Visualización de datos y funciones de distribución

Francisco Paz

16/8/2019

Visualización de datos

Utilizaremos la paqueteria de **ggplot2** (Hadley Wickham)también contenido en la librería **tidyverse** para mostrar algunas ideas de visualización de datos. La idea es dar una pequeña guía de buenas prácticas para creación de gráficas.

Utilizaremos el conjunto de datos (¡muy conocido!) *iris*

```
library(tidyverse)
library(pander)
library(gridExtra)
pander(head(iris), caption = 'Un viztaso a la base de datos')
```

Table 1: Un viztaso a la base de datos

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Veamos una explicación acerca de las variables.

```
knitr::include_graphics("iris2.png")
```

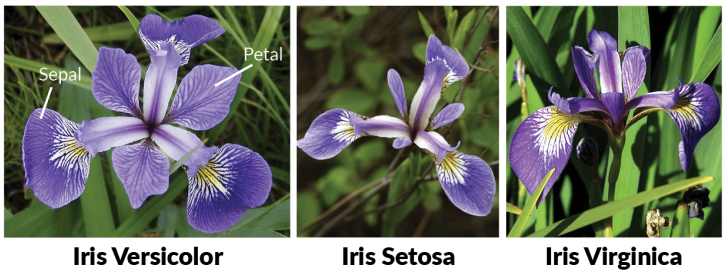
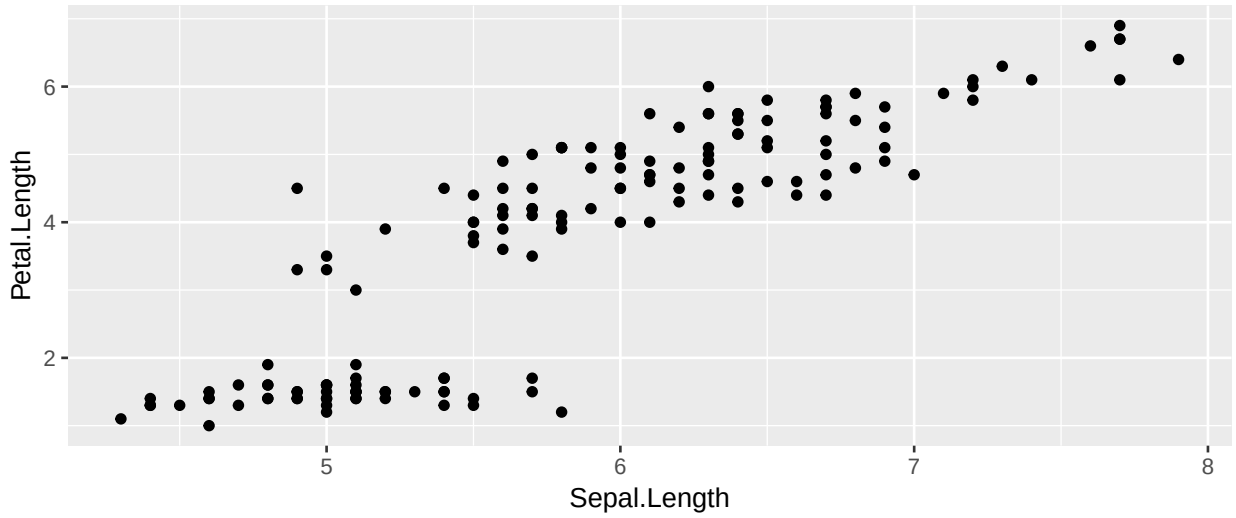


Figure 1: Iris data set https://medium.com/@srishtisawla/iris-flower-classification-fb6189de3fff

A continuación presentaremos distintas gráficas, la idea es nombrar todos aquellos errores/mejoras que podemos encontrar aquí.

