraste para una proporción: Prueba de conformidad

<i>Contraste de Hipótesis</i>	<i>Estadístico del contraste</i>
$H_0 : p = p_0$ $H_1 : p \neq p_0$	$T = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \equiv_{H_0} \mathcal{N}(0, 1)$

`prop.test(x, n, p=NULL, alternative=c("two.sided", "less", "greater"), conf.level=0.95, correct=T)`
`binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)`

Ejemplo 3.13 Supongamos que la secuenciación revela que un cierto microARN tiene 18 purinas de un total de 22. Suponiendo que la distribución binomial se mantiene, contrastar:

$H_0: p_0 = 0.7$ vs. $H_1: p_0 \neq 0.7$

`prop.test(18, 22, p=0.7)`

1-sample proportions test with continuity correction

```
data: 18 out of 22, null probability 0.7
X-squared = 0.95455, df = 1, p-value = 0.3286
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
 0.5899288 0.9400827
sample estimates:
              p
0.8181818
```

`binom.test(18, n=22, p = 0.7)`

Exact binomial test

```
data: 18 and 22
number of successes = 18, number of trials = 22, p-value = 0.351
alternative hypothesis: true probability of success is not equal to 0.7
95 percent confidence interval:
 0.5971542 0.9481327
sample estimates:
probability of success
      0.8181818
```

Contraste para dos proporciones

<i>Contraste de Hipótesis</i>	<i>Estadístico del contraste</i>
$H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$	$T = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \equiv_{H_0} \mathcal{N}(0, 1)$

Si al testar dos nuevos fármacos, fármaco 1 y fármaco 2, en 150 y 125 unidades experimentales, respectivamente, se tiene que 14 unidades no han reaccionado bien con el fármaco 1 y 15 unidades no han reaccionado bien con el fármaco 2. ¿Hay una evidencia estadística que nos permita asegurar que el porcentaje de reacciones adversas con ambos fármacos es distinto?

prop.test(prop, n, alt="two.sided")

2-sample test for equality of proportions with continuity correction

```
data:  prop out of n
X-squared = 0.27016, df = 1, p-value = 0.6032
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.10756913  0.05423579
sample estimates:
   prop 1   prop 2 
0.09333333 0.12000000
```

Datos extraídos de Almoguera (2011) accesible en <https://repositorio.uam.es/handle/10486/6784> sobre el estudio del alelo ATA como factor de susceptibilidad a la esquizofrenia. Se agruparon los genotipos según el número de alelos ATA (2, 1 ó 0) y las distribuciones del recuento de mujeres esquizofrénicas (casos) y mujeres de la muestra objeto de estudio (muestra), portadoras de 2, 1 o ningún alelo ATA, respectivamente, son:
muestra = c(20, 197, 156) y casos = c(13, 27, 48).

```
muestra=c(20, 197, 156)
casos=c(13, 27, 48)
prop.test(casos, muestra)
```

3-sample test for equality of proportions without continuity correction

```
data:  casos out of muestra
X-squared = 34.163, df = 2, p-value = 3.816e-08
alternative hypothesis: two.sided
sample estimates:
  prop 1    prop 2    prop 3
0.6500000 0.1370558 0.3076923
```



Warning message:

In prop.test(casos, muestra) : Chi-squared approximation may be incorrect

1º Opción: Colapsando de clases

```
muestra2=c(20+197, 156); muestra2  
[1] 217 156
```

```
casos2=c(13+27, 48); casos2  
[1] 40 48
```

```
prop.test(casos2, muestra2)
```

2-sample test for equality of proportions with continuity correction

```
data:  casos2 out of muestra2  
X-squared = 6.9925, df = 1, p-value = 0.008185  
alternative hypothesis: two.sided  
95 percent confidence interval:  
-0.21779195 -0.02892907  
sample estimates:  
prop 1      prop 2  
0.1843318 0.3076923
```

2º Opción: Test exacto de Fisher (Prueba hipergeométrica)

`fisher.test(m)`

Fisher's Exact Test for Count Data

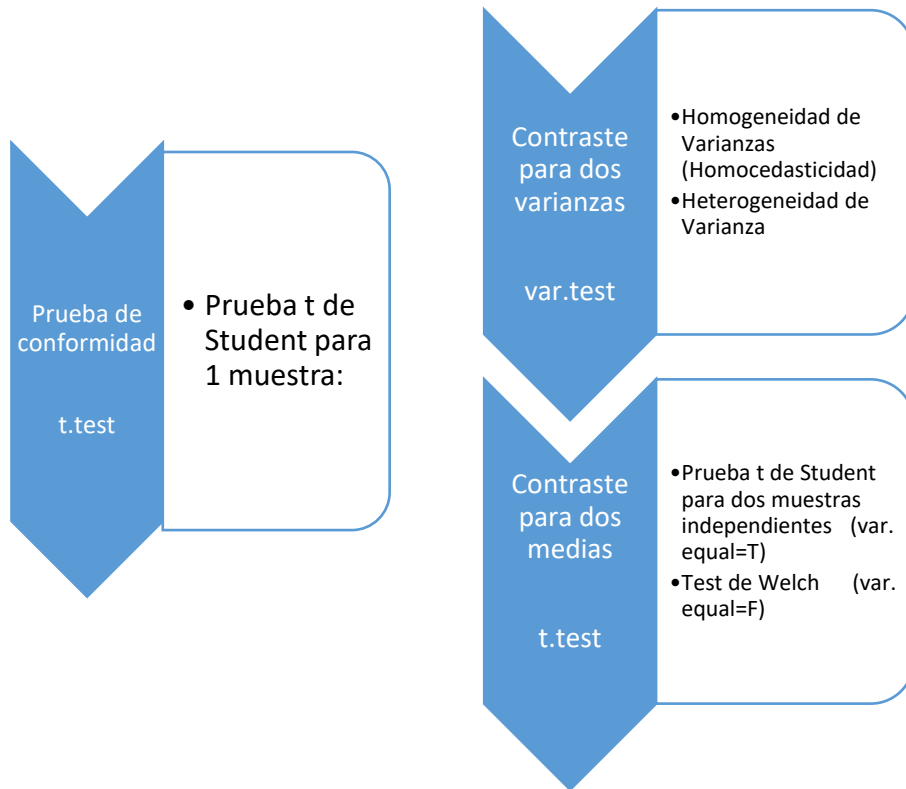
```
data: m
p-value = 8.493e-08
alternative hypothesis: two.sided
```

