UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Herramientas estadísticas para Big Data

Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

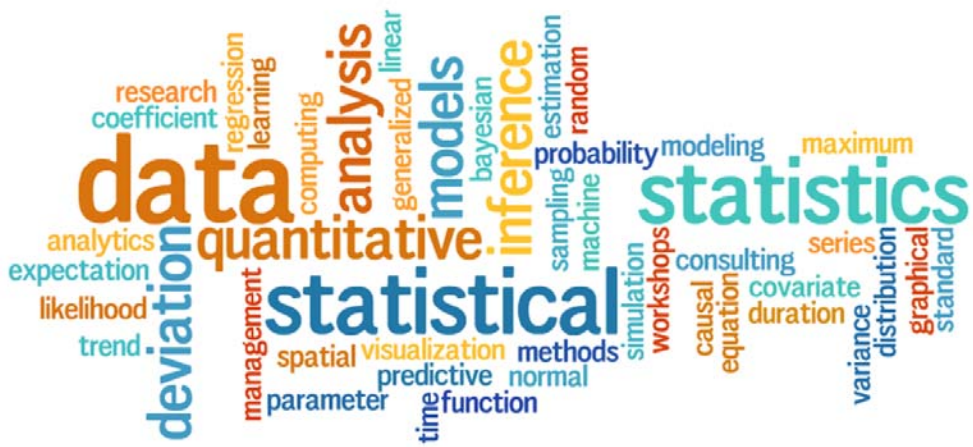
Máster **Big Data** Analytics

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

Valencia, Octubre 2016

Elena Vázquez

Contenidos



research
coefficient
regression
learning
computing
analysis
generalized
linear
models
bayesian
inference
estimation
random
probability
modeling
maximum
data
statistics
analytics
expectation
quantitative
sampling
machine
likelihood
trend
deviation
statistical
management
spatial
visualization
methods
simulation
workshops
consulting
series
covariate
duration
causal
equation
predictive
normal
parameter
time
function
variance
distribution
graphical
standard

3 Variables aleatorias y distribuciones

1. Introducción
2. Distribuciones de probabilidad y variables aleatorias
   - Distribuciones de probabilidad
   - Esperanza Matemática
   - La Distribución Normal
   - La Distribución Binomial
   - Teorema Central del Límite
   - Aproximaciones normales

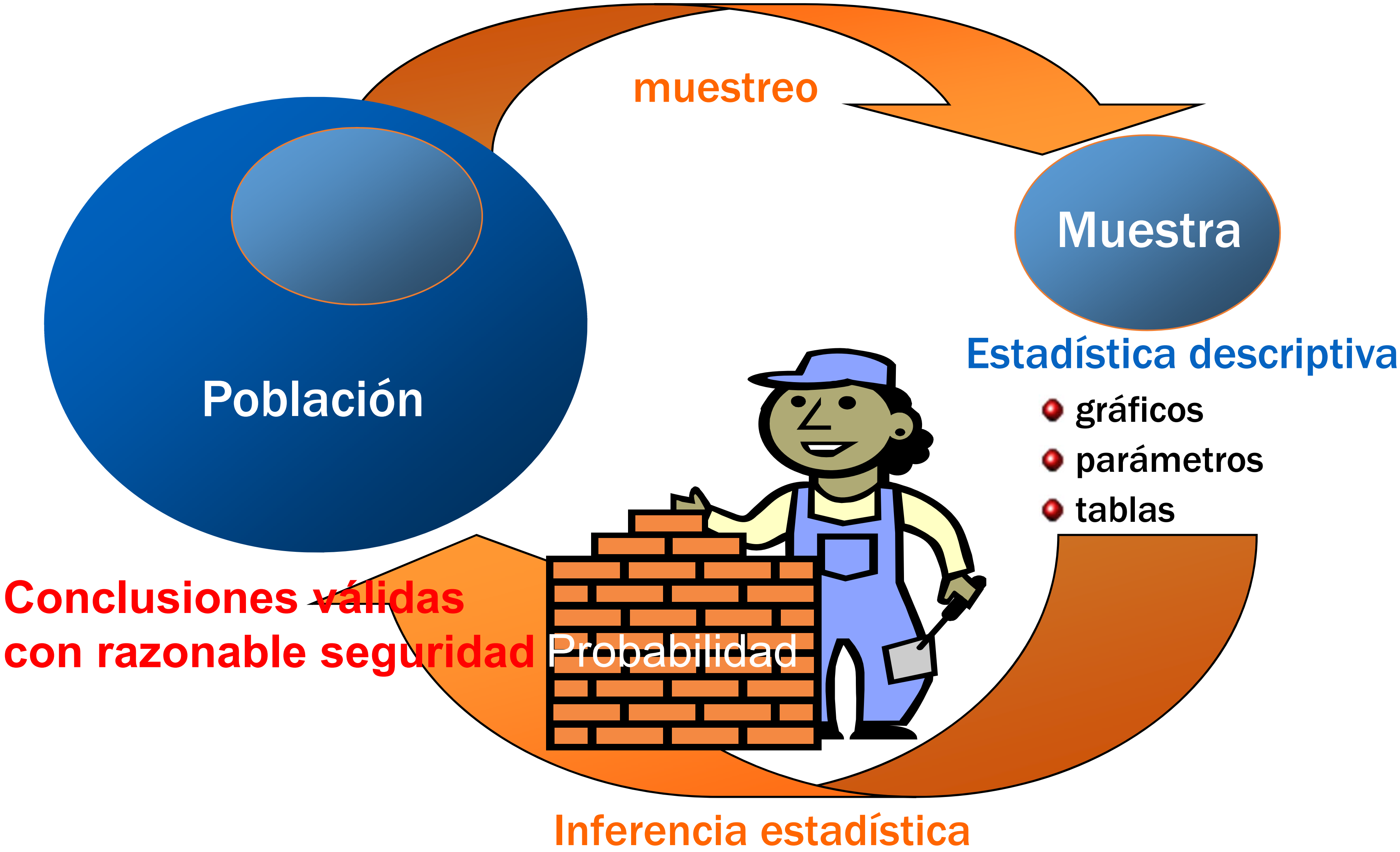
3. Ecuación fundamental

Introducción

- Los analistas necesitan responder a preguntas que se plantean sobre fenómenos del mundo real.
- Para ello:
  - Planteamos **hipótesis**
  - Recogemos **datos** el mundo real
  - **Verificamos** las hipótesis
- Verificar dichas hipótesis implica construir **modelos estadísticos** del fenómeno estudiado

La **inferencia estadística** permite extrapolar las conclusiones obtenidas con los datos observados sobre los fenómenos al mundo real mediante modelos

Introducción

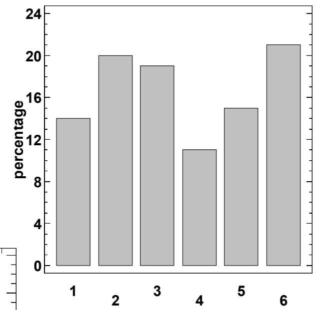


Distribuciones de probabilidad

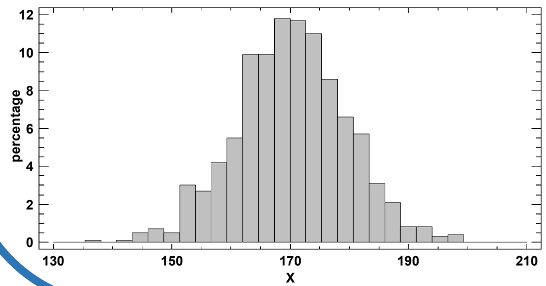
Muestra: Lo que tenemos

Frecuencia relativa (%)
% de veces que sale un 5 al lanzar el dado

Parámetros muestrales
$\overline{X}$
S^2
S
Cuartiles, ...



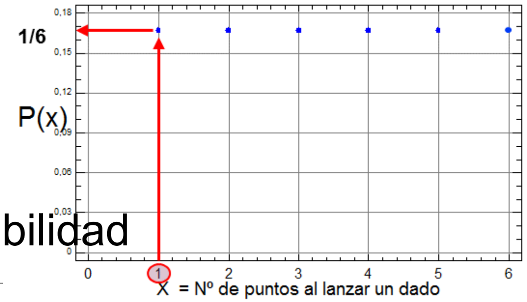
Distribución frecuencias



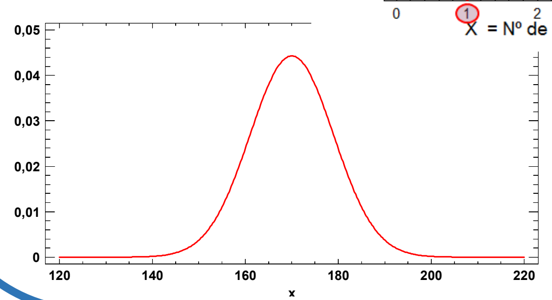
Población: Lo teórico, lo ideal

Probabilidad (tanto por uno)
P(A) = P(X=5)

Parámetros poblacionales
m o μ o E(X)
σ^2
σ
Cuartiles, …

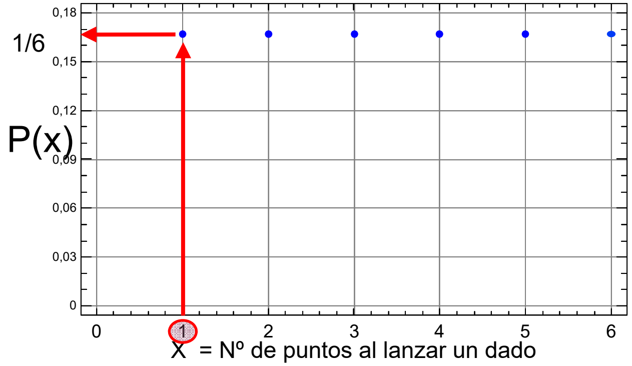


Distribución probabilidad

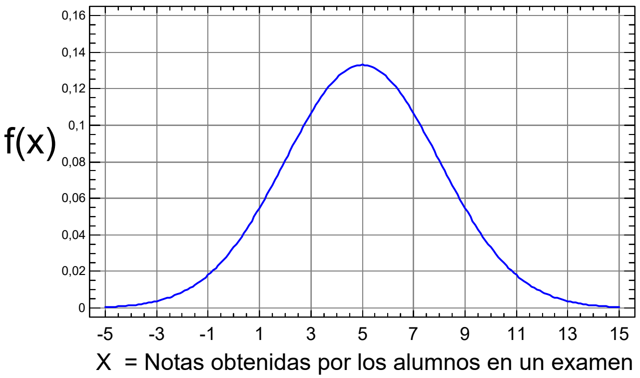


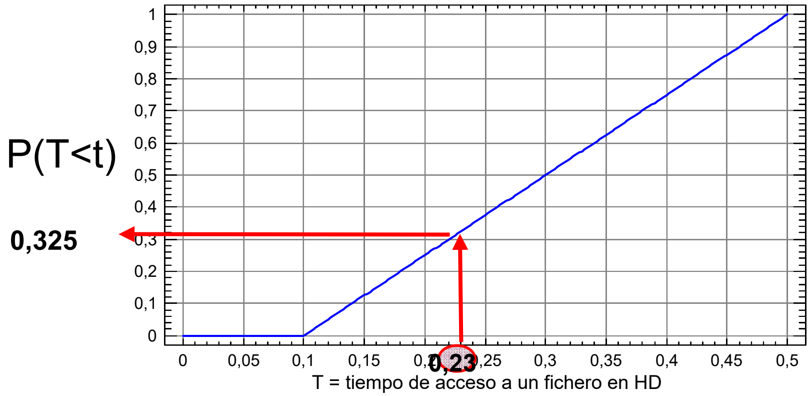
Distribuciones de Probabilidad y v.a.

Todas las técnicas de inferencia se basan en conocer lo probable o improbable que es un determinado suceso, o lo que es lo mismo, lo probable que es un determinado valor de una característica aleatoria.



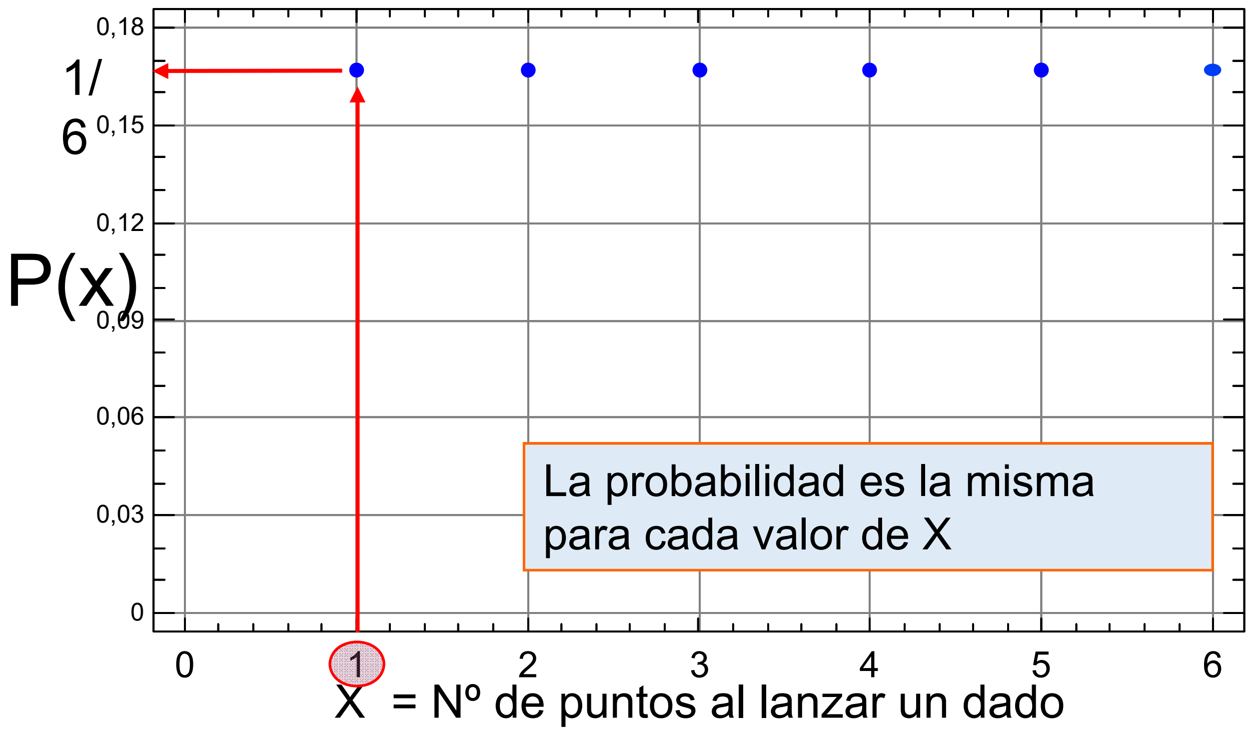
Esto implica que conocemos la probabilidad de cada valor y en la mayoría de los casos se recurre a modelos.