Para empezar tenemos la siguiente gráfica de dispersión donde vemos el largo del sepal vs el largo del petal (claro que esta no es la única configuración posible).

```
ggplot(iris,aes(x= Sepal.Length,y=Petal.Length)) + geom_point()
```



Aplicando algunas de las sugerencias

```
ggplot(iris,aes(x= Sepal.Length,y=Petal.Length)) +
  geom_point(aes(col = Species)) + labs(x = 'Sepal', y = 'Petal',
      title = 'Sepal length vs petal length')
```

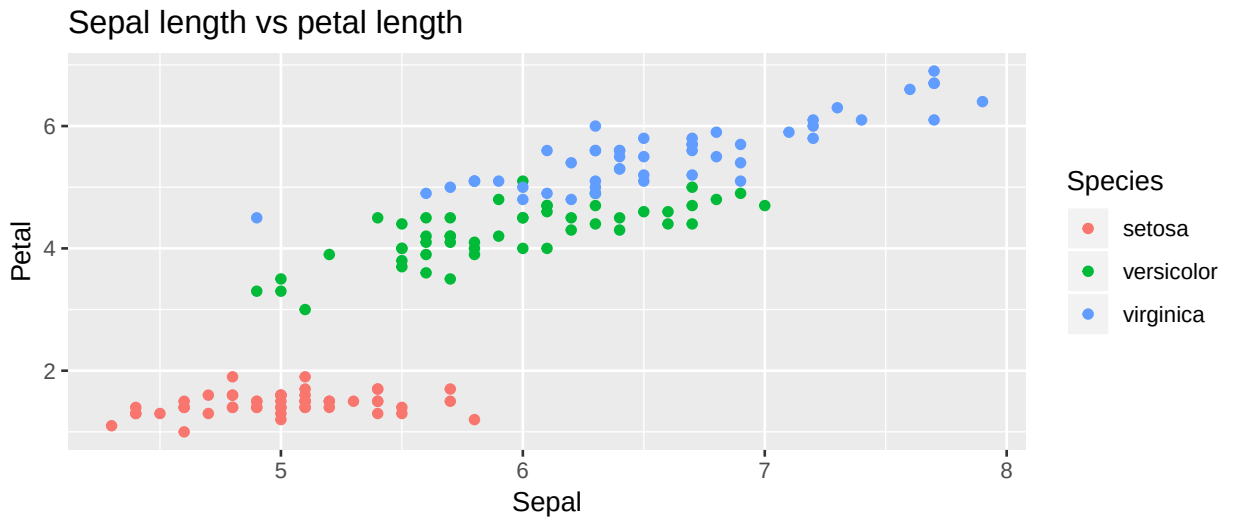


Figure 2: Gráfica sepal length vs petal length para los 3 tipos de flores del data set iris

Siempre es una buena idea que una gráfica este autocontenida, es decir, que podamos extraer toda la información (si somos los que estamos reportando, al menos la información acerca del fenómeno que queremos mostrar) sin necesidad de recurrir a la base de datos.

Empezamos a ver mucha mejora respecto a la gráfica anterior, pero, ¿hay algo más que podamos hacer para mejorar la gráfica?. Aunque parezca poco común problemas visuales se presentan en muchas personas, e incluso sin ir tan lejos, podemos pensar en los defectos que nuestra gráfica puede tener al ser vista en distintos dispositivos (inlcuso en el proyector)

```
ggplot(iris,aes(x= Sepal.Length,y=Petal.Length,shape=Species)) +
  geom_point(aes(col = Species)) + labs(x = 'Sepal', y = 'Petal',
```

```
      title = 'Sepal length vs petal length')
```

