# Práctica 5

Alejandro Cáceres
UPC - Statistics 2019/2020

# Objetivos

Variables aleatorias continuas:

- Distribución Normal
- Distribuciones generales

# Distribución Normal

**Definición**

Una variable aleatoria X definida sobre los números reales tiene una densidad de distribución **Normal** si tiene la forma

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

Cuando X sigue una distribución Normal escribimos

$$X \to N(\mu, \sigma^2)$$

# Distribución Normal

Cuando

$$X \to N(\mu, \sigma^2)$$

Entonces

$$E(X) = \mu$$

y

$$V(X) = \sigma^2$$

$\mu$ y $\sigma^2$ son los parametros de la distribución y coinciden con su valor esperado y vairanza

# Distribución Normal

En R existe la función **dnorm** que es

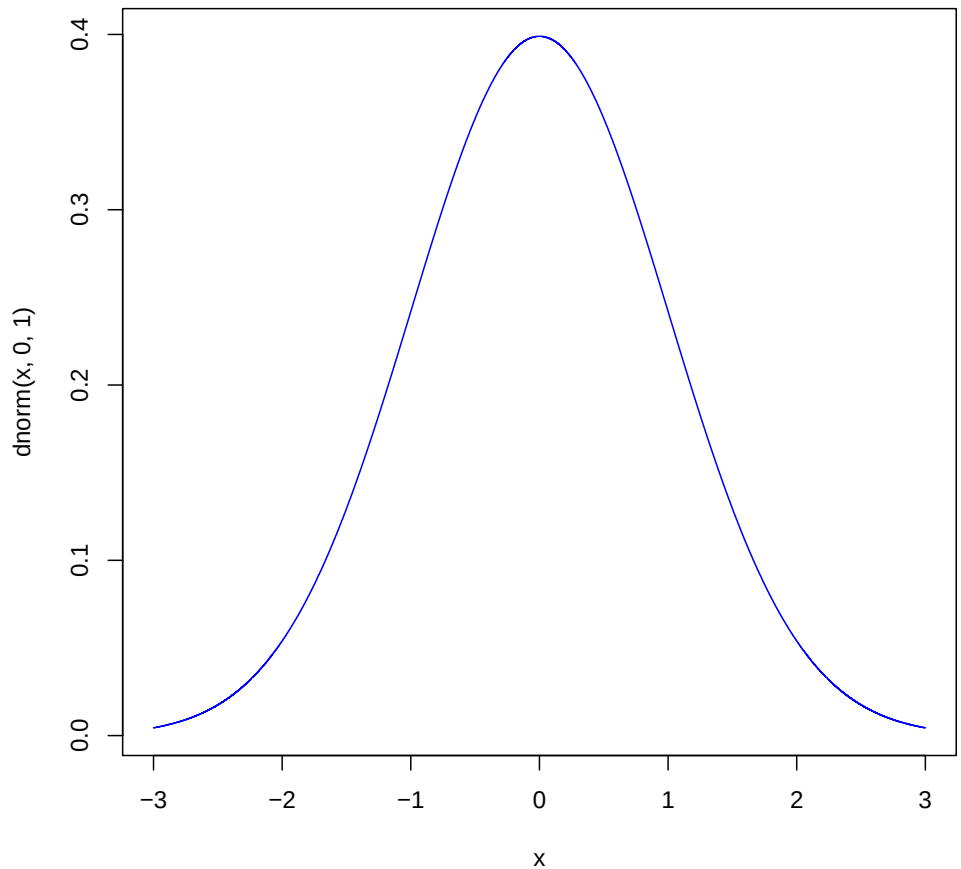
$$dnorm(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma}}$$

Grafiquemos $dnorm(x;\mu = 0, \sigma = 1)$ en el intervalo

```
> x <- seq(-3,3,0.0001)
> head(x)
[1] -3.0000 -2.9999 -2.9998 -2.9997 -2.9996 -2
```

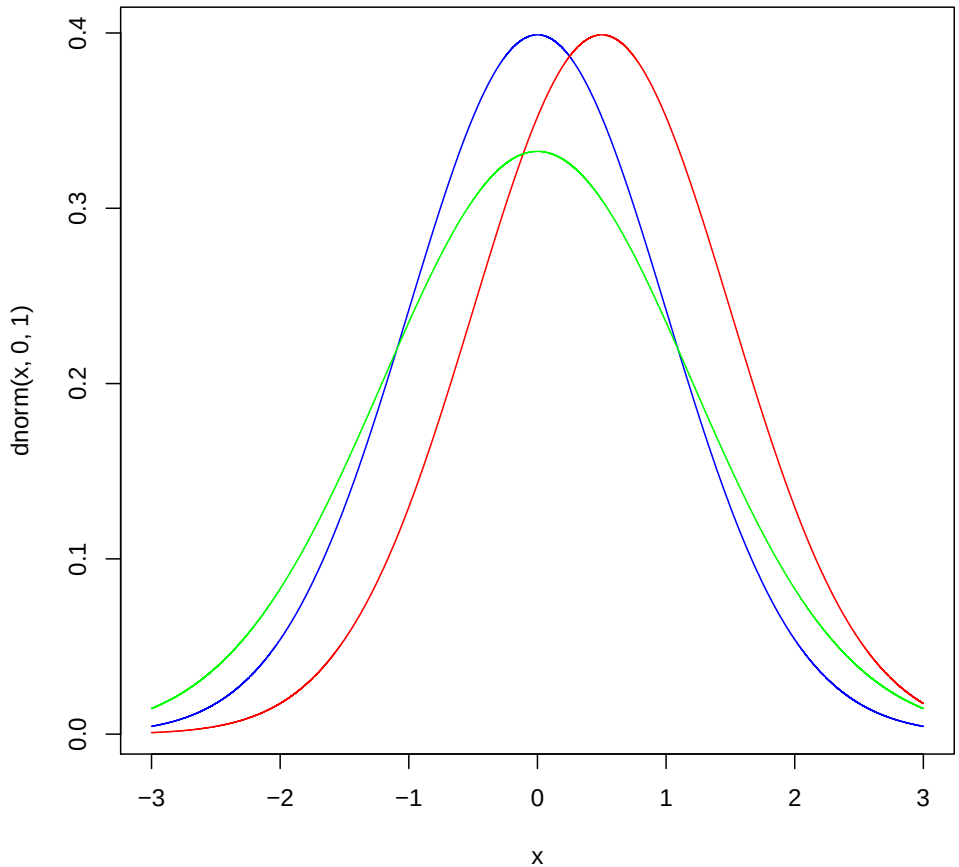
# Distribución Normal

```
> plot(x, dnorm(x,0,1), type="l", col="blue")
```