Modelos de distribuciones. Ejemplo 1: var discreta

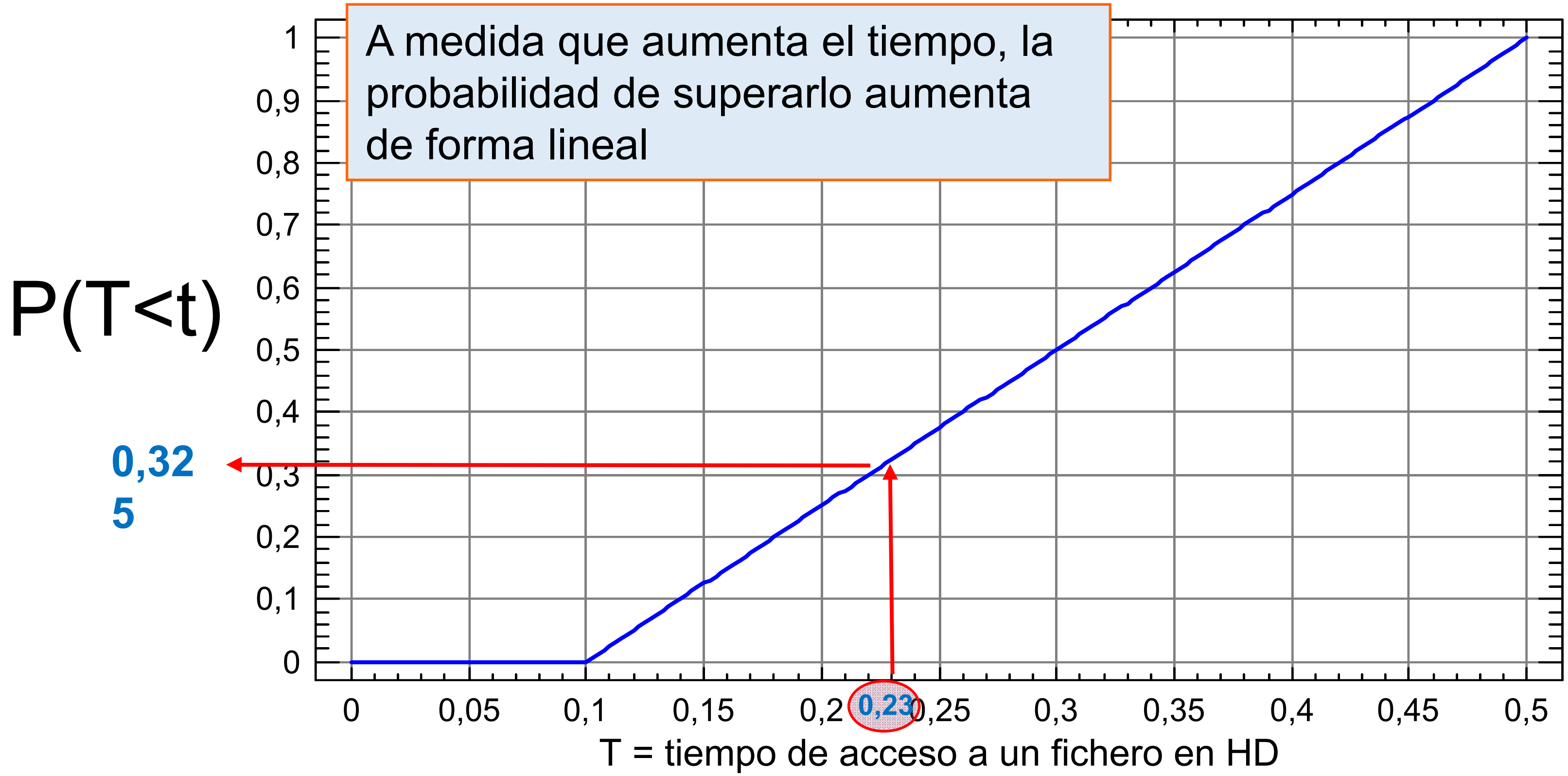
Variable aleatoria **X**: nº de puntos al lanzar el dado



La probabilidad de que al lanzar un dado salga un 1 es 1/6

Modelos de distribuciones. Ejemplo 2: var continua

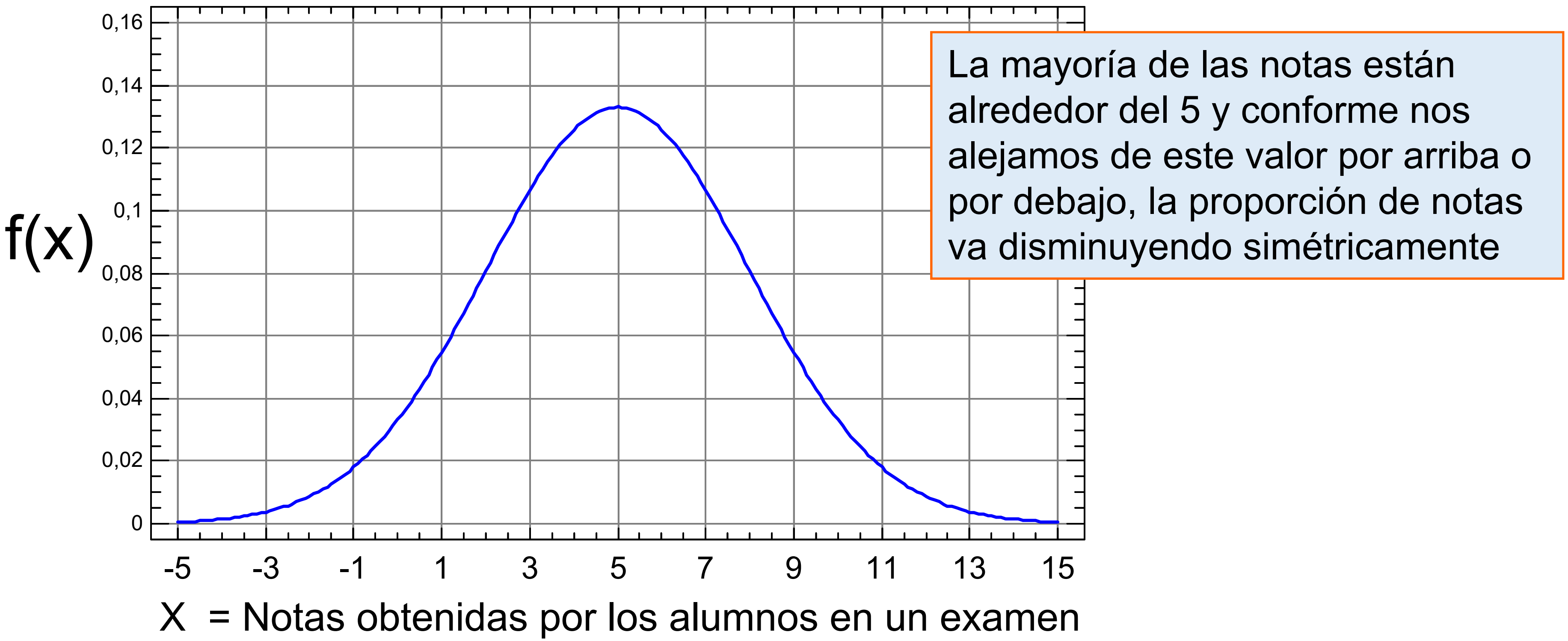
Variable aleatoria T: tiempo de búsqueda de ficheros en HD



Probabilidad de que el tiempo de búsqueda de un fichero en HD sea inferior a 0,23

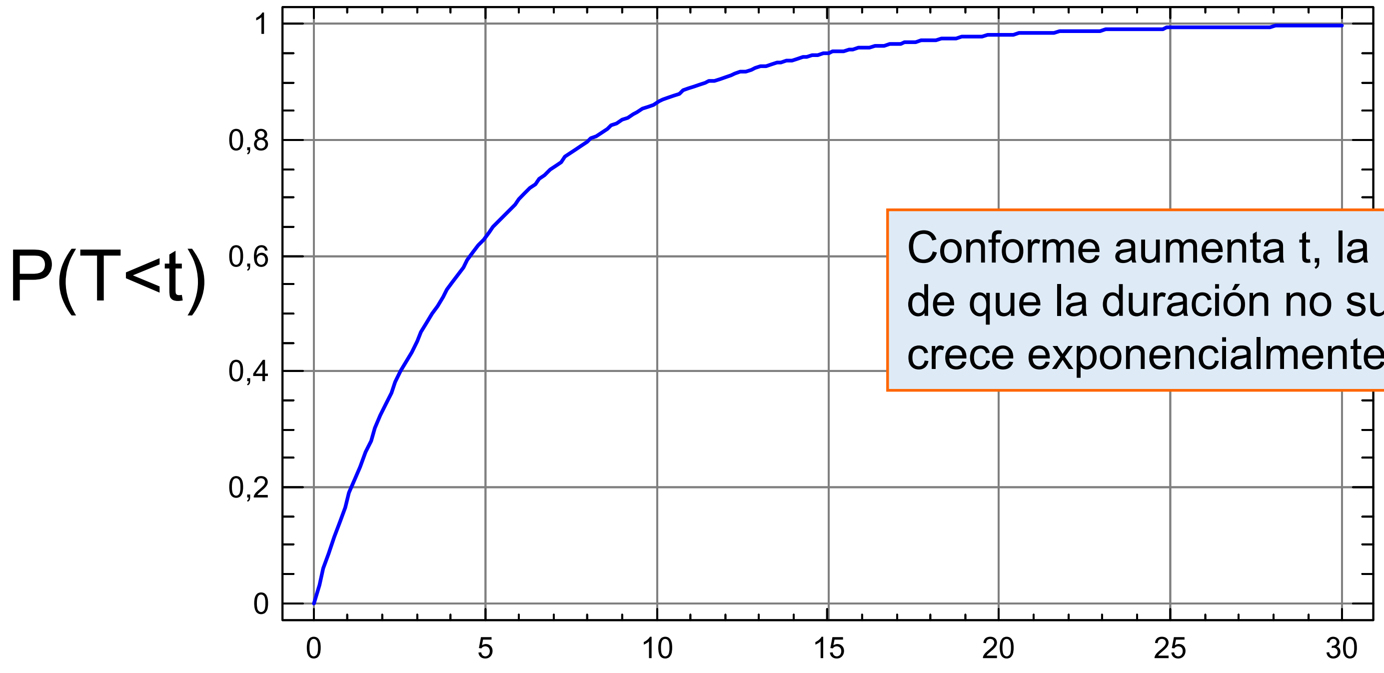
Modelos de distribuciones. Ejemplo 3: var continua

Variable aleatoria **X**: notas obtenidas por los alumnos en un examen



Modelos de distribuciones. Ejemplo 4: var continua

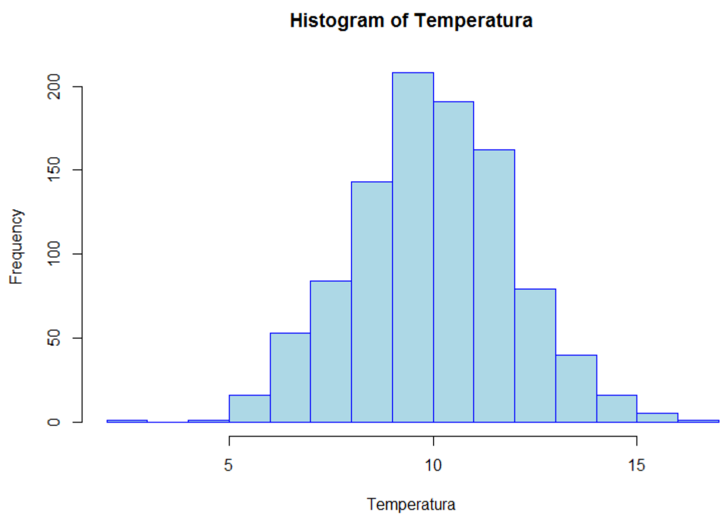
Variable aleatoria **T**: tiempo de duración de un comp. electrónico



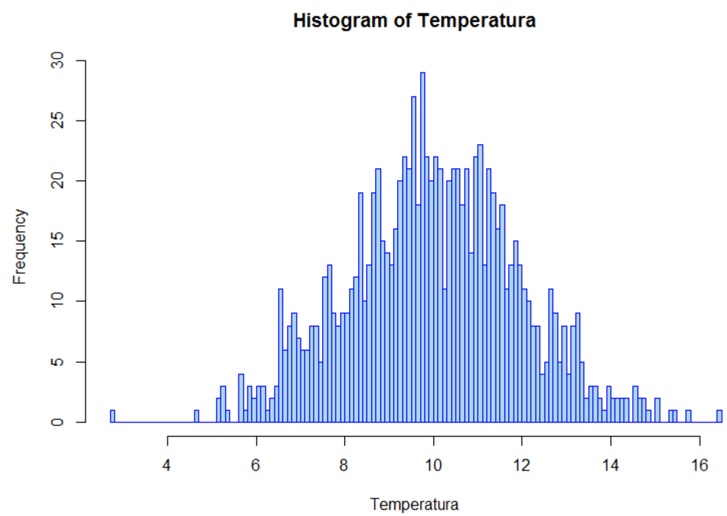
Conforme aumenta t, la probabilidad de que la duración no supere ese t crece exponencialmente

Distribuciones de probabilidad y v.a.

En el mundo "imaginario" de la *población,* a toda v.a X se le puede asociar una expresión matemática que se adecue a su pauta de variabilidad



Frecuencia para los días con temperatura de X ºC



Frecuencia para los días con temperatura de X ºC

Distribuciones de Probabilidad y v.a.

- Existen **modelos** (expresiones matemáticas) que se adecuan a las diferentes pautas de variabilidad de las variables aleatorias.

- Se agrupan en dos grandes bloques porque su tratamiento matemático es diferente:

v.a. discretas
- Binomial
- Poisson
- ...

v.a. continuas
- Normal
- Exponencial
- Uniforme
- …

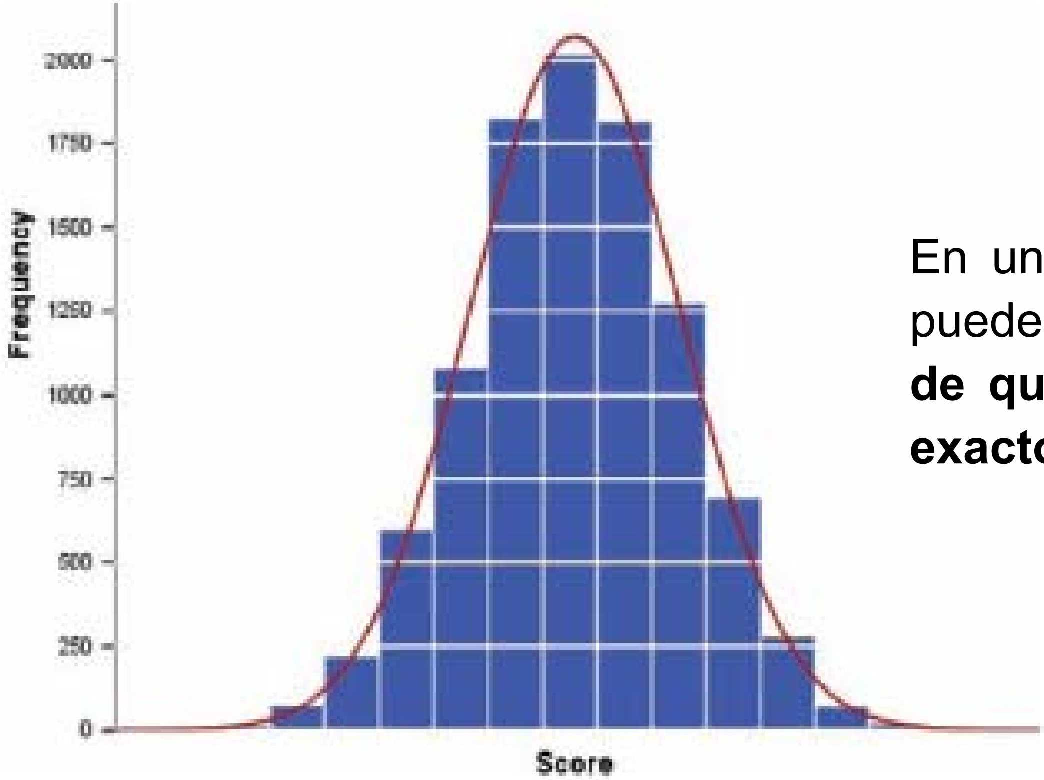
Distribuciones de Probabilidad y v.a.

