# Inferencia estadística básica

María Guzmán Martínez

2025-11-19

## Contents



```r
library(ggplot2)
library(tidyverse)
library(dplyr)
library(GGally)
library(tinytex)
library(moments) # skewness
library(DescTools) # Mode

#library(agricolae)
#library(nortest)
#library(car) # powerTransform
```

# 1 Introducción

La inferencia estadística se ocupa de los métodos relacionados con estimación de parámetros, estimación que se verifica mediante juegos de hipótesis e intervalos de confianza.

La estimación de parámetros puede estar relacionada con los parámetros de una función de distribución o con los parámetros de un modelo estadístico.

Así el planteamiento de un juego de hipótesis está en función de los parámetros de una función de distribución o con los parámetros de un modelo estadístico.

Para realizar inferencias sobre un conjunto de parámetros se puede utilizar un juego de hipótesis o bien un intervalo de confianza.

# 2 Funciones de distribución

En estadística existe un conjunto de funciones de distribución continuas; las cuales permiten modelar fenómenos sociales, naturales y económicos, entre otros. Estas funciones de distribución se encuentran relacionadas a una variable aleatoria.

En un espacio de probabilidad dado $(\Omega, S, P(\cdot))$ una variable aleatoria $X(\cdot)$ es una función con que:

$$X : \Omega \to \mathbb{R}$$

Observaciones:

- Si $\Omega$ es finito (numerable o contable), entonces
$$X(\omega) : \Omega \to \mathcal{X} \subset \mathbb{Z}$$
en este caso $\Omega$ es un conjunto discreto, y por lo tanto $X$ es una variable aleatoria discreta.
- Si $\Omega$ es no numerable, entonces
$$X(\omega) : \Omega \to \mathcal{X} \subset \mathbb{R}$$
en este caso $\Omega$ es conjunto continuo, y por lo tanto $X$ es una variable aleatoria continua.

## 2.1 Normal

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

$-\infty < \mu < \infty, \ \sigma > 0$

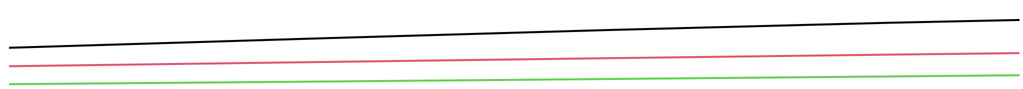
Si $X$ tiene distribución $N(\mu, \sigma^2)$, entonces

$$\begin{aligned} E(X) &= \mu \\ V(X) &= \sigma^2 \\ m_X(t) &= e^{\mu t + \frac{(\sigma t)^2}{2}} \end{aligned}$$

Distribución normal estándar

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

```r
plot.new()
mu<-c(2,3,5)
sd<-c(2,3,sqrt(4+9))

for (i in 1:3) curve(dnorm(x, mu[i], sd[i]), from=-15, to=15, col=i, add = TRUE, ylim=c(0,0.2),
                     ylab = expression(f[X](x)))

leg.txt<-c(expression(paste("N(",mu==2,  ", ",sigma^2==4,")")),
           expression(paste("N(",mu==3,  ", ", sigma^2==9,")")),
           expression(paste("N(",mu==5,  ", " , sigma^2==15,")")))
color<-1:3
legend(-15,.15, leg.txt, col=color, lwd=1, lty=1, bty="n")
```

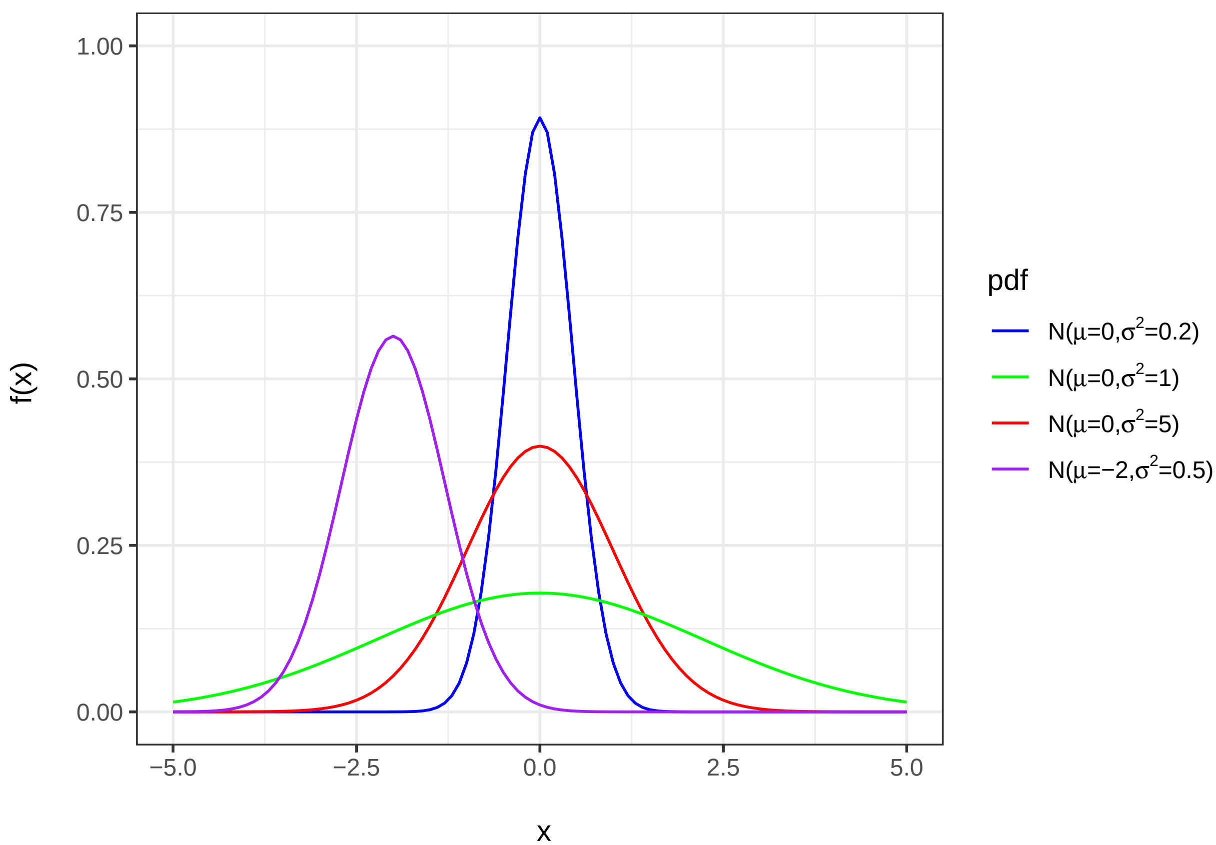


```r
dev.off()
```

```
## null device
##           1
```

```r
x_lower <- -5
x_upper <- 5

ggplot(data.frame(x = c(x_lower, x_upper)), aes(x = x)) + xlim(x_lower, x_upper) +
     ylim(0, 1) +
     stat_function(fun = dnorm, args = list(mean=0, sd=sqrt(0.2)), aes(colour = "1")) +
```

```r
    stat_function(fun = dnorm, args = list(mean=0, sd=sqrt(1)),   aes(colour = "3")) +
    stat_function(fun = dnorm, args = list(mean=0, sd=sqrt(5)),   aes(colour = "2")) +
    stat_function(fun = dnorm, args = list(mean=-2, sd=sqrt(0.5)),   aes(colour = "5")) +
    scale_color_manual("pdf", values = c("blue", "green", "red", "purple"),
                       labels=c(expression(paste("N(", mu, "=0,", sigma^2,"=0.2)")),
                                expression(paste("N(", mu, "=0,", sigma^2,"=1)")),
                                expression(paste("N(", mu, "=0,", sigma^2,"=5)")),
                                expression(paste("N(", mu, "=-2,", sigma^2,"=0.5)"))) ) +
    labs(x = "\n x", y = "f(x) \n") +
    theme(plot.title = element_text(hjust = 0.5),
        legend.position = "right")+
  theme_bw()
```




## 2.2 t de Sdudent

$$f(x;v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi}}\left(1+\frac{x^2}{v}\right)^{-\left(\frac{v+1}{2}\right)} I_{(-\infty,\infty)}(x)$$