# Distribución Normal

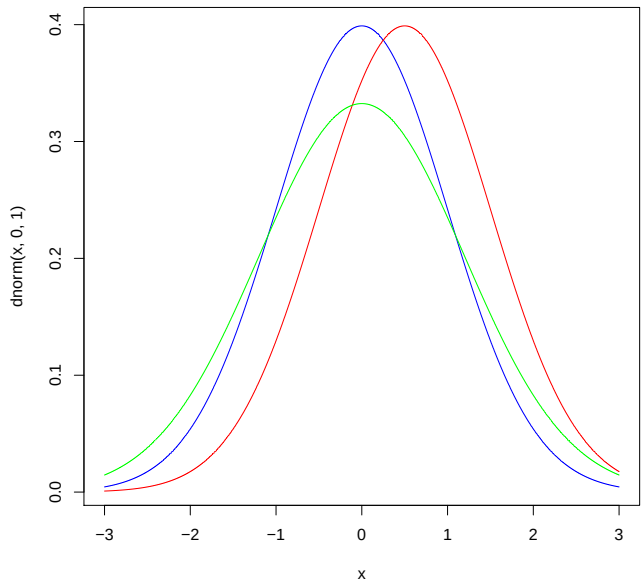
Cómo harías el siguiente gráfico?



# Distribución Normal

Pistas:

- la linea verde es una normal con $\mu = 0.5$ y $\sigma = 1$
- la linea roja es una normal con $\mu = 0$ y $\sigma = 1.2$
- para añadir una curva a un gráfico existente usamos la función **lines**



# Distribución Normal

```
> plot(x, dnorm(x,0,1), type="l", col="blue", ylab="f(x)")
> lines(x, dnorm(x,0.5,1), type="l", col="red")
> lines(x, dnorm(x,0,1.2), type="l", col="green")
```

con los parametros *xlab* y *ylab* nombramos los ejes x e y

# Distribución Normal

En R existe la función **rnorm** que genera una muestra aleatoria de una distribución normal con media $\mu$ y desviación normal $\sigma$

# Distribución Normal

Si las mujeres tienen una altura media de 165cm con desviación típica de 8cm, que se distribuye **normalemente**

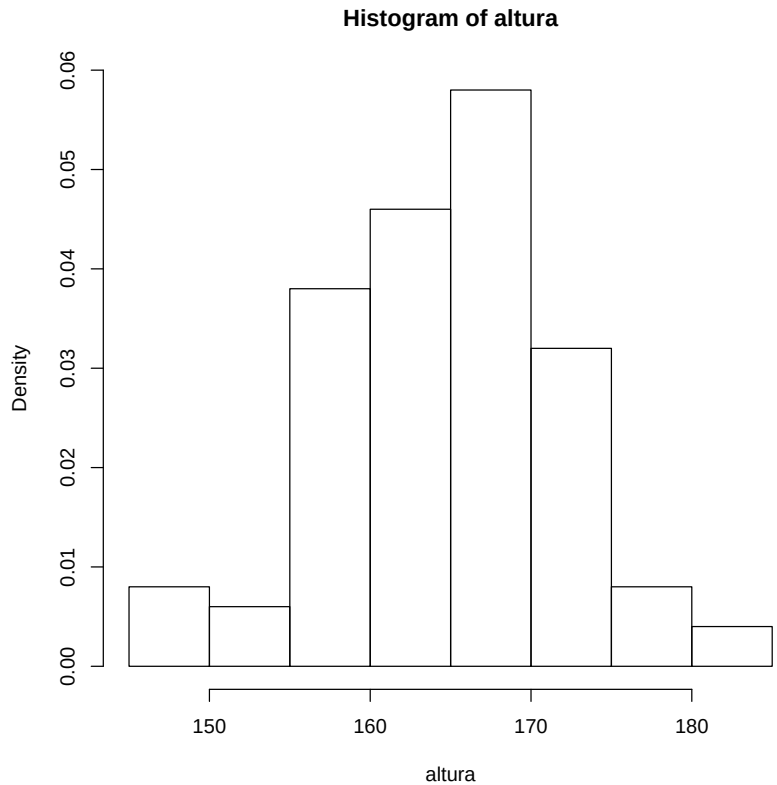
Toma una muestra aleatoria de la altura de 3 mujeres

```
> rnorm(n=3,mean=165,sd=8)
[1] 156.2405 168.2088 163.3979
```