Aplicación al análisis de enriquecimiento funcional/análisis de sobreexpresión

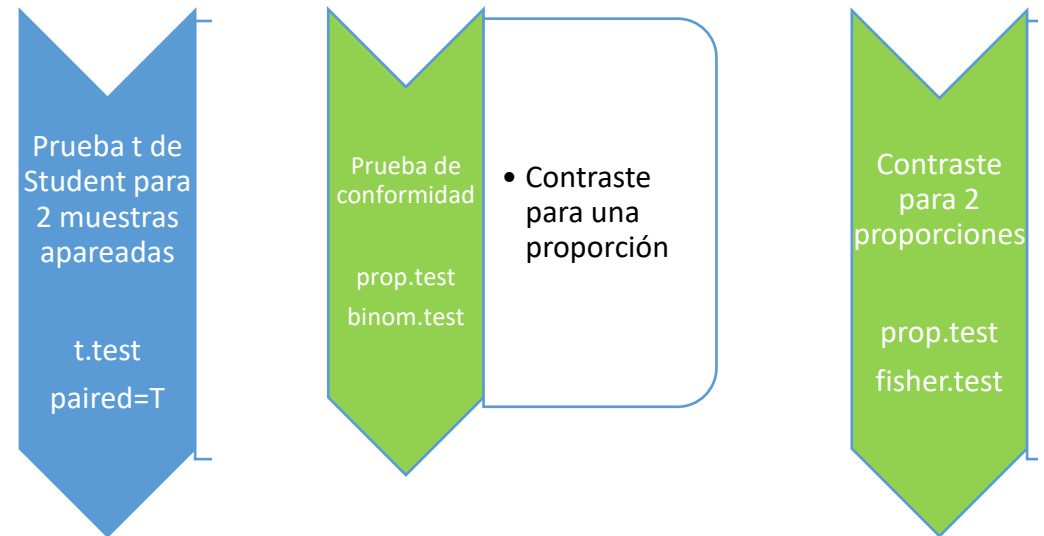
Supongamos que tenemos de datos de expresión de 25 genes (con 10 diferencialmente expresados) de los que 7 pertenecen a nuestro grupo objeto de estudio. Si 3 están diferencialmente expresados, obtener el p-valor que nos permita concluir si este grupo de 7 genes está enriquecido significativamente en genes diferencialmente expresados.

Hasta aquí hemos visto: Pruebas Paramétricas

Contrastes de Medias



Contrastes de Proporciones



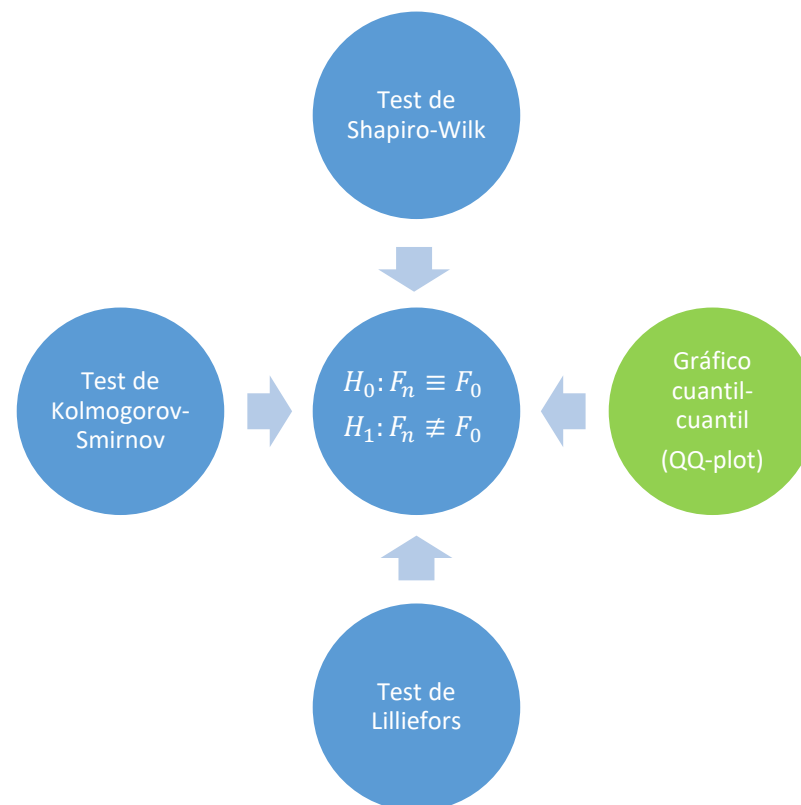
Ahora vamos a ver: Pruebas No Paramétricas

1. Pruebas de Bondad de Ajuste: Contraste de normalidad
2. Pruebas Chi-cuadrado
3. Contrastes de Medianas

Pruebas de Bondad de Ajuste: Contraste de normalidad

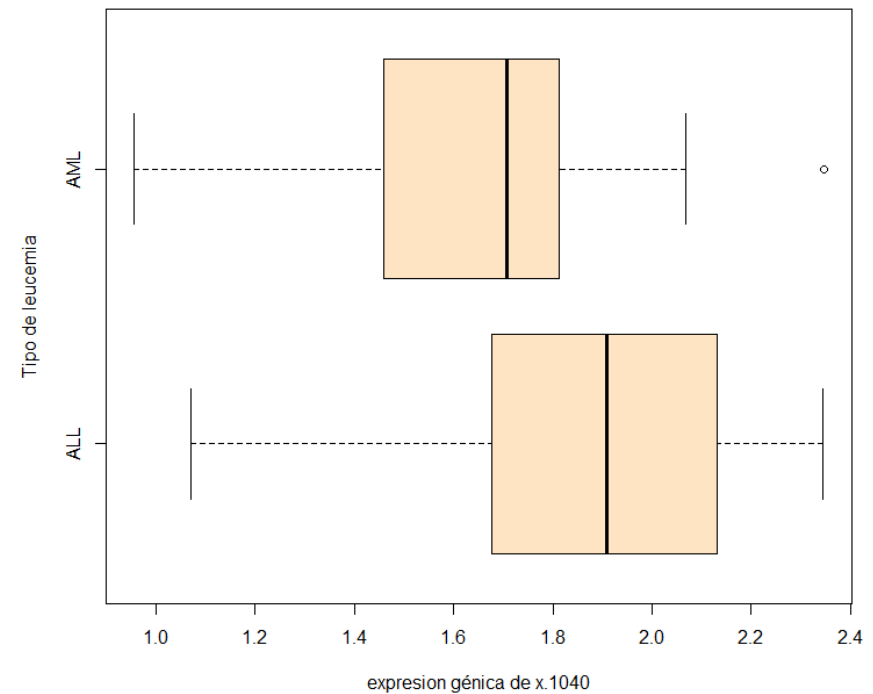
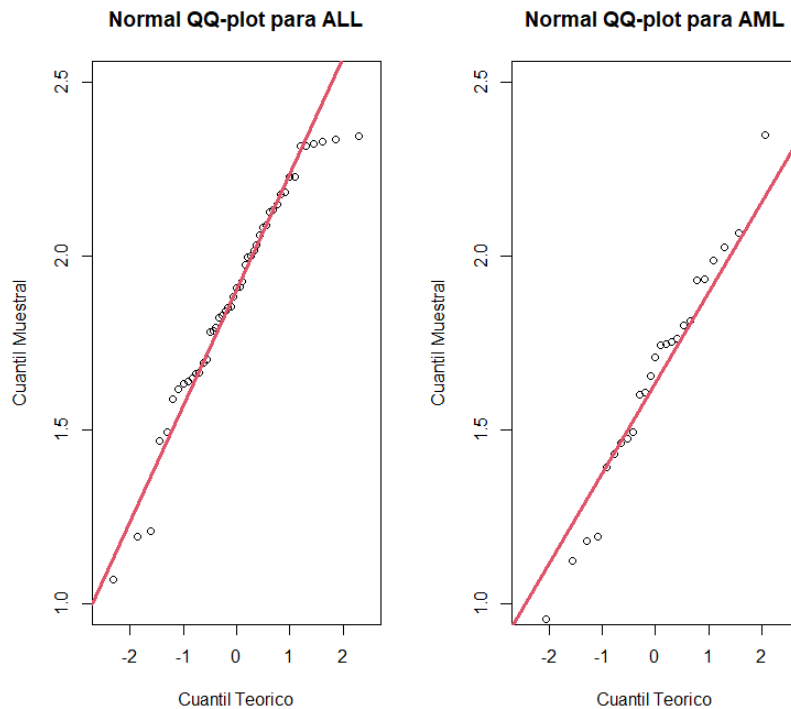
H_0 : la muestra sigue una distribución F_0

