Análisis de Datos

Métodos Numéricos y Simulación.

Segundo de Grado en Física.

Objetivo del tema

1) En ocasiones, cuando uno hace una medida de una magnitud X no espera obtener el mismo resultado aunque la medida se repita exactamente en las mismas condiciones.

Se dice entonces que X constituye una variable aleatoria

Aprenderemos a describir series de datos de variables aleatorias y a sacar conclusiones de ellas.

2) En otras ocasiones, se observa que los valores que toma una magnitud Y, aunque tienen cierta aleatoriedad, se ven influidos por los valores de otras magnitudes X, Z, T.., y se dispone de un modelo f(X,Z,T,...) para esa posible influencia que depende de una serie de parámetros.

Aprenderemos a comprobar si la influencia de X,Z,T,.. es significativa y a encontrar los valores óptimos de dichos parámetros.

Importar datos a Matlab

El primer paso es importar los datos a Matlab.

- Copiando y pegando.
- Usando el *Import Data Wizard* (Menú File/Import Data)
- Mediante comandos en un script:

  fid = fopen(NombreArchivo): Abre el archivo *NombreArchivo* y le asigna el identificador *fid*
  ***fclose(fid)*:** Cierra el archivo con identificador *fid*.
  C=textscan(fid,format): lee del archivo con identificador *fid* mientras los contenidos se ajustan al formato especificado en *format*. En caso contrario, se detiene. Los datos leídos se almacenan en una matriz de celdas *C*.
  tline=fgetl(fid): lee una línea de texto del archivo con identificador *fid* y la almacena en la cadena de caracteres *tline*.
  otros: textread, csvread, fread, fscanf, etc…

Ejemplo (importar datos a Matlab)

Importar los datos del radio y carga eléctrica de partículas de un medio granular en suspensión.

```
4 -  fid=fopen('Serie-ejemplo.txt');
5 -  TextoInicial = fgetl(fid);  % Lee la primera línea del archivo
6 -  linea1 = fgetl(fid);
7 -  linea2 = fgetl(fid);
8 -  linea3 = fgetl(fid); % Leen las tres líneas de texto antes de cada serie
9 -  serie1 = textscan(fid,'%f %f'); % Lee mientras encuentre sólo los caracteres que aparecen entre corchetes
10 -  linea1 = fgetl(fid);
11 -  linea2 = fgetl(fid);
12 -  linea3 = fgetl(fid); % Leen las tres líneas de texto antes de cada serie
13 -  serie2 = textscan(fid,'%f %f'); % Lee mientras encuentre sólo los caracteres que aparecen entre corchetes
14 -  linea1 = fgetl(fid);
15 -  linea2 = fgetl(fid);
16 -  linea3 = fgetl(fid); % Leen las tres líneas de texto antes de cada serie
17 -  serie3 = textscan(fid,'%f %f'); % Lee mientras encuentre sólo los caracteres que aparecen entre corchetes
18 -  linea1 = fgetl(fid);
19 -  linea2 = fgetl(fid);
20 -  linea3 = fgetl(fid); % Leen las tres líneas de texto antes de cada serie
21 -  serie4 = textscan(fid,'%f %f'); % Lee mientras encuentre sólo los caracteres que aparecen entre corchetes
22 -  linea1 = fgetl(fid);
23 -  linea2 = fgetl(fid);
24 -  linea3 = fgetl(fid); % Leen las tres líneas de texto antes de cada serie
25 -  serie5 = textscan(fid,'%f %f'); % Lee mientras encuentre sólo los caracteres que aparecen entre corchetes
26 -  fclose(fid);
27 -  clear('TextoInicial','linea','linea1','linea2','linea3');
28 -  r1=serie1{1};
29 -  r2=serie2{1};
30 -  r3=serie3{1};
31 -  r4=serie4{1};
32 -  r5=serie5{1};
33 -  q1=serie1{2};
34 -  q2=serie2{2};
35 -  q3=serie3{2};
36 -  q4=serie4{2};
37 -  q5=serie5{2};
38 -  clear('serie1','serie2','serie3','serie4','serie5','fid');
```

ImportarSE.m × CumulRadio.m ×

Algunas definiciones

• **Población:** el conjunto de todos los posible valores que puede tomar la magnitud objeto de nuestro estudio.

En nuestro caso, la población es el conjunto de las cargas y los radios de todas las partículas de polvo del bote.

• **Muestra:** el conjunto de datos que nosotros hemos medido.

➢ Para que lo que estudiemos en este tema sea válido, la muestra debe ser **aleatoria**.

En nuestro ejemplo, que la muestra sea aleatoria significa que todas las partículas del bote han tenido la misma probabilidad de participar en la muestra.

➢ Cuando una muestra no es aleatoria, se dice que está sesgada.

Descripción de una población

Cuando no es factible dar todos los valores de una población, ésta se describe dando:

- Su distribución cumulativa F(x)

  $F(x_o)$: Fracción del total de medidas N en los que $x<x_o$.

- O bien su distribución diferencial f(x)

  $f(x_o)\Delta x$ da la probabilidad de obtener un valor de x comprendido entre $x_o-\Delta x$ u $x_o+\Delta x$.