Distribución

- Bernoulli
- Binomial
- Uniforme Discreta
- Geométrica
- Hypergeométrica
- Binomial Negativa
- Poisson
- Beta
- Beta (4 parámetros)
- Birnbaum-Saunders
- Cauchy
- Chi-Cuadrada
- Erlang
- Exponencial
- Exponencial (2 parámetros)
- Potenciación Exponencial
- F (Razón de Varianzas)
- Normal Plegada
- Gamma
- Gamma (3 parámetros)
- Gamma Generalizada
- Logística Generalizada
- Mitad Normal (2 parámetros)
- Gaussiana Inversa
- Laplace
- Valor Extremo Más Grande
- Logística
- Loglogística
- Loglogística (3 parámetros)
- Lognormal
- Lognormal (3 parámetros)
- Maxwell (2 parámetros)
- Chi-Cuadrada No-Central
- F No-Central
- t No-Central
- Normal
- Pareto
- Pareto (2 parámetros)
- Rayleigh (2 parámetros)
- Valor Extremo Más Chico
- t de Student
- Triangular
- U
- Uniforme
- Weibull
- Weibull (3 parámetros)

Distribuciones de probabilidad continuas

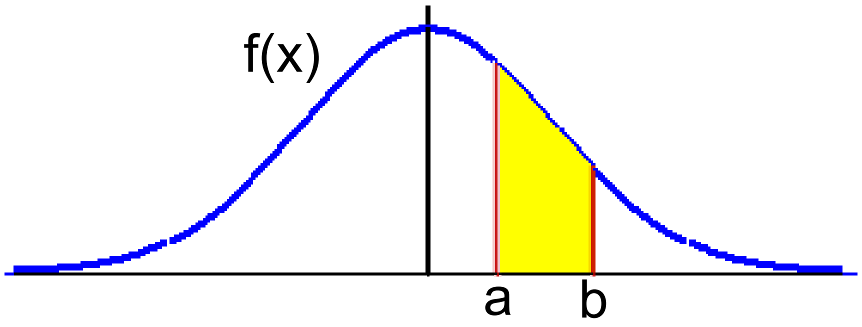
Para las **variables continuas**, la idealización del histograma de frecuencias, se denomina **función de densidad f(x)**



En una **v.a. continua: no** se puede calcular la **probabilidad de que la v.a. tome un valor exacto** (siempre es 0)

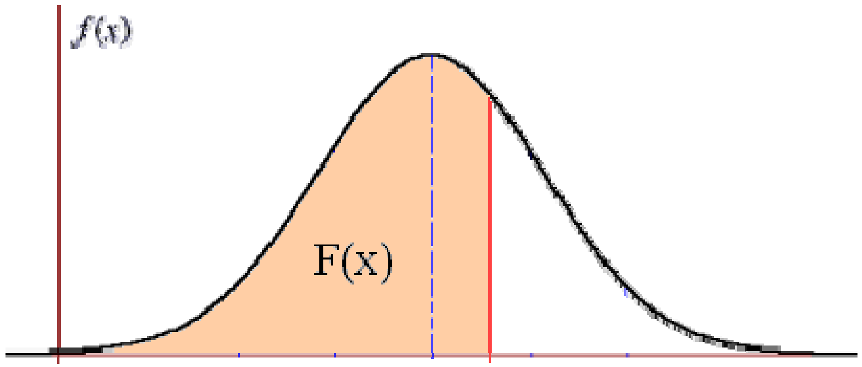
Distribuciones de probabilidad continuas

El área bajo la curva de la **función de densidad f(x)** nos da la probabilidad de que la variable tome valores en un determinado intervalo



$$\int_a^b f(x)dx = P(a < X \leq b)$$

O también las probabilidades acumuladas P(X ≤ a) o **función de distribución F(x)**



$$\int_{-\infty}^a f(x)dx = P(X \leq a)$$

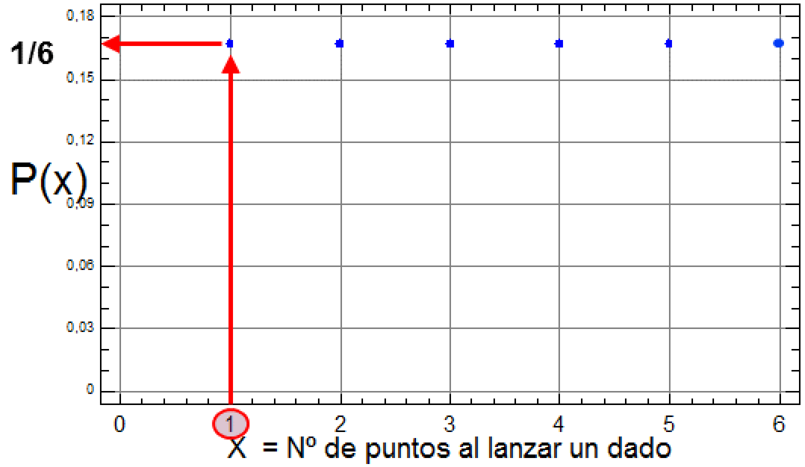
a

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

Distribuciones de probabilidad discretas

Para las **variables discretas**, se tiene **función de probabilidad P(x),** que da el valor de la probabilidad para cada valor posible de la variable.

Probabilidad de obtener el 1 al lanzar un dado



Y también las probabilidades acumuladas $P(X \leq a)$ o **función de distribución F(x)**

$$\sum_{\forall x_i / x_i \leq a} P(X = X_i) = \mathrm{P}(\mathrm{X} \leq \mathrm{a})$$

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

Función cuantil

- $Q(p)=a$ / $P(X \leq a) = p$

- Dada una probabilidad p, Q(p) devuelve el valor de la variable aleatoria (**a**), de forma que la probabilidad de que la v.a. **X** sea menor o igual que **a** es, precisamente, **p**
- En contextos de inferencia a a se le suele denominar **valor crítico**

Esperanza Matemática

Sea:

X: variable aleatoria (Tiempo, Estatura, Presión,...)

h(X) → función de la variable aleatoria:

- **h(Tiempo)** = Tiempo*100
- **h(Estatura)** = $(\text{Estatura})^2$
- **h(Tiempo)** = Σ(Tiempo)/N → *media*
- ...

Esperanza Matemática

- **Esperanza matemática → E(h(X))**
  - Si **X** es **discreta**:

$$E(h(X)) = \sum_{\forall x_i} h(x_i)\, P(X = x_i)$$

Función de Probabilidad

  - Si **X** es **continua**:

$$E(h(X)) = \int_{-\infty}^{+\infty} h(x) f(x) dx$$

Función de Densidad

Valor medio de la distribución

Si $h(x)=X^1 \Leftrightarrow E(h(\mathbf{x})) = E(X^1) = E(X)$

- Si **X** es **discreta**:

$$E(X) = \sum_{\forall x_i} x_i \, P(X = x_i) = m_X = \mu_X$$

La Esperanza Matemática de **X** es la Media de la distribución

- Si **X** es **continua**:

$$E(X) = \int_{-\infty}^{+\infty} X \, f(x)dx = m_X = \mu_X$$

Concepto intuitivo:
idealización del concepto de media aritmética de un conjunto de datos

Valor medio de la distribución

- Si la v.a. **Y** es una combinación lineal de **n** v.a.:

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

$$\boxed{\begin{aligned} E(Y) &= E(a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n) = \\ &= a_0 + a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) \end{aligned}}$$

Varianza de la distribución

Si **h**(x)=(**X**-**m**)**²** $\Leftrightarrow$ E(**h**(**x**))=E((**X**-**m**)**²**)

Varianza de la distribución

- Si **X** es **discreta**:

$$E((X-m)^2) = \sum_{\forall x_i} (x_i - m)^2 P(X = x_i) = \sigma_X^2$$

- Si **X** es **continua**:

$$E((X-m)^2) = \int_{-\infty}^{\infty} (X-m)^2 f(x)dx = \sigma_X^2$$

Varianza de la distribución

- Si la v.a. **Y** es una combinación lineal de 2 v.a.:

$$Y = a_0 + a_1X_1 + a_2X_2$$

$$\boxed{\begin{aligned}&\sigma^2(Y) = \sigma^2(a_0 + a_1X_1 + a_2X_2) = \\ &= a_1^2\sigma^2(X_1) + a_2^2\sigma^2(X_2) + 2a_1a_2\mathrm{Cov}_{X_1,X_2}\end{aligned}}$$

Esperanza Matemática

Resumiendo

Lo que hay que saber…

- **Concepto**: entender e identificar
- **Propiedades**: aplicar
- **Nomenclatura**: escribir correctamente

Media población (teórica) = m = μ = E(X) = Esperanza matemática de X

Varianza población (teórica) = σ^2 = $E(X-m)^2$ = Esperanza matemática de …

Ejemplos Esperanza Matemática

1) Si **X** (Estatura alumnos UPV (m)) ~ m_x=1,70, σ^2_x=1

Y = Estatura alumnos UPV (cm) = 100**X**

→ m_y= **100**x**1,70 y** σ_y^2 **= 100**^{2} x**1)**

2) Si **X**, **Y** y **Z** son unidades vendidas de 3 productos al mes e independientes y m_x=9, σ^2_x=3, m_y =8, σ^2_y=1, m_z=5 y σ^2_z =2

Si el beneficio mensual W = 10X + 20Y+ 15Z →

$m_w = m_x + m_y + m_z = 10x9 + 20x8 + 15x5 = 325$

$\sigma_w^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2 = 10^2x3 + 20^2x1 + 15^2x2 = 1150$

Distribuciones de probabilidad o “modelos”

- Distintos modelos que se corresponden con el comportamiento de ciertas variables y con diferentes formas características:
  - Normal
  - Uniforme
  - Binomial
  - ….
- La **Normal** es la distribución de probabilidad que con más frecuencia aparece, ya que, hay muchas variables asociadas a fenómenos naturales que siguen este modelo.
- La mayor parte de las técnicas de **Inferencia Estadística paramétrica** para variables continuas asumen que las poblaciones muestreadas son normales.

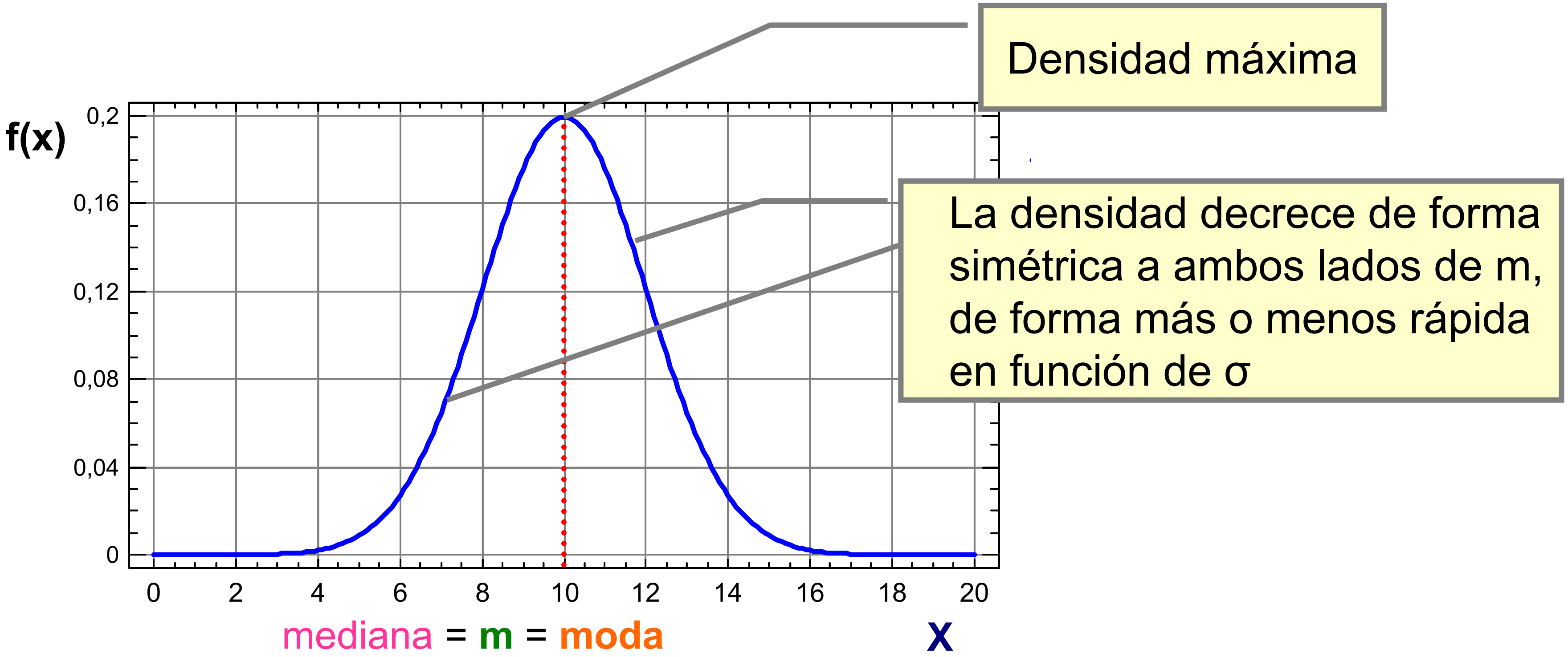
Distribuciones de probabilidad

Distribución Y parámetros	**f(x) si continuas o P(x) si discretas**	**F(x)**	**Q(x)**	**Muestras aleatorias (simulación)**
Binomial (n, p)	dbinom(x, n, p)	pbinom(x, n, p)	qbinom(x, n, p)	rbinom(x, n, p)
Poisson (λ)	dpois(x, λ)	ppois(x, λ)	qpois(x, λ)	rpois(x, λ)
Exponencial (α)	dexp(x, α)	pexp(x, α)	qexp(x, α)	rexp(x, α)
Normal(m, σ)	dnorm(x, m, σ)	pnorm(x, m, σ)	qnorm(x, m, σ)	rnorm(x, m, σ)
…..	…..	…..	…..	…..

Distribución normal

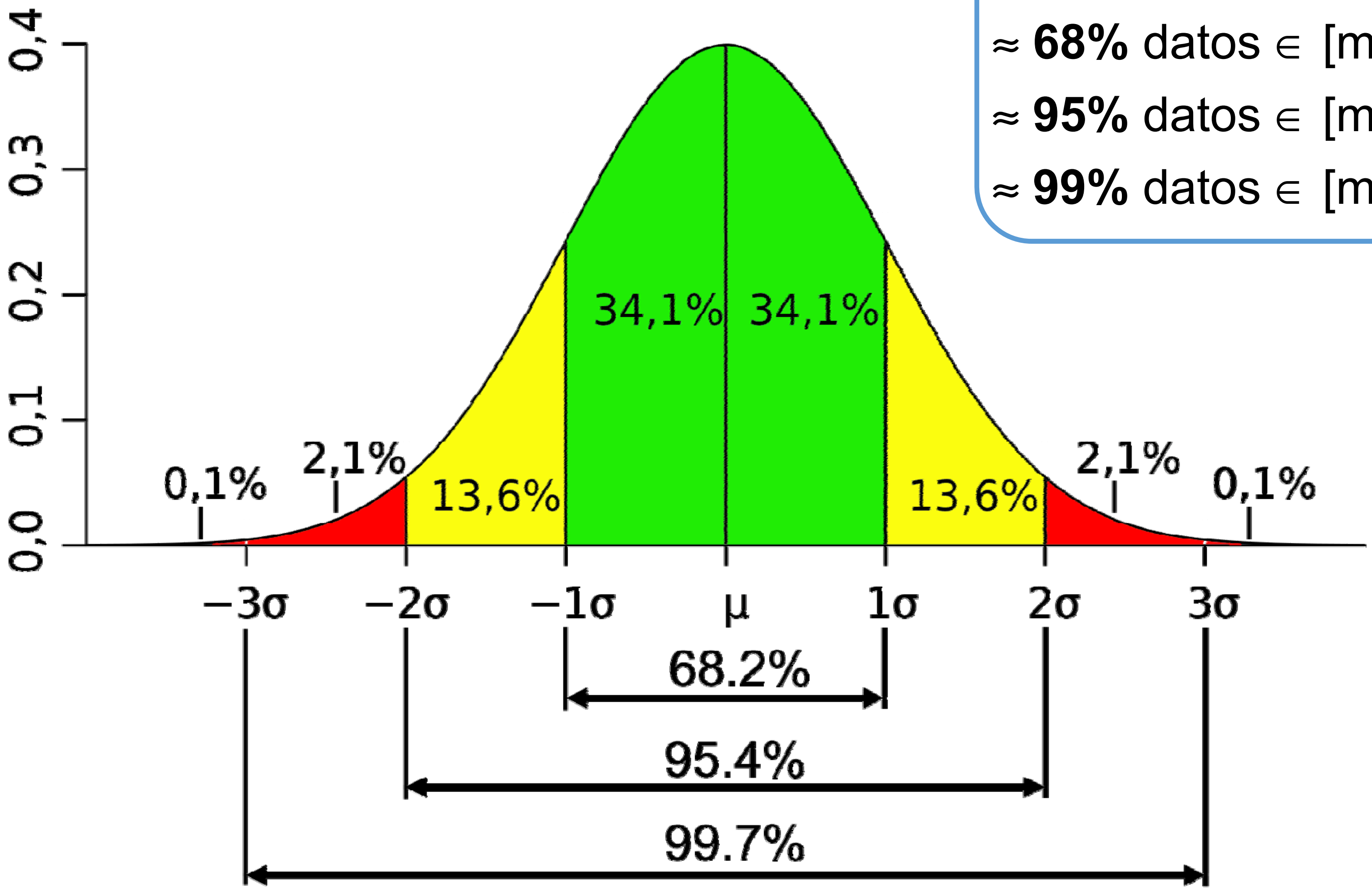
$$X \sim N(m, \sigma)$$

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



Coeficiente de asimetría=0
Coeficiente de curtosis=0

Propiedades de la Normal



Si X ~ N(m,σ) →

≈ **68%** datos ∈ [m-σ, m+σ]