H_1 : la muestra sigue la distribución F_0



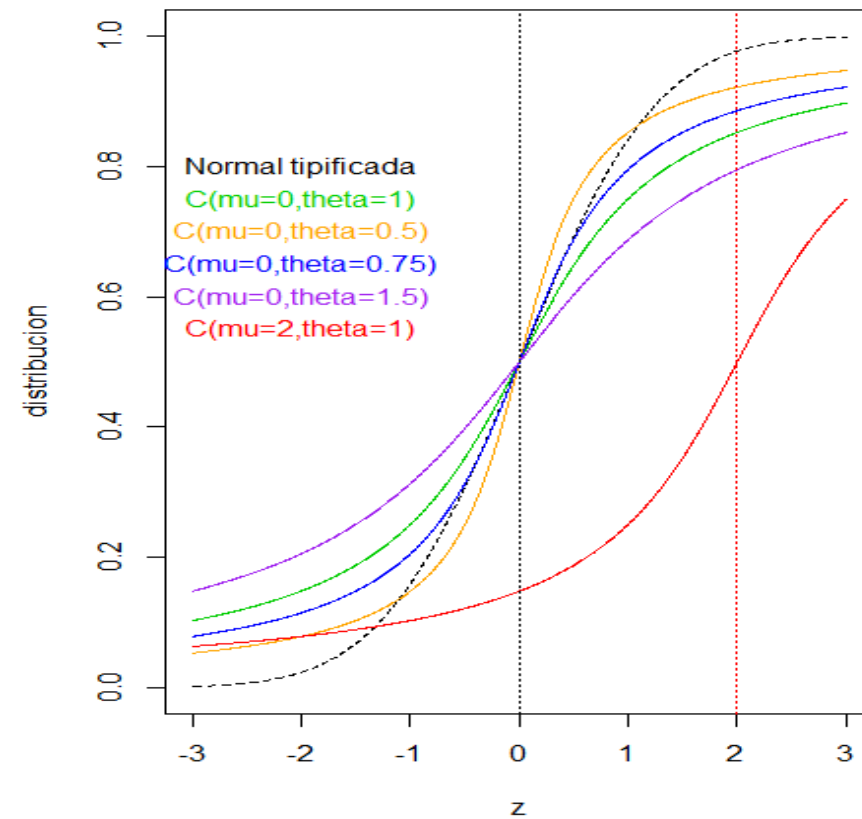
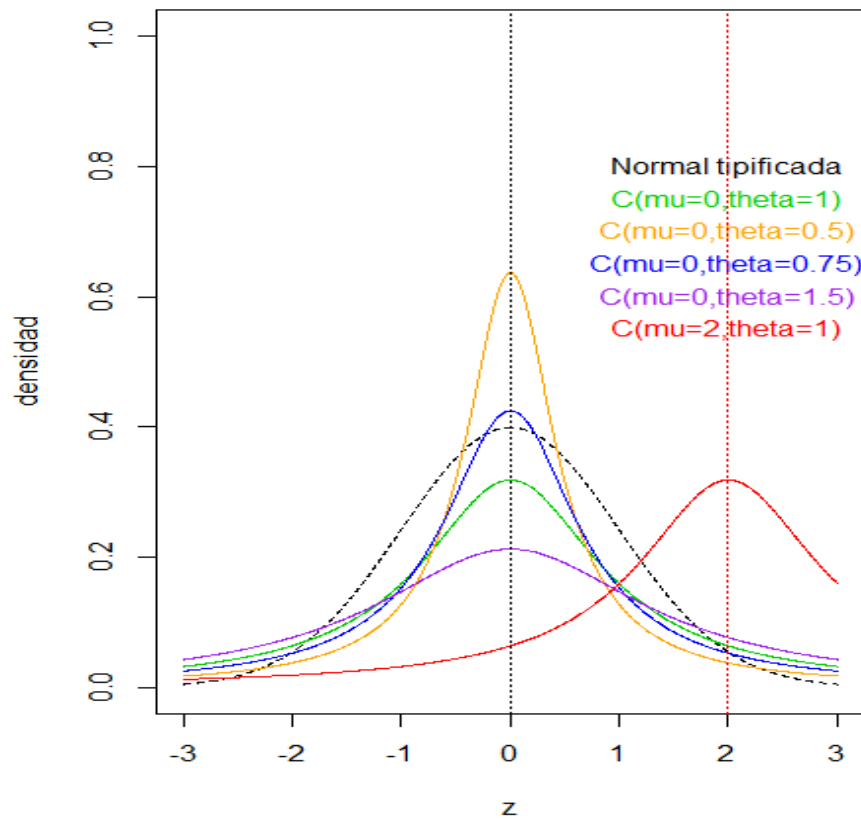
Técnica visual de bondad de ajuste: QQ-plot

1. Dado un entero k , calcular los $k-1$ cuantiles de
 - a. $F_0: \{p_1, \dots, p_{k-1}\}$
 - b. $F: (\{q_1, \dots, q_{k-1}\})$
2. Representar los puntos (p_i, q_i) , para cada j entre 1 y $k-1$.
3. Si las distribuciones son iguales, los puntos representados están todos en la diagonal.