Propiedades:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \qquad F(x_o) = \int_{-\infty}^{x_o} f(x)dx \qquad f(x) = \frac{dF}{dx}$$

Descripción de una muestra

- No podemos determinar F(x) y f(x), porque eso significa medir para toda la población (en nuestro caso, para todos los granos del bote).

- Lo que podemos hacer es aproximar F(x) y f(x) a partir de las medidas experimentales.

➢Voy a llamar $F_{exp}(x)$ a la distribución cumulativa y $f_{exp}(x)$ a la distribución diferencial obtenidas a partir de los datos experimentales

➢$F_{exp}(x)$ y $f_{exp}(x)$ no se extienden entre $-\infty$ y $+\infty$, sino entre $x_{min}<x<x_{max}$, porque el número de valores de x que tenemos es finito)

Distribución cumulativa F_{exp}

- Supongamos que tengo una muestra $x_1, x_2, .., x_N$ y quiero representar su distribución cumulativa.

1) Ordeno los x_j de menor a mayor: a la muestra así ordenada la llamo $x'_1, x'_2, \ldots, x'_j, \ldots, x'_N$

- Para ordenar una serie se usa la función **B =sort(A)**.
  A: vector desordenado.
  B: vector ordenado en sentido ascendente.

2) Represento 1/N, 2/N, …, j/N, …, (N-1)/N, 1 frente a x'_1, $x'_2, \ldots, x'_j, \ldots, x'_{N-1}, x'_N$

- Se usa la función **stairs(X,Y)**.
  Representa el vector Y frente al vector X haciendo cambiar Y sólo donde toma valores el vector X.

Ejemplo (distribución cumulativa)

Editor - F:\docencia\clases\materiales recurrentes\MNS\análisis\Cumulativas.m

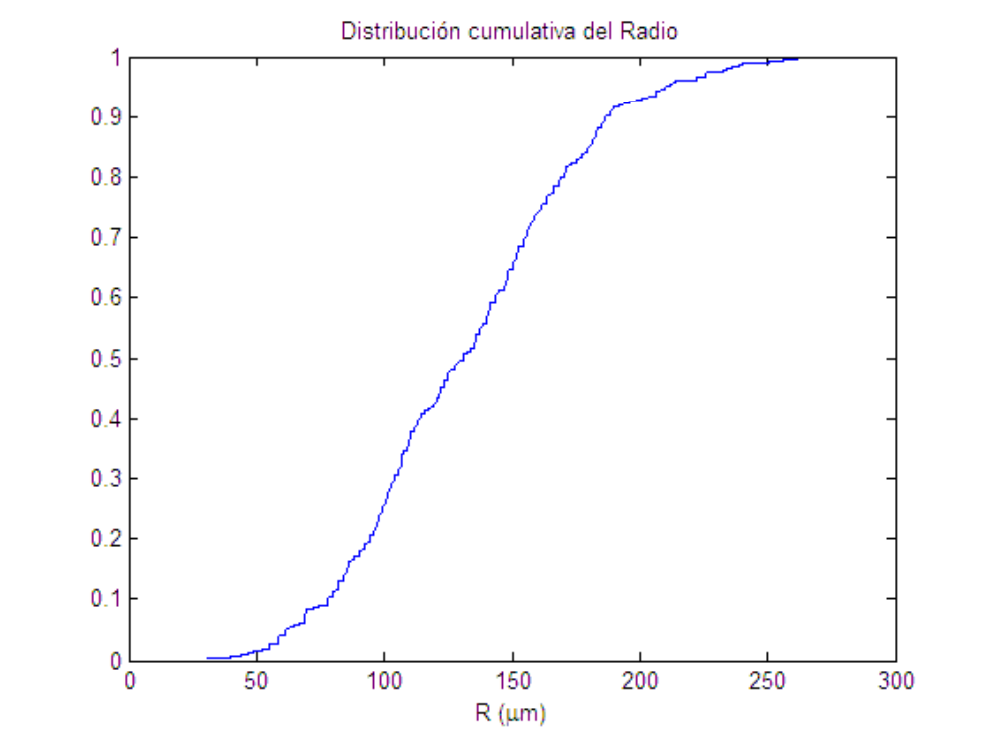
File Edit Text Go Cell Tools Debug Desktop Window Help