≈ **95%** datos ∈ [m-2σ, m+2σ]

≈ **99%** datos ∈ [m-3σ, m+3σ]

Probabilidades distribución Normal

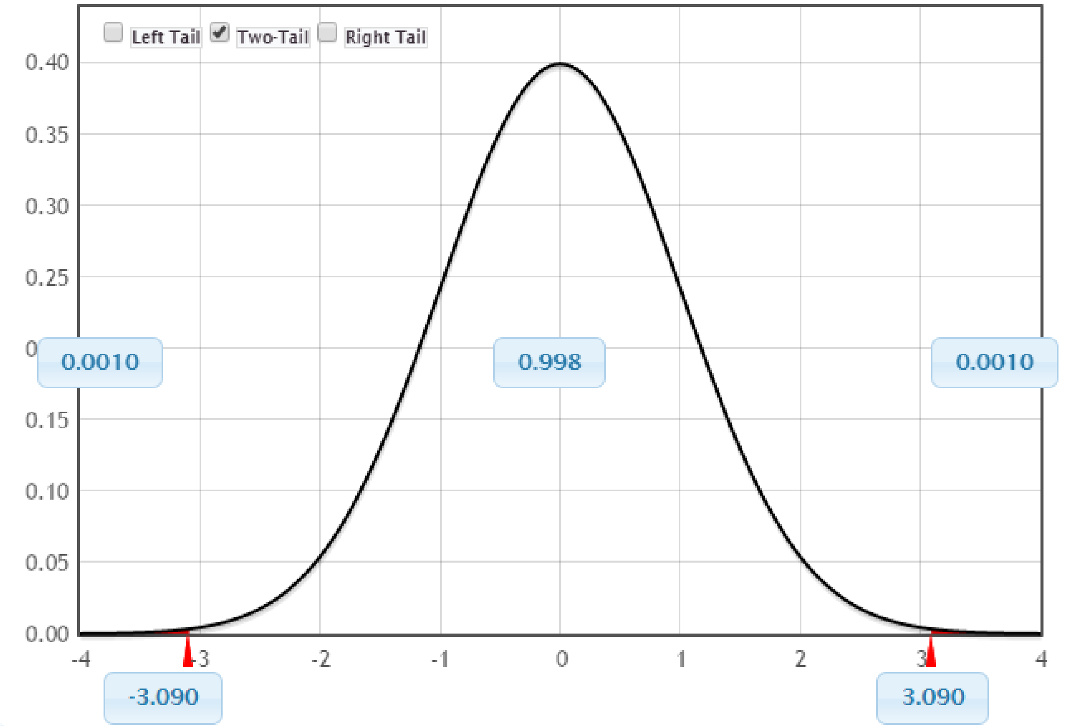
```
## N(m=5, S=2). P(X < 2)?
> pnorm(2,mean= 5, sd=2)
[1] 0.0668072


## N(m=5, S=2). P(X > 2)?
> 1-pnorm(2,5,2)
[1] 0.9331928
## O bien
> pnorm(2,5,2, lower.tail = F)
[1] 0.9331928
```

Probabilidades distribución Normal

```
> ## ¿Entre que valores está el 95% (aproximadamente)
de los datos de una v.a.  N(m=0, S=4.5)?
> ## x1, x2 / P(x1 < X < x2) = 0.95
> ## x2 / P(X > x2) = (1-0.95)/2
> x2 <- qnorm(((1-0.95)/2), 0, 4.5, lower.tail=F)
>
> ## x1 / P(X < x1) = (1-0.95)/2
> x1 <- qnorm(((1-0.95)/2), 0, 4.5)
>
> x1
[1] -8.819838 ## m - 2*S
> x2
[1] 8.819838 ## m - 2*S
```

Normal tipificada o estandarizada

- Parámetros:
  - su media es 0
  - su desviación típica es 1
- Se simboliza como: $\boldsymbol{Z \sim N(0,1)} \equiv N(m_Z{=}0, \sigma_Z{=}1)$



Normal tipificada o estandarizada

- Cualquier v.a. N(m, σ) puede estandarizarse, de este modo obtenemos valores comparables (estándar) entre variables normales con diferentes m y σ

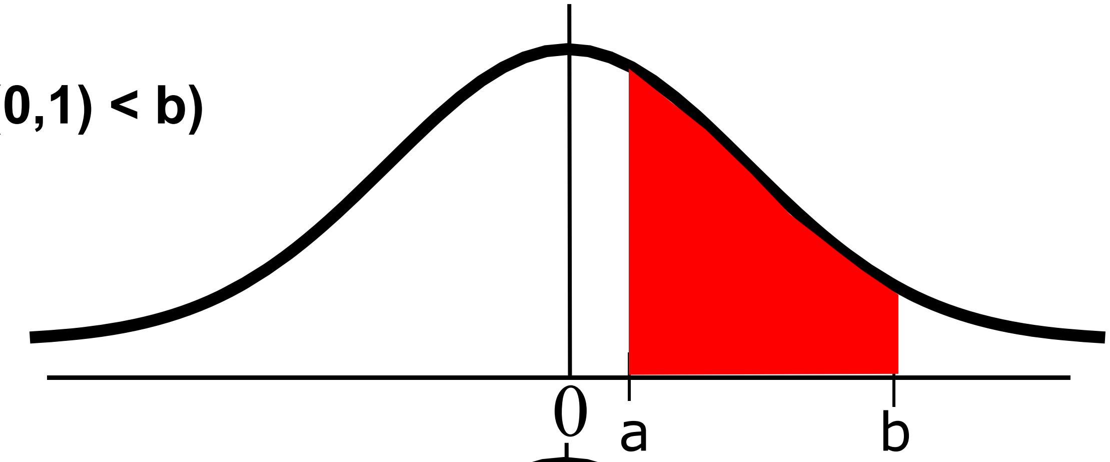
- Es un tipo de **transformación**

- A los valores de una v.a. normal tipificada también se les llama ***z-scores*** y expresan el valor de una v.a. en términos del número de desviaciones típicas con respecto a la media

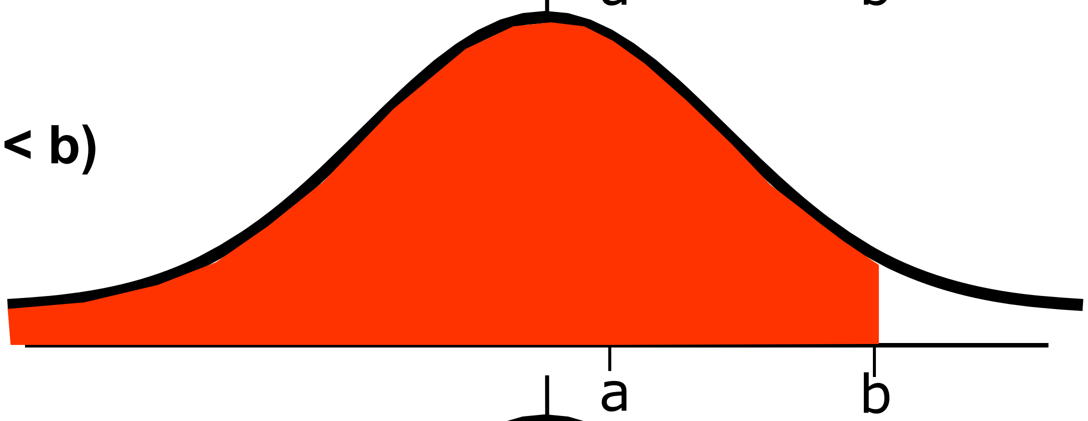
$$X \sim N(m, \sigma) \xrightarrow[\dfrac{X-m}{\sigma}]{\text{Tipificar}} Z \sim N(0,1)$$

Probabilidad y áreas bajo la curva de f(x)

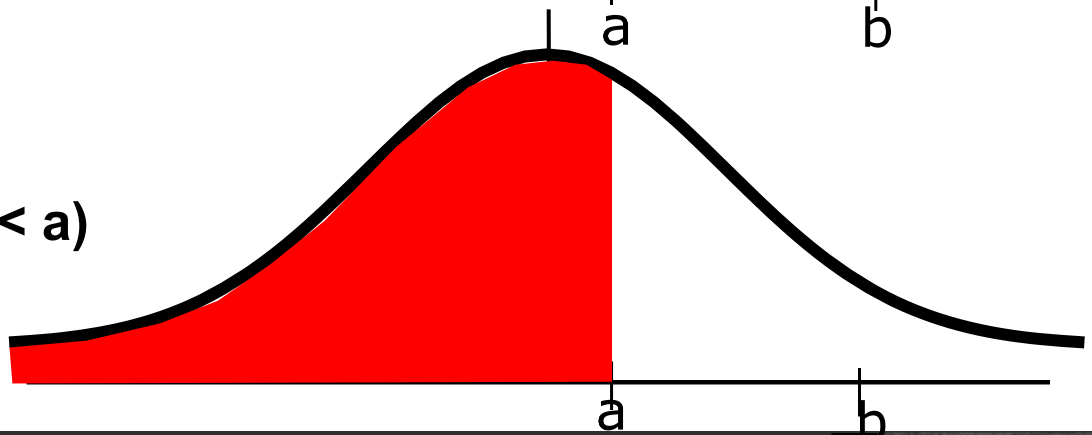
= P (a < N(0,1) < b)



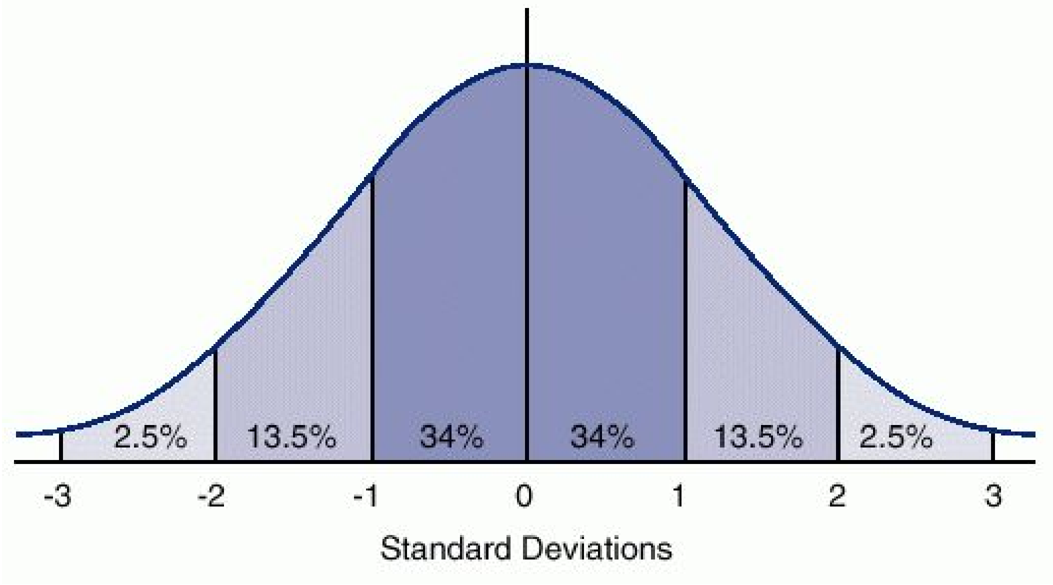
= P (N(0,1) < b)

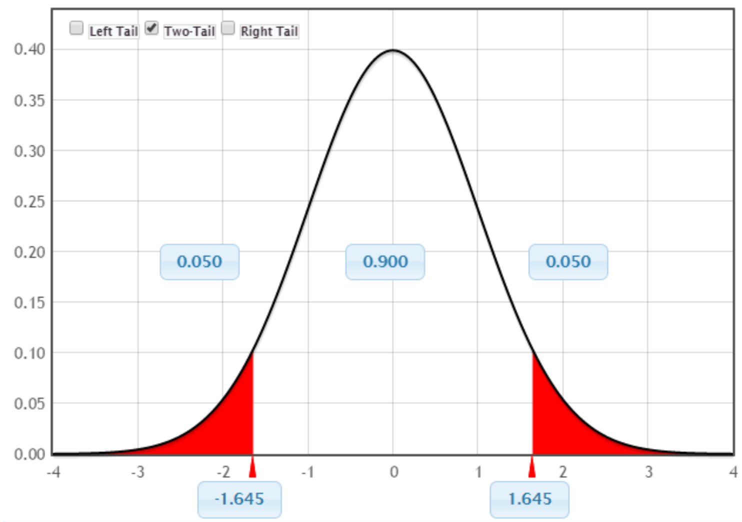


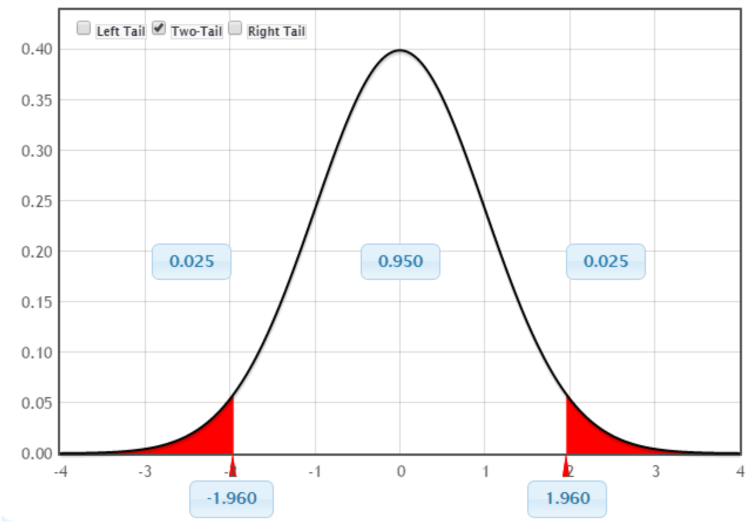
= P (N(0,1) < a)



Propiedades de los *z*-scores







Probabilidades distribución N(0,1)

```
## P(X<=2)
> pnorm(2, mean=0, sd=1, lower.tail=T)
[1] 0.9772499

## P(X>=2)
> pnorm(2, mean=0, sd=1, lower.tail=F)
[1] 0.02275013

## O bien 1-pnorm(2, mean=0, sd=1, lower.tail=T)


## P(-2<= X <=2)
> pnorm(2, mean=0, sd=1, lower.tail=T)- pnorm(-2,
mean=0, sd=1, lower.tail=T)
[1] 0.9544997
```

Probabilidades distribución Normal

```
## x / P(X <= x) = 0,99
> qnorm(0.99, 0, 1)
[1] 2.303598

## P(x1 ≤ X ≤ x2)= 0,95
> x1<-qnorm(0.025, mean=0, sd=1, lower.tail=T)
> x2<-qnorm(0.025, mean=0, sd=1, lower.tail=F)
> x1
[1] -1.959964
> x2
[1] 1.959964
```