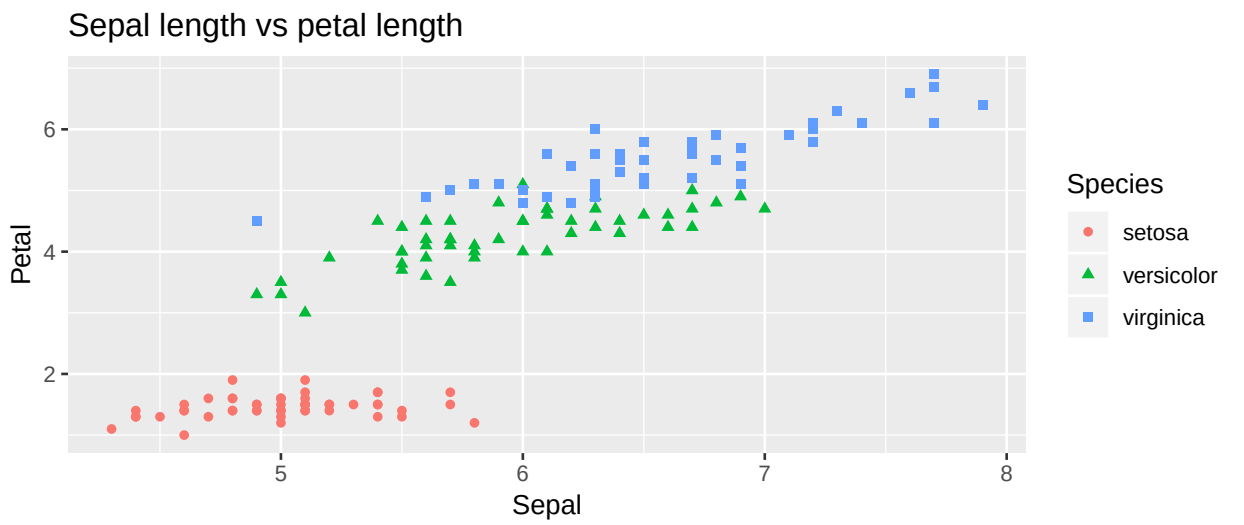


Figure 3: Gráfica sepal length vs petal length para los 3 tipos de flores del data set iris, todas las variables numéricas están en cm

Otra buena idea para mejorar esto, es jugar con la escla de blancos y negros.

¿Cuántas dimensiones tiene nuestra gráfica?

¿Podemos agregar más?

```
ggplot(iris,aes(x=Sepal.Length,y=Petal.Length,
                size=Petal.Width, shape=Species)) +  geom_point(aes(col = Sepal.Width)) +
  labs(x = 'Sepal', y = 'Petal', title = 'Sepal length vs petal length')
```