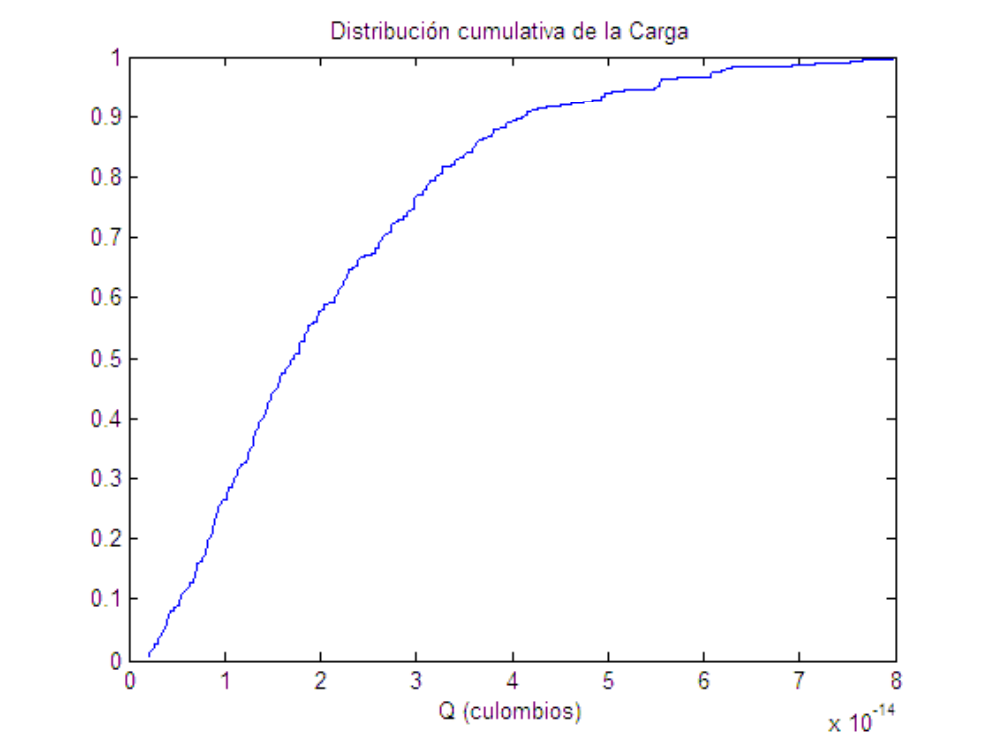
Stack: Base

1.0 1.1

```
1   % Dibuja la distribución cumulativa del radio y de la carga.
2
3 - r = [r1 ; r2 ; r3 ; r4 ; r5];    % Concateno todas las series del radio
4 - q = [q1 ; q2 ; q3 ; q4 ; q5];    % Y también de la carga
5 - R = sort(r);              % Ordeno los radios de menor a mayor
6 - Q = sort(q);
7 - [N , M] = size(R);
8 - DistVal = 1:N;
9 - DistVal = DistVal/N;
10 - DistVal= DistVal';              % Transpongo para obtener un vector columna
11 - figure(1);
12 - stairs(R,DistVal);
13 - title('Distribución cumulativa del Radio');
14 - xlabel('R (\mum)');
15 - figure(2);
16 - stairs(Q,DistVal);
17 - title('Distribución cumulativa de la Carga');
18 - xlabel('Q (culombios)');
```

Ejemplo (distribución cumulativa)





Estas curvas representan $F_{exp}(R)$ y $F_{exp}(Q)$. Las funciones F(R) y F(Q) para nuestra población nos son desconocidas

Momentos de la distribución

Los momentos de una distribución nos ayudan a describir su forma numéricamente.

Son: la mediana, la media, la desviación estándar, la skewness y la curtosis.

Pueden calcularse tanto para la población como para la muestra que hemos medido.

Normalmente sólo podemos medir sus valores en una muestra: a partir de los valores de la muestra tenemos que estimar los de la población.

Media

Cálculo para la población: $$< x >= \int_{-\infty}^{+\infty} x f(x) dx$$

Normalmente no podemos calcularla de esta forma porque no conocemos f(x). La estimamos a partir de una muestra de la forma:

$$< x >= \frac{1}{N} \sum_{i=1}^{N} x_i$$

En Matlab, se usa la función:
M = mean(x): M es la media de los elementos del vector x.

Mediana

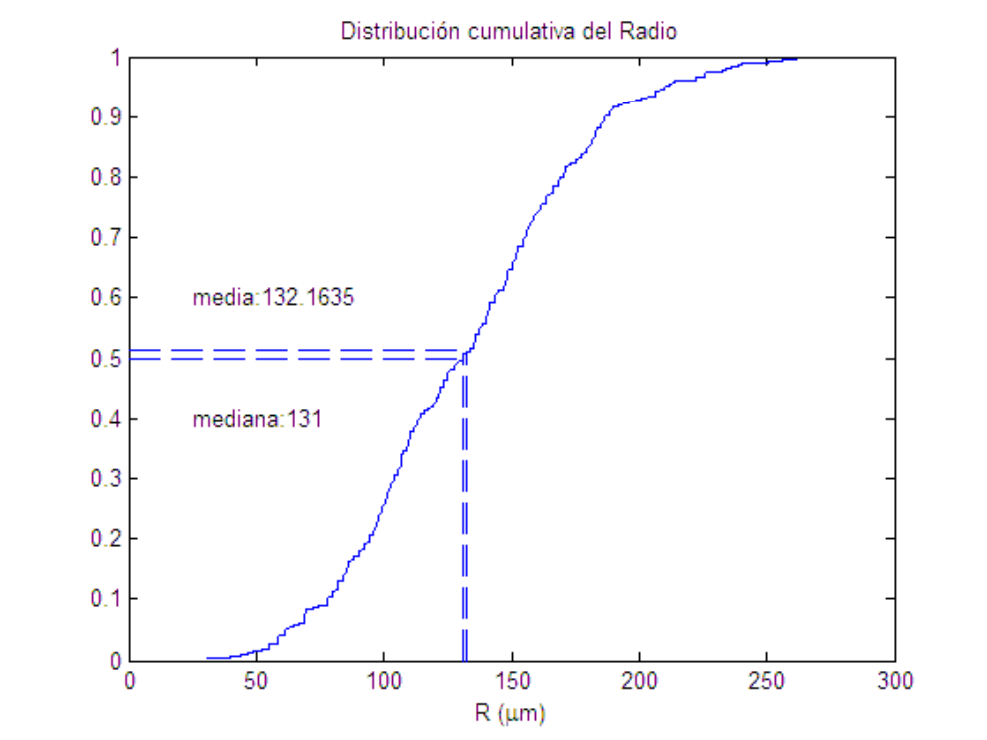
Es el valor x_o para el cual $F(x_o)=1/2$, es decir, el 50% de los datos están por debajo de la mediana y 50% está por encima.