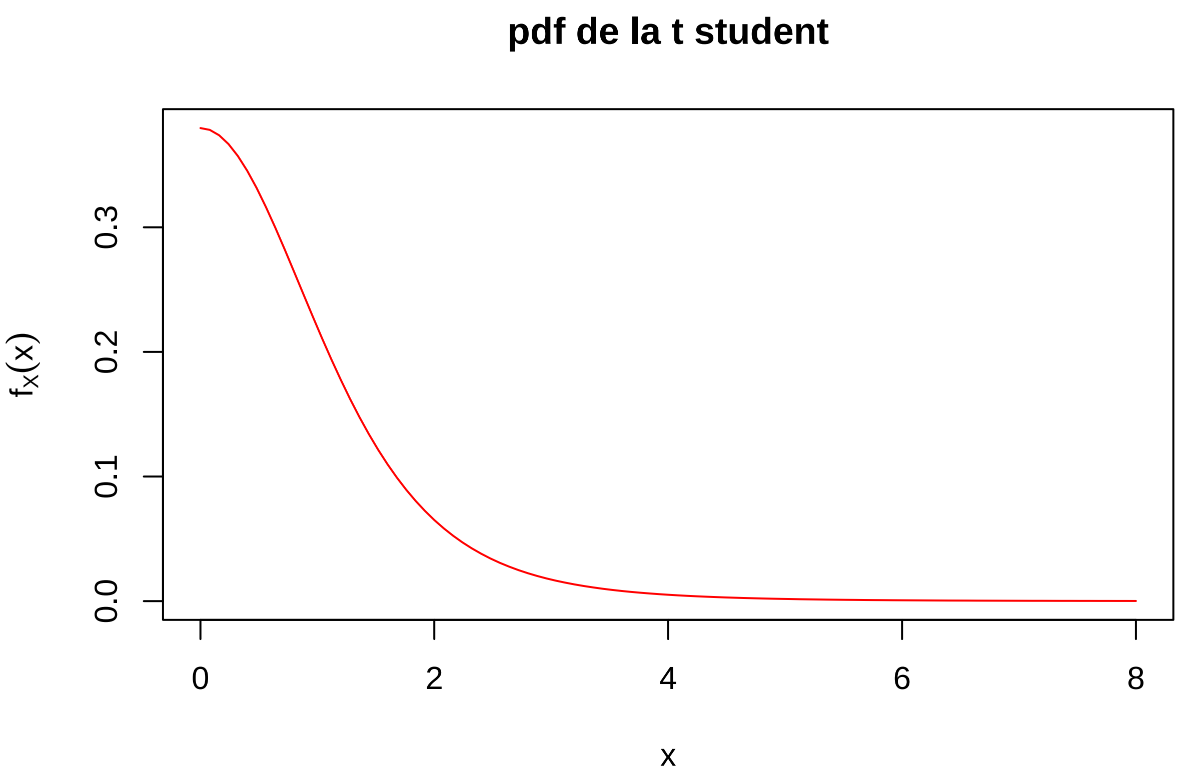
$v > 0$

Si $X$ tiene distribución $t(v)$, entonces

$$\begin{aligned} E(X) &= 0,\ v > 1 \\ V(X) &= \tfrac{v}{v-2},\ v > 2 \end{aligned}$$

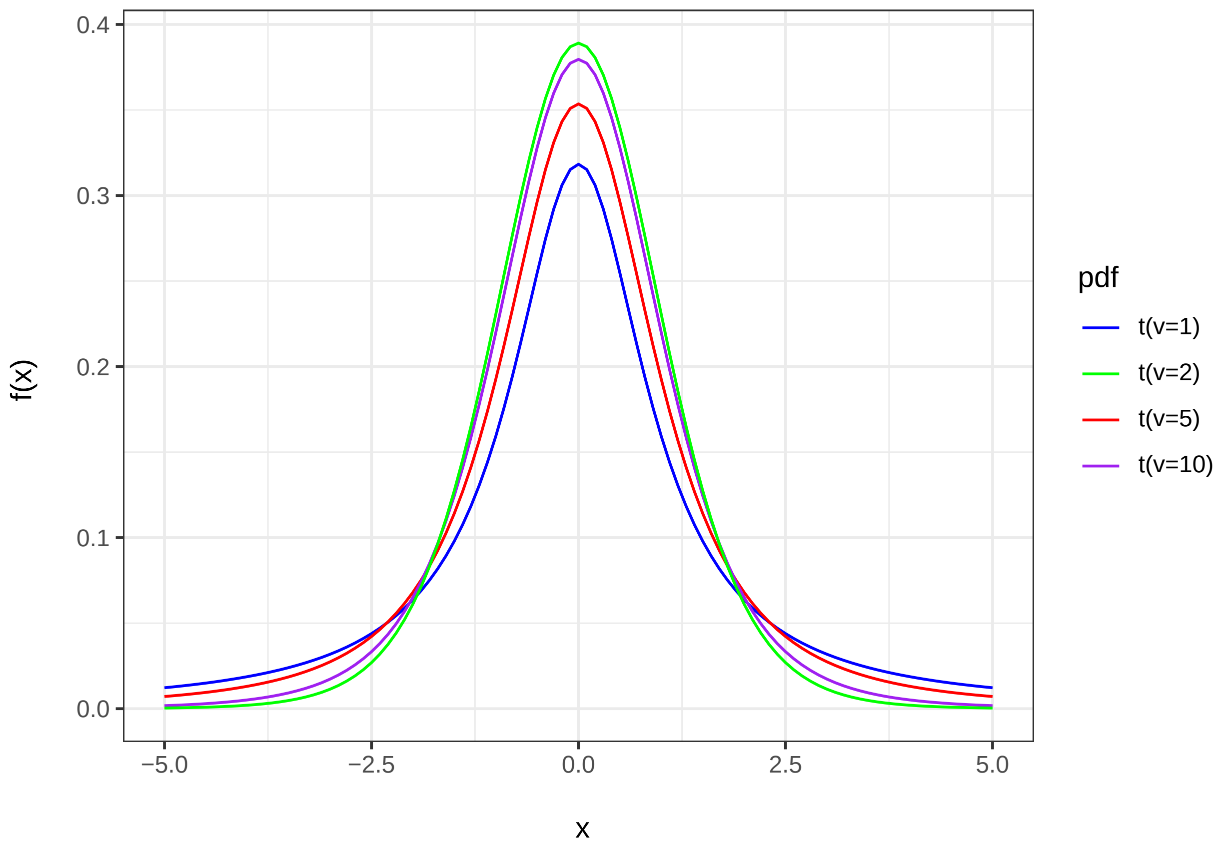
La fgm no existe.

```r
curve(dt(x, df=5), from=0, to=8,
      ylab = expression(f[X](x)),
      col="red",main="pdf de la t student")
```




```r
x_lower <- -5
x_upper <- 5

max_height2 <- max( dt(x_lower:x_upper, df = 1, log = FALSE),
                    dt(x_lower:x_upper, df = 2, log = FALSE),
                    dt(x_lower:x_upper, df = 5, log = FALSE),
                    dt(x_lower:x_upper, df = 10, log = FALSE)
                    )


ggplot(data.frame(x = c(x_lower, x_upper)), aes(x = x)) + xlim(x_lower, x_upper) +
     ylim(0, max_height2) +
     stat_function(fun = dt, args = list(df = 1),  aes(colour = "1")) +
     stat_function(fun = dt, args = list(df = 2),  aes(colour = "2")) +
     stat_function(fun = dt, args = list(df = 5),  aes(colour = "5")) +
     stat_function(fun = dt, args = list(df = 10), aes(colour = "10")) +
     scale_color_manual("pdf", values = c("blue", "green", "red", "purple"),
                       labels=c(expression(paste("t(", v, "=1",")")),
                                expression(paste("t(", v, "=2", ")")),
                                expression(paste("t(", v, "=5",")")),
                                expression(paste("t(", v, "=10",")"))) ) +
```

```
    labs(x = "\n x", y = "f(x) \n") +
    theme(plot.title = element_text(hjust = 0.5),
        legend.position = "right")+
  theme_bw()
```




## 2.3 Cauchy

$$f(x;\alpha,\beta) = \frac{1}{\pi\beta\left[1+\left(\frac{x-\alpha}{\beta}\right)^2\right]} I_{(-\infty,\infty)}(x)$$

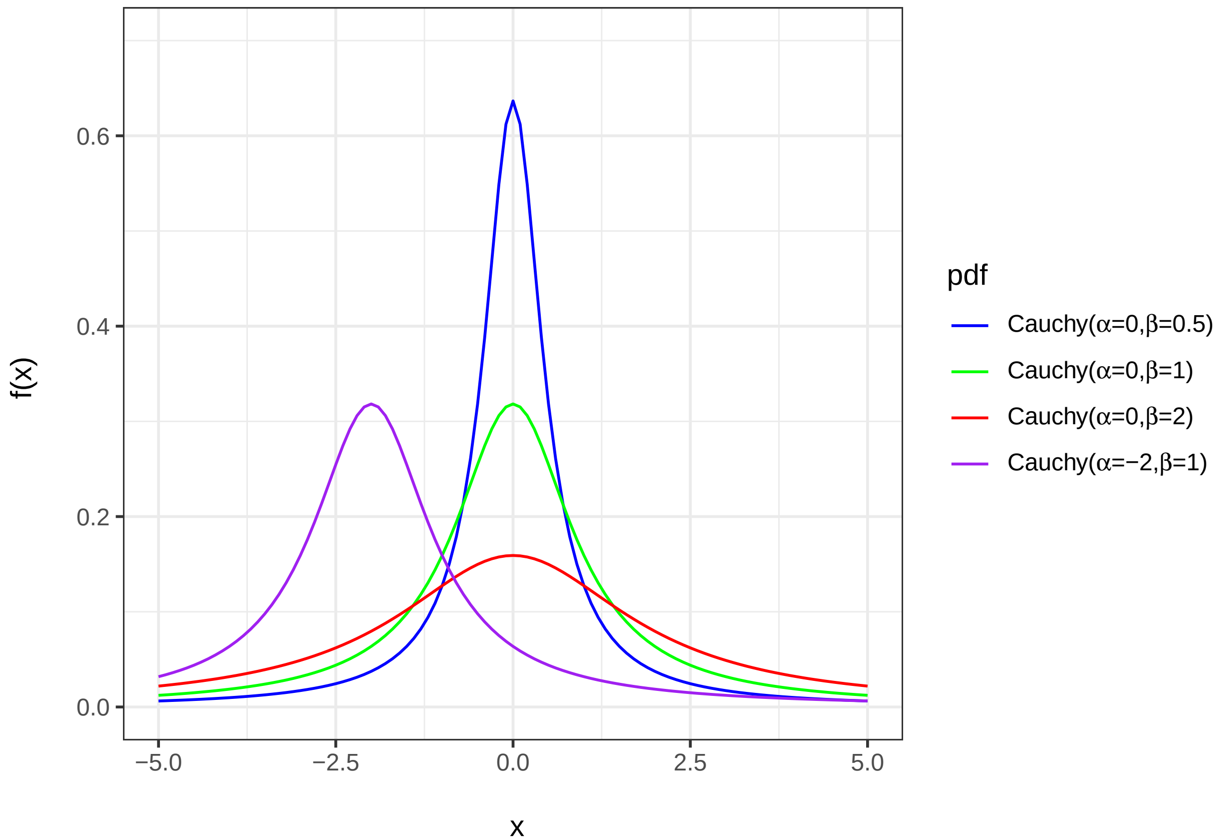
$-\infty < \alpha < \infty, \ \beta > 0$