Haz el histograma (**hist**) de 100 alturas aleatorias

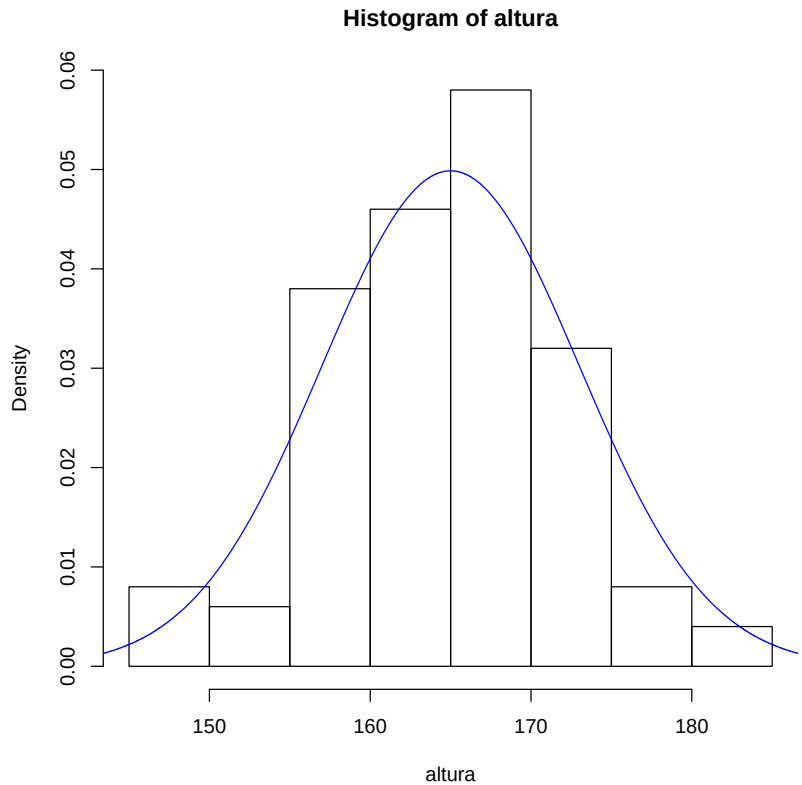
# Distribución Normal

```
> altura <- rnorm(n=100,mean=165,sd=8)
> hist(altura ,freq=FALSE)
```



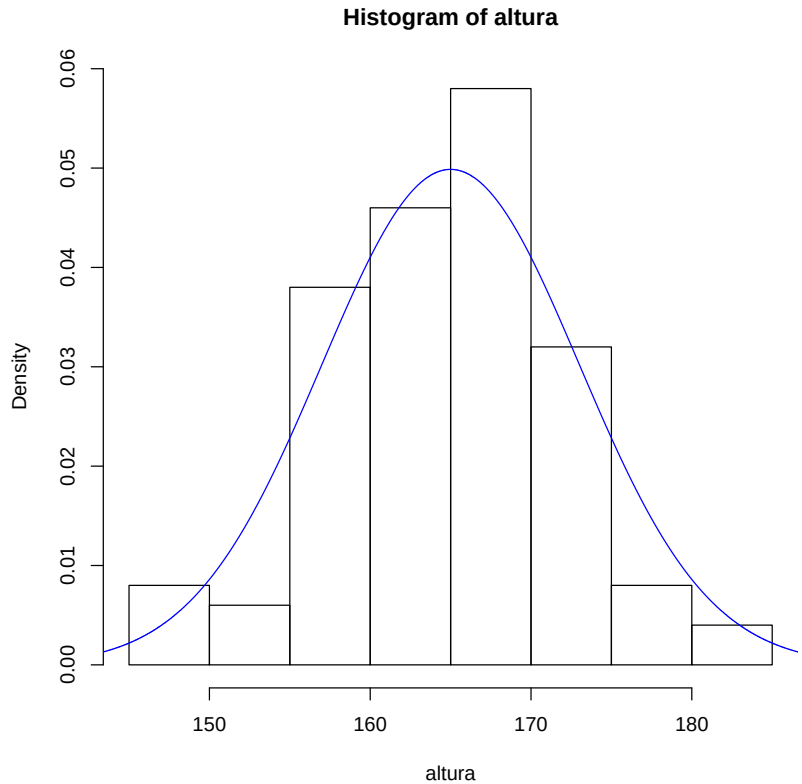
# Distribución Normal

Cómo harías este gráfico?



# Distribución Normal

```
> hist(altura ,freq=FALSE)
> x <- seq(130,200,0.001)
> lines(x, dnorm(x,mean=165,sd=8), type="l", col="blue")
```



# Distribución Normal

En R existe la función **pnorm** que da la función de acumulación de probabilidad de una normal con media $\mu$ y desviación estandard $\sigma$

$$pnorm(x, \sigma, \mu) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma}} dt$$

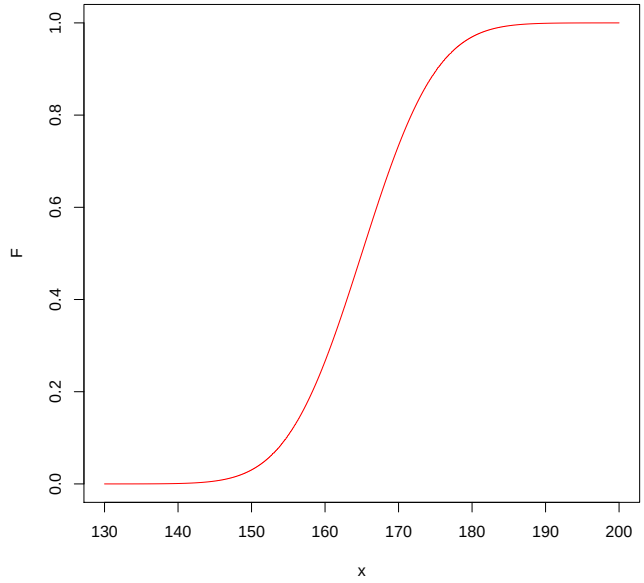
para la distribución estandard (en nuestro intervalo x) estos son sus valores

```
> F <- pnorm(x,mean=165,sd=8)
> F
[1]  6.071624e-06 6.075104e-06 6.078586e-06 ...
```

Grafica la función!