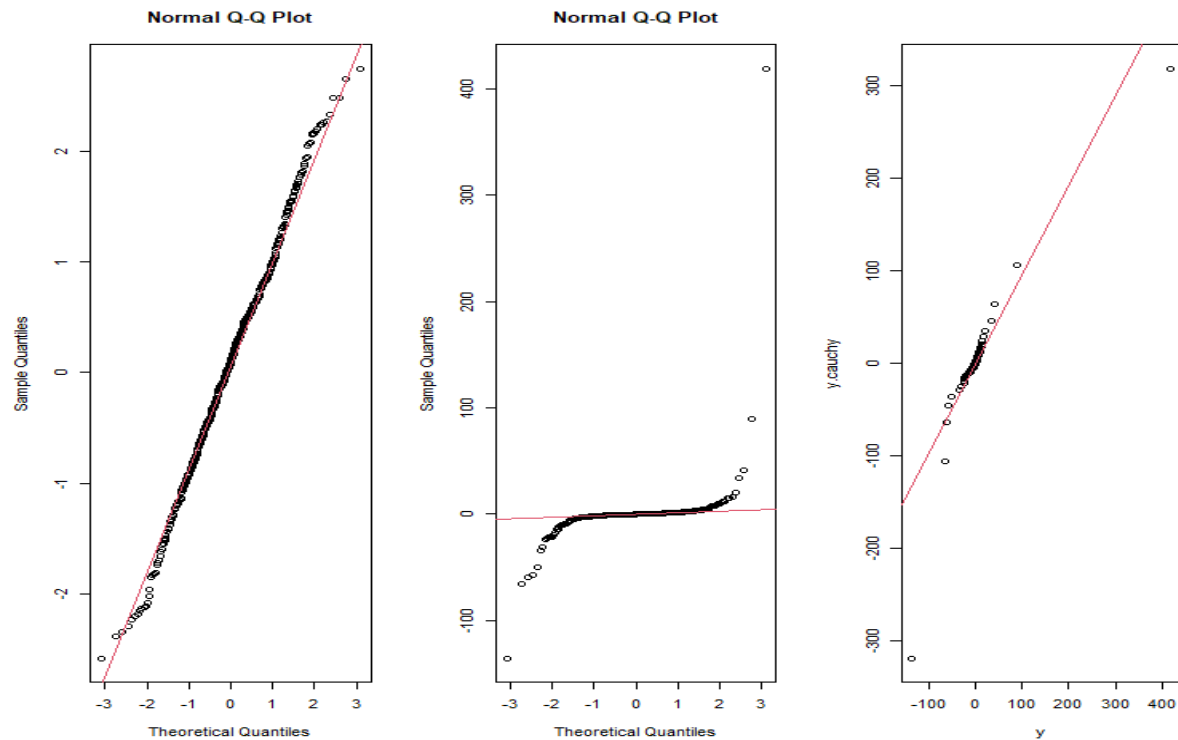
Ejemplo 3.16 Obtener los gráficos QQ-plot con muestras generadas de una distribución normal tipificada y de una distribución de Cauchy de parámetros $\mu=0$ y $\theta=1$, $X \sim \mathcal{C}(\mu, \theta)$.

En primer lugar, obsérvese que la distribución de Cauchy es simétrica respecto a μ , pero con colas más pesadas que el modelo normal.



En segundo lugar, generamos dos muestras aleatorias, una procedente de un modelo $N(0,1)$ y otra de $C(0,1)$, y representamos sus gráficos normal QQ-plot, y un cauchy QQ-plot solo para la segunda muestra.

```
n <- 500
set.seed(12345); x <- rnorm(n)
set.seed(12345); y <- rcauchy(n)
op <- par(mfrow=c(1, 3))
qqnorm(x, xlab="Theoretical Quantiles", ylab="Sample Quantiles"); qqline(x, col=2)
qqnorm(y, xlab="Theoretical Quantiles", ylab="Sample Quantiles"); qqline(y, col=2)
y.cauchy <- qcauchy(ppoints(length(y))); qqplot(y, y.cauchy)
qqline(y, col=2, distribution = qcauchy)
par(op)
```



Test de Kolmogorov-Smirnov

El estadístico de este test viene dado por: $D = \max|F_n(x) - F_0(x)|$, donde $F_n(x)$ es la función de la distribución muestral y $F_0(x)$ es la función teórica. Contrastar la hipótesis de normalidad a través de este test requiere que los parámetros poblacionales sean conocidos.

H_0 : Las observaciones provienen de una población distribuida normalmente

H_1 : Las observaciones provienen de una población no distribuida normalmente

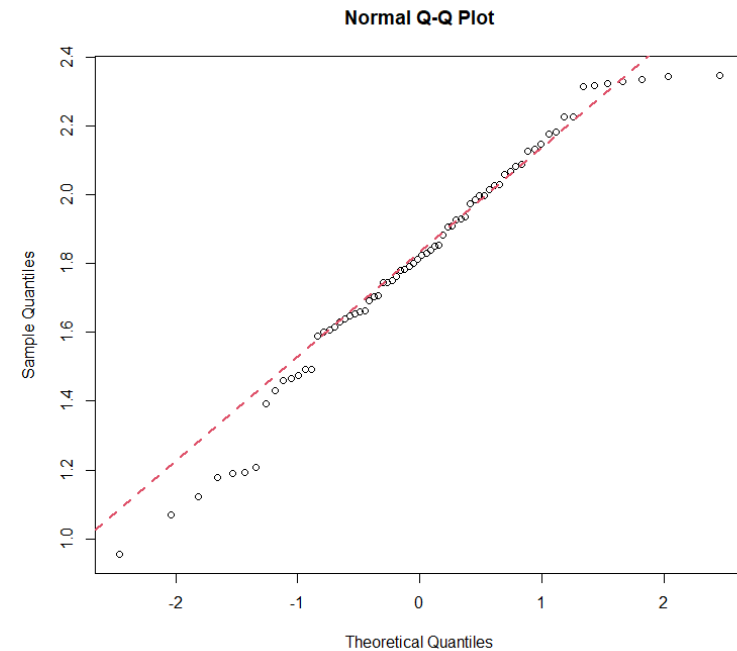
```
ks.test(leukemi a$x. 1040, pnorm, mean(leukemi a$x. 1040), sd(leukemi a$x. 1040))
```

One-sample Kolmogorov-Smirnov test

data: leukemi a\$x. 1040

D = 0.064357, p-value = 0.9079

alternative hypothesis: two-sided



H_0 : La muestra sigue una distribución t de Student
 H_1 : La muestra no sigue una distribución t de Student



```
ks.test(leukemia$x.1040, pt, df=4)
```

One-sample Kolmogorov-Smirnov test

```
data: leukemia$x.1040  
D = 0.81365, p-value = 4.441e-16  
alternative hypothesis: two-sided
```



¿Contrastes de normalidad para:

```
leukemia$x.1040[leukemia$Y==0]  
leukemia$x.1040[leukemia$Y==1]?
```

```
> summary(leukemia$factor)  
ALL  AML  
47   25
```

Para muestras de tamaño inferior a 50, el contraste de normalidad adecuado es....