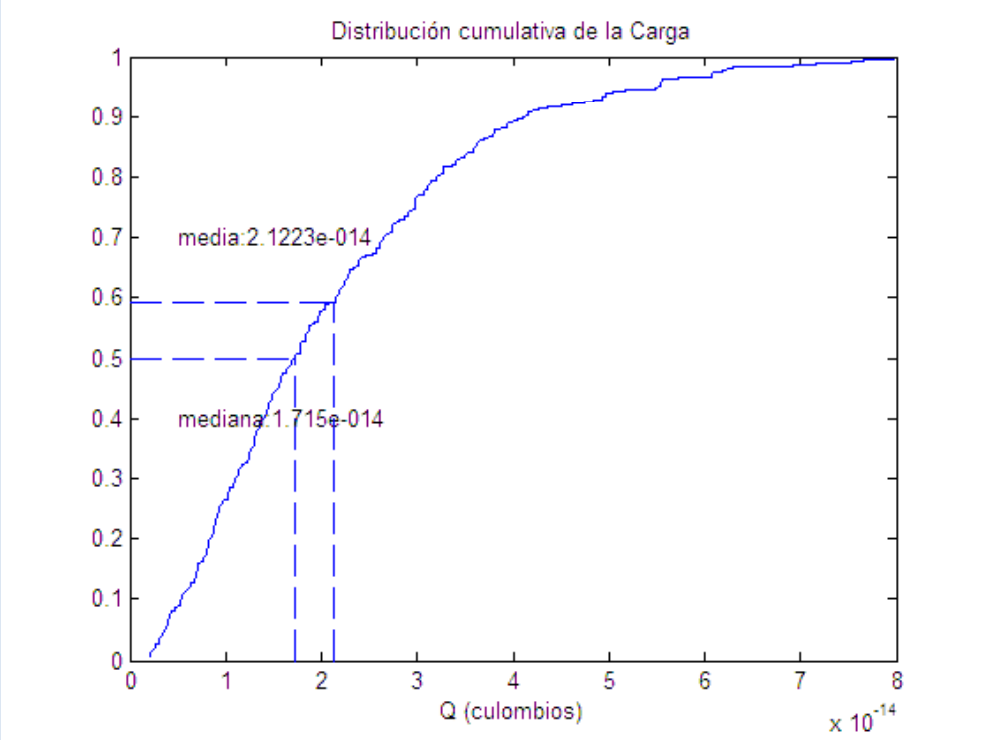
Para estimar su valor a partir de una muestra, buscamos $F_{exp}(x_o)=1/2$.

en Matlab, se usa:
M = median(x): M es la mediana de los valores en el vector X

Tiene la ventaja sobre la media de que resulta menos afectada por puntos erróneos (outliers)

Ejemplo (media y mediana)





Cuando una distribución no es simétrica, la mediana y la media no coinciden.

Varianza y desviación estándard

- Cálculo para la población:

$$\mathrm{Var}_x = \int_{-\infty}^{+\infty} \left(x - < x >\right)^2 f(x)dx$$

- Cálculo para una muestra:

$$\mathrm{Var}_x = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j - < x >\right)^2$$

A partir de la varianza, se define la desviación estándar como:

$$s(x), \sigma(x) = \sqrt{\mathrm{Var}_x}$$

var(X,flag), std(X,flag): varianza y desviación estándar de los valores contenidos en el vector *X*. *flag* = 0 si los valores de *X* son una muestra de una población. *flag* = 1 si los valores de *X* corresponden a una población completa.

La desviación estándar da una medida de cuánto de agrupada está una distribución alrededor de su valor medio.

Histogramas de frecuencias (I)

Una forma cualitativa de dar $f_{exp}(x)$ es presentarla en la forma de un histograma de frecuencias. Para ello:

1) Defino una serie de M intervalos de anchura Δx:

$$\left(x_1^* - \frac{\Delta x}{2}, x_1^* + \frac{\Delta x}{2}\right), \left(x_2^* - \frac{\Delta x}{2}, x_2^* + \frac{\Delta x}{2}\right), \ldots, \left(x_M^* - \frac{\Delta x}{2}, x_M^* + \frac{\Delta x}{2}\right)$$

2) Cuento el número de valores h_k de x que caen en cada intervalo centrado en x^*_k.

Estos dos pasos los hacen las funciones:

[n,bin]= hist(X,M): dado un número de intervalos M indica cuántos elementos del vector *X* caen en cada intervalo centrado en los valores del vector bin.

[n,bin]=histc(X,edges): indica cuántos elementos del vector X caen en cada intervalo limitado por los valores del vector edges.