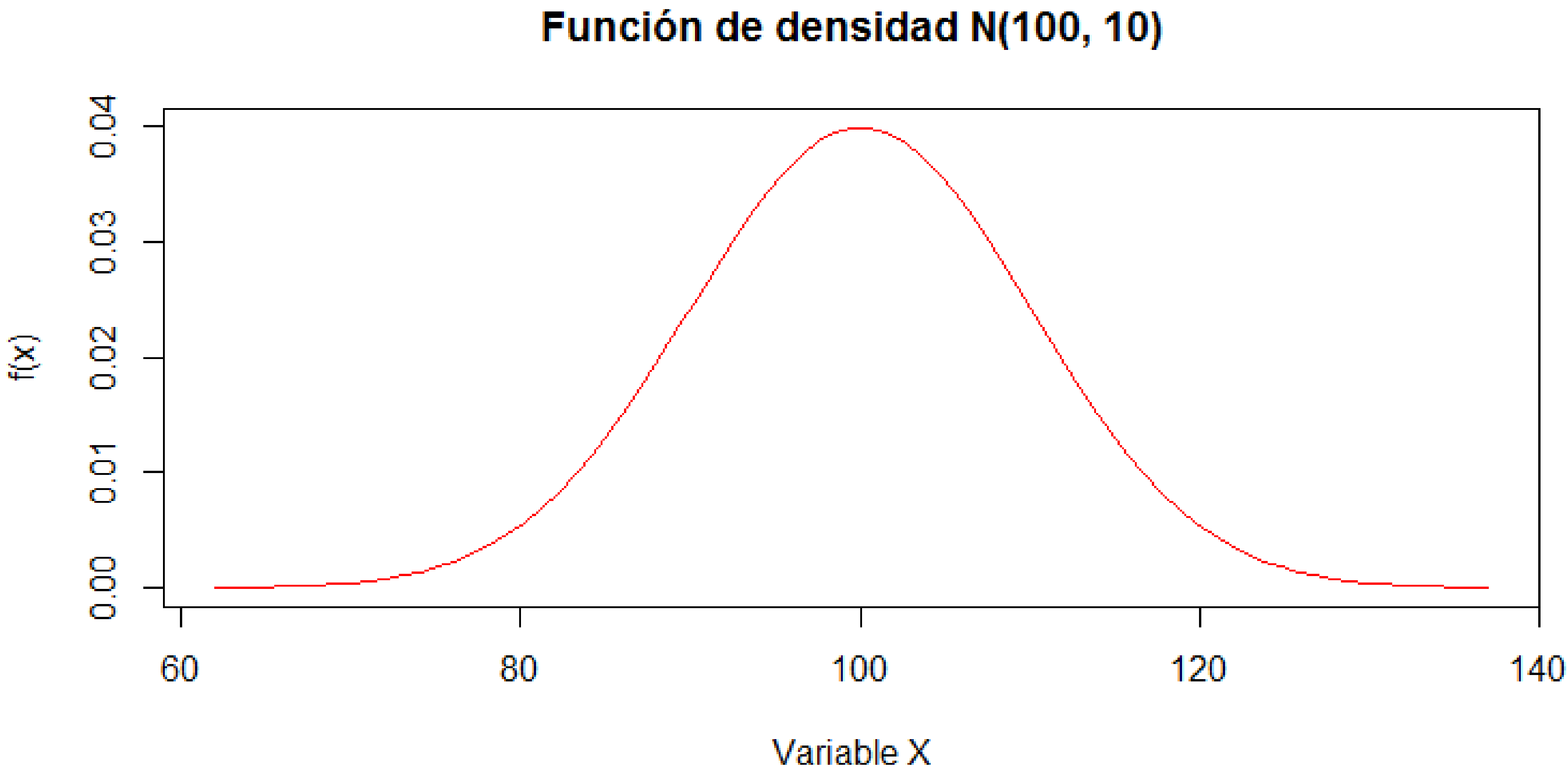
Probabilidades distribución Normal

```
#### GRÁFICOS f(x), F(x) y Q(x)
## Función de densidad N(100, 10)

# Valores de X. 99% valores están en [m-3sigma,m+3sigma]
# Límites. [m-3sigma,m+3sigma]
m<-100
sigma<-10
li<-m-3*sigma
ls<-m+3*sigma

# Otra forma de obtener los límites útil para cualquier distribución
# obtener x1 y x2 tal que P(x1 <= X <= x2)=0.999
li<-trunc(qnorm(0.0001, m, sigma))
ls<-trunc(qnorm(0.9999, m, sigma))
```

Probabilidades distribución Normal

```
# Número de puntos a dibujar
npuntos<-ls-li
x<-seq(li, ls, length.out=npuntos)


# Valores de y
y<-dnorm(x, m, sigma)


# Dibujar f(x)
fnorm<-plot(x, y, type="l", xlab="Variable X", ylab="f(x)",
main="Función de densidad N(100, 10)", col="red")
```

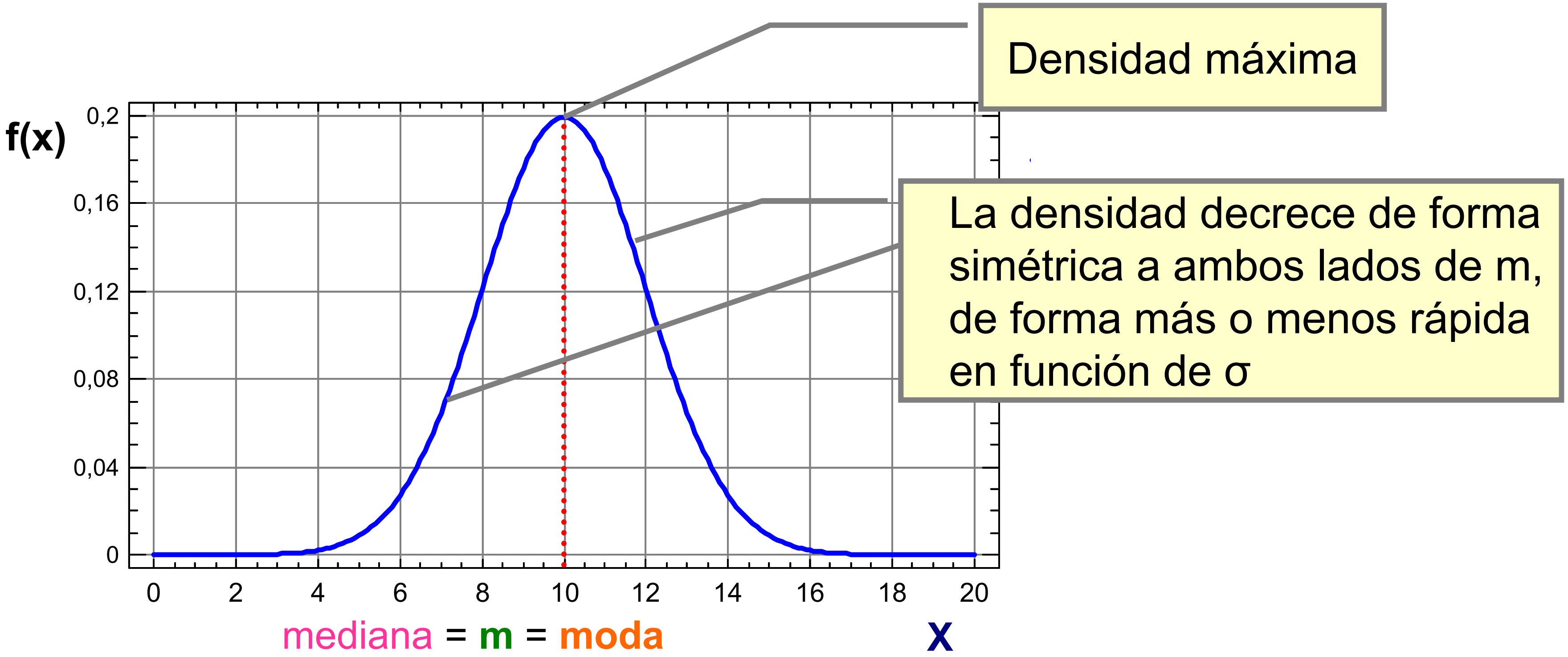
Probabilidades distribución Normal



Distribución uniforme

$$X \sim N(m, \sigma)$$

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



Coeficiente de asimetría=0
Coeficiente de curtosis=0

Distribución Uniforme

- Se utiliza en v.a. continuas en las que la única información de la que se dispone sobre su comportamiento es que toman valores en un intervalo, y que la densidad de probabilidad para todos los valores de ese intervalo es la misma.
- La distribución uniforme tiene una aplicación muy importante en **simulación**.

Ejemplos:

- Tiempo de acceso a un archivo en un disco duro ~ 1 y 3 ms
- Tamaño de un tipo de archivo ~ entre 100 y 1000 Kb
- Distancia entre origen y destino recorrida por un mensaje en una red regular tipo toro
- etc

Distribución Uniforme: Definición

- Una v.a continua **X** tiene una **Distribución Uniforme en (a,b)** si su **función de densidad** es:
  - constante en un intervalo (a,b)
  - nula fuera de dicho intervalo

- Se simboliza como:

$$\boxed{X \sim U(a,b)}$$

Funciones de densidad y probabilidad acumulada

$$X \sim U(a,b)$$

Función de densidad f(x)

$$f(x) = K = \frac{1}{b-a} \qquad x \in [a, b]$$

$$f(x) = 0 \qquad x \notin [a, b]$$

P(X≤x) = P(X<x)

$$P(X \leq x) = 0 \qquad X < a$$

$$P(X \leq x) = \frac{x-a}{b-a} \qquad a \leq X \leq b$$

$$P(X \leq x) = 1 \qquad X > b$$

Media

$$E(X) = \frac{a+b}{2}$$

Varianza

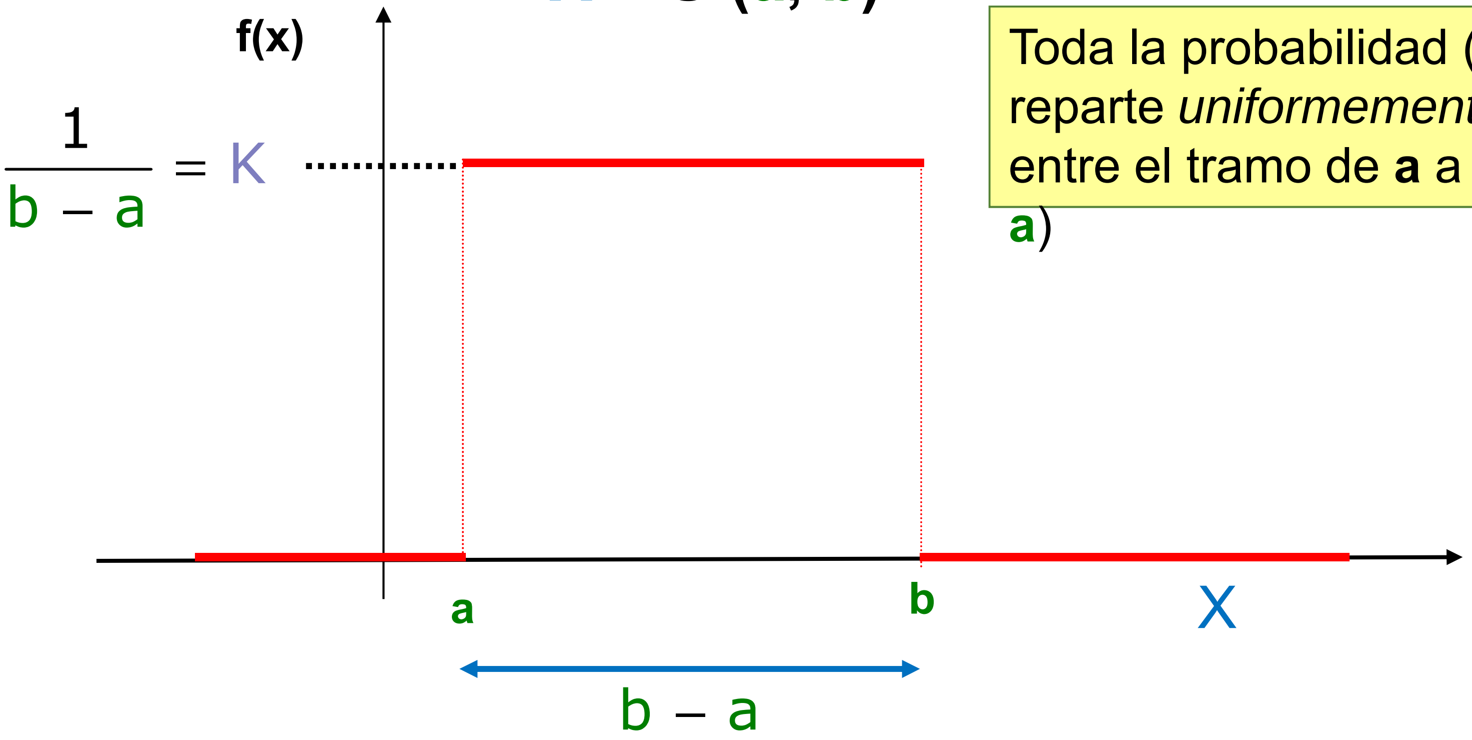
$$\sigma^2(X) = \frac{(b-a)^2}{12}$$

¡Recordar! UD4-Parte 1

Cómo se calcula la Esperanza matemática

Función de densidad

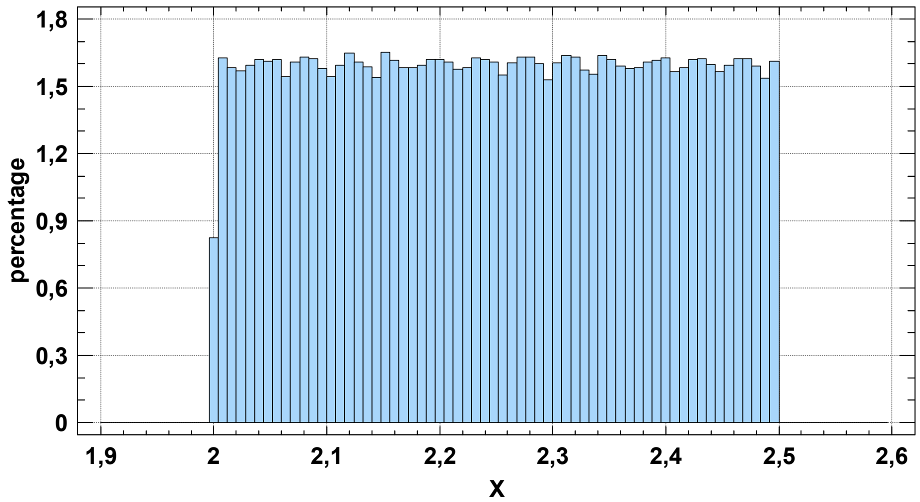
X ~ U (a, b)



Toda la probabilidad (**1**) se reparte *uniformemente* entre el tramo de **a** a **b** (**b - a**)