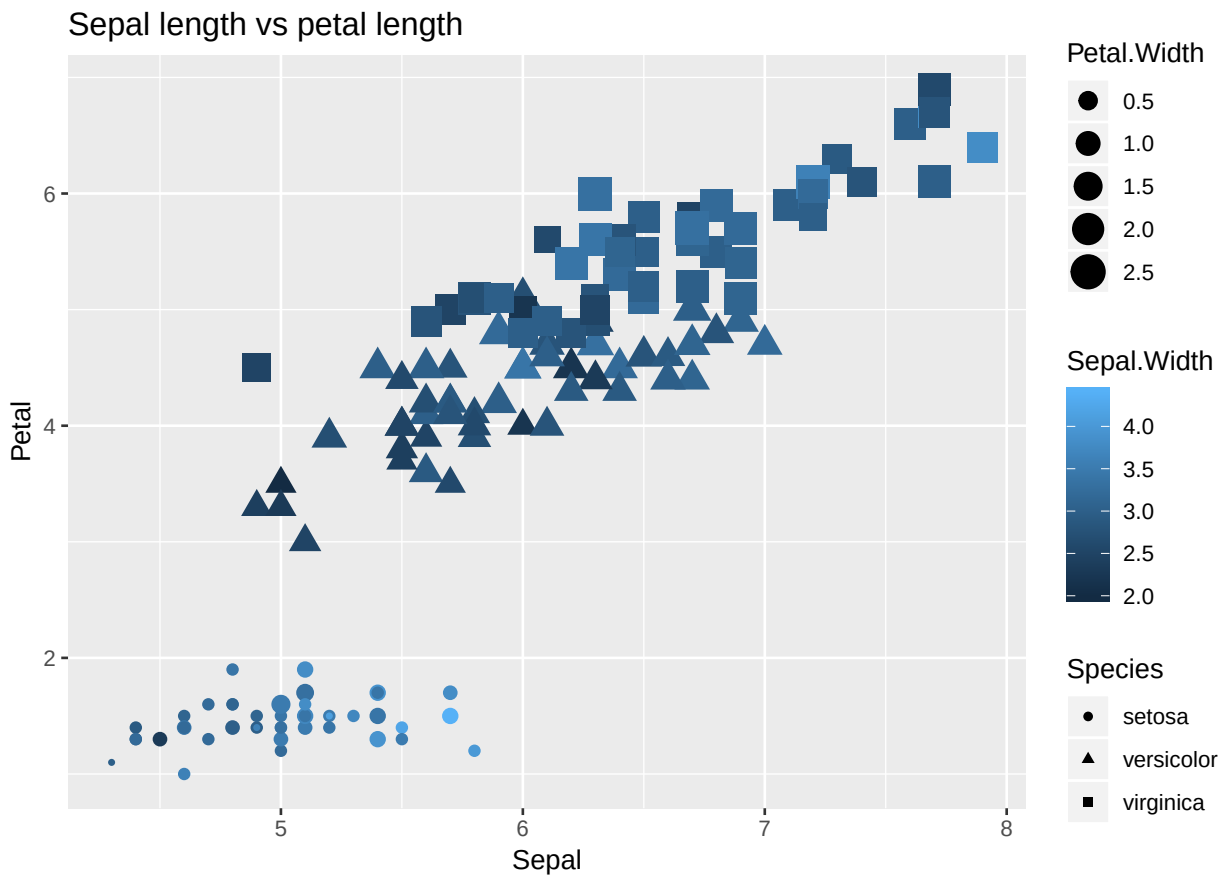


Figure 4: Gráfica sepal length vs petal length para los 3 tipos de flores del data set iris, todas las variables numéricas están en cm

La visualización de datos nos ayuda a entender en una primera instancia una base de datos, pero en mi opinión, de lo más importante es el poder plantearnos ciertas conjeturas/hipótesis acerca del problema ¿qué estoy viendo?¿qué necesito saber?¿cuál es el fin de este trabajo?

¿Cuál sería una buena pregunta para esta base?

¿Hay algo que te gustaría saber?

Algunas funciones de distribución

Complementando lo visto en clase, veremos algunos ejemplos de funciones de distribución acumulada (FDA) y la función de distribución tanto discretas como continuas.

Distribución binomial

Función de distribución

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

Función de distribución acumulada

$$F(x) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

```r
densidades <- ggplot(data.frame(x = -1:20)) +
  geom_point(aes(x = x, y = dbinom(x, size = 20, prob = 0.5), color = "n=20;p=0.5"),
      show.legend = FALSE) +
  geom_path(aes(x = x, y = dbinom(x, size = 20, prob = 0.5), color = "n=20;p=0.5"),
      alpha = 0.6, linetype = "dashed", show.legend = FALSE) +
  geom_point(aes(x = x, y = dbinom(x, size = 20, prob = 0.1), color = "n=20;p=0.1"),
      show.legend = FALSE) +
  geom_path(aes(x = x, y = dbinom(x, size = 20, prob = 0.1), color = "n=20;p=0.1"),
      alpha = 0.6, linetype = "dashed", show.legend = FALSE) +
  labs(color = "", y  = "", title = "Distribución binomial")

dists <- ggplot(data_frame(x = -1:20), aes(x)) +
    stat_function(fun = pbinom, args = list(size = 20, prob = 0.5),
        aes(colour = "n=20;p=0.5"), alpha = 0.8) +
    stat_function(fun = pbinom, args = list(size = 20, prob = 0.1),
        aes(colour = "n=20;p=0.1"), alpha = 0.8) +
    labs(y = "", title = "FDA", color = "")

grid.arrange(densidades, dists, ncol = 2, newpage = FALSE)
```