Histograma de frecuencias (II)

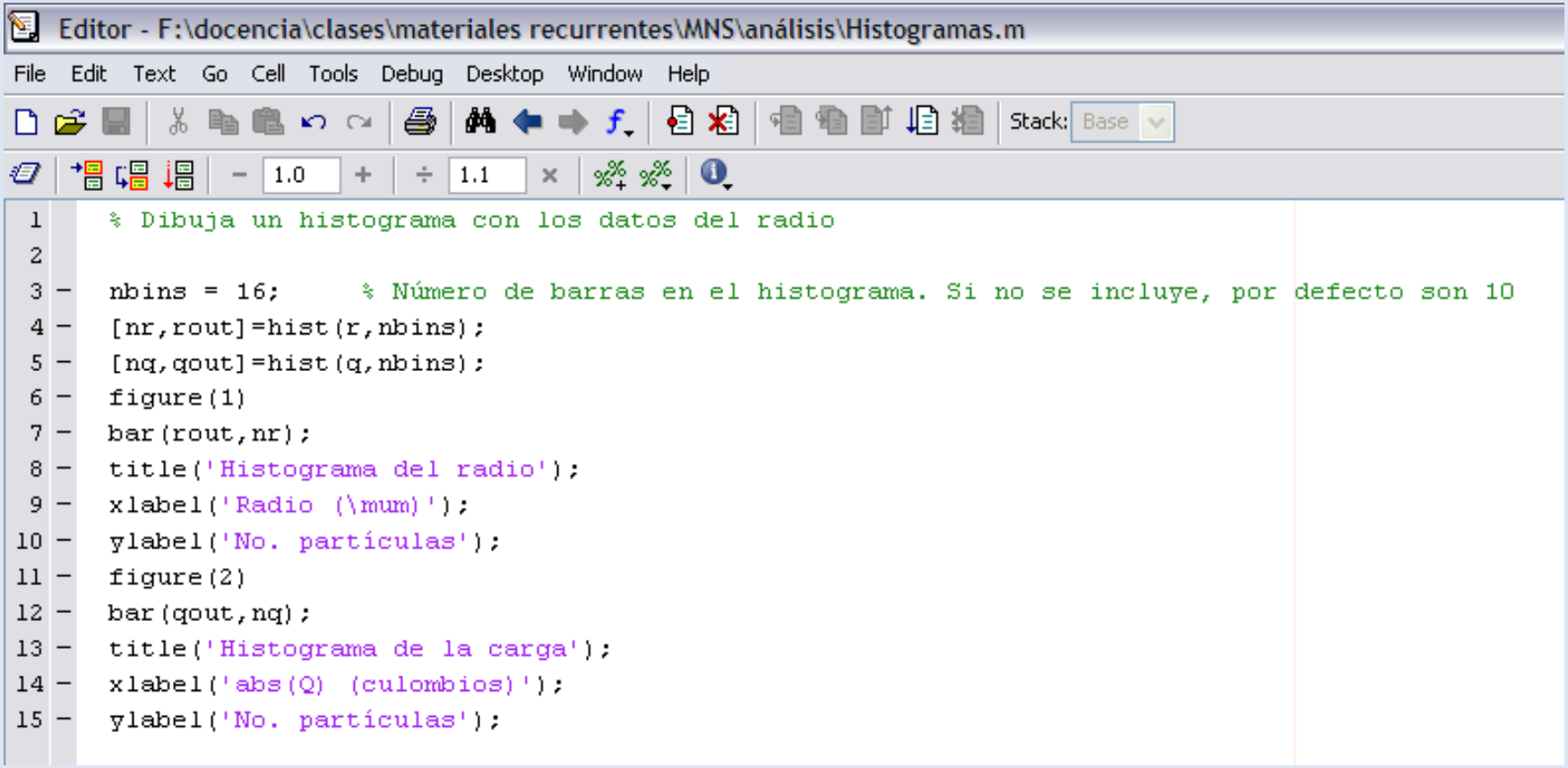
3) Represento h_k frente a x^*_k

De eso se encargan las funciones:

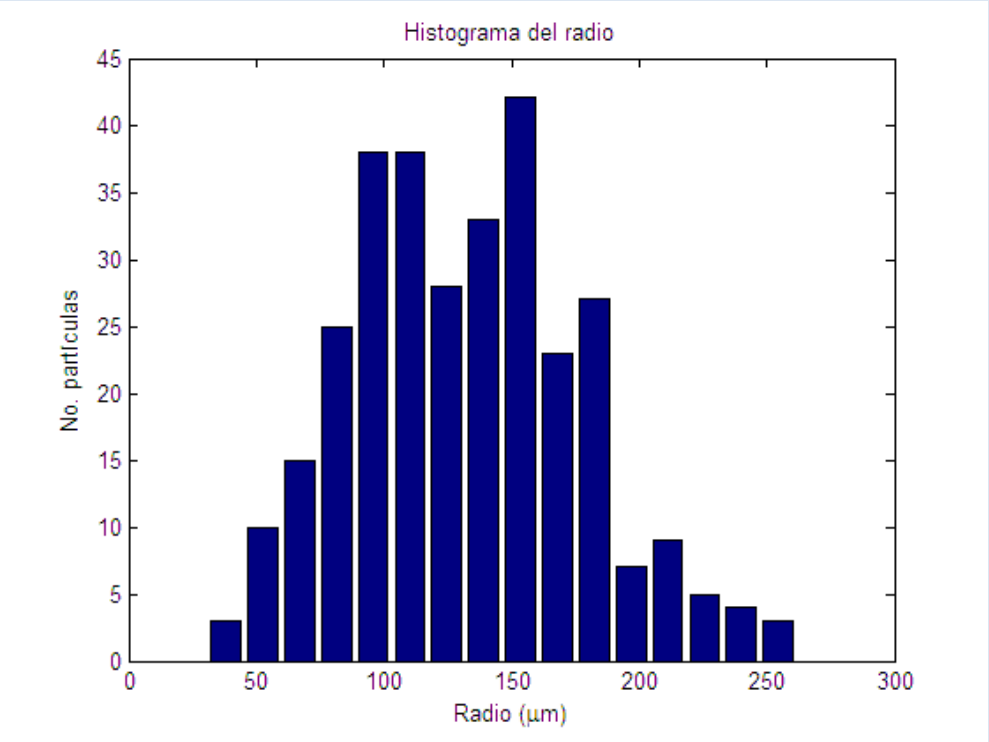
hist(X,bin): dibuja el histograma de los valores de X que caen en cada uno de los intervalos centrados en los valores del vector bin.