# Distribución Normal

```
plot(x, F, type="l", col="red")
```




```
> pnorm(150,mean=165,sd=8)
[1] 0.03039636
```

La probabilidad acumulada hasta 150cm es de 0.03. O el 3% de las mujeres tienen altura menor de 150cm

# Distribución Normal

Usa **pnorm** y **qnorm** (la función inversa de **pnomr**: $F^{-1}(p) = x$) para responder

- la probabilidad de que un hombre mida **por lo menos** 165cm, si el promedio de la población es de 175cm y desviación típica de 10cm.
- la probabilidad de que un hombre mida **entre** 165cm y 180cm.
- Cuál es el la altura que define al 5% de hombres mas bajos en la población?

# Distribución Normal

```
> pnorm(165, mean=175, sd=10)
[1] 0.1586553

> pnorm(180, mean=175, sd=10) - pnorm(165, mean=175, sd=10)
[1] 0.5328072

> qnorm(0.05, mean=175, sd=10)
[1] 158.5515
```

# Material adicional: Distribución normal

- Los datos de fallecimiento por COVID19 son publicados diariamente en la página del ministerio de salud: `https://covid19.isciii.es/`

Pregunta:

- Suponiendo que la probabilidad de fallecer en una fecha dada durante pandemia sigue una **distribución normal** en Cataluña, según los datos publicados por el ministerio hasta el 10 de abril, Cual es la probabilidad de fallecimiento por la enfermedad a partir del día martes 13 de abril, cuando las actividades no escenciales están de vuelta al trabajo?

# Distribución normal

## Cargemos los datos

```
> covid <- read.table("Covid19_9abril.csv", header=TRUE, sep=","
```

```
> head(covid)
  CCAA      FECHA CASOS Hospitalizados UCI Fallecidos Recuperados
1   AN 20/2/2020    NA              NA  NA         NA          NA
2   AR 20/2/2020    NA              NA  NA         NA          NA
3   AS 20/2/2020    NA              NA  NA         NA          NA
4   IB 20/2/2020     1              NA  NA         NA          NA
5   CN 20/2/2020     1              NA  NA         NA          NA
6   CB 20/2/2020    NA              NA  NA         NA          NA
```

# Distribución normal

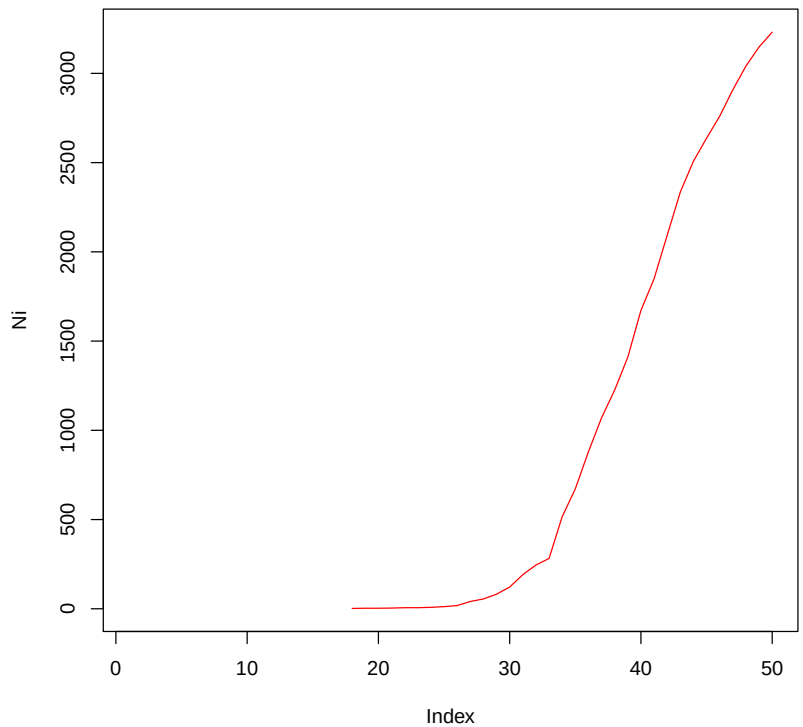
Seleccionemos los datos para Cataluña: líneas en las cuales la variable CCAA toma valor CT

```
> cat <- covid$CCAA=="CT"
> covidCAT <- covid[cat,]
> head(covidCAT)
    CCAA      FECHA CASOS Hospitalizados UCI Fallecidos Recuperad
9     CT 20/2/2020    NA              NA  NA          NA
28    CT 21/2/2020    NA              NA  NA          NA
47    CT 22/2/2020    NA              NA  NA          NA
66    CT 23/2/2020    NA              NA  NA          NA
85    CT 24/2/2020    NA              NA  NA          NA
104   CT 25/2/2020     1              NA  NA          NA
>
```

# Distribución normal

Seleccionemos lo variable Fallecidos, que es la frecuencia absoluta acumulada de fallecimientos para cada uno de los días

```
> Ni <- covidCAT$Fallecidos
> plot(Ni, type="l", col="red")
```