Test de Shapiro-Wilk

El estadístico del test de Shapiro-Wilk viene dado por $W = \frac{D^2}{nS^2}$ donde D es la suma de las diferencias corregidas.

```
shapiro.test(leukemia$x.1040[leukemia$Y==0])
```

Shapiro-Wilk normality test

```
data: leukemia$x.1040[leukemia$Y == 0]  
W = 0.95519, p-value = 0.06937
```

```
shapiro.test(leukemia$x.1040[leukemia$Y==1])
```

Shapiro-Wilk normality test

```
data: leukemia$x.1040[leukemia$Y == 1]  
W = 0.98287, p-value = 0.9354
```

Ejemplo 3.18 Contrastar la hipótesis de que los valores de expresión del gen Gdf5 (x.3395) para pacientes ALL y AML (respectivamente) están normalmente distribuidos.

Ejemplo 3.20 Contrastar la normalidad de la variable Peso.kg para no corredores, $n = 35$.

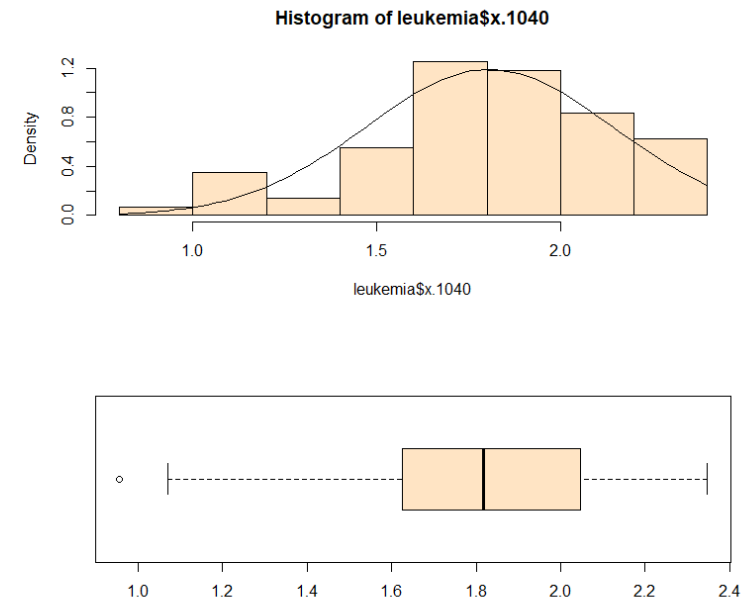
Ejemplo 3.21 Desde el fichero de datos “calorie”, 20 observaciones y 4 variables, realizar un contraste de la normalidad de la variable wgt usando el test de Shapiro-Wilk, puesto que $n = 20$.

Test de Lilliefors

A diferencia de `ks.test`, la función `lillie.test` no requiere que los parámetros de la distribución sean conocidos. El test de Lilliefors se lleva a cabo fácilmente una vez instalado el paquete `nortest`.

```
lillie.test(leukemia$x.1040)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test  
data: leukemia$x.1040  
D = 0.064357, p-value = 0.6507
```

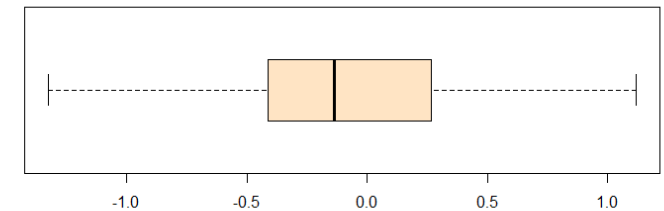
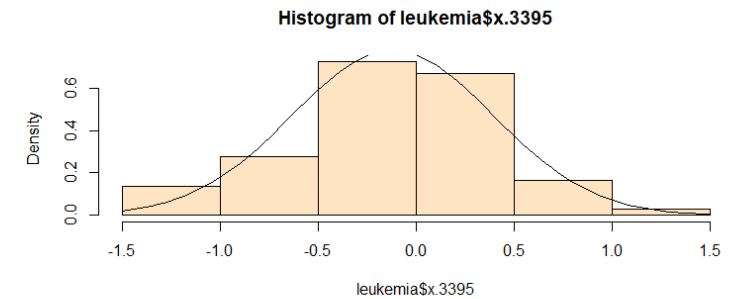


```
lillie.test(leukemia$x.3395)
```

Lilliefors (Kolmogorov-Smirnov) normality test

data: leukemia\$x.3395

D = 0.053839, p-value = 0.8721



Ejercicio. Contrastar la normalidad de las variables continuas del fichero ElPulso.

Pruebas Chi-cuadrado

- **Test Chi-Cuadrado de Bondad de Ajuste**, para contrastar si una distribución de frecuencias tiene una determinada función de probabilidad.

$$H_0: (O_1, O_2, \dots, O_K) \equiv \mathcal{M}(n, p_1, p_2, \dots, p_k)$$

$$H_1: (O_1, O_2, \dots, O_K) \not\equiv \mathcal{M}(n, p_1, p_2, \dots, p_k)$$

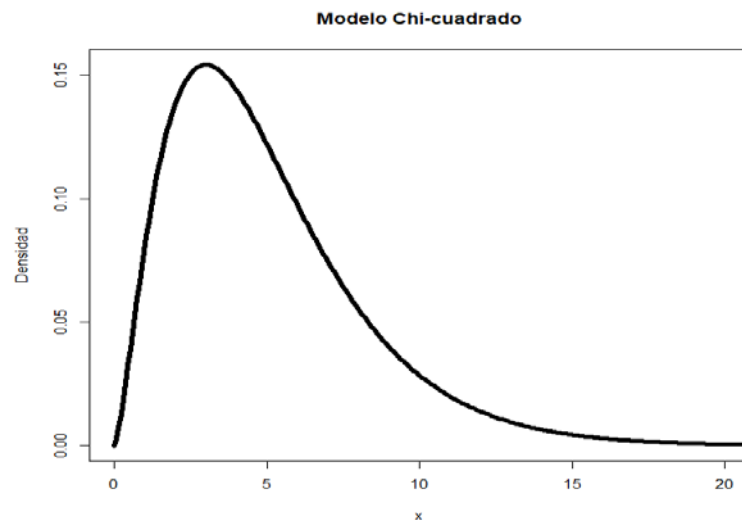
$$T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^k \frac{(O_i - np_i)^2}{np_i} \equiv \chi_{k-1}^2$$

Test Chi-Cuadrado de Independencia, para contrastar si dos variables son estadísticamente independientes.

$$H_0: X \text{ e } Y \text{ son independientes}$$

$$H_1: X \text{ e } Y \text{ no son independientes}$$

$$T = \sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - np_{ij})^2}{np_{ij}} \equiv \chi_{ab-1}^2$$



Importante: $E \geq 5$

Ejemplo Test Chi-Cuadrado de Bondad de Ajuste

Ejemplo 3.15 (Resuelto en apartado “Contraste de dos proporciones”)