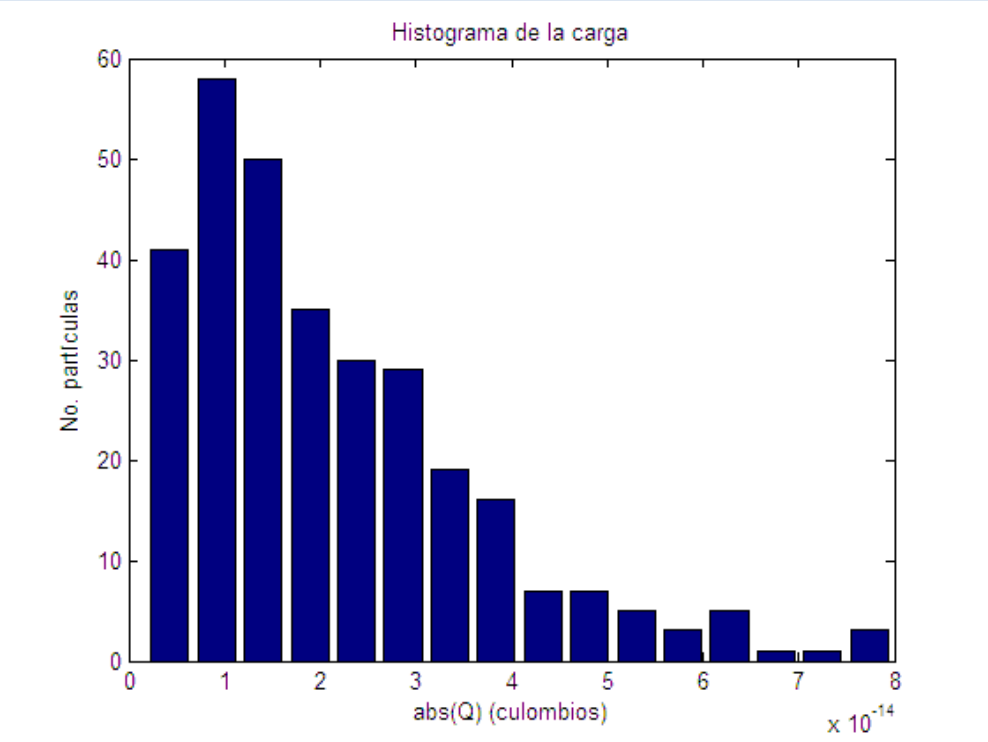
bar(n,bin): dibuja el histograma de los valores del vector n frente a los valores del vector bin.

Ejemplo (histograma)



Ejemplo (histogramas)





Una pregunta recurrente es: ¿cuántas barras es razonable poner?

Volveremos a considerar esta pregunta más adelante

Distribuciones diferenciales a partir del histograma

El histograma no es lo mismo que la función de distribución diferencial.

Recuerda que: $$f(x_o) = \left.\frac{dF}{dx}\right|_{x=x_o} \approx \frac{F\left(x_o + \Delta x/2\right) - F\left(x_o - \Delta x/2\right)}{\Delta x}$$

El número de datos x_j comprendidos entre los valores $x^*_o + \Delta x/2$ y $x^*_o - \Delta x/2$ viene dado por:

$$N\left[F\left(x_o^* + \Delta x/2\right) - F\left(x_o^* - \Delta x/2\right)\right]$$

Eso significa que si queremos estimar la distribución diferencial $f_{exp}(x)$, tenemos que dividir cada columna del histograma por su anchura y por el número total de datos.