# Distribución normal

Obtengamos la frecuencia absoluta de fallecimientos por día **ni**. **ni** para el día k es la diferencia de la frecuencia acumulada en k+1 menos la frecuencia acumulada en k:

$$ni(k) = Ni(k+1) - Ni(K)$$

esto se hace con la funcion **diff**

# Distribución normal

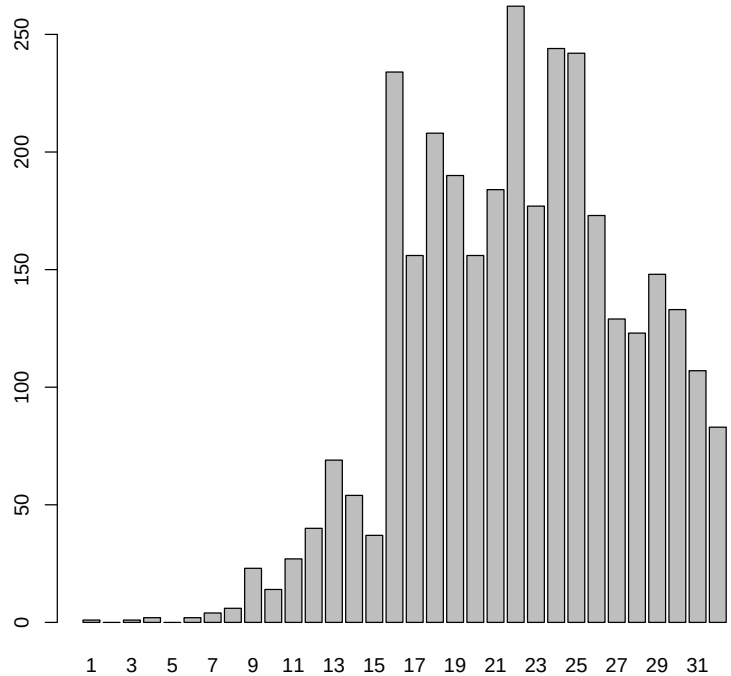
esto se hace con la funcion **diff**

```
> ni <- diff(Ni)
#quitamos missings con complete.cases
> ni <- ni[complete.cases(ni)]
#revisamos para cuantos dias tenemos datos
> length(ni)
[1] 32
#creamos el rango de observaciones (dias)
> x <- 1:32
> names(ni) <- x
> head(ni)
 1  2  3  4  5  6  7  8  9 10
 1  0  1  2  0  2  4  6 23 14
```

# Distribución normal

garfiquemos con **barplot**

```
> barplot(ni)
```



# Distribución normal

Calculemos la frecuencia relativa **fi** de ni

```
> fi <- ni/sum(ni)
> names(fi) <- x
> head(fi)
           1            2            3            4            5
0.0003096934 0.0000000000 0.0003096934 0.0006193868 0.0000000000
           6
0.0006193868
```

Cual es la media y la varianza muestral de fi?

# Distribución normal

Recordemos fi es una funcion de distribución (experimental), que tiene media y desviación típica

```
> mui <- sum(fi*x)
> mui
[1] 22.17002
> sdi <- sqrt(sum(fi*(x -mui)^2))
> sdi
[1] 5.35589
```

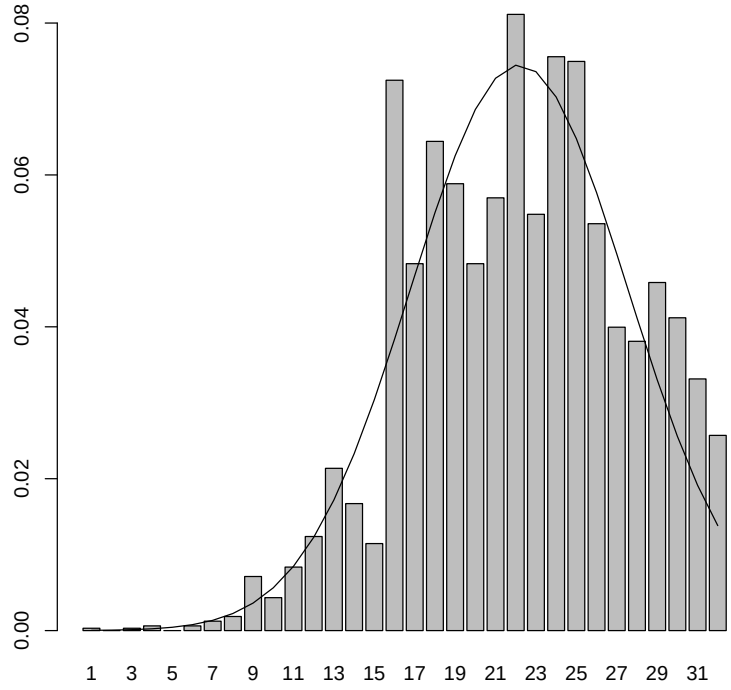
Nota importante: fi no son observaciones de la variable aleatoria X (día de fallecimiento para cada individuo), fi son observaciones de la frecuencia de X. Para x usamos las funciones **mean** o **sd**, para fi usamos las definiciones de media y varianza para distribuciones.

# Distribución normal

Asumimos que fi proviene de una distribución normal con media **mui** y desviación típica **sdi**. Grafiquemos fi con barplot y superpongamos la distribucion normal

# Distribución normal

```
> xparaplot <- barplot(fi)
> lines(xparaplot, dnorm(x,mui,sdi))
```




La variable **xparaplot** se usa para ubicar los puntos de x en el barplot

# Distribución normal

El día 12 de abril es el día 36 de la pandemia en Cataluña por lo que la probabilidad acumulada de fallecimientos hasta ese día sería

```
> pnorm(36,mui,sdi)
[1] 0.9950914
```

y la fracción de fallecimientos esperada después de ese día

```
> 1 - pnorm(36,mui,sdi)
[1] 0.004908634
```

Todo asumiendo que la distribución es normal.

# Cuidado!

- Este es un ejercicio académico y tiene muchos problemas.
- Seguramente la distribución no es normal, no sube y baja al mismo ritmo.
- Es un proceso dinámico, las intervenciones afectan la forma de la distribución.
- Si el modelo no bueno las autoridades pueden cometer errores graves.
- Los modelos dependen críticamente de los datos, si no contamos los fallecimientos bien los modelos no son buenos.
- Nadie puede decir si un modelo es bueno solo hasta que pase la pandemia.