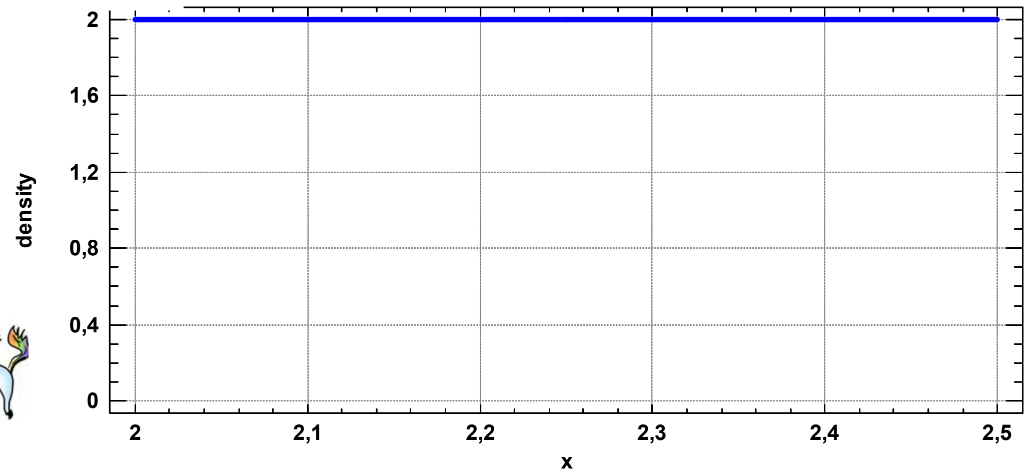
Histograma y Función de densidad

muestra

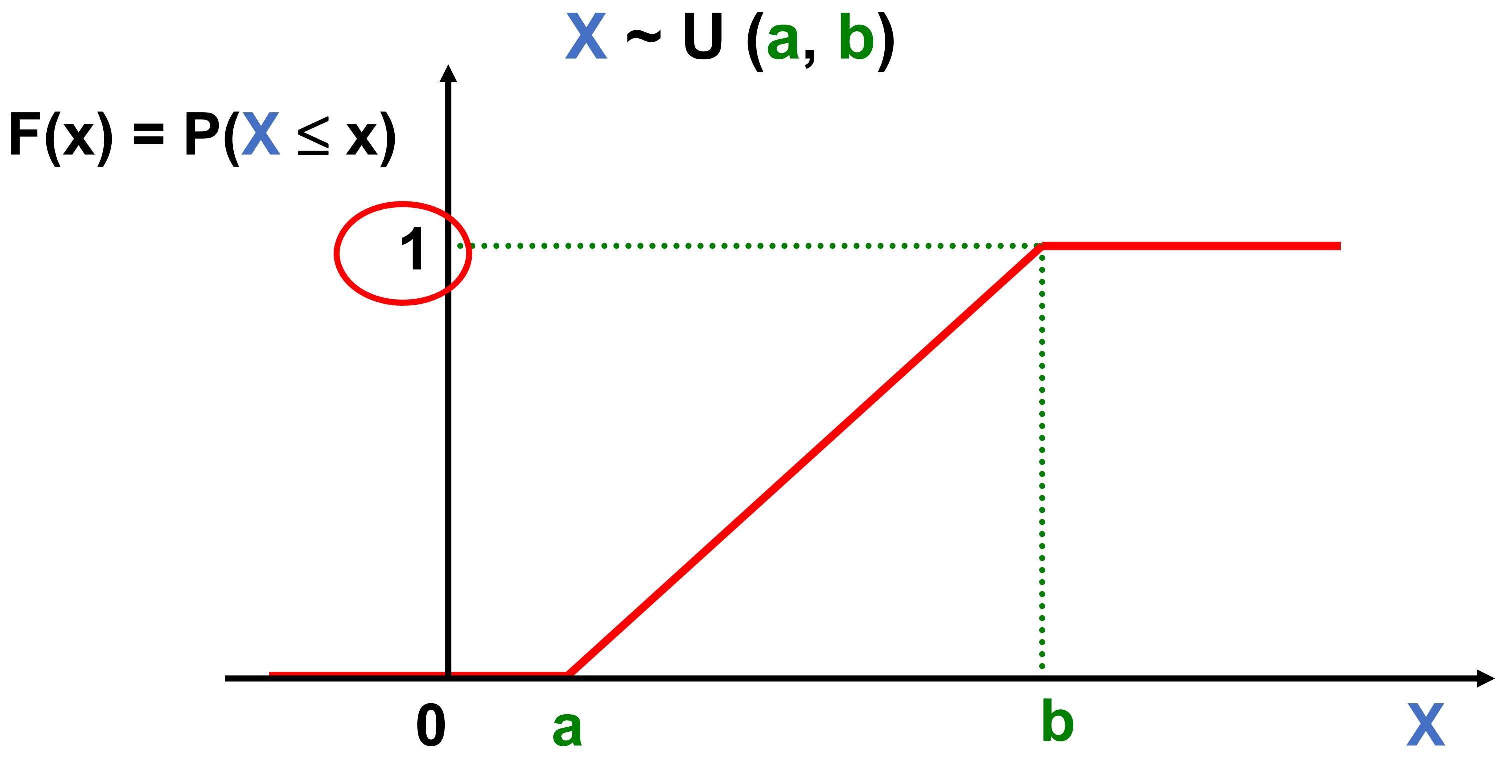


población

$X \sim U\,(2,\ 2{,}5)$

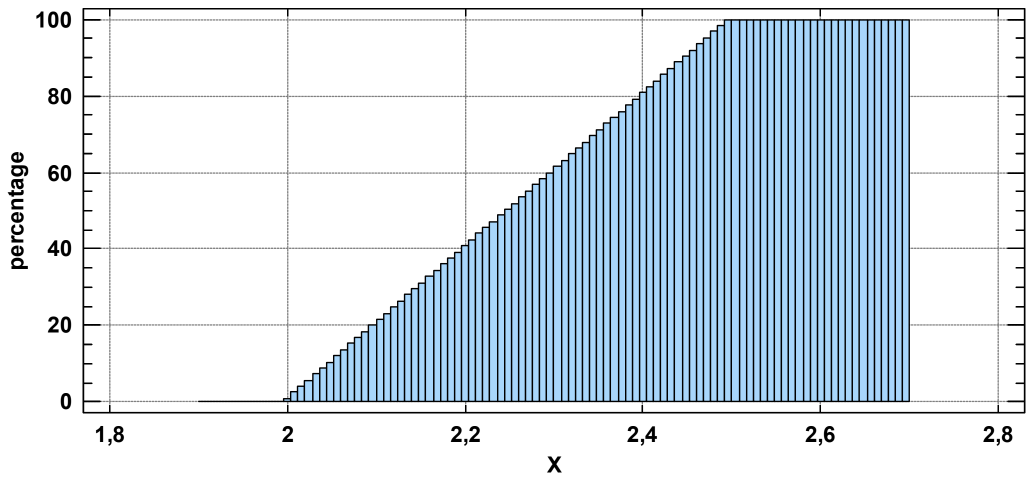


Probabilidad acumulada



Frecuencias y Probabilidades acumuladas

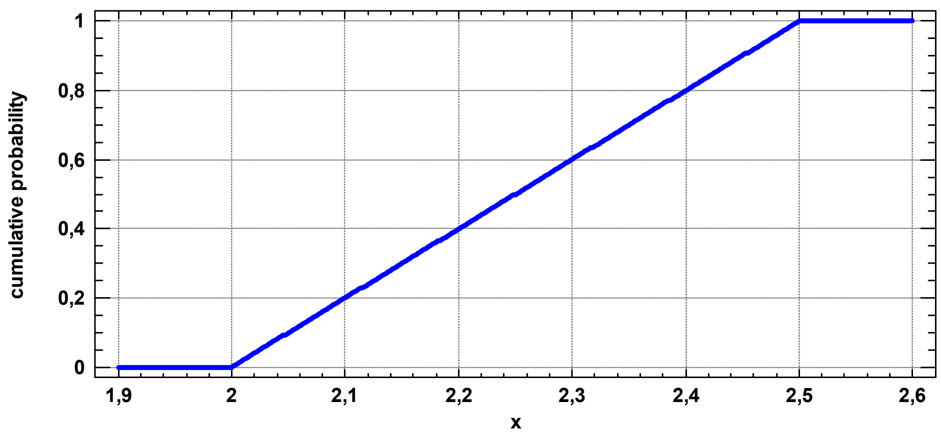
muestra



$X \sim U\,(2,\ 2{,}5)$

población

$P(X \leq x)$



La distribución Binomial

Dado un suceso **A** (**cliente compra en la primera visita a la web**) de probabilidad **p** (se sabe que el **10%** de los clientes, por ejemplo, compra en la primera visita) asociado a un determinado experimento aleatorio. Se llevan a cabo **n** repeticiones independientes del experimento (**comprobar, en los 131 clientes, si éste compra en la primera visita o no**), y sea **X** el número de veces que se presenta el suceso **A**

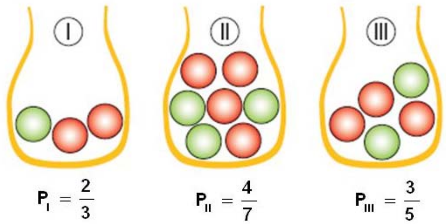
La variable **X** así definida "**nº de clientes que compran en la primera visita a la web**" sigue una **distribución Binomial** que depende de los parámetros **n** y **p**

$$\boxed{X \sim B(n,p)}$$

Los valores posibles de **X** son: 0, 1, 2,.........., n

Ejemplos

- Nº de chicos de un grupo de 20 estudiantes de 1º de Informática.
- Nº de piezas defectuosas extraídas de una partida de 50.
- Nº de personas que responden “sí” a una pregunta, de entre un grupo de 100.
- Nº de caras que se obtienen al lanzar 200 veces una moneda
- Nº de libros extraídos de una partida de 10 que pertenecen a una determinada categoría.



Función de probabilidad Binomial

Se demuestra:

Función de probabilidad

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

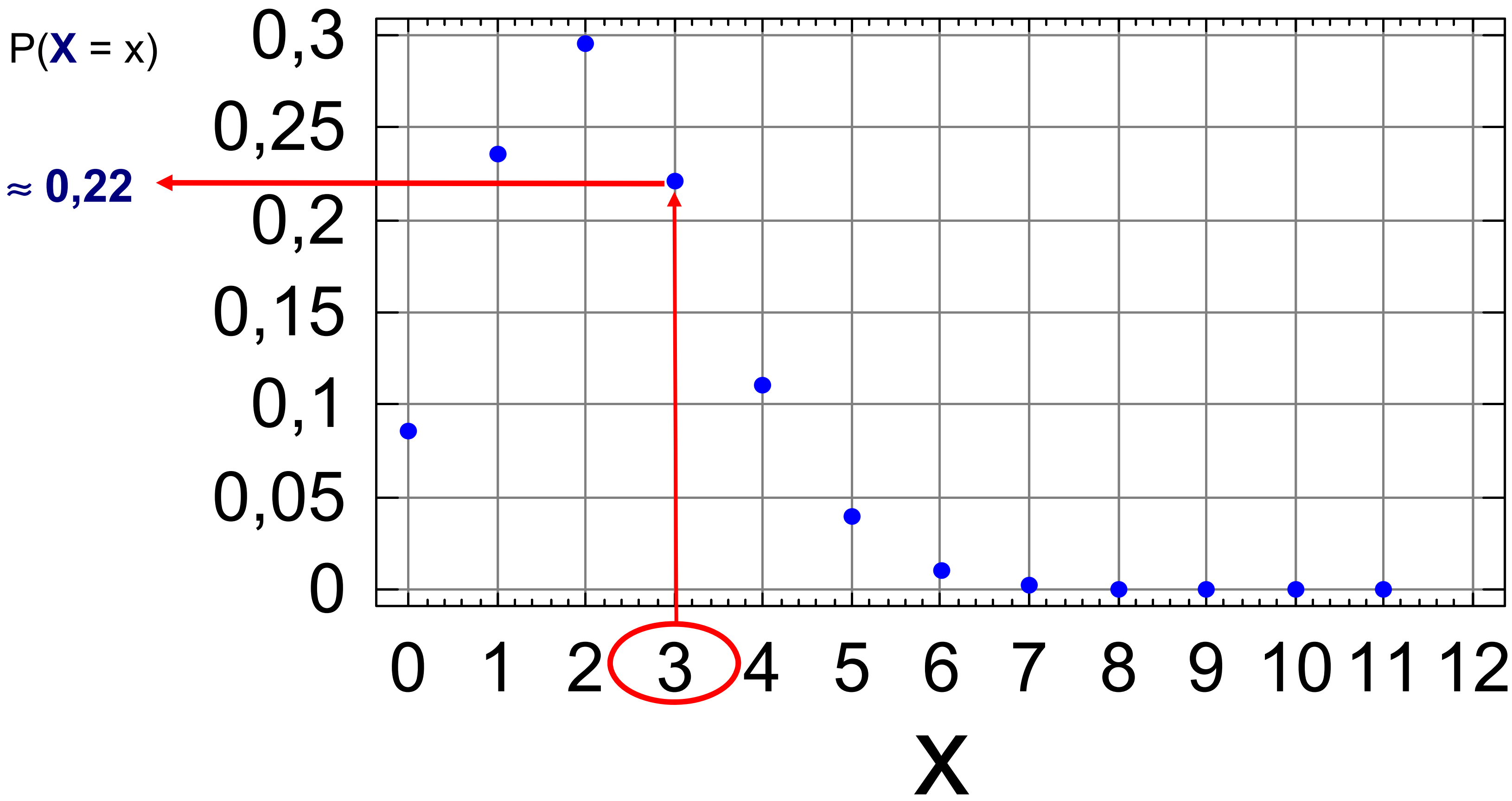
Media y Varianza (Esperanza matemática)

$$E(X) = np$$

$$\sigma^2(X) = np(1-p)$$

Gráficamente: Función de Probabilidad

X ~ Binomial (**n**=11, **p**=0,2) → P(**X** = 3) ≈ **0,22**



Probabilidades distribución Binomial

B(n=10, p=0.25). ¿P(X = 3)?

```
# FUNCIÓN DE PROBABILIDAD, CUANTÍA O MASA P(x)
dbinom(3, 10, 0.25)
```

```
## [1] 0.2502823
```

B(n=10, p=0.25). ¿P(X = 0), P(X = 1), P(X = 2) y P(X = 3)?

Probabilidades distribución Binomial

```
# FUNCIÓN DE DISTRIBUCIÓN F(x)
dbinom(0:3, 10, 0.25)

## [1] 0.05631351 0.18771172 0.28156757 0.25028229
```

B(n=10, p=0.25). ¿P(X < 4)?

```
# Sumando las probabilidades
sum(dbinom(0:3, 10, 0.25))

## [1] 0.7758751

# Utilizando la F(x). P(X <= 3)
pbinom(3, 10, 0.25)

## [1] 0.7758751

# Utilizando la F(x) y las propiedades de la probabilidad
# P(X < 4) = 1 - P(X >= 4) = 1 - P(X > 3)
1 - pbinom(3, 10, 0.25, lower.tail = F)

## [1] 0.7758751
```

Probabilidades distribución Binomial

B(n=10, p=0.25). ¿P(X > 2)?

```
# Utilizando la F(x). P(X > 2)
pbinom(2, 10, 0.25, lower.tail = F)

## [1] 0.4744072

# Utilizando la F(x) y las propiedades de la probabilidad
# P(X > 2) = 1 - P(X <= 2)
1 - pbinom(2, 10, 0.25)

## [1] 0.4744072
```

B(n=10, p=0.25). ¿P(2 < X <= 4)?

```
# P(2 < X <= 4) = P(X <= 4) - P(X <= 2)
F4 <- pbinom(4, 10, 0.25)
F2 <- pbinom(2, 10, 0.25)
F4

## [1] 0.9218731

## [1] 0.5255928

## [1] 0.3962803
```

Probabilidades distribución Binomial

B(n=10, p=0.25). ¿P(2 <= X <= 4)?

```
# P(2 <= X <= 4) = P(1 < X <= 4) = P(X <= 4) - P(X <= 1)
F4 <- pbinom(4, 10, 0.25)
F1 <- pbinom(1, 10, 0.25)
F4

## [1] 0.9218731

F1

## [1] 0.2440252

F4 - F1

## [1] 0.6778479
```

La distribución de Poisson

- En algunas situaciones es necesario utilizar variables aleatorias binomiales con un valor muy elevado de **n** y un valor muy bajo de **p**.
- En estos casos, la mayoría de las veces resulta casi imposible conocer **n** y **p** con exactitud, pero se tiene cierta idea de su valor medio **m = np**
- La variable **X** así definida sigue una distribución denominada **distribución de Poisson** que depende sólo del parámetro λ (λ= **np**)

$$\boxed{X \sim \text{Poisson}(\lambda)}$$

- Los valores posibles de **X** son: 0, 1, 2,.......... (como la Binomial)

Ejemplos de variables de Poisson

- Nº de registros dañados en una base de datos documental a lo largo de 365 días.
- Nº de coches que pasan por minuto por un control durante un fin de semana.
- Nº de errores de compilación en un programa.
- Nº de fallos en un sistema informático de consultas a lo largo de un mes.