Si $X$ tiene distribución $Cauchy(\alpha,\beta)$, entonces

$$\begin{array}{l} E(X) \text{ no existe} \\ V(X) \text{ no existe} \\ \text{La fgm no existe} \end{array}$$

```
x_lower <- -5
x_upper <- 5


ggplot(data.frame(x = c(x_lower, x_upper)), aes(x = x)) + xlim(x_lower, x_upper) +
    ylim(0, 0.7) +
```

```r
    stat_function(fun = dcauchy, args = list(location=0, scale=0.5), aes(colour = "0.5")) +
    stat_function(fun = dcauchy, args = list(location=0, scale=1),   aes(colour = "1")) +
    stat_function(fun = dcauchy, args = list(location =0, scale=2),   aes(colour = "2")) +
    stat_function(fun = dcauchy, args = list(location =-2, scale=1), aes(colour = "5")) +
    scale_color_manual("pdf", values = c("blue", "green", "red", "purple"),
                       labels=c(expression(paste("Cauchy(", alpha, "=0,", beta,"=0.5)")),
                                expression(paste("Cauchy(", alpha, "=0,", beta,"=1)")),
                                expression(paste("Cauchy(", alpha, "=0,", beta,"=2)")),
                                expression(paste("Cauchy(", alpha, "=-2,", beta,"=1)"))) ) +
    labs(x = "\n x", y = "f(x) \n") +
    theme(plot.title = element_text(hjust = 0.5),
        legend.position = "right")+
  theme_bw()
```



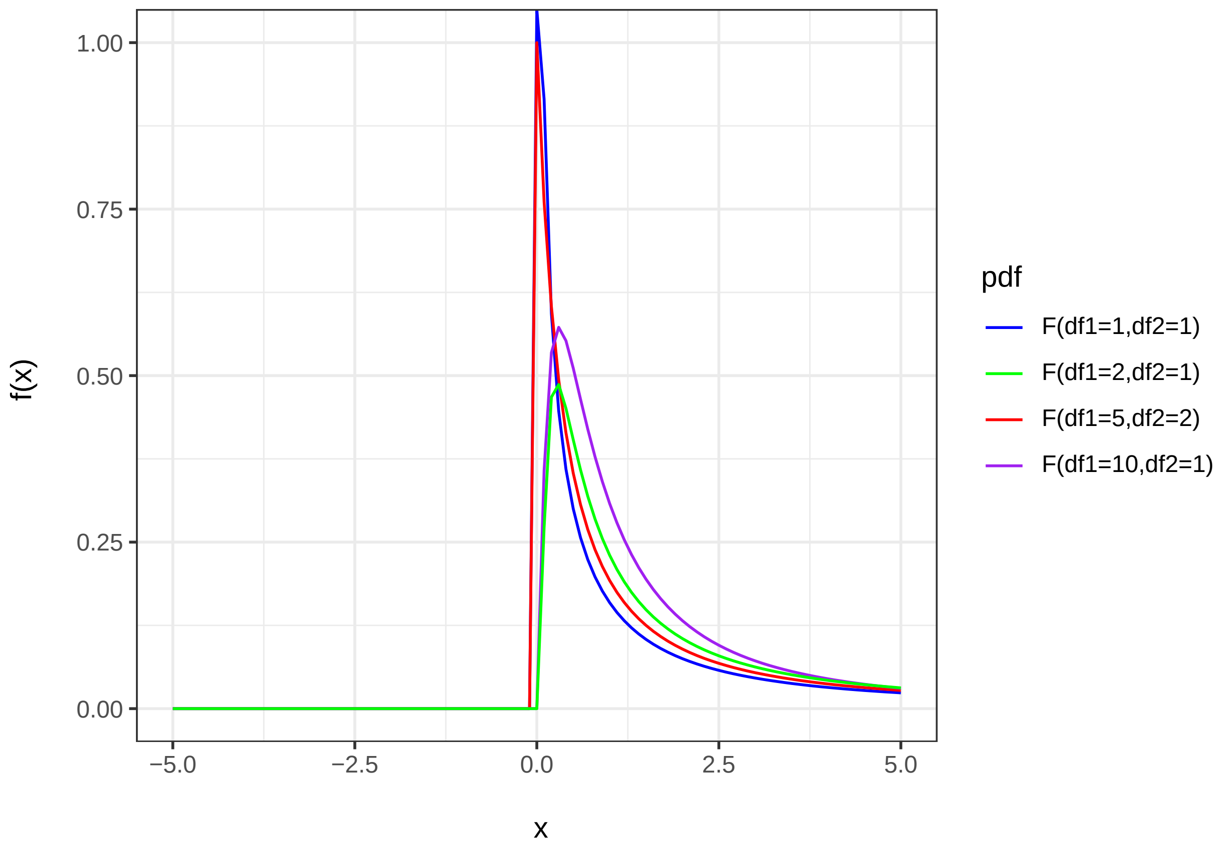

## 2.4 Distribución F

$$f(x;m,n) = \frac{\Gamma\left((m+n)/2\right)}{\Gamma\left(m/2\right)\Gamma\left(n/2\right)} \left(\frac{m}{n}\right)^{m/2} \frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}} I_{(0,\infty)}(x)$$

$m, n = 1, 2, ...$

$$\begin{aligned} E(X) &= \tfrac{n}{n-2},\ n>2 \\ V(X) &= \tfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)},\ n>4 \end{aligned}$$

```r
x_lower <- -5
x_upper <- 5


ggplot(data.frame(x = c(x_lower, x_upper)), aes(x = x)) + xlim(x_lower, x_upper) +
     #ylim(0, 0.7) +
     stat_function(fun = df, args = list(df1=1,  df2=1),  aes(colour = "1")) +
     stat_function(fun = df, args = list(df1=2,  df2=1),   aes(colour = "2")) +
     stat_function(fun = df, args = list(df1=5,  df2=2),   aes(colour = "5")) +
     stat_function(fun = df, args = list(df1=10, df2=1),   aes(colour = "10")) +
     scale_color_manual("pdf", values = c("blue", "green", "red", "purple"),
                        labels=c(expression(paste("F(", df1, "=1,", df2,"=1)")),
                                 expression(paste("F(", df1, "=2,", df2,"=1)")),
                                 expression(paste("F(", df1, "=5,", df2,"=2)")),
                                 expression(paste("F(", df1, "=10,", df2,"=1)"))) ) +
     labs(x = "\n x", y = "f(x) \n") +
     theme(plot.title = element_text(hjust = 0.5),
         legend.position = "right")+
    theme_bw()
```



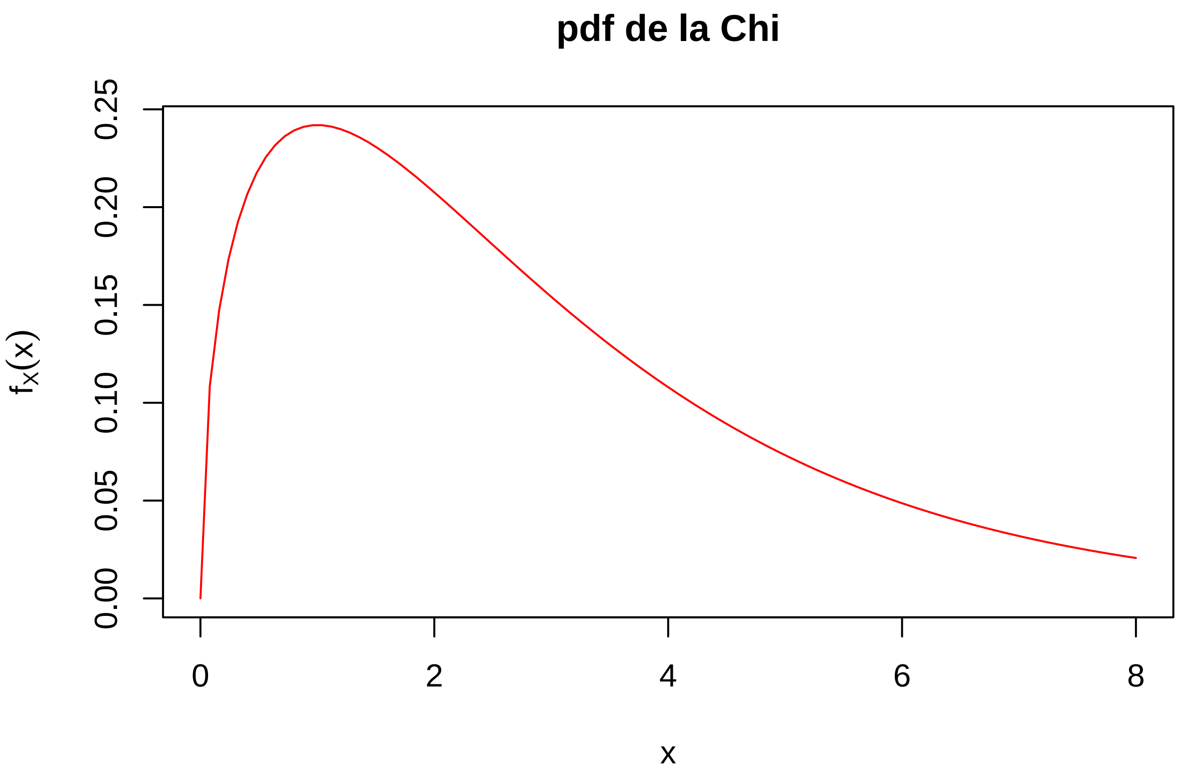

## 2.5 Chi-Cuadrada

$$f(x,k) = \frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{k/2} x^{k/2-1} e^{-(1/2)x} I_{(0,\infty)}(x)$$