# Distribuciones generales

Definamos una función de distribución

$$f(x) = \begin{cases} 1/8 + 3/8 * x, & \text{if } x \in (0, 2) \\ 0, & \textit{resto} \end{cases}$$

Queremos generar una muestra aleatoria que siga esta distribución.

# Distribuciones generales

Para la distribución

$$f(x) = \begin{cases} 1/8 + 3/8 * x, & \text{if } x \in (0, 2) \\ 0, & \textit{resto} \end{cases}$$

definamos una secuencia de puntos en el intervalo donde la distribución no es cero

```
x <- seq(0,2,0.0001)
```

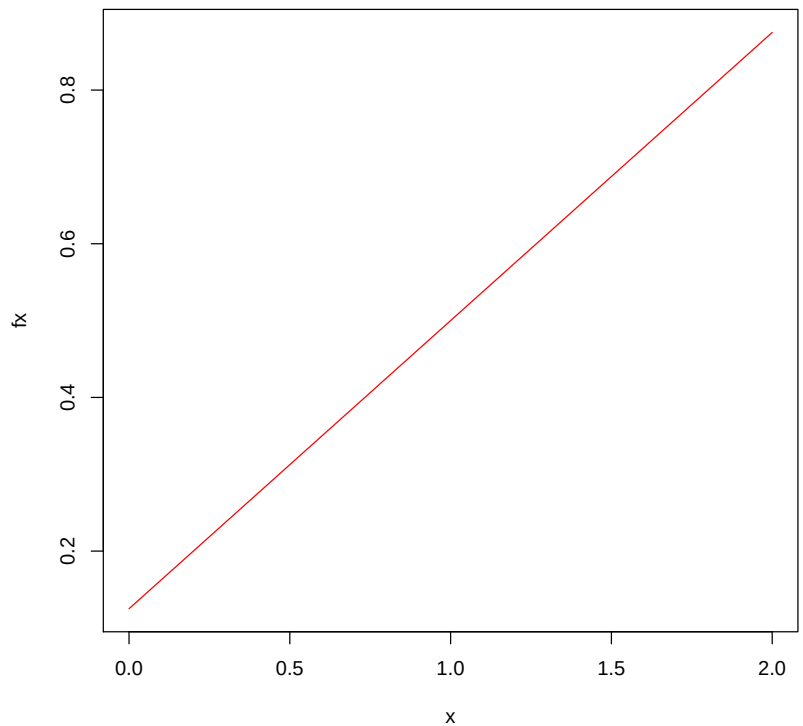
En este intervalo la distribución es simplemente

```
fx <- 1/8+3/8*x
```

Grafícala con **plot(..., type="l")**!

# Distribuciones generales

```
> plot(x,fx,type="l", col="red")
```



# Distribuciones generales

Ahora tomemos 100 muestras aletorias con **sample**

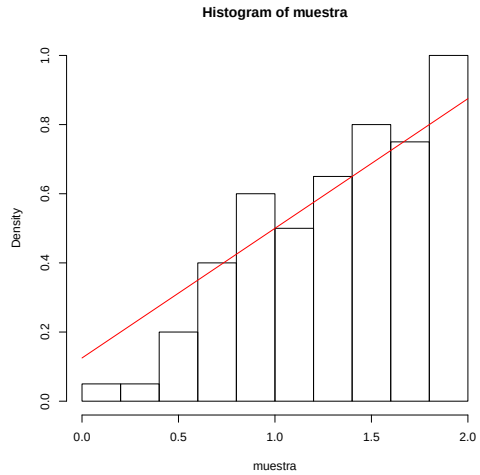
```
> muestra <- sample(x, size=100,replace=TRUE, prob=fx)
```

En nuestro intervalo, cada x tiene una probabilidad de ser obtenido que es dada por nuestro vector de probabilidades fx

- pinta el histograma de `muestra` y añádele la distribucion fx con **lines(...)**
- cuales son la media y el error estandard de `muestra`?

# Distribuciones generales

```
> hist(muestra,freq=FALSE)
> lines(x, fx, col="red")
```




```
> mean(muestra)
[1] 1.339012
> sd(muestra)
[1] 0.4540648
```

# Distribuciones generales

Definamos la función de distribución

$$f(x) = \begin{cases} 1/8 + 3/8 * x, & \text{if } x \in (0, 2) \\ 0, & \textit{resto} \end{cases}$$

sobre todos los valores de x, inclusive cuando $x \notin (0, 2)$

# Distribuciones generales

Definamos la distribución como una función

```
> f<-function(x)
 {
    out <- 1/8+3/8*x
    out
  }


> f(0.5)
[1] 0.3125
```

La función es vectorial:

```
> f(c(-1,0.5,3))
[1] -0.2500  0.3125  1.2500
```

# Distribuciones generales

Pero ahora queremos que haga

```
> f(-1)
[1] 0.0000
> f(0.5)
[1] 0.3125
> f(3)
[1] 0.0000
```