Datos extraídos de Almoguera (2011) accesible en <https://repositorio.uam.es/handle/10486/6784> sobre el estudio del alelo ATA como factor de susceptibilidad a la esquizofrenia. Se agruparon los genotipos según el número de alelos ATA (2, 1 ó 0) y las distribuciones del recuento de mujeres esquizofrénicas (casos) y mujeres de la muestra objeto de estudio (muestra), portadoras de 2, 1 o ningún alelo ATA, respectivamente, son:
muestra = c(20, 197, 156) y casos = c(13, 27, 48).

muestra = c(20, 197, 156)

casos = c(13, 27, 48)

m <- matrix(c(casos, muestra-casos), ncol=2); m

	[, 1]	[, 2]
[1,]	13	7
[2,]	27	170
[3,]	48	108

colnames(m) <- c("caso", "control"); rownames(m) <- c("2", "1", "0"); m

	caso	control
2	13	7
1	27	170
0	48	108

`chi sq. test (m)`

Pearson' s Chi - squared test

data: m

X-squared = 34.163, df = 2, p-value = 3.816e-08

Warning message:

In `chi sq. test (m)` : Chi-squared approximation may be incorrect

`chi sq. test (m) $expected`

	caso	control
2	4.718499	15.2815
1	46.477212	150.5228
0	36.804290	119.1957

Warning message:

In `chi sq. test (m)` : Chi-squared approximation may be incorrect

```
muestra2=c(20+197, 156); muestra2
[1] 217 156
```

```
casos2=c(13+27, 48); casos2
[1] 40 48
```

```
m2 <- matrix(c(casos2, muestra2- casos2), ncol=2); m2
      [, 1] [, 2]
[1, ]    40   177
[2, ]    48   108
```

```
colnames(m2) <- c("caso", "control")
rownames(m2) <- c("presencia", "ausencia"); m2
```

```
      caso control
presencia    40    177
ausencia     48    108
chisq.test(m2)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: m2
X-squared = 6.9925, df = 1, p-value = 0.008185
```

```
chisq.test(m2)$expected
```

```
      caso control
presencia 51.19571 165.8043
ausencia  36.80429 119.1957
```

Comparaciones post-hoc

Residuos = Diferencias de las frecuencias observadas y las esperadas.

`chi sq. test(m2)$residuals`

	caso	control
presencia	- 1. 564714	0. 8694683
ausencia	1. 845451	- 1. 0254662

Residuos estandarizados = Residuos divididos por la raíz cuadrada de las frecuencias esperadas

`chi sq. test(m2)$stdres`

	caso	control
presencia	- 2. 767955	2. 767955
ausencia	2. 767955	- 2. 767955

Ejercicio. Test Chi-Cuadrado de bondad de ajuste para las proporciones 9:3:3:1 de una F2 dihíbrida mendeliana con valores fenotípicos 430:139:166:41.

`chi sq. test(c(430, 139, 166, 41), p=c(9, 3, 3, 1)/16)`

Chi-squared test for given probabilities

data: c(430, 139, 166, 41)
X-squared = 4. 4353, df = 3, p-value = 0. 2181

Ejemplo Test Chi-Cuadrado de Independencia

Ejemplo 3.14 Ejemplo basado en el estudio de las propiedades codificantes de los exones alternativos:
<http://bioinformatica.upf.edu/2005/projectes05/3.7.1/>

H_0 : Independencia de uso de codones entre los exones constitutivos y skipped-exons

H_1 : No Independencia de uso de codones entre los exones constitutivos y skipped-exons

¿Existen diferencias significativas en el uso de codones entre los exones constitutivos y alternativos?

```
Alani na <- as.table(cbind(c(9413, 12012, 4143, 10283), c(3598, 4389, 1067, 3861)))
```

```
dimnames(Alani na) <- list(codon = c("GCA", "GCC", "GCG", "GCT"), exon = c("CONST", "ALTERNA"))
```

```
XsqAla <- chisq.test(Alani na)
```

```
attributes(XsqAla)
```

```
XsqAla$observed
```

```
XsqAla$expected
```

```
XsqAla$residuals
```

```
XsqAla$stdres
```

```
Alani na <- as.table(cbind(c(9413, 12012, 4143, 10283), c(3598, 4389, 1067, 3861)))
dimnames(Alani na) <- list(codon = c("GCA", "GCC", "GCG", "GCT"), exon = c("CONST", "ALTERNA"))
XsqAl a <- chi sq. test (Alani na)
```

Pearson's Chi-squared test

```
data: Alani na
X-squared = 111.06, df = 3, p-value < 2.2e-16
```

attributes(XsqAl a)

```
$names
[1] "statistic" "parameter" "p.value" "method" "data.name" "observed" "expected"
[8] "residuals" "stdres"
```

```
$class
[1] "htest"
```

XsqAl a\$observed

	exon	
codon	CONST	ALTERNA
GCA	9413	3598
GCC	12012	4389
GCG	4143	1067
GCT	10283	3861

XsqAl a\$expected

exon

codon	CONST	ALTERNA
GCA	9565. 217	3445. 783
GCC	12057. 422	4343. 578
GCG	3830. 204	1379. 796
GCT	10398. 157	3745. 843

XsqAl a\$resi dual s

exon

codon	CONST	ALTERNA
GCA	- 1. 5563780	2. 5930956
GCC	- 0. 4136575	0. 6891986
GCG	5. 0541809	- 8. 4208166
GCT	- 1. 1293113	1. 8815558

XsqAl a\$stdres

exon

codon	CONST	ALTERNA
GCA	- 3. 5319670	3. 5319670
GCC	- 0. 9866726	0. 9866726
GCG	10. 3919420	- 10. 3919420
GCT	- 2. 6043990	2. 6043990


```
Glutamina <- as.table(rbind(c(14272, 5527), c(15087, 6066)))
dimnames(Glutamina) <- list(codon = c("GAA", "GAG"), exon = c("CONST", "ALTERNA"))
(XsqGlu <- chisq.test(Glutamina))
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: Glutamina
X-squared = 2.8826, df = 1, p-value = 0.08954
```