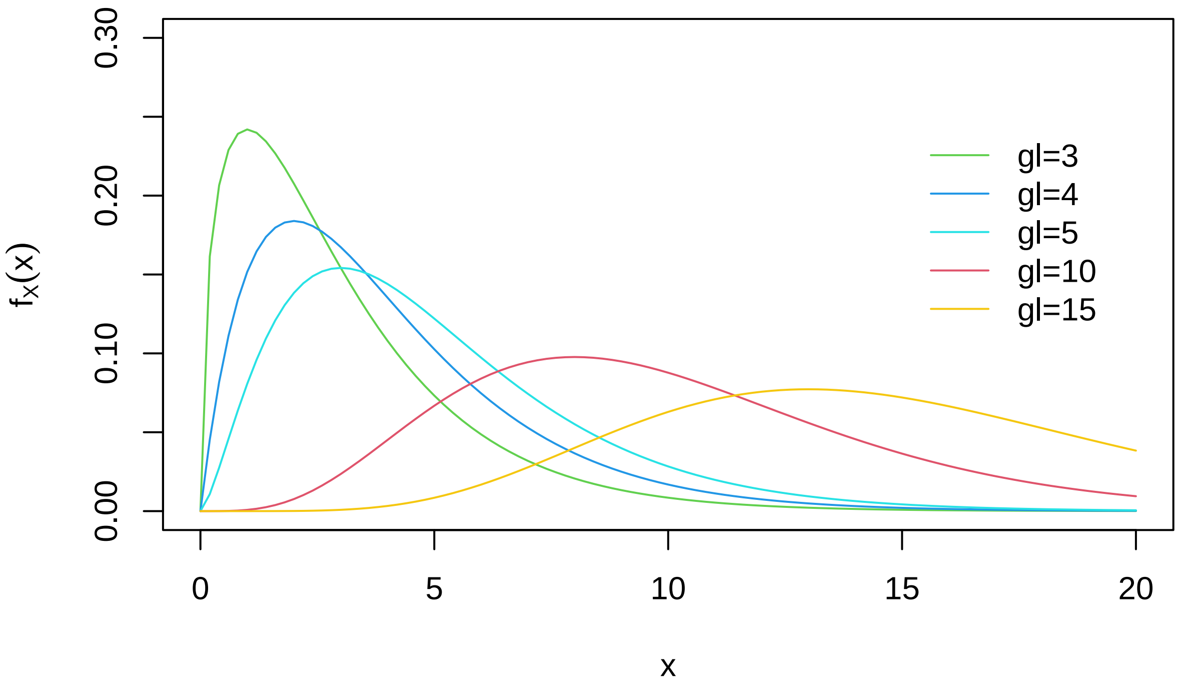
$k = 1, 2, 3..$

$$\begin{aligned} E(X) &= k \\ V(X) &= 2k \\ m_X(t) &= \left(\frac{1}{1-2t}\right)^{\frac{k}{2}}, \;\; t < 1/2 \end{aligned}$$

```
gl<-3

curve(dchisq(x, df=gl), from=0, to=8,
      ylab = expression(f[X](x)),
      col="red",main="pdf de la Chi")
```

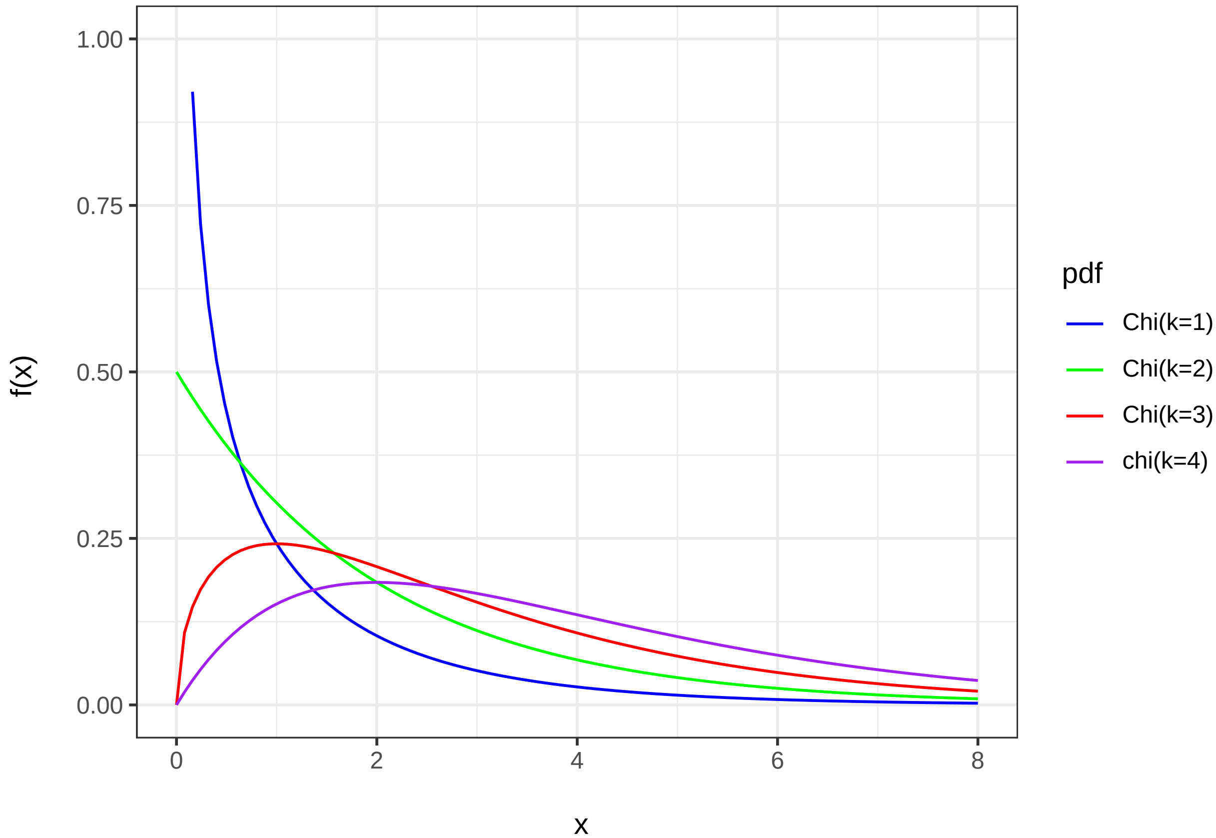



```
# Grupo de Chicuadradas
ind <- c(3,4, 5, 10, 15)

x1<-seq(0, 20,length=10)
y1<-seq(0,0.30,length=10)
plot(x1,y1,type = "n",ylab = expression(f[X](x)),xlab="x")

for (i in ind) curve(dchisq(x, df = i), from=0, to=20, col=i,
                     add =TRUE,ylim=c(0,.3))

leg.txt <- c("gl=3","gl=4","gl=5","gl=10","gl=15")
color<-c(3,4, 5, 10, 15)
legend(15,0.25, leg.txt, col=color, lwd=1, lty=1, bty="n")
```




```r
#title("PDF")
#leg.txt <- paste(expression(X^{2}(2)))
```

```r
x_lower <- 0
x_upper <- 8

ggplot(data.frame(x = c(x_lower, x_upper)), aes(x = x)) + xlim(x_lower, x_upper) +
     ylim(0, 1) +
     stat_function(fun = dchisq, args = list(df = 1),  aes(colour = "1")) +
     stat_function(fun = dchisq, args = list(df = 2),  aes(colour = "2")) +
     stat_function(fun = dchisq, args = list(df = 3),  aes(colour = "3")) +
     stat_function(fun = dchisq, args = list(df = 4),  aes(colour = "4")) +
     scale_color_manual("pdf", values = c("blue", "green", "red", "purple"),
                        labels=c(expression(paste("Chi(", k, "=1",")")),
                                 expression(paste("Chi(", k, "=2", ")")),
                                 expression(paste("Chi(", k, "=3",")")),
                                 expression(paste("chi(", k, "=4",")"))) ) +
     labs(x = "\n x", y = "f(x) \n") +
     theme(plot.title = element_text(hjust = 0.5),
         legend.position = "right")+
    theme_bw()
```




## 2.6 Momentos poblacionales

Una variable aleatoria puede ser caracterizada por sus momentos.

Momento central respecto al origen: Si $X$ es una v.a el r-ésimo momento de $X$, está dado por

$$\mu_r' = E(X^r)$$

si la esperanza, existe.

Momento central respecto a una constante: Si $X$ es una v.a el r-ésimo momento central de $X$ respecto a una constante $a$ es:

$$\mu_r = E[(X-a)^r]$$

Note que si $a = \mu_X$, entonces

$$\mu_r = E[(X-\mu_X)^r]$$

es el r-ésimo momento central de $X$ respecto a $\mu_X$.

Entonces el r-ésimo momento central de $X$ respecto a $\mu_X$ está dado por (Momento central respecto a $\mu$)

$$\mu_r = E[(X-\mu_X)^r]$$

Si $X$ es una v.a, y $E(X^2)$ existe, entonces

$$\sigma_X^2 = V(X) = E[(X-E(X))^2] = E(X^2) - (E(X))^2$$

La varianza es una medida de la dispersión de la variable aleatoria.