o sea que sea cero para valores de $x \leq 0$ o $x \geq 2$

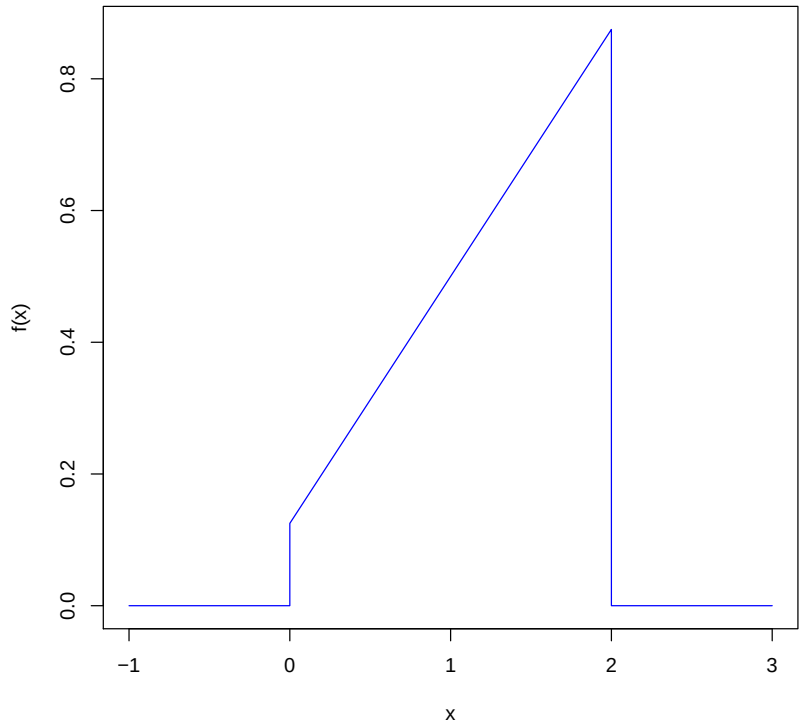
# Distribuciones generales

```
> f<-function(x)
 {
    out<-1/8+3/8*x
    out[x<=0]<-0
    out[x>=2]<-0
    out
  }
> f(c(-1,0.5,3))
[1] 0.0000 0.3125 0.0000
```

Grafica f en el intervalo donde la distribución sí toma varores de cero

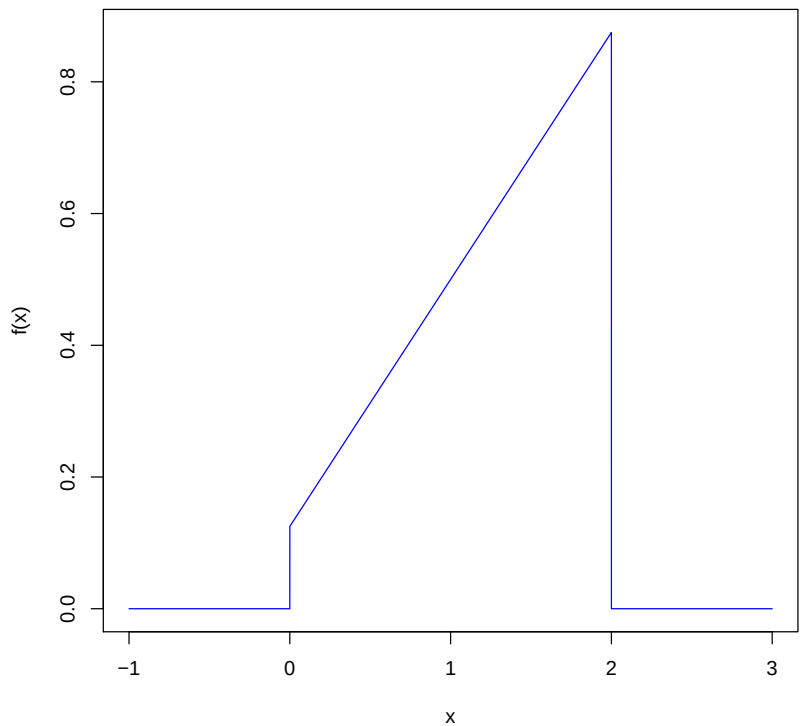
```
> x <- seq(-1,3,0.0001)
```

# Distribuciones generales



# Distribuciones generales

```
plot(x, f(x), type="l", col="blue")
```



# Distribuciones generales

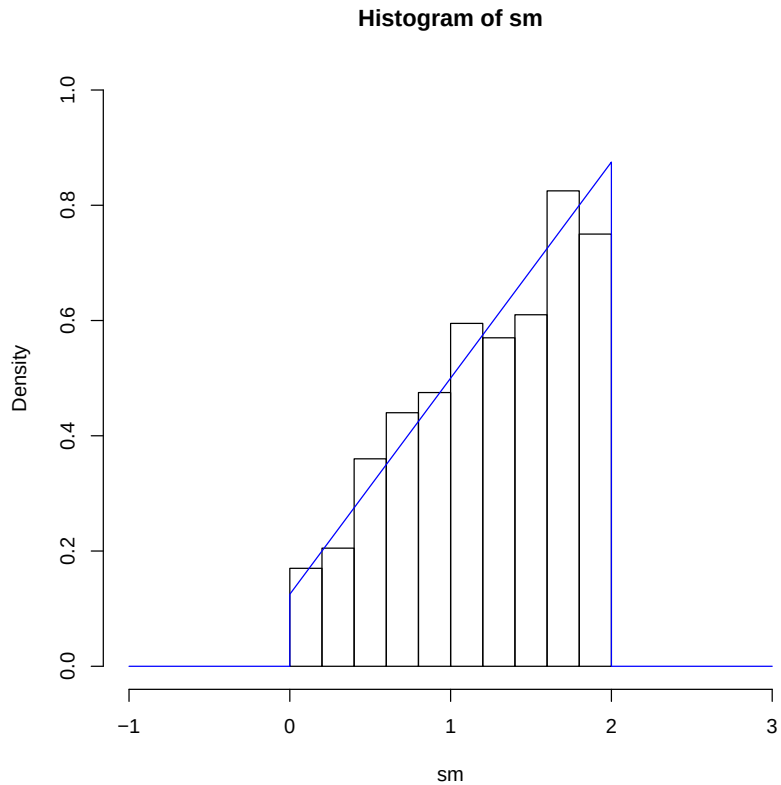
Ahora generemos 1000 muestras aleatorias de la distribucion **f**

```
> fx <- f(x)
> sm <- sample(x, size=1000, replace=TRUE, prob=fx)
[1] 1.4275 0.4275 1.6052 ...
```