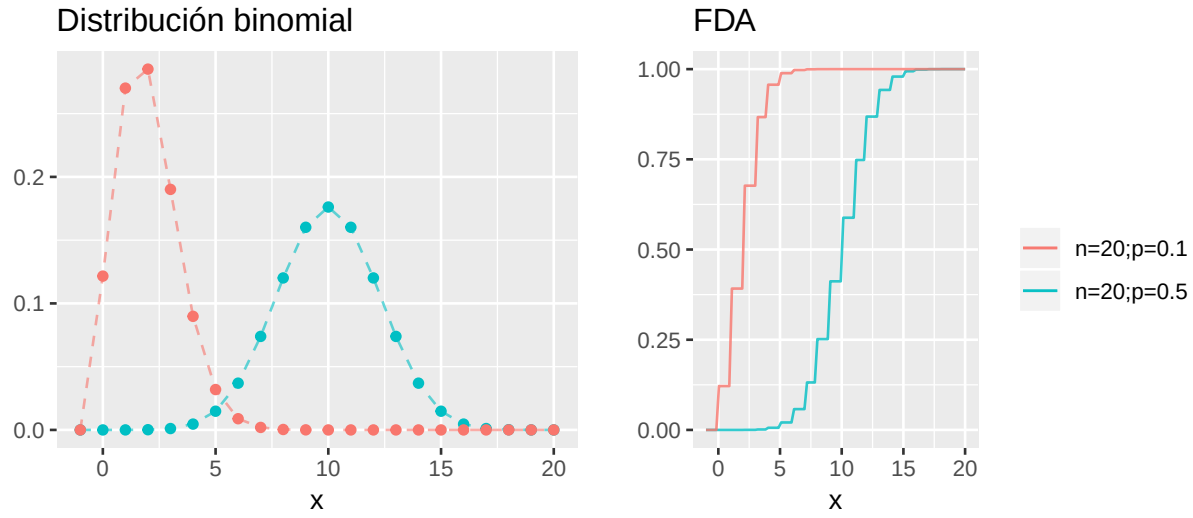


Figure 5: Distribución binomial y su FDA

Distribución Geométrica

Función de distribución

$$f(x) = (1-p)^{k-1}p$$

Función de distribucin acumulada

$$F(x) = \sum_{k=1}^{\infty}(1-p)^{k-1}p = 1-(1-p)^{k}$$

```r
densidades <- ggplot(data.frame(x = -1:20)) +
  geom_point(aes(x = x, y = dgeom(x, p = 0.5), color = "p=0.5"), show.legend = FALSE) +
  geom_path(aes(x = x, y = dgeom(x, p = 0.5), color = "p=0.5"), show.legend = FALSE,
```

```
      alpha = 0.6, linetype = "dashed") +
  geom_point(aes(x = x, y = dgeom(x, p = 0.1), color = "p=0.1"), show.legend = FALSE) +
  geom_path(aes(x = x, y = dgeom(x, p = 0.1), color = "p=0.1"),
      show.legend = FALSE, alpha = 0.6, linetype = "dashed") +
  labs(title = "Distribución geométrica", y = "")

dists <- ggplot(data_frame(x = -1:20), aes(x)) +
    stat_function(fun = pgeom, args = list(p = 0.5),
        aes(colour = "p=0.5"), alpha = 0.8) +
    stat_function(fun = pgeom, args = list(p = 0.1),
        aes(colour = "p=0.1"), alpha = 0.8) +
    labs(y = "", title = "FDA", color = "")

grid.arrange(densidades, dists, ncol = 2, newpage = FALSE)
```