Función de probabilidad de Poisson

Límite la distribución **Binomial** cuando:

- **n** es grande (**n** $\rightarrow \infty$) y
- **p** es pequeño (**p** $\rightarrow$ **0**)

Se demuestra:

Función de probabilidad

$$\boxed{P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}}$$

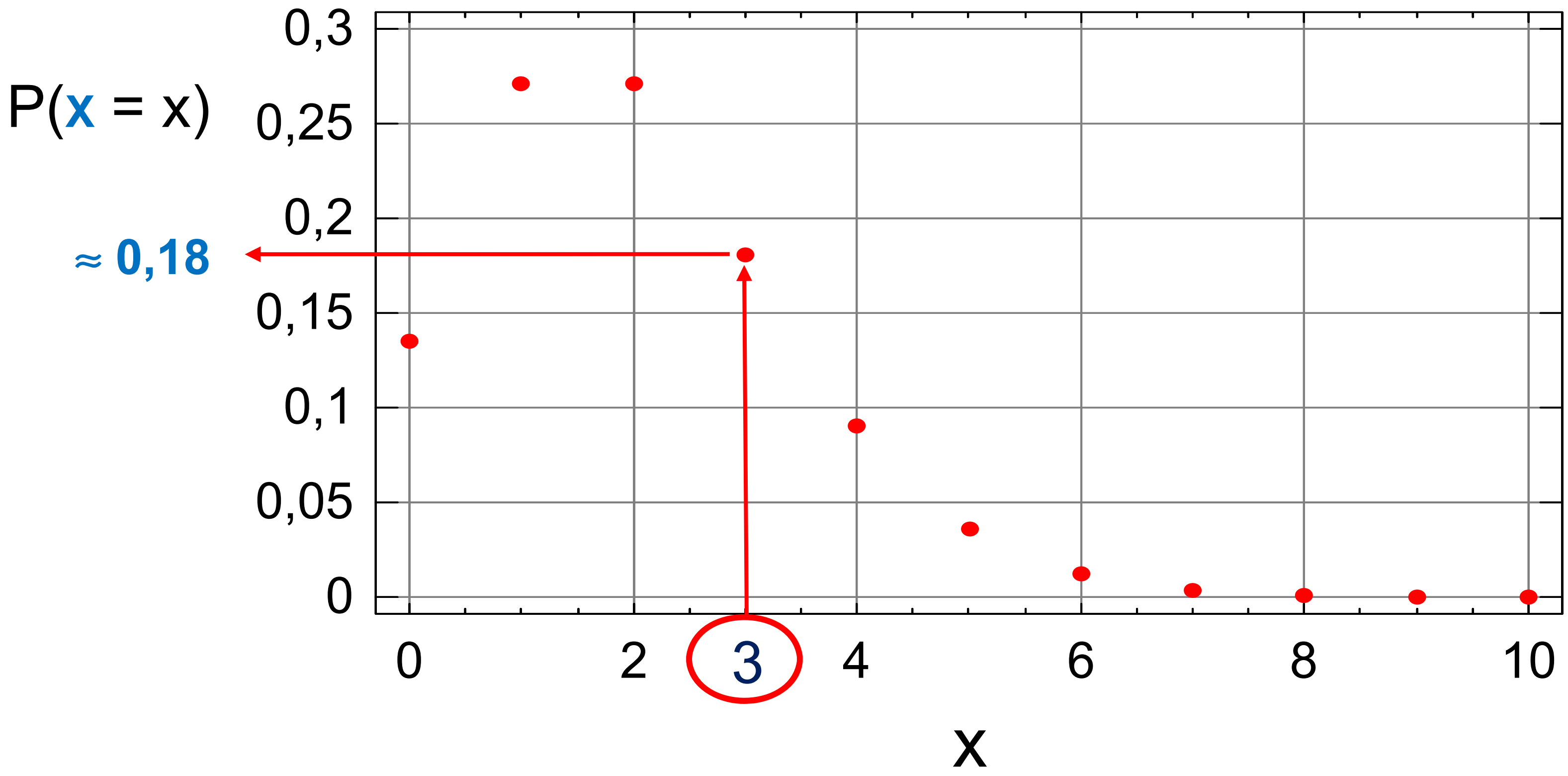
$$P(X \leq x) = \sum_{0}^{x} e^{-\lambda} \frac{\lambda^x}{x!}$$

Media y Varianza (Esperanza matemática)

$$\boxed{E(X) = \lambda} \qquad \boxed{\sigma^2(X) = \lambda}$$

Gráficamente: Función de Probabilidad

- **X** ~ Poisson(λ=2) → P(**X** = 3) ≈ **0,18**



Teorema Central del Límite

En condiciones muy generales, la suma de variables aleatorias independientes tiende a distribuirse normalmente, a medida que aumenta el número de sumandos.

Sean $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_n}$ variables aleatorias que se distribuyen según una **distribución <u>cualquiera</u>** e **independientes** :

$$Y = \mathbf{X_1} + \mathbf{X_2} + \ldots + \mathbf{X_n} \approx N(m_Y, \sigma_Y) \qquad n \rightarrow \infty$$

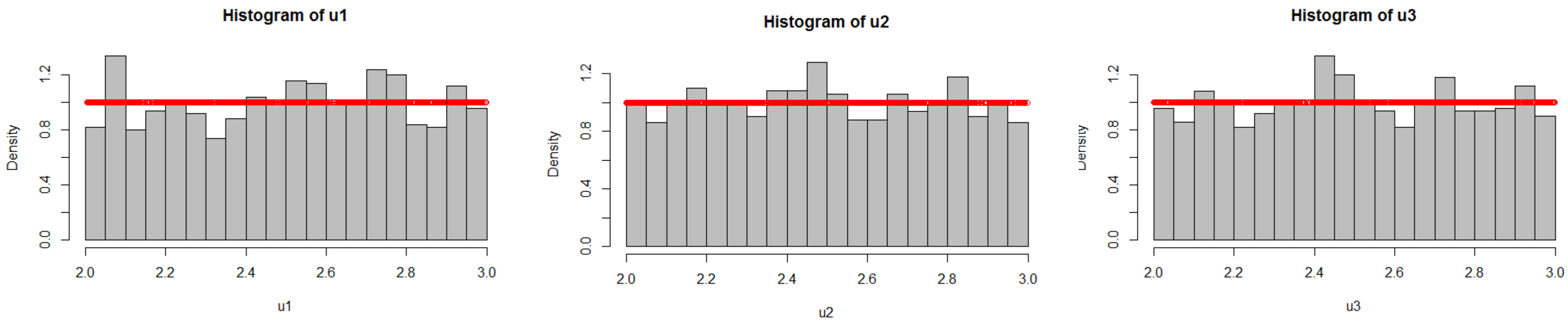
Este teorema justifica el hecho de que la mayoría de las distribuciones de las variables en problemas reales sean normales

Teorema Central del límite

```
#### Teorema central del límite
## Generar 10 v.a uniformes U(2,3)con 1000 datos cada una
a<-2
b<-3
n<-1000
u1<-runif(n, a, b)
u2<-runif(n, a, b)
u3<-runif(n, a, b)
u4<-runif(n, a, b)
u5<-runif(n, a, b)
u6<-runif(n, a, b)
u7<-runif(n, a, b)
u8<-runif(n, a, b)
u9<-runif(n, a, b)
u10<-runif(n, a, b)
```

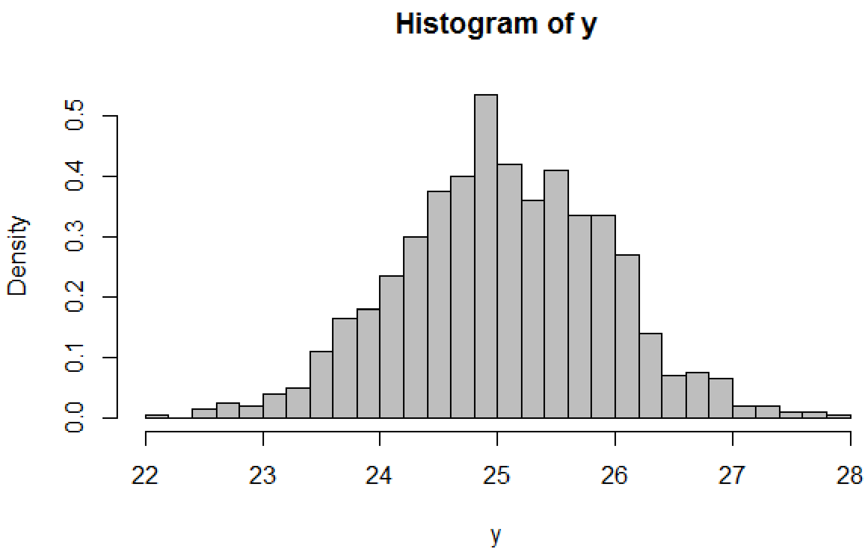
Teorema Central del límite

```
## Representar el histograma correspondiente y la función de
densidad teórica
hist(u1, breaks=trunc(sqrt(n)), freq=F, col="gray")
lines(u1, dunif(u1, a, b), type="p", col="red")
hist(u2, breaks=trunc(sqrt(n)), freq=F, col="gray")
lines(u2, dunif(u2, a, b), type="p", col="red")
hist(u3, breaks=trunc(sqrt(n)), freq=F, col="gray")
lines(u3, dunif(u3, a, b), type="p", col="red")
```



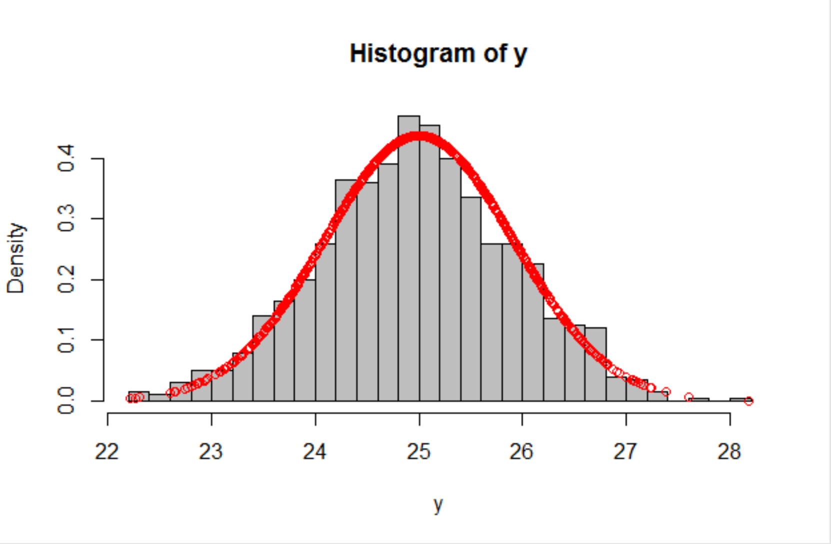
Teorema Central del límite

```
## Crear una nueva v.a como suma de las anteriores
## y=u1 + u2 + ... + u10
y<-u1+u2+u3+u4+u5+u6+u7+u8+u9+u10


## Representar el histograma correspondiente
hist(y, breaks=trunc(sqrt(n)), freq=F, col="gray")
```



Teorema Central del límite

```
> ## ¿Se ajusta a la distribución Normal?
> ## Media y desviación típica de y
> m<-10*((a+b)/2)
> sigma<-sqrt(10*((b-a)^2)/12)
> m
[1] 25
> sigma
[1] 0.9128709
>
> ## función de densidad normal
> lines(sort(y), dnorm(sort(y), m, sigma), type="l", col="red")
```



Aproximaciones normales

Dado que una variable **Binomial** es la suma de los resultados obtenidos en **n** repeticiones independientes de un experimento, su distribución se irá aproximando a la de una **Normal** a medida que aumente **n**

$$\boxed{\begin{array}{c} n \geq 20 \; y \; p \leq 0,05 \\ n \geq 100 \\ \left(\sigma_X^2 = np(1-p)\right) \geq 9 \end{array}}$$

$$\boxed{X \sim B(n,p)} \longrightarrow \boxed{X \sim N\left(m = np, \sigma^2 = np(1-p)\right)}$$

$$\frac{X - \overbrace{np}^{\mathbf{m}}}{\underbrace{\sqrt{np(1-p)}}_{\boldsymbol{\sigma}}} \longrightarrow Z \sim N(0,1)$$

Aproximaciones normales

Dado que una variable **Poisson** es la suma de los resultados obtenidos en **n** repeticiones independientes de un experimento, su distribución se irá aproximando a la de una **Normal** a medida que aumente **n**

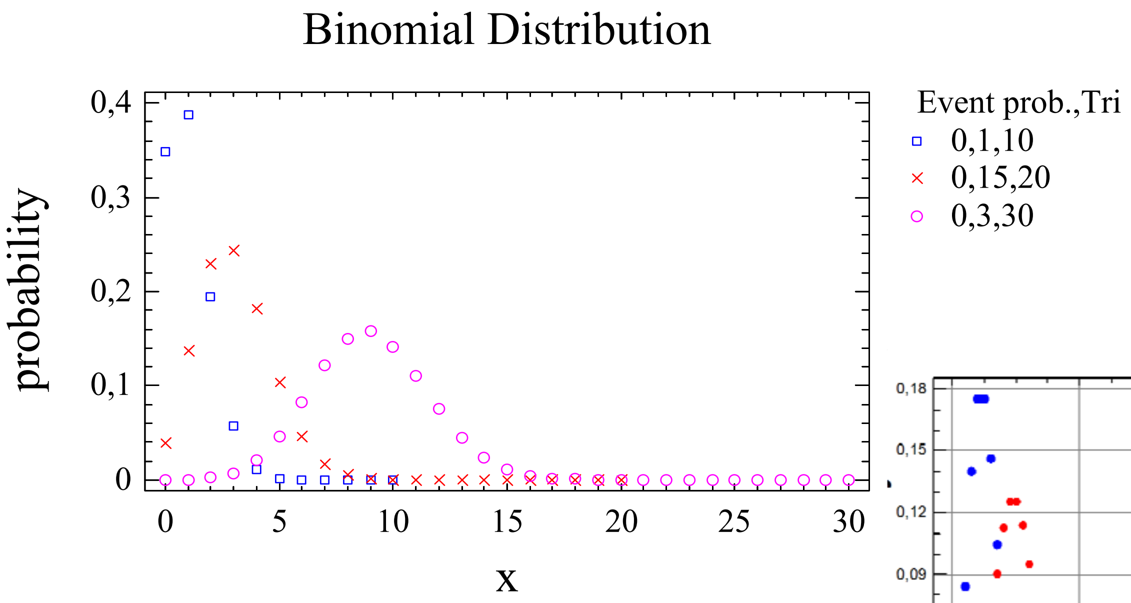
$$\boxed{\lambda \geq 9}$$

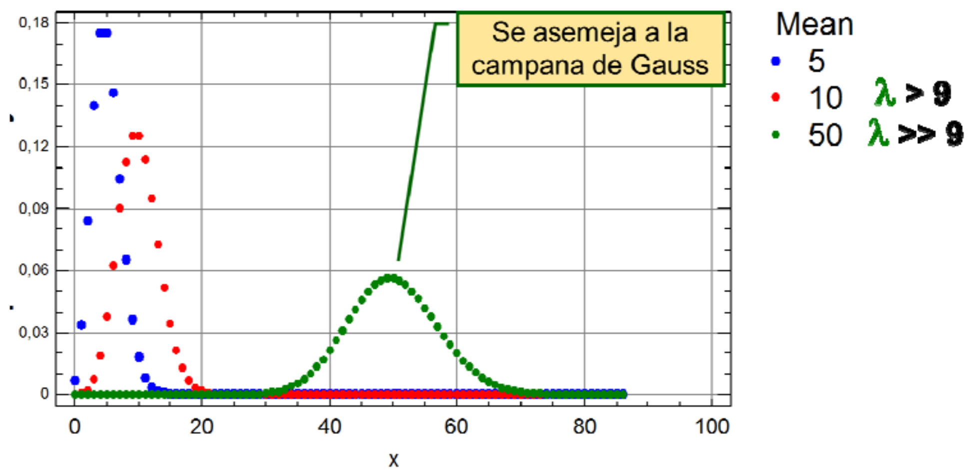
$$\boxed{X \sim Ps(\lambda)} \longrightarrow \boxed{X \sim N\left(m = \lambda, \sigma^2 = \lambda\right)}$$

$$\frac{X - \overbrace{\lambda}^{\mathbf{m}}}{\underbrace{\sqrt{\lambda}}_{\boldsymbol{\sigma}}} \longrightarrow Z \sim N(0,1)$$

Aproximación Binomial, Poisson - Normal

Propiedad muy **útil**: cuando $\sigma^2 \geq 9$ las probabilidades correspondientes a una variable binomial pueden también aproximarse usando las tablas de la distribución.