Haz el histograma (hist) de **sm** y añade la función f(x) al histograma con **lines(...)**

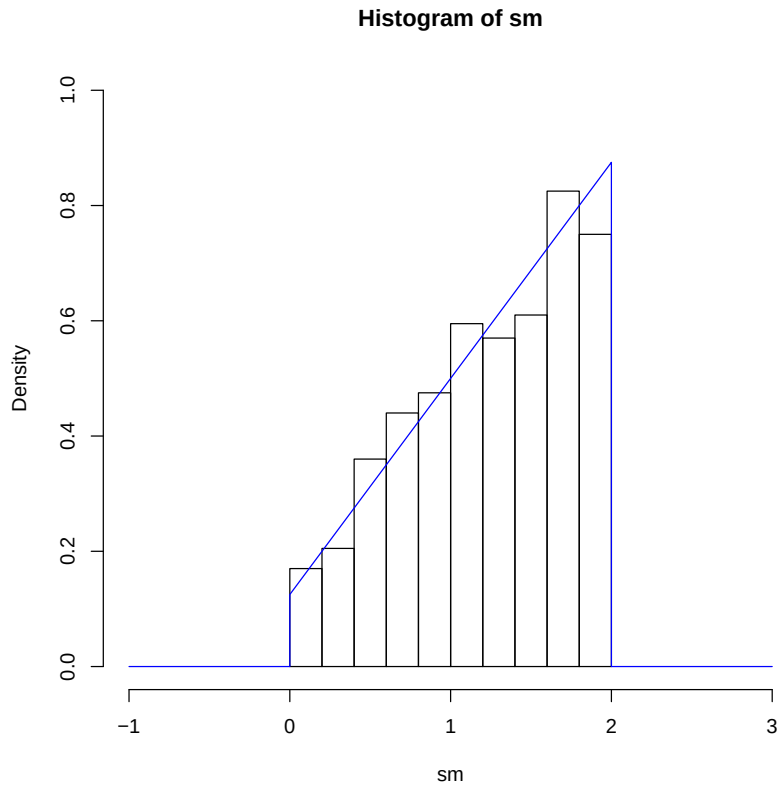
# Distribuciones generales

Sólo necesitamos sm, x y fx



# Distribuciones generales

```
> hist(sm,freq=FALSE, ylim=c(0,1),xlim=c(-1,3))
> lines(x, fx, col="blue")
```



# Distribuciones generales

La función de acumulación de probabilidad para nuestra distribución es

$$F(x) = P(X \leq x) = \int_0^t 1/8 + 3/8 * t dt$$

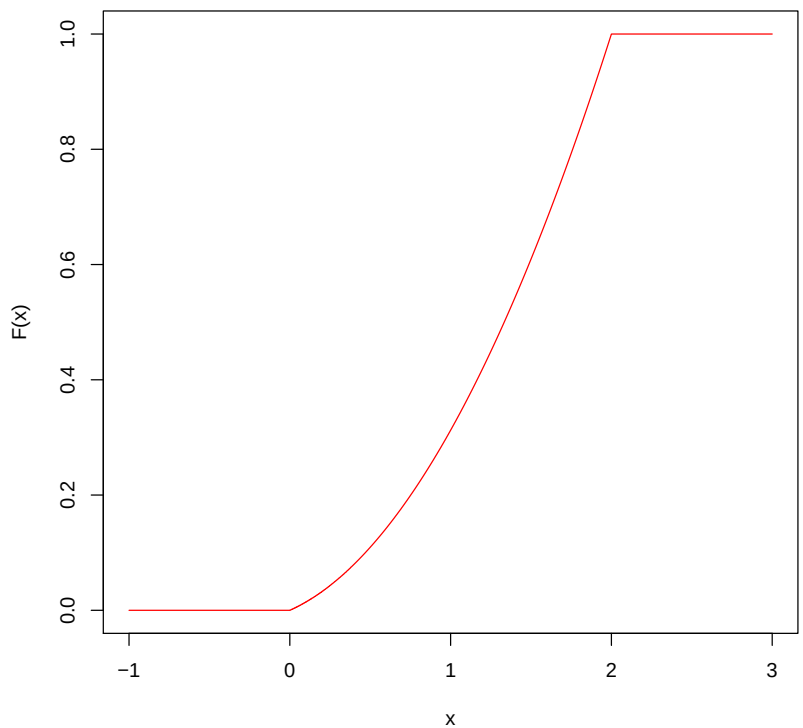
O sea

$$F(x) = \begin{cases} 1/8 * x + 3/16 * x^2, & \text{if } x \in (0, 2) \\ 0, & \textit{resto} \end{cases}$$

Haz la función F y grafícala

# Distribuciones generales

$$F(x) = \begin{cases} 1/8 * x + 3/16 * x^2, & \text{if } x \in (0, 2) \\ 0, & \textit{resto} \end{cases}$$



# Distribuciones generales

```
F<-function(x)
{
    out<-1/8*x+3/16*x^2
    out[x<=0]<-0
    out[x>=2]<-1
    out
}

plot(x, F(x), type="l", col="red")
```

# Distribuciones generales

Grafica la función de distribución

$$f(x) = \begin{cases} \frac{4}{\pi(1+x^2)}, & \text{if } x \in (0,1) \\ 0, & \textit{resto} \end{cases}$$

- Toma 10000 muestras aleatorias y compara el histograma con la función de distribución
- Grafica la función de acumulación de probabilidad
- Cuál es la probabilidad $P(0.4 \leq X \leq 0.6)$?