XsqGlu\$observed

	exon	
codon	CONST	ALTERNA
GAA	14272	5527
GAG	15087	6066

XsqGlu\$expected

	exon	
codon	CONST	ALTERNA
GAA	14194.15	5604.85
GAG	15164.85	5988.15

XsqGlu\$residuals

	exon	
codon	CONST	ALTERNA
GAA	0.6534352	-1.0398609
GAG	-0.6321762	1.0060299

XsqGlu\$stdres

	exon	
codon	CONST	ALTERNA
GAA	1.70881	-1.70881
GAG	-1.70881	1.70881

Ejemplo. Analiza si la reacción alérgica a un compuesto y una determinada mutación en un gen están relacionadas, con los resultados de un test alérgico de una muestra aleatoria y el genotipado del estado del gen ¿Existen diferencias significativas en la ocurrencia de la mutación entre los individuos?

```
> datos <- data.frame( sujeto = c("No alérgico", "No alérgico", "No alérgico",
                                "No alérgico", "alérgico", "No alérgico",
                                "No alérgico", "alérgico", "alérgico",
                                "No alérgico", "alérgico", "alérgico",
                                "alérgico", "alérgico", "alérgico",
                                "No alérgico", "No alérgico", "No alérgico",
                                "No alérgico", "alérgico", "alérgico",
                                "alérgico", "alérgico", "No alérgico",
                                "alérgico", "No alérgico", "No alérgico",
                                "alérgico", "alérgico", "alérgico"),
                      mutacion = c(FALSE, FALSE, FALSE, FALSE, TRUE, FALSE,
                                   FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE,
                                   TRUE, FALSE, FALSE, TRUE, FALSE, TRUE, FALSE,
                                   TRUE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE,
                                   TRUE, FALSE, TRUE))
```

```
Xsq. datos=chi sq. test(datos$sujeto, datos$mutacion)
```

```
Xsq. datos$observed
```

	datos\$mutacion	
datos\$sujeto	FALSE	TRUE
alérgico	6	10
No alérgico	11	3

```
col1 <- c(6, 11)
```

```
col2 <- c(10, 3)
```

```
tabla <- as.table(cbind(col1, col2))
```

```
dimnames(tabla)=list(sujeto=c("Alérgico", "No alérgico"), mutacion=c("False", "True"))
```

```
tabla
```

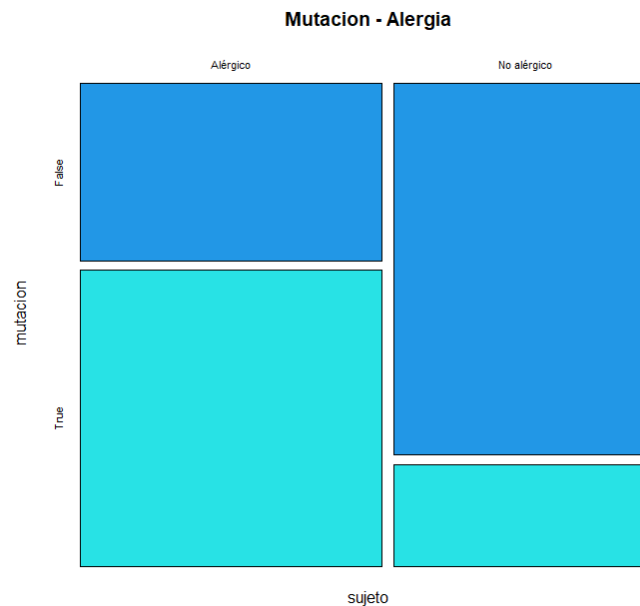
sujeto	mutacion	
	False	True
Alérgico	6	10
No alérgico	11	3

> `chi sq. test(table)`

Pearson's Chi-squared test with Yates' continuity correction

data: table

X-squared = 3.593, df = 1, p-value = 0.05802



Contrastes de Medianas

En ausencia de normalidad y muestras con tamaño suficiente los contrastes de medianas ([wilcox.test](#)) se presentan como alternativa robusta a los contrastes de medias ([t.test](#))

Test de Wilcoxon para una muestra

Se trata de un contraste de centralidad de una población de distribución simétrica

$$H_0: Me = Me_0$$

$$H_1: Me \neq Me_0$$

```
summary(leukemia$x.3395[leukemia$Y==0])
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.1563891	-0.3796605	-0.0055636	0.0000758	0.3201224	1.1204422

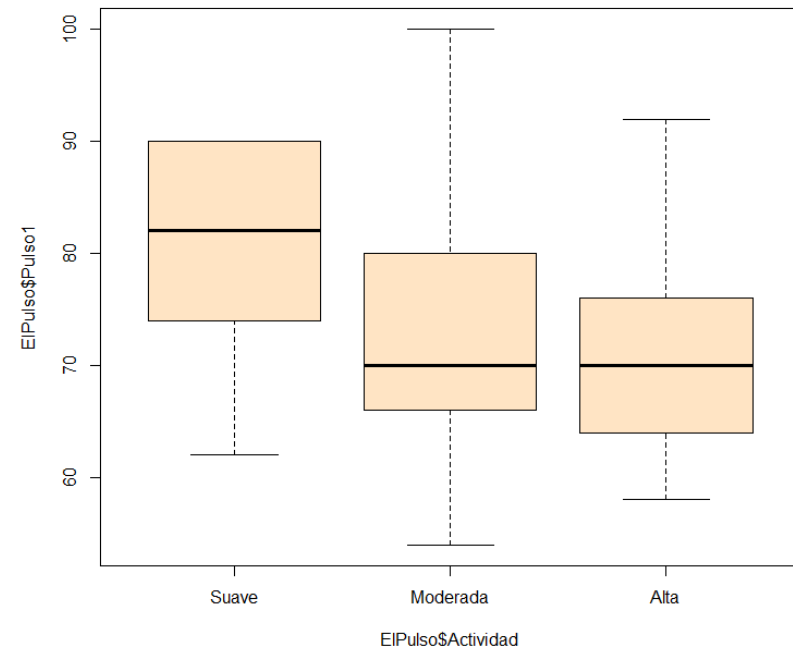
```
wilcox.test(leukemia$x.3395[leukemia$Y==0])
```

Wilcoxon signed rank exact test