Ejemplo (distribución diferencial)

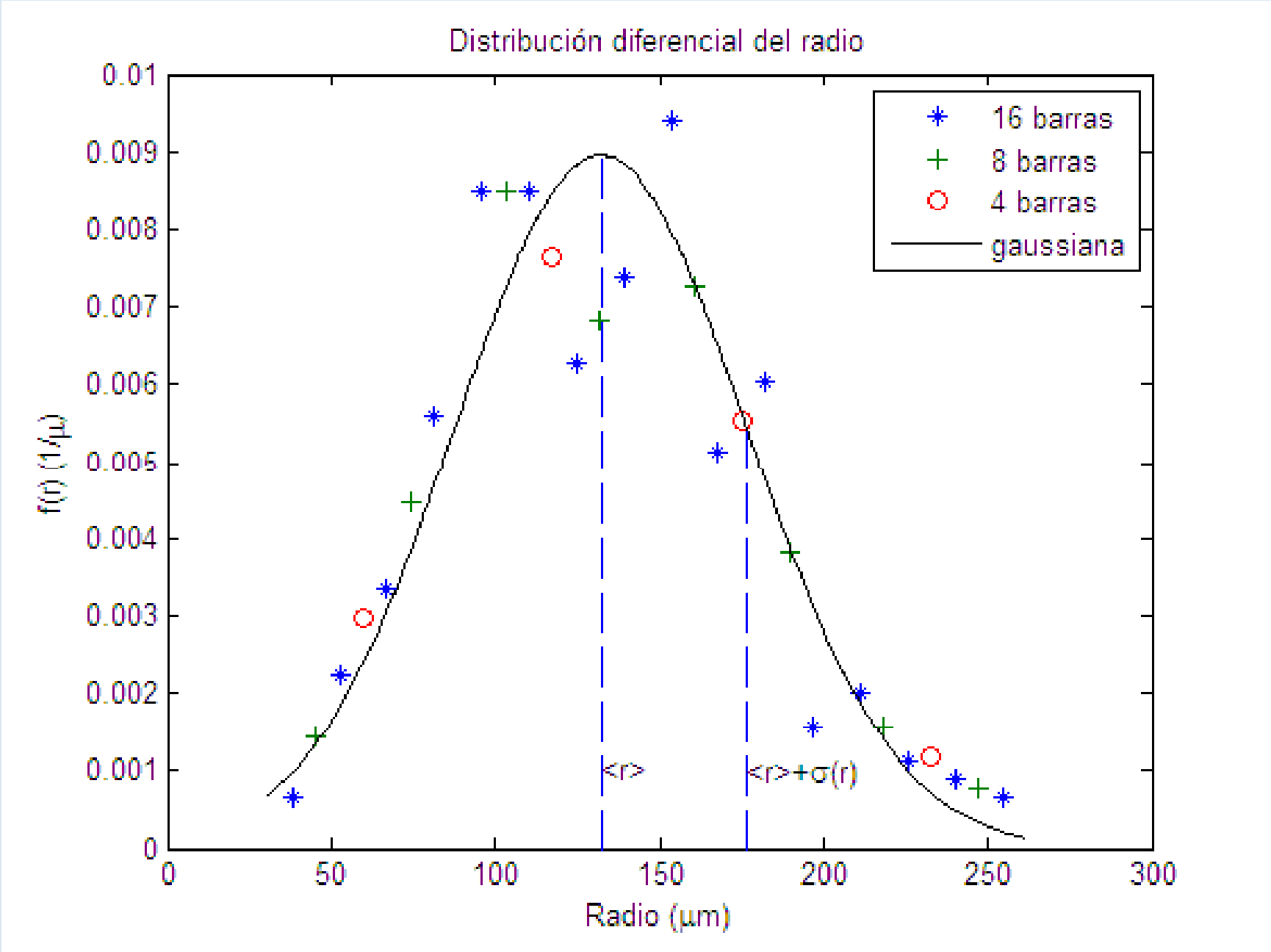
Editor - F:\docencia\clases\materiales recurrentes\MNS\análisis\EjemploDiffR.m

File Edit Text Go Cell Tools Debug Desktop Window Help

```
% Estima la distribución diferencial con los datos del radio
% a partir del histograma usando distinto número de intervalos.

r = [r1 ; r2 ; r3 ; r4 ; r5];       % Concateno todas las series del radio
sr = std(r,0);
mr = mean(r);
R = sort(r);
[N M] = size(r);
rmin= min(r);
rmax= max(r);
[n16,rout16]=hist(r,16);             % Datos para el histograma usando 16 barras
[n8,rout8]=hist(r,8);               % Datos para el histograma usando 8 barras
[n4,rout4]=hist(r,4);               % Datos para el histograma usando 4 barras
dr16 = rout16(2)-rout16(1);
fr16 = n16/(N*dr16);                % Estimación de la densidad de probabilidad usando 16 barras
dr8 = rout8(2)-rout8(1);
fr8 = n8/(N*dr8);                  % Estimación de la densidad de probabilidad usando 16 barras
dr4 = rout4(2)-rout4(1);
fr4 = n4/(N*dr4);                  % Estimación de la densidad de probabilidad usando 16 barras
Gdist = Gauss(R,mr,sr);
figure(1);
plot(rout16,fr16,'*',rout8,fr8,'+',rout4,fr4,'o',R,Gdist,'k-');
title('Distribución diferencial del radio');
xlabel('Radio (\mum)');
ylabel('f(r)');
legend('16 barras','8 barras','4 barras','gaussiana');
```

Ejemplo (distribución gausiana)



La distribución gausiana

• La distribución que hemos pintado en la figura anterior recibe el nombre de distribución gausiana o distribución normal.