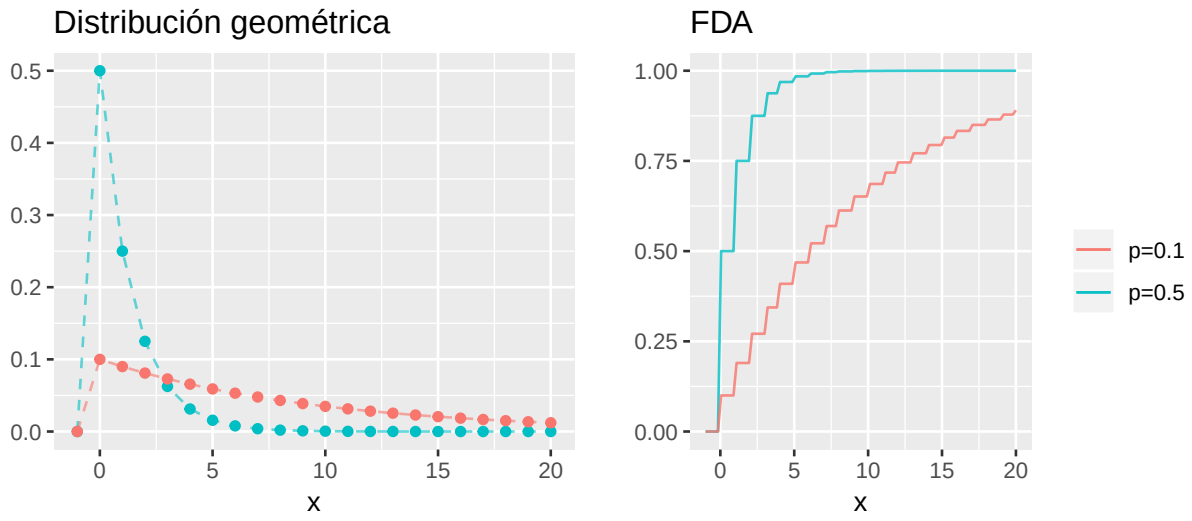


Figure 6: Distribución geométrica y su FDA

Distribución Gamma

Función de distribución

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

Función de distribución acumulada

$$F(x) = \int_0^{\infty} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx$$

```
densidades <- ggplot(data_frame(x = c(0 , 12)), aes(x)) +
    stat_function(fun = dgamma, args = list(shape = 1), aes(colour = "a=1;b=1"), show.legend = FALSE) +
    stat_function(fun = dgamma, args = list(scale = 0.5, shape = 2), aes(colour = "a=2;b=0.5"), show.leg
    stat_function(fun = dgamma, args = list(scale = 3, shape = 4), aes(colour = "a=4,b=3"), show.legend
    labs(y = "", title = "Distribución Gamma", colour = "")

dists <- ggplot(data_frame(x = c(0 , 12)), aes(x)) +
    stat_function(fun = pgamma, args = list(shape = 1), aes(colour = "a=1;b=1")) +
```

```
    stat_function(fun = pgamma, args = list(scale = 0.5, shape = 2), aes(colour = "a=2;b=0.5")) +
    stat_function(fun = pgamma, args = list(scale = 3, shape = 4), aes(colour = "a=4,b=3")) +
    labs(y = "", title = "FDA", color="")

grid.arrange(densidades, dists, ncol = 2, newpage = FALSE)
```