```
data: leukemia$x.3395[leukemia$Y == 0]
```

```
V = 550, p-value = 0.8875
```

```
alternative hypothesis: true location is not  
equal to 0
```



Test U de Mann-Witney para dos muestras independientes

$$H_0: Me_1 = Me_2$$

$$H_1: Me_1 \neq Me_2$$

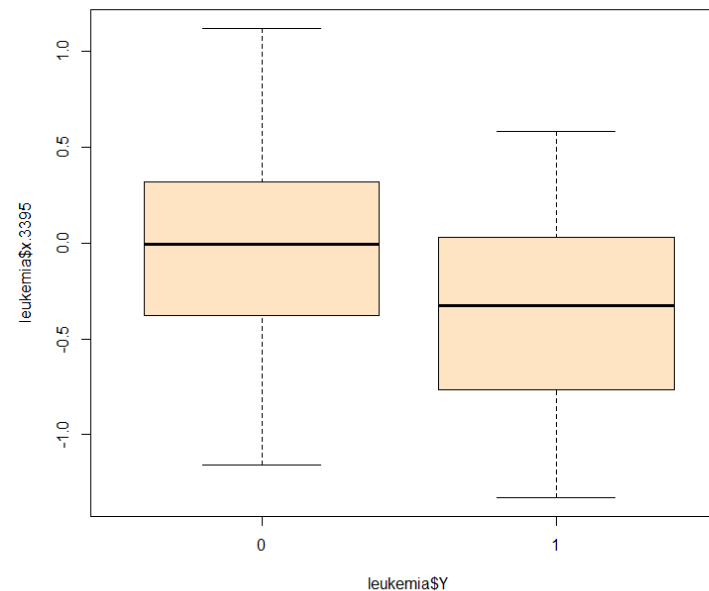
`wilcox.test(leukemia$x.3395~leukemia$a$Y)`

Wilcoxon rank sum exact test

data: leukemia\$x.3395 by leukemia\$a\$Y

W = 793, p-value = 0.01459

alternative hypothesis: true location shift is not equal to 0



Test de Wilcoxon para 2 muestras apareadas

$$H_0: Me_{antes} = Me_{despues}$$

$$H_1: Me_{antes} \neq Me_{despues}$$

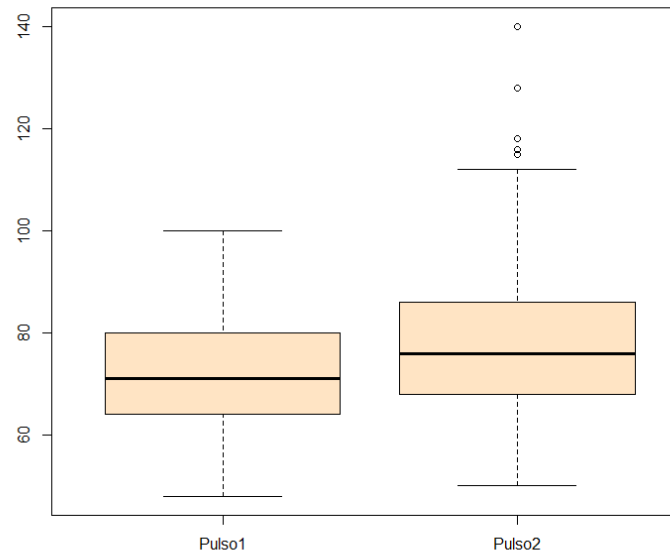
`wilcox.test(El Pulso$Pulso1, El Pulso$Pulso2, paired=T)`

Wilcoxon signed rank test with continuity correction

data: El Pulso\$Pulso1 and El Pulso\$Pulso2

V = 601, p-value = 1.27e-05

alternative hypothesis: true location shift is not equal to 0



Más pruebas no paramétricas

$$H_0: Me_1 = Me_2 = \dots = Me_k$$

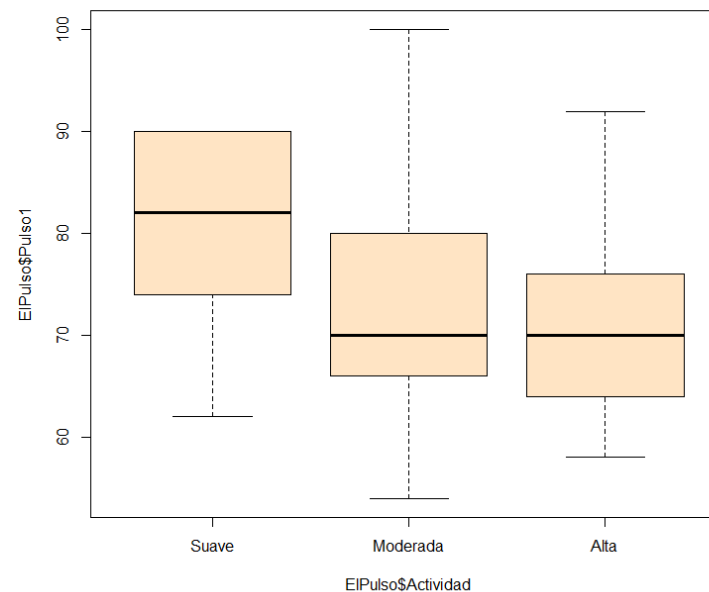
H_1 : No todas iguales

`kruskal.test(ElPulso$Pulso1~ElPulso$Actividad)`

Kruskal - Wallis rank sum test

data: ElPulso\$Pulso1 by ElPulso\$Actividad

Kruskal - Wallis chi-squared = 3.6546, df = 2, p-value = 0.1608



Muestreando con R

El diseño del experimento o procedimiento de muestreo son de vital importancia para no introducir sesgos.

Muestreo aleatorio simple

Este proceso de extracción probabilístico garantiza:

1º cada elemento tiene la misma probabilidad de ser seleccionado.

2º las extracciones se realizan con reposición, de manera que la población es idéntica en todas las extracciones.

Adecuado: Los elementos de la población son homogéneos respecto a la variable de estudio.

```
x<- 1:15; x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
set.seed(12); sample(x, replace=T) #Muestra aleatoria con reposición n=15
[1] 2 10 7 11 14 5 5 12 2 8 11 2 13 2 9
set.seed(12); sample(x) #Muestra aleatoria sin reposición de tamaño 15
[1] 2 10 7 11 5 12 15 8 3 9 14 6 1 4 13
set.seed(12); sample(x, 7) # Muestras de tamaño 7 sin reposición
[1] 2 10 7 11 5 12 15
set.seed(12); sample(x, 7, replace=T) # Muestras tamaño 7 con reposición
[1] 2 10 7 11 14 5 5
```


Aplicación práctica 1

#Muestra aleatoria sin repetición de tamaño 60 de la base de datos iris

```
> indices<- sample(1:nrow(iris), 60); iris.muestreado<- iris[indices, ]
> head(iris.muestreado)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
79	6.0	2.9	4.5	1.5	versicolor
72	6.1	2.8	4.0	1.3	versicolor
73	6.3	2.5	4.9	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
131	7.4	2.8	6.1	1.9	virginica
37	5.5	3.5	1.3	0.2	setosa

#Muestra aleatoria con repetición de tamaño 60 de la base de datos iris

```
indices<- sample(1:nrow(iris), 60, replace=T); iris.muestreado<-
iris[indices, ]
head(iris.muestreado)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
65	5.6	2.9	3.6	1.3	versicolor
108	7.3	2.9	6.3	1.8	virginica
96	5.7	3.0	4.2	1.2	versicolor
72	6.1	2.8	4.0	1.3	versicolor
53	6.9	3.1	4.9	1.5	versicolor
80	5.7	2.6	3.5	1.0	versicolor

Aplicación práctica 2

Muestra aleatoria con reemplazamiento de 100 elementos de una Bernoulli

```
sample(c(0, 1), 100, replace = TRUE)
```

```
[1] 0 1 1 0 1 1 1 0 1 0 1 0 0 0 1 0 0 1 0 0 0 0 0 1 1 1 1 1 1 0 1 0 1 0 1 0 0 0 0 1 1 1 0 1 0 1 1
[48] 1 0 1 0 1 0 0 1 1 0 0 1 0 1 1 0 0 1 0 1 1 1 1 0 0 1 1 0 1 0 1 0 1 1 1 0 0 0 1 0 1 0 1 0 0 0
[95] 1 1 1 1 1 0
```