## 2.7 Momentos muestrales

## 2.8 Medida de asimetría

El tercer momento central de $X$ respecto a $\mu_X$ dado por

$$\mu_3 = E[(X - \mu_X)^3]$$

es una medida de asimetría o falta de asimetría (sesgo, skewness) de la distribución de la v.a $X$.

El sesgo de una pdf puede ser medido en términos de $\mu_3$:

- Si $\mu_3 < 0$, la pdf (pmf) es sesgada a la izquierda.
- Si $\mu_3 = 0$, la pdf (pmf) es simétrica respecto a $\mu_X$.
- si $\mu_3 > 0$, la pdf (pmf) es sesgada a la derecha.

En la práctica la simetría (o falta de simetría) de una pdf se puede calcular con el coeficiente de simetría dado por

$$\gamma = \frac{\mu_3}{\sigma^3}$$

$\gamma$ es una medida adimensional.

La mediana de una v.a $X$, denotada por $Mediana(X)$ o $x_{0.50}$ es el 0.50-ésimo cuantil.

La moda de una v.a $X$ ($Moda(X)$) es donde la $f_X(x)$ alcanza su valor máximo.

De las medidas se centralidad se puede establecer la siguiente relación:

- Simetría de la pdf (pmf):

$$\mu = Media(X) = Moda(X)$$

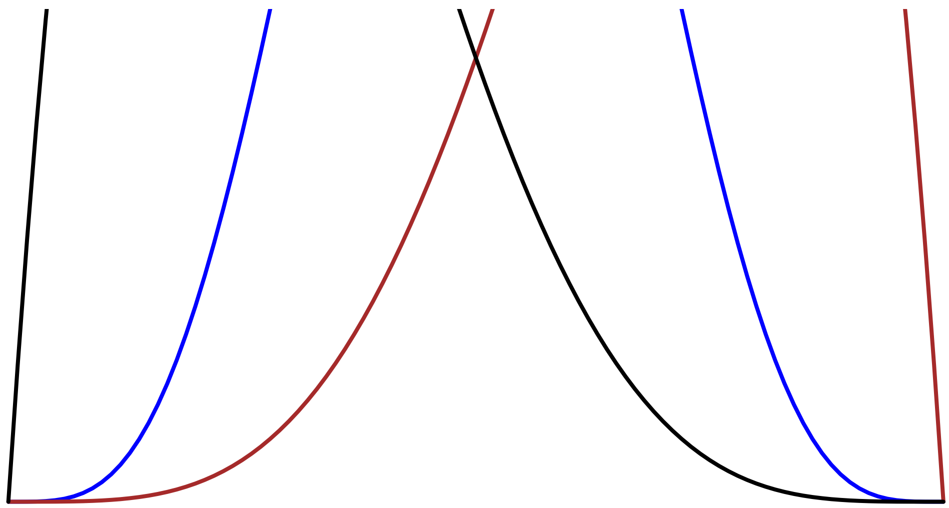
- Asimetría hacia la izquierda de la pdf (pmf):

$$\mu_X < Media(X) < Moda(X)$$

- Asimetría hacia la derecha de la pdf (pmf):

$$\mu_X > Media(X) > Moda(X)$$

```
plot.new()
curve(dbeta(x, 5, 5), from=0, to=1, col="blue",add = TRUE, ylim=c(0,5),lwd=2, ylab = expression(f[X](x)
curve(dbeta(x, 5, 2), from=0, to=1, col="brown", add = TRUE, ylim=c(0,5),lwd=2, ylab = expression(f[X](
curve(dbeta(x, 2, 5), from=0, to=1, col="black", add = TRUE, ylim=c(0,5),lwd=2, ylab = expression(f[X](
leg.txt<-c("Distribución simétrica",
           "Distribución sesgada a la izquierda (Asimetría negativa)",
           "Distribución sesgada a la derecha (Asimetría positiva)"
            )
color<-c("blue","brown","black")
legend(0.2,5, leg.txt, col=color, lwd=2, lty=1, bty="n")
```



# 3 Muestras aleatorias

*Definición (Población objetivo)*: La totalidad de los elementos que están bajo discusión y de los cuales se desea tener una información específica.

*Definición (Muestra aleatoria)*: Las variables aleatorias $X_1, X_2, ..., X_n$ es una muestra aleatoria de tamaño $n$ si

$$f(x_1, ..., x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

ES decir, si las variables aleatorias $X_1, X_2, ..., X_n$ son independientes y tienen la misma función de densidad, entonces se dice que la muestra, es *aleatoria*.

Las observaciones $x_1, x_2, ..., x_n$ se dice que constituyen una muestra aleatoria si $x_1, x_2, ..., x_n$ son los valores de $X_1, X_2, ..., X_n$, y estas variables son una muestra aleatoria.

La siguiente definición también nos puede ayudar a entender lo que es una muestra aleatoria.

*Definición*: Las variables aletorias $X_1, X_2, ..., X_n$ son llamadas una muestra aleatoria de tamaño $n$ de la población $f(x)$ si las $X_1, X_2, ..., X_n$ son variables aleatorias mutuamente independientes y la función de densidad es la misma $f(x)$. Alternativamente $X_1, X_2, ..., X_n$ son llamadas variables aleatorias independientes e idénticamente distribuidas con función de densidad $f(x)$.

Sea

$$X\,|\theta \sim f(\cdot)$$

a partir de $x = (x_1, ..., x_n)^t$, se quiere hacer inferencia sobre $\theta$.

## 3.1 Medidas de centralidad

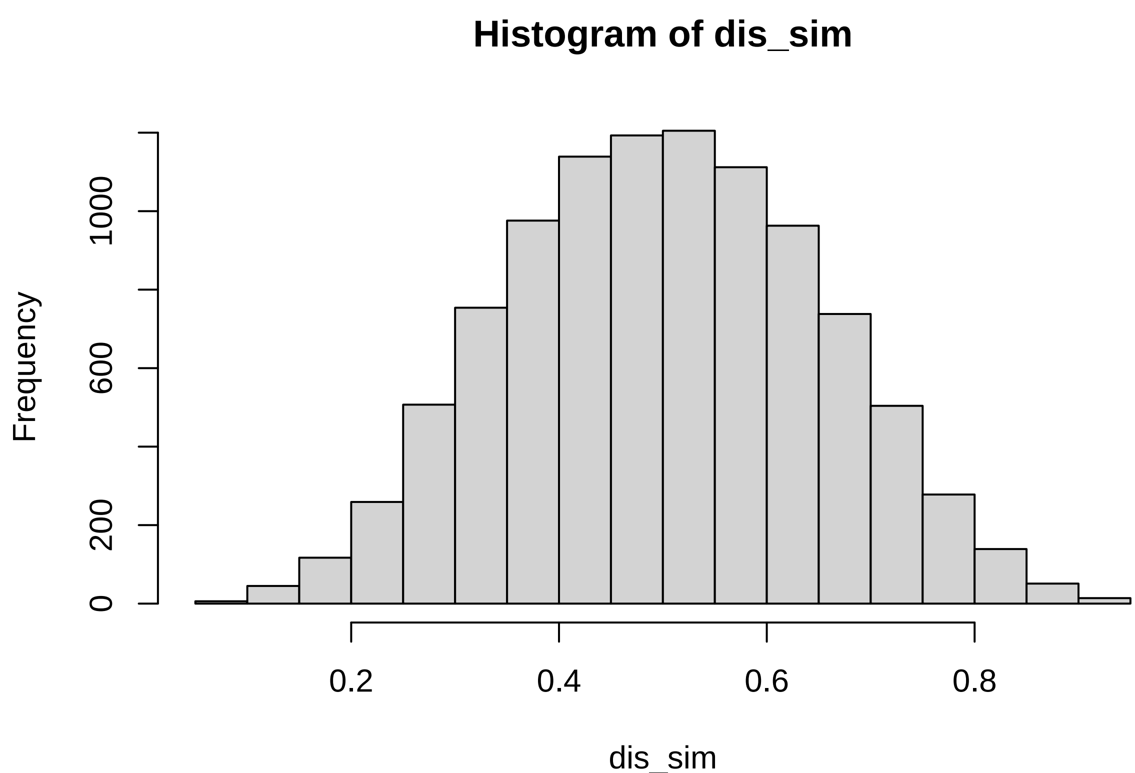
Dada una muestra aleatoria, $x_1, ..., x_n$, el *promedio* se define como

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

La *Media* de una muestra: Ordenando los datos, y el la observación que queda en el medio es la media sie el número de datos es impar, si es par entonces es el promedio de los dos datos centrales.