Aproximación normal

```
#### Aproximación normal
### Poisson(lambda)

# Qué media (lambda)?
lambda<-2.5

# Cuántos números?
n<-250

## Generación de n números aleatorios
muestra<-rpois(n,lambda)

## HISTOGRAMA
# Intervalos?
# Calculados para que los enteros delimiten las barras (discreta)
c1<-0
c2<-trunc(qpois(0.9999, lambda))
```

Aproximación normal

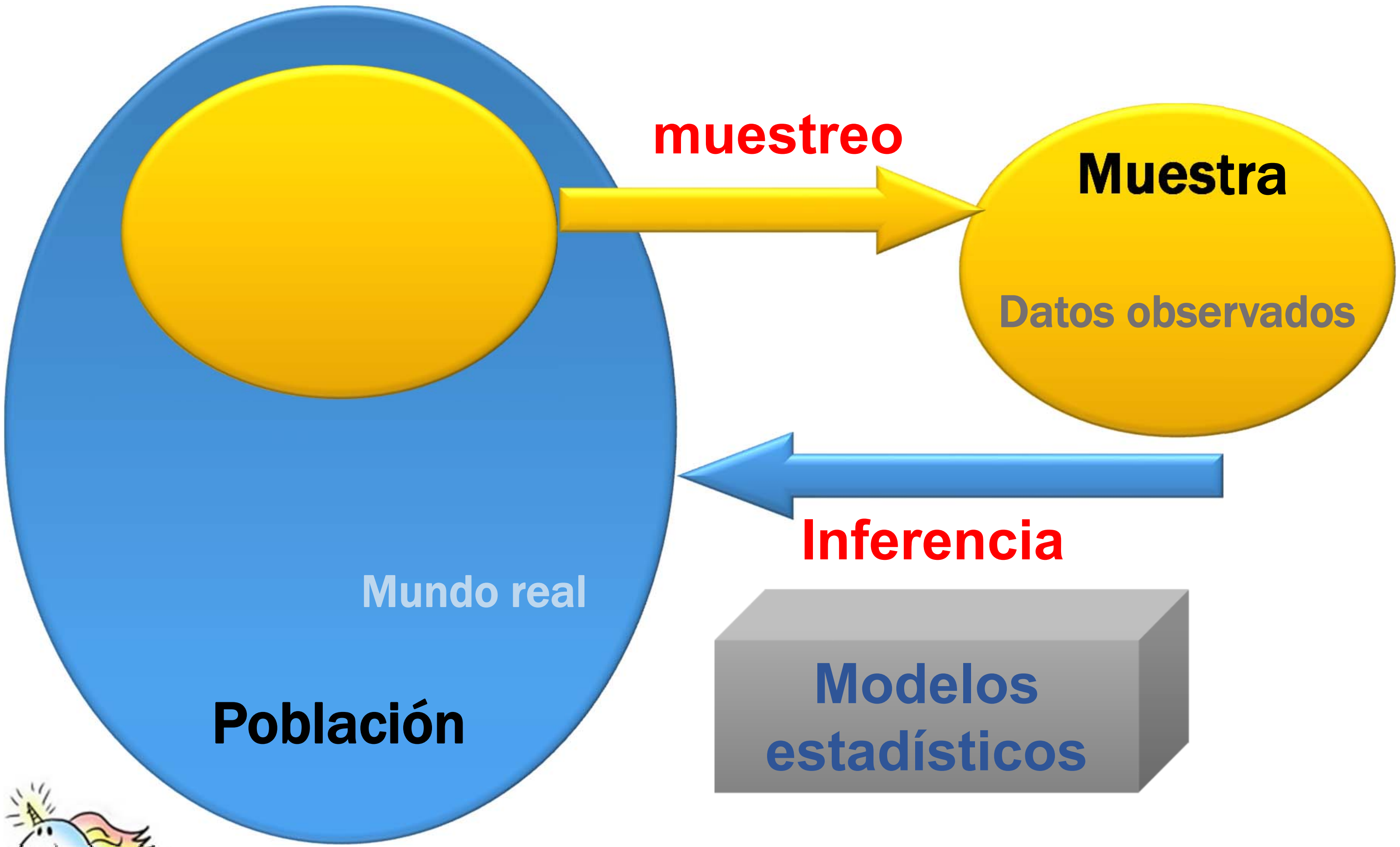
```
# Dibujar histograma con frecuencias absolutas
hist(muestra, breaks=c1:c2, freq=T, xlab="muestra",
ylab="´Densidad", main="Histograma", col="lightblue", border="blue")
# Para que el valor de la variable discreta esté en el centro de la
barra
hist(muestra, breaks=(c1-0.5):(c2+0.5), freq=T, xlab="muestra",
ylab="´Densidad", main="Histograma", col="lightblue", border="blue")


# Dibujar la f(x)
lines(c1:c2, dpois(c1:c2, lambda)*n, type="p", col="red", xpd=T)


# Ahora cambiar el lamda por 100, por ejemplo
```

Inferencia y modelos estadísticos

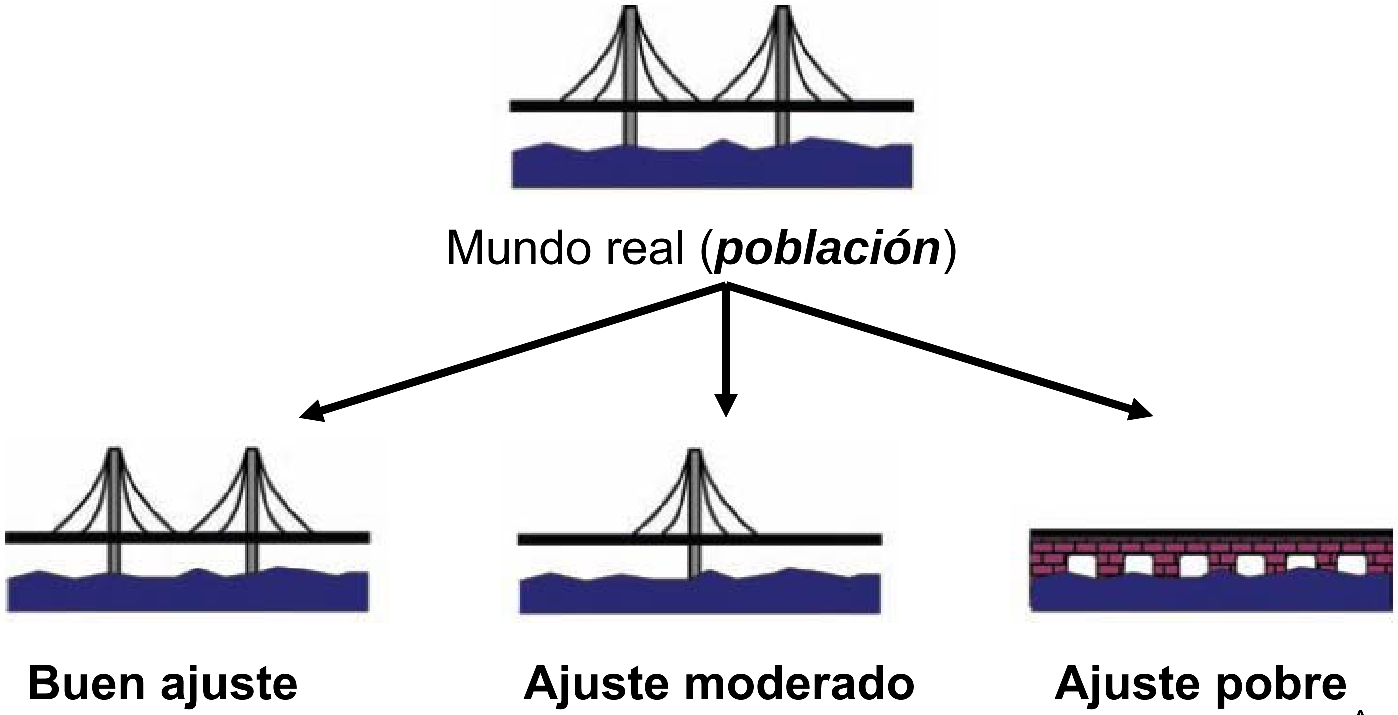
- Necesitamos extrapolar o **predecir** el comportamiento de los fenómenos bajo diferentes condiciones, pero…
- No tenemos toda la información del mundo real, …
- Sólo podemos **inferir** dicho comportamiento a partir de los modelos construidos.
- Para estar seguros de que las predicciones sobre el mundo real son acertadas, el modelo estadístico construido debe representar fielmente los **datos observados** (recogidos)

El grado con el que un modelo estadístico es capaz de representar los datos observados se denomina **ajuste del modelo** (bueno, moderado o pobre)

Inferencia, Modelos, Población y Muestra



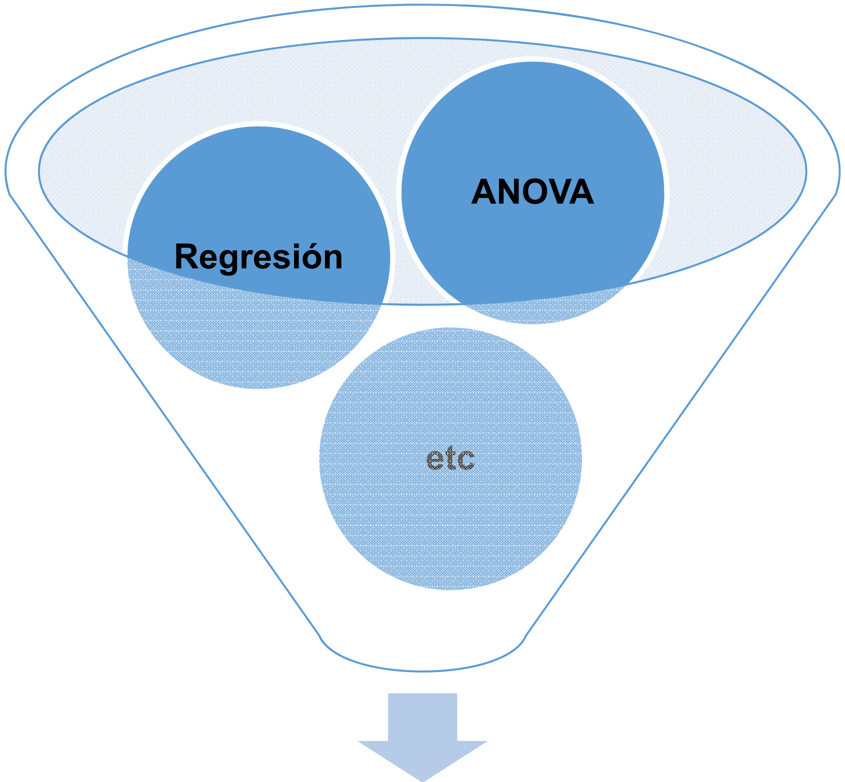
Inferencia y modelos estadísticos



Andy Field (2013)

Siguiendo con esta analogía, si el modelo tiene un *ajuste pobre* a nuestros datos observados, las predicciones sobre el mundo real serán igualmente *pobres*.

Ecuación fundamental



Todos los modelos estadísticos se pueden reducir a una sola ecuación

$$\textbf{Observado}_i = \textbf{(modelo)} + \textbf{error}_i$$

Ecuación fundamental

$\text{Observado}_i = (\text{modelo}) + \text{error}_i$

- Cada dato observado en la muestra puede reproducirse a partir del modelo elegido para el ajuste más cierta cantidad de error.
- Cada modelo dependerá de:
  - Objetivo de la inferencia
  - Diseño del estudio
  - Tipo de variables
- El modelo está compuesto por **parámetros** y **variables**.

Construcción del modelo

Parámetros (b_j)

- Son habitualmente constantes que representan alguna "verdad fundamental" sobre las variables o las relaciones entre las variables en el mundo real (población)→ **parámetros poblacionales**
- No son observables, se estiman **a partir de los datos observados** (muestra): media, S, Coeficiente de regresión, ….

$$\boxed{\textbf{(modelo): } f(b_j, X_j)}$$

$$(\text{modelo})_i = (b) + \text{error}_i = \text{media}$$

$$(\text{modelo})_i = (bX_i) + \text{error}_i$$

$$(\text{modelo})_i = (b_1X_{1i} + b_2X_{2i}) + \text{error}_i$$

Evaluación del ajuste de un modelo

- Comparando las diferencias entre los observado y lo obtenido por el modelo, los **residuos**

error = desviación = residuo

- El **error** cometido **al utilizar el modelo** para un individuo de la muestra en particular se puede obtener como:

$$\textbf{Observado}_i - \textbf{modelo}_i = \textbf{error}_i$$

Evaluación del ajuste de un modelo

- El **error total** cometido por todos los individuos de la muestra se podría obtener como suma de los errores al cuadrado (**Suma de Cuadrados**):

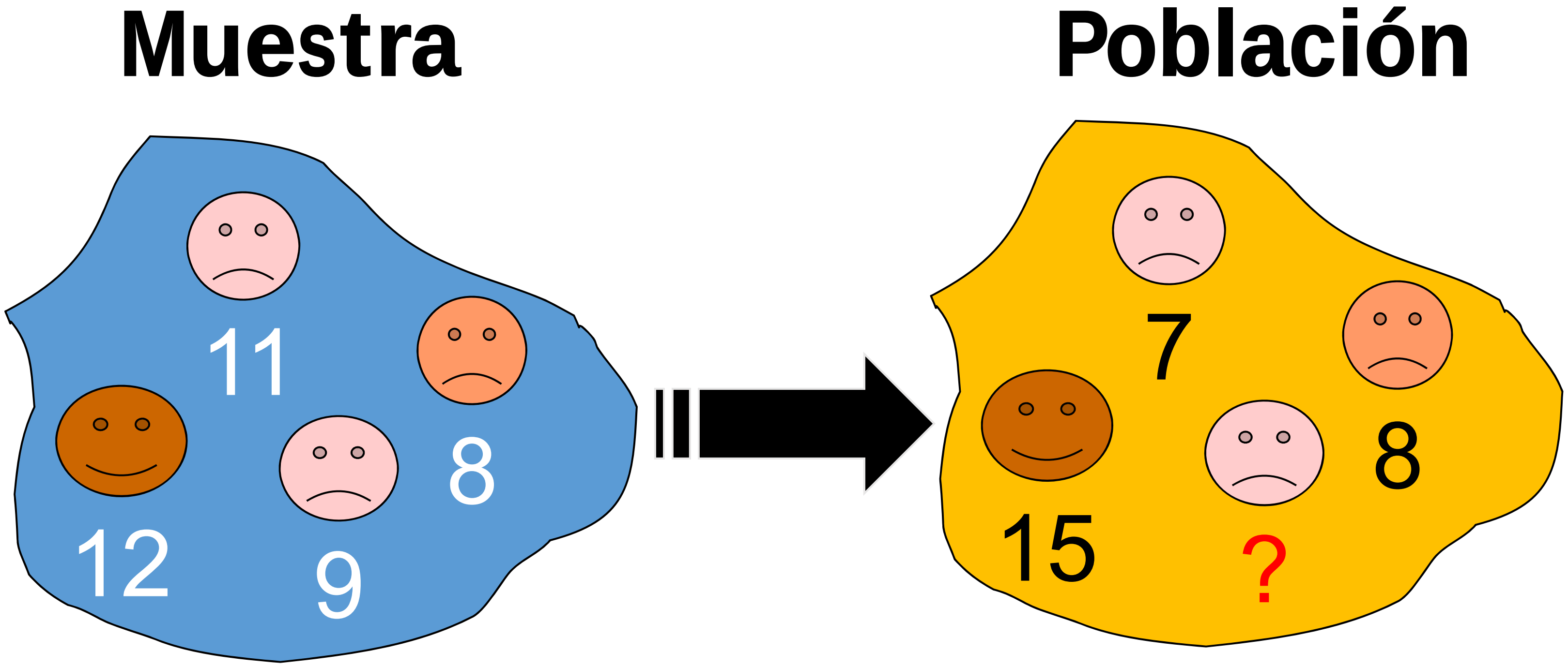
$$SC = \sum_{i=1}^{N} (observado_i - modelo_i)^2$$

- La SC depende del tamaño de la muestra, cuantos más valores observados, mayor será la SC, por lo que resulta más conveniente usar el **error medio** o **Cuadrado Medio del error**:

$$\boxed{CM = \frac{SC}{gl} = \frac{\sum_{i=1}^{n} (observado_i - modelo_i)^2}{N-1}}$$

Grados de Libertad

Grados de libertad *(degrees of freedom)*



$$\overline{X} = 10 \qquad \mu = 10$$

Glosario

Distribución de probabilidad o distribución
Distribución de frecuencias
Distribución de una v.a.
Parámetros poblacionales
Parámetros muetrales
Distribuciones discretas
Distribución Binomial o Binomial
Distribución de Poisson o Poisson
Distribución Normal o de Gauss
Función de Probabilidad P(x)
Probabilidad acumulada
Probabilidad de un intervalo
Función de Densidad f(x)
Esperanza Matemática
Esperanza Matemática de X E(X)
Media de la distribución E(X), m o μ
Media de la v.a X E(X), m o μ
Varianza de la distribución $E(X\text{-}m)^2$ o σ^2
Varianza de la v.a X $E(X\text{-}m)^2$ o σ^2
Modelo
Ajuste

Herramientas Estadísticas para Big Data

Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

3- Variables aleatoria y distribuciones

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

www.upv.es

E. Vázquez

Dto. De Estadística e Investigación Operativa, Aplicadas y Calidad