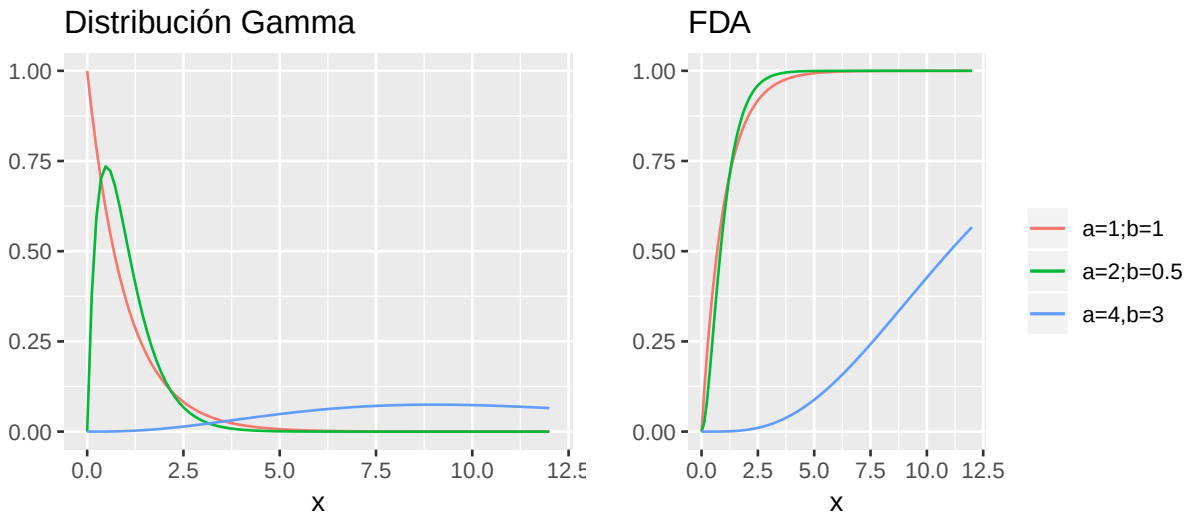


Figure 7: Distribución gamma y su FDA

Distribución Beta

Función de distribución

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Función de distribución acumulativa

$$F(x) = \int_0^{\infty} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} dx$$

```
densidades <- ggplot(data_frame(x = c(0 , 1)), aes(x)) +
    stat_function(fun = dbeta, args = list(shape1 = 2, shape2 = 2),
        aes(colour = "a=2; b=2"), show.legend = FALSE) +
    stat_function(fun = dbeta, args = list(shape1 = 5, shape2 = 2),
        aes(colour = "a=5; b=2"), show.legend = FALSE) +
    stat_function(fun = dbeta, args = list(shape1 = .5, shape2 = .5),
        aes(colour = "a=.5; b=.5"), show.legend = FALSE) +
    labs(y = "", title = "Distribución Beta", colour = "")

dists <- ggplot(data_frame(x = c(0 , 1)), aes(x)) +
    stat_function(fun = pbeta, args = list(shape1 = 2, shape2 = 2),
        aes(colour = "a=2; b=2")) +
    stat_function(fun = pbeta, args = list(shape1 = 5, shape2 = 2),
        aes(colour = "a=5; b=2")) +
    stat_function(fun = pbeta, args = list(shape1 = .5, shape2 = .5),
        aes(colour = "a=.5; b=.5")) +
    labs(y = "", title = "FDA", color="")
```

```
grid.arrange(densidades, dists, ncol = 2, newpage = FALSE)
```