La *Moda* de una muestra es la observación con mayor frecuencia

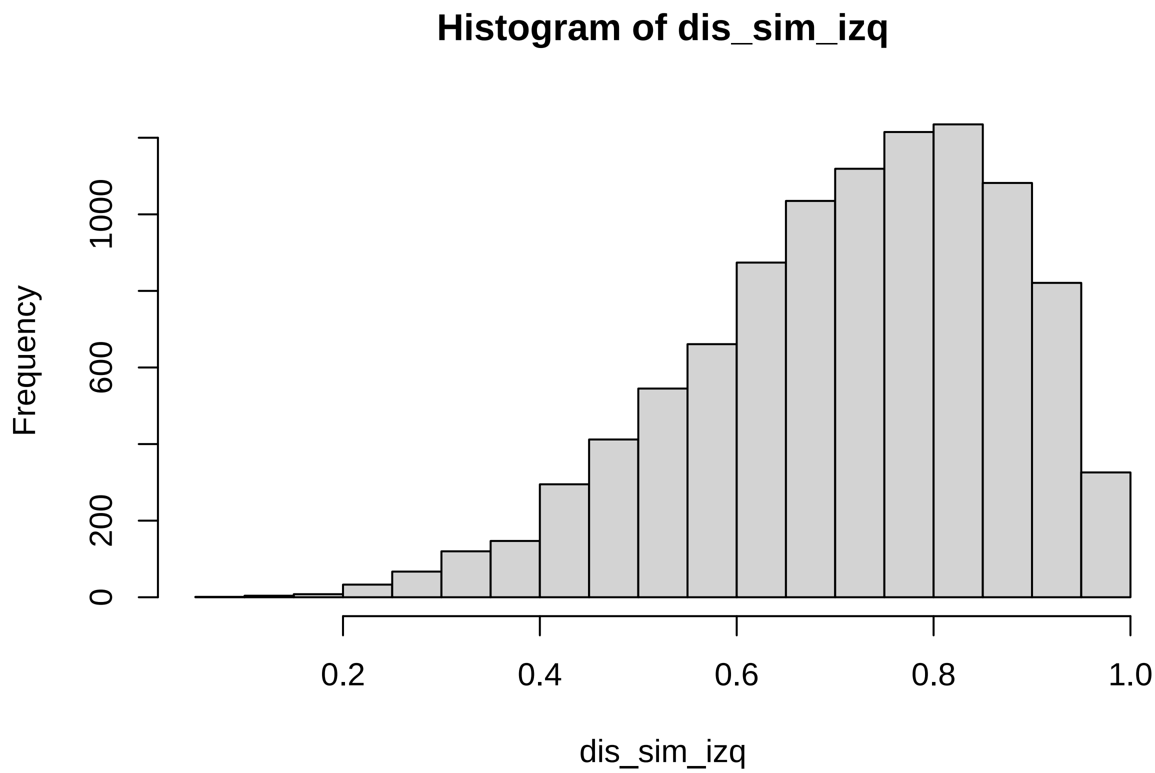
```
set.seed(100)
dis_sim<-rbeta(10000, 5,5) # simétrica
hist(dis_sim)
```




```
skewness(dis_sim)    # Sesgo
```

```
## [1] 0.03041756
```

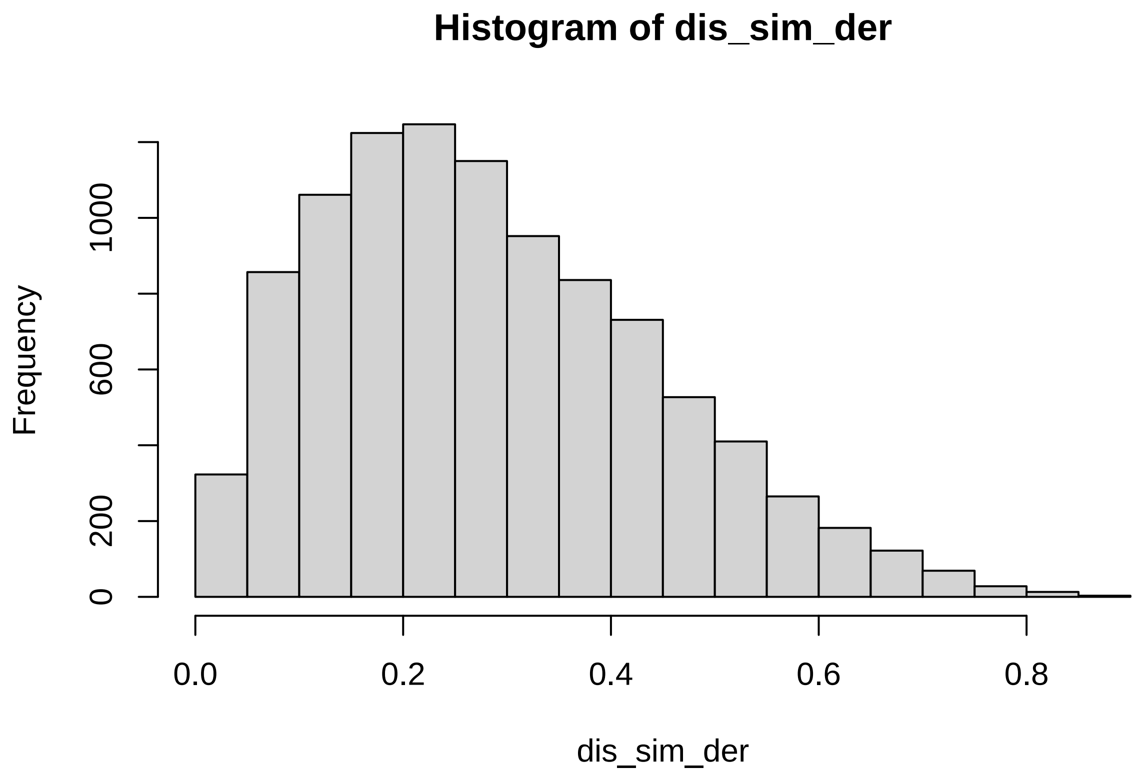
```
dis_sim_izq<-rbeta(10000, 5,2)  # sesgada a la izquierda (Asimetría negativa)
hist(dis_sim_izq)
```




```r
skewness(dis_sim_izq)    # Sesg
```

```
## [1] -0.6044371
```

```r
dis_sim_der<-rbeta(10000, 2,5)  # sesgada a la derecha (Asimetría positiva)
hist(dis_sim_der)
```




```r
skewness(dis_sim_der)     # Sesgo
```

```
## [1] 0.6036976
```

Cálculo de la media, moda y mediana

```r
mean(dis_sim)
```

```
## [1] 0.501277
```

```r
median(dis_sim)                  #B) Mediana
```

```
## [1] 0.5000674
```

```r
#quantile(m.a_x, probs=0.5)     #C) Mediana o Segundo cuartil

Mode(dis_sim)    # Ningun dato se repita
```

```
## [1] NA
## attr(,"freq")
## [1] NA
```

## 3.2 Medidas de dispersión

Dada una muestra aleatoria $x_1, , , , x_n$, un estimador insesgado de la varianza $\sigma^2$, está dado por

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

## 3.3 Prueba de hipótesis

Existen en la literatura lo que se conoce como juego de hipótesis, las cuales están diseñadas para probar la significancia de parámetros.

Así dado un parámetro $\theta$, de la función de densidad $f(x,\theta)$ con $\theta \in \Theta$

Se puede formular los siguientes juegos de hipótesis, para $\theta_0,\ \theta_1 \in \Theta$:

Simple vs Simple

$$H_0 : \theta = \theta_0 \ \ vs \ \ H_1 : \theta = \theta_1$$

Simple vs Compuestas de una cola

$$H_0 : \theta = \theta_0 \ \ vs \ \ H_1 : \theta < \theta_0$$

$$H_0 : \theta = \theta_0 \ \ vs \ \ H_1 : \theta > \theta_0$$

Simple vs Compuestas de dos colas

$$H_0 : \theta = \theta_0 \ \ vs \ \ H_1 : \theta \neq \theta_0$$

Compuesta vs Compuestas de una cola

$$H_0 : \theta < \theta_0 \ \ vs \ \ H_1 : \theta > \theta_0$$

En un juego de hipótesis se puede cometer dos tipos de errores:

1. $$\alpha = P(Error\ Tipo\ I) = P(Rechazar\ H_0\ cuando\ es\ cierta)$$

2 $$\beta = P(Error\ Tipo\ II) = P(No\ rechazar\ H_0\ cuando\ es\ falsa)$$

## 3.4 Comparación de medias de muestras independientes

Para la comparación de medias, se va a utilizar pruebas paramétricas, cuando los supuestos de normalidad y homogeneidad de varianzas no se cumple se usan pruebas no paramétricas:

- Prueba de U de Mann-Whitney: para dos muestras independientes
- Prueba de Wilcoxon: para dos muestras relacionadas
- Prueba de Kruskal-Wallis: para tres o más muestras independientes

### 3.4.1 Una muestra

Dada $X_1, .., X_n$ una muestra aleatoria tal que

$$X_i \sim N(\mu, \sigma^2)$$

Note que la muestra viene de v.a con distribución normal. Para hacer inferencia sobre el parámetro $\mu$, se puede plantear los siguientes juegos de hipótesis:

1. $$H_0 : \mu \geq \mu_0 \;\; vs \;\; H_1 : \mu < \mu_0$$
2. $$H_0 : \mu \leq \mu_0 \;\; vs \;\; H_1 : \mu > \mu_0$$
3. $$H_0 : \mu = \mu_0 \;\; vs \;\; H_1 : \mu \neq \mu_0$$

Si $\sigma^2$ es conocida, entonces el estadístico de prueba está dado por

$$Z_0 = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$$

Bajo $H_0$, $Z_0$ tiene distribución:

$$Z_0 \sim N(0, 1)$$

Si $\sigma^2$ es desconocida, entonces el estadístico de prueba está dado por

$$t_0 = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$$

bajo $H_0$

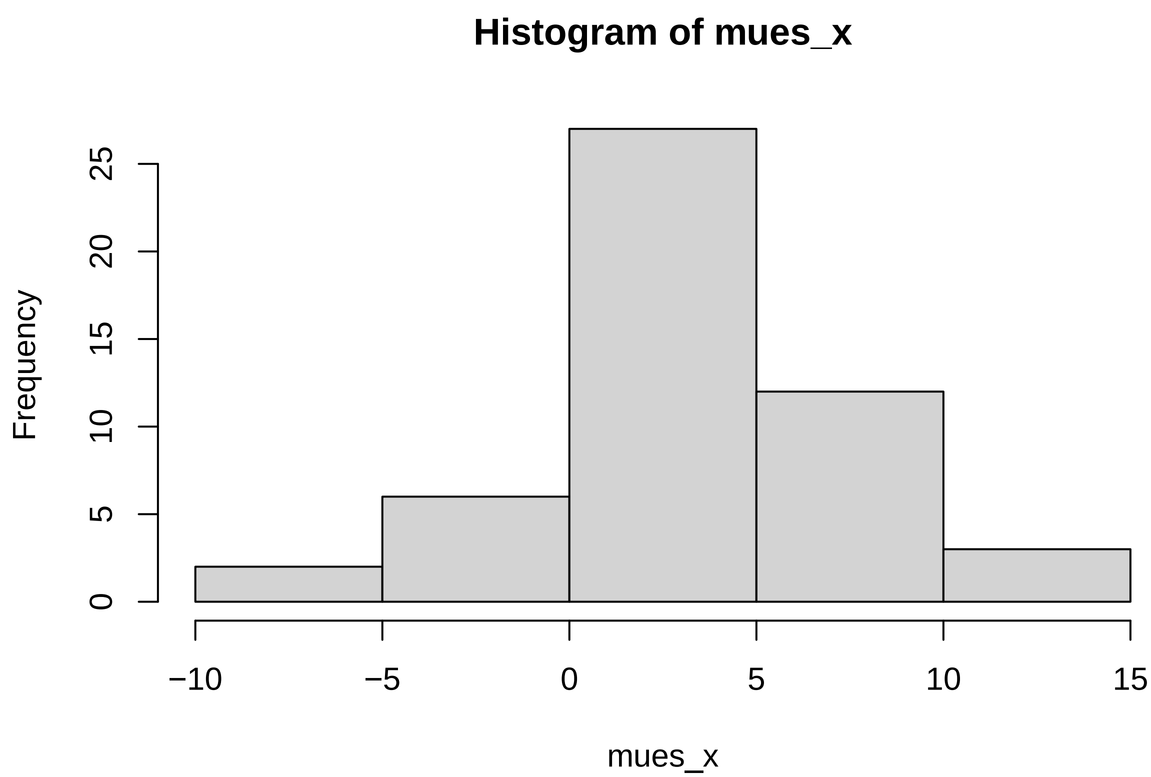
$$t_0 \sim t_{n-1}$$

Luego

1. Se rechaza $H_0$ si $t_0 < -t_{\alpha,n-1}$, lo cual es equivalente a rechazar $H_0$, si
$$p - valor = P(t_{\alpha,n-1} < -t_0)$$
2. Se rechaza $H_0$ si $t_0 > t_{\alpha,n-1}$, lo cual es equivalente a rechazar $H_0$, si
$$p - valor = P(t_{\alpha,n-1} > t_0)$$
3. Se rechaza $H_0$ si $|t_0| \geq t_{\alpha/2,n-1)}$, lo cual es equivalente a rechazar $H_0$, si
$$p - valor = P(t_{\alpha,n-1} \leq -t_0) + P(t_{\alpha,n-1} \geq t_0)$$

En R, la función t.test{stats} implementa el estadístico de prueba Student's t

```r
set.seed(100)

mu<-3

sigma<-5

mues_x<-rnorm(50, mu, sigma)

hist(mues_x)
```




```
mean(mues_x)
```

```
## [1] 3.407545
```

```
sd(mues_x)
```

```
## [1] 4.095056
```

Supongamos que $\mu_0 = 1$, con

$$H_0 : \mu \geq 1 \ \ vs \ \ H_1 : \mu < 1$$

```
t.test(x=mues_x,
       alternative = "less",   #  H_1: mu<1
       mu = 1,
       paired = FALSE,
       var.equal = FALSE,
       conf.level = 0.95)   # alpha=0.05
```

```
##
##  One Sample t-test
##
## data:  mues_x
```

```
## t = 4.1572, df = 49, p-value = 0.9999
## alternative hypothesis: true mean is less than 1
## 95 percent confidence interval:
##      -Inf 4.378483
## sample estimates:
## mean of x
##  3.407545
```

De los resultados se observa que $t_0 = 4.1572$, ya que realizando los cálculos se tiene

$$t_0 = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} = \frac{\sqrt{50}(3.407545 - 1)}{4.095056} = 4.157187$$

Por otra parte, a un nivel de significancia $\alpha = 0.05$, el $t$ de tablas está dado por

```r
sqrt(50)*( 3.407545-1)/4.095056
```

```
## [1] 4.157187
```

```r
# Cálculo del t de tablas (cuantil)

qt(0.95, 49)   # alpah=0.05, 50-1=49  t_(alpha,n-1)
```

```
## [1] 1.676551
```

Se rechaza $H_0$ si $t_0 < -t_{\alpha,n-1}$ como

$$t_0 = 4.1572 \nless -t = -1.676551$$

entonces no se rechaza $H_0$, luego $\mu \geq 1$.

Usando el $p - valor$, se tiene

$$p - valor = P(t_{\alpha,n-1} < -t_0) = P(t_{\alpha,n-1} < -4.1572) = 0.9999$$

Cómo la distribución t es simétrica, entonces

```r
# Cálculo de p-valor
pt(4.1572,49)  # 50-1=49
```

```
## [1] 0.9999355
```

es decir $p - valor = 0.9999$, se sabe que se rechaza $H_0$ si el $p - valor < \alpha = 0.05$, como no se cumple la desigualdad, entonces no se rechaza $H_0$.

### 3.4.2 Dos muestras

Comparación de medias de dos poblaciones mediante dos muestras aleatorias independientes:

Sea $x_1, ..., x_n$ una muestra aleatoria de una población de $X \sim N(\mu_X, \sigma^2)$; y sea $y_1, ..., y_m$ una muestra aleatoria de una población de $Y \sim N(\mu_Y, \sigma^2)$; ambas muestras independientes; con varianzas iguales además de tener independencia entre las observaciones dentro de cada muestra.

Juegos de hipótesis:

1. $$H_0 : \mu_X \geq \mu_Y \;\; vs \;\; H_1 : \mu_X < \mu_Y$$
2. $$H_0 : \mu_X \leq \mu_Y \;\; vs \;\; H_1 : \mu_X > \mu_Y$$
3. $$H_0 : \mu_X = \mu_Y \;\; vs \;\; H_1 : \mu_X \neq \mu_Y$$

   Bajo $H_0$ y $\sigma^2$ conocida,

se tiene que el estadístico de prueba esta dado por

$$Z_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}$$

donde

$$Z_0 \sim N(0,1)$$

Generalmente, no se conoce $\sigma^2$, entonces se usa el estimador

$$t_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

Que bajo $H_0$, la distribución del estadístico de prueba está dado por

$$t_0 \sim t(n+m-2)$$

Luego

1. Se rechaza $H_0$ si $t_0 < -t_{\alpha,(n+m-2)}$; lo cual es equivalente a rechazar $H_0$, si
$$p-valor = P(t_{\alpha,(n+m-2)} < -t_0)$$
2. Se rechaza $H_0$ si $t_0 > t_{\alpha,(n+m-2)}$, lo cual es equivalente a rechazar $H_0$, si
$$p-valor = P(t_{\alpha,(n+m-2)} > t_0)$$
3. Se rechaza $H_0$ si $|t_0| \geq t_{\alpha/2,(n+m-2))}$, lo cual es equivalente a rechazar $H_0$, si
$$p-valor = P(t_{\alpha,(n+m-2)} \leq -t_0) + P(t_{\alpha,(n+m-2)} \geq t_0)$$

```
set.seed(100)

mu_x<-3
sigma_x<-5
mues_x<-rnorm(30,mu_x,sigma_x)



mu_y<-0
sigma_y<-5
mues_y<-rnorm(50,mu_y,sigma_y)


var(mues_x)
```

```
## [1] 12.3185
```

```r
var(mues_y)
```

```
## [1] 32.57887
```

```r
t.test(x=mues_x, y=mues_y,

       alternative = "less",  #  H_1: mu_x < mu_y
       paired = FALSE,
       var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  mues_x and mues_y
## t = 2.9847, df = 77.927, p-value = 0.9981
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 4.791756
## sample estimates:
##  mean of x  mean of y
## 3.14431988 0.06819164
```

De los resultados se observa que $t_0 = 2.9847$, ya que realizando los cálculos se tiene

$$t_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} = \frac{3.14431988 - 0.06819164}{\sqrt{\frac{12.3185}{30} + \frac{32.57887}{50}}} = 2.984712$$

```r
(3.14431988-0.06819164)/sqrt(25/30 + 25/50) # valor t_0; con varianzas iguales
```

```
## [1] 2.664005
```

```r
(3.14431988-0.06819164)/sqrt(12.3185/30 + 32.57887/50) # valor t_0, con varianzas diferentes
```

```
## [1] 2.984712
```

Por otra parte, a un nivel de significancia $\alpha = 0.05$, el $t$ de tablas está dado por

```r
# Cálculo del t de tablas (cuantil)

qt(0.95, 77.927)   # alpah=0.05, gl=77.927
```

```
## [1] 1.664643
```

Se rechaza $H_0$ si $t_0 < -t_{\alpha,n+m-2}$ como

$$t_0 = 2.9847 \not< -t = -1.664643$$

entonces no se rechaza $H_0$, luego $\mu_x \geq \mu_y$.

Usando el $p-valor$, se tiene

$$p-valor = P(t_{\alpha,n+m-2} < -t_0) = P(t_{\alpha,n+m-2} < -2.9847) = 0.9981047$$

Cómo la distribición t es simétrica, entonces

```r
# Cálculo de p-valor
pt(2.9847,77.927)
```

```
## [1] 0.9981047
```

es decir $p - valor = 0.9981047$, se sabe que se rechaza $H_0$ si el $p - valor < \alpha$, como no se cumple la desigualdad, entonces no se rechaza $H_0$.