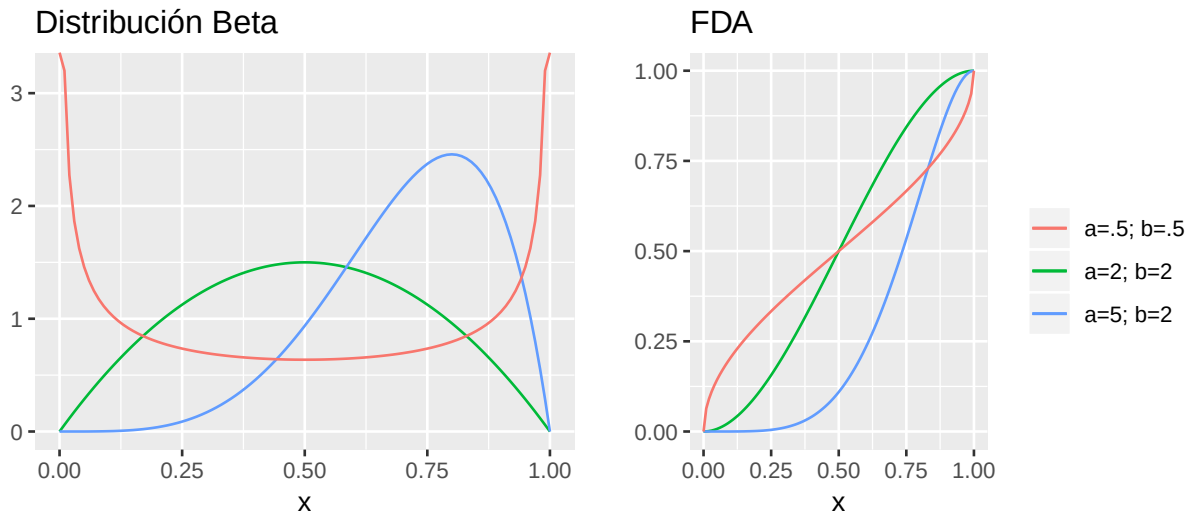


Figure 8: Distribución beta y su FDA

Distribución Normal

Función de distribución

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

Función de distribución acumulada

$$F(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{(\frac{-1}{2\sigma^2}(x-\mu)^2)} dx$$

```
densidades <- ggplot(data_frame(x = c(-5 , 5)), aes(x)) +
    stat_function(fun = dnorm, aes(colour = "m=0; s=1"), show.legend = FALSE) +
    stat_function(fun = dnorm, args = list(mean = 1), aes(colour = "m=1; s=1"), show.legend = FALSE) +
    stat_function(fun = dnorm, args = list(sd = 2), aes(colour = "m=1; s=2"), show.legend = FALSE) +
    labs(y = "", title = "Distribución Normal", colour = "")

dists <- ggplot(data_frame(x = c(-5 , 5)), aes(x)) +
    stat_function(fun = pnorm, aes(colour = "m=0; s=1")) +
    stat_function(fun = pnorm, args = list(mean = 1), aes(colour = "m=1; s=1")) +
    stat_function(fun = pnorm, args = list(sd = 2), aes(colour = "m=1; s=2")) +
    labs(y = "", title = "FDA")

grid.arrange(densidades, dists, ncol = 2, newpage = FALSE)
```

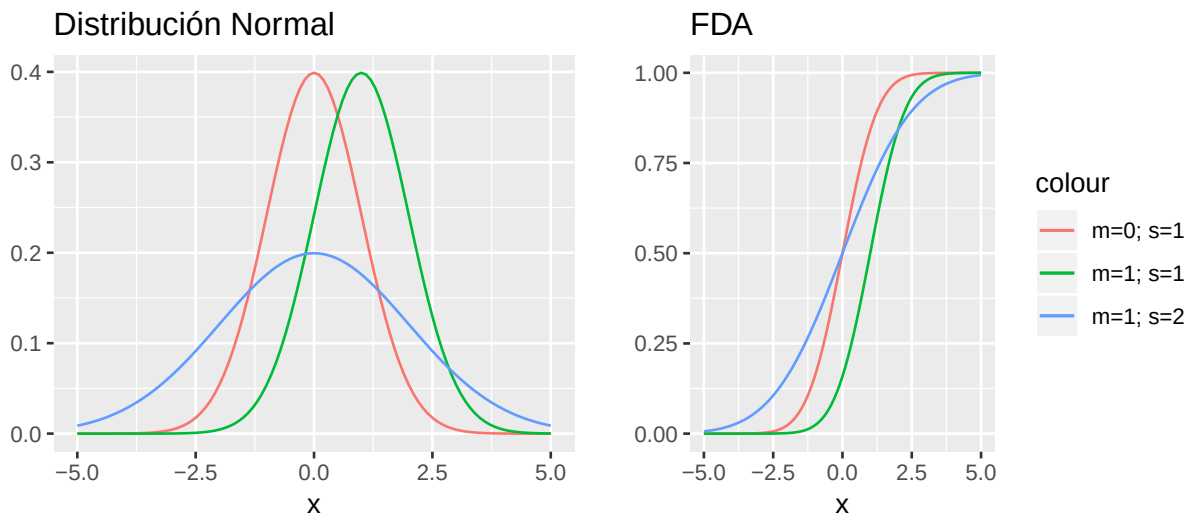


Figure 9: Distribución normal y su FDA

Como ejemplo del uso de distribuciones, regresemos al ejemplo de la base de datos *iris* (discusión).

```
ggplot(data=iris, aes(Sepal.Length,fill=Species)) +
geom_density(alpha=0.7)
```