La comparación de medias en muestras relacionadas, también se define un juego de hipótesis y un estadístico de prueba para probar el juego de hipótesis.

## 3.5 Coeficiente de correlación

El coeficiente de correlación para dos variables continuas (cunatitatias) dadas, $X$ y $Y$, está dada por

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

El coeficiente de correlación:

- Mide la relación lineal que existe entre $X$ y $Y$.
- Está entre: $-1 \leq \rho \leq 1$.
- No depende de la escala de las variables $X$ y $Y$.
- $\rho$ es invariante bajo transformaciones de escala y localidad.
- A mayor asociación lineal entre las v.a $X$ y $Y$, más cercano estará $\rho$ de $-1$ o 1.
- Si $X$ y $Y$ son v.a con distribución normal y $\rho = 0$, entonces $X$ y $Y$ son independientes, en las variables puede existir otro tipo de relaciones.
- Si $\rho = 0$ entonces no existe asociación lineal entre $X$ y $Y$.
- Si $\rho = -1$ se tiene una relación lineal perfecta negativa entre $X$ y $Y$.
- Si $\rho = 1$ se tiene una relación lineal perfecta positiva entre $X$ y $Y$.

Dado $x_1, ..., x_n$ una muestra aleatoria de $X$ y $y_1, ..., y_n$ una muestra aleatoria de $Y$, se tiene lo que se conoce como coeficiente de correlación muestral de Pearson:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

a partir de este coeficiente se pude identificar si existe correlación entre las variables $X$ y $Y$:

Con los siguientes juegos de hipótesis:

1. $H_0 : r = 0$ vs $H_1 : r > 0$
2. $H_0 : r = 0$ vs $H_1 : r < 0$
3. $H_0 : r = 0$ vs $H_1 : r \neq 0$

A un nivel de significancia $\alpha = 0.05$, se rechaza $H_0$ si $p - valor < \alpha$.

```r
cor.test(iris$Petal.Length,iris$Petal.Width)
```

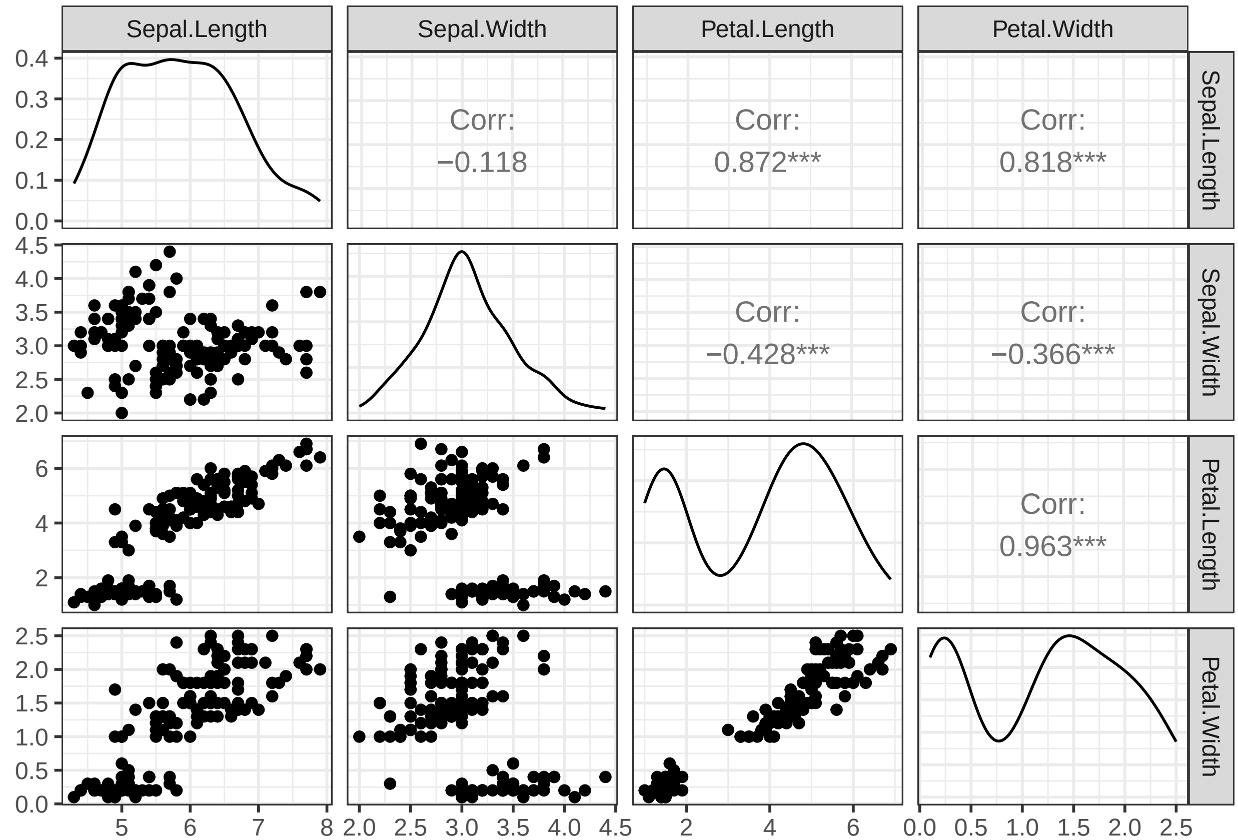
```
##
##  Pearson's product-moment correlation
##
## data:  iris$Petal.Length and iris$Petal.Width
## t = 43.387, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9490525 0.9729853
## sample estimates:
##       cor
## 0.9628654
```

```r
p_s.w_p.w <- cor.test(iris$Sepal.Width,iris$Petal.Width)

p_s.w_p.w$p.value
```

```
## [1] 4.073229e-06
```

De acuerdo al juego de hipótesis $H_0 : r = 0$ vs $H_1 : r \neq 0$, se rechaza $H_0$ si el $p - valor < \alpha$, donde $\alpha = 0.05$, como p-valor= 0 y es menor al $\alpha$, entonces se rechaza $H_0$, es decir el coeficiente de correlación es distinto de cero, por lo tanto existe una relación lineal entre Sepal.Width y Petal.Width.

```r
iris%>%
  select_if(is.numeric)%>%
  ggpairs()+
  theme_bw()
```




Códigos de significancia.:

- $-$ para $\alpha = 0.05$
- ** para $\alpha = 0.01$
- *** para $\alpha = 0.001$

Variables Sepal.Width y Sepal.Length.

```r
cor.test(iris$Sepal.Width ,iris$Sepal.Length) # -0.1175698
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Sepal.Width and iris$Sepal.Length
## t = -1.4403, df = 148, p-value = 0.1519
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.27269325  0.04351158
## sample estimates:
##        cor
## -0.1175698
```

Para las variables Sepal.Width y Sepal.Length existe una correlación de $r = -0.1175698$, con un $p-valor = 0.1519$. De acuerdo con el la regla de decisión; se rechaza $H_0$, si p-valor$< \alpha = 0.05$.

Así en este caso no hay correlación entre las variables Sepal.Width y Sepal.Length.

Variables Petal.Length y Sepal.Length:

```
cor.test(iris$Petal.Length ,iris$Sepal.Length) # 0.8717538
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Petal.Length and iris$Sepal.Length
## t = 21.646, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8270363 0.9055080
## sample estimates:
##       cor
## 0.8717538
```

Para las variables Petal.Length y Sepal.Length correlación 0.8717538, con un p-valor=2.2e-16;

Regla: Se rechaza $H_0$ si p-valor$< \alpha = 0.05$

En este caso se rechaza $H_0 : \rho = 0$, luego existe una correlación lineal positiva entre estas variables.

# 4 Pruebas de normalidad para una muestra

Dada una muestra independiente, $x_1, ..., x_n$ de una variable aleatoria, se quiere saber si dicha muestra proviene de una distribución Normal.

En general, dependiendo del tamaño y del grado de simetría que presenta la muestra , se puede utilizar una u otra prueba estadística para probar normalidad.

Pruebas para evaluar la prueba de normalidad en una muestra aleatoria:

1. Kolmogorov-Smirnov (K-S) (Para muestras grandes)
2. Lilliefors corrected K-S
3. Shapiro-Wilk test (para muestras pequeñas)
4. Anderson-Darling
5. Cramer-von Mises
6. D'Agostino skewness
7. Anscombe-Glynn kurtosis
8. D'Agostino-Pearson omnibus
9. Jarque-Bera

Cada una de estas pruebas tiene asociada un estadístico de prueba, cuya distribución está determinada por la hipótesis nula del juego de hipótesis que se busca probar.

Nivel de significancia Juego de hipótesis Estadístico de prueba regla de decisión

Funciones para probar normalidad

```
#--- H_0: Los errores tienen distribución Normalidad

shapiro.test(res)  # Shapiro-Wilk  Muestras n<=50
ad.test(res)       # Anderson-Darling

cvm.test(res)      # Cramer-von Mises
sf.test(res)       # Shapiro-Francia
lillie.test(res)   # Lilliefors (Kolmogorov-Smirnov) Los errores tienen distribución normal y Muestras

pearson.test(res)
```

Transformación de Box-Cox ayuda con la normalización de los datos y homogeneidad de varianzas, para implementarla se puede utilizar la función *powerTransform* del paquete *car*.

```
lam<-powerTransform(dat$Densidad_basica_mad)$lambda

dat$Y<-((dat$Densidad_basica_mad^lam)-1)/lam
```