• Su ecuación es:

$$f_{gauss}(x;<x>,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-<x>)^2}{2\sigma^2} \right]$$

σ: desviación estándar de la variable x

• Y para la distribución cumulativa:

$$F_{gauss}(x;<x>,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(x_o-<x>)^2}{2\sigma^2} \right] dx_o$$

Gausiana cumulativa: Implementación en matlab

Matlab tiene las siguientes funciones

y = erf(x): función de error.

$$y = \frac{2}{\sqrt{\pi}} \int_{o}^{x} \exp\left(-t^2\right) dt$$

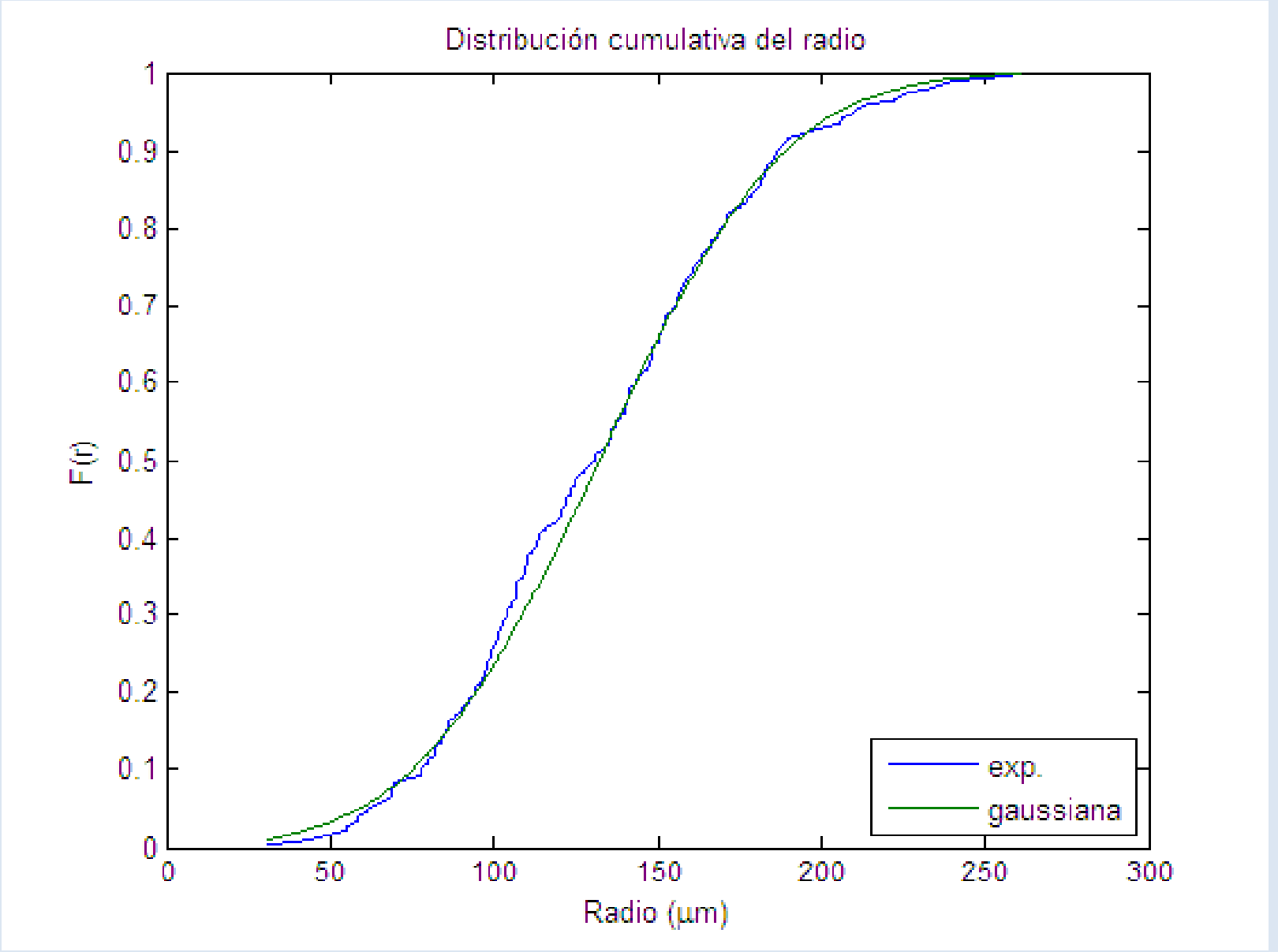
y = erfc(x): función de error complementaria

$$y = \frac{2}{\sqrt{\pi}} \int_{x}^{+\infty} \exp\left(-t^2\right) dt = 1 - \operatorname{erf}(x)$$

Para la distribución gausiana cumulativa:

$$F_{gauss}(x; < x >, \sigma) = 1 - 0.5 \operatorname{erfc}\left(\frac{z}{\sqrt{2}}\right) \qquad z = \frac{x - < x >}{\sigma}$$

Ejemplo (gausiana cumulativa)



¿Cuántos intervalos escogemos?

•Cuando queremos representar un histograma, se nos presenta un dilema:

- Con muchos intervalos, la distribución tiene muchos picos espúreos.
- Con pocos intervalos, es difícil ver la forma de $f_{exp}(x)$.

• Si $f_{exp}(x)$ es una distribución gaussiana, el valor óptimo del intervalo viene dado por:

$$\Delta x \approx 3.49 \frac{s(x)}{N^{1/3}}$$

s(x): desviación estándar de la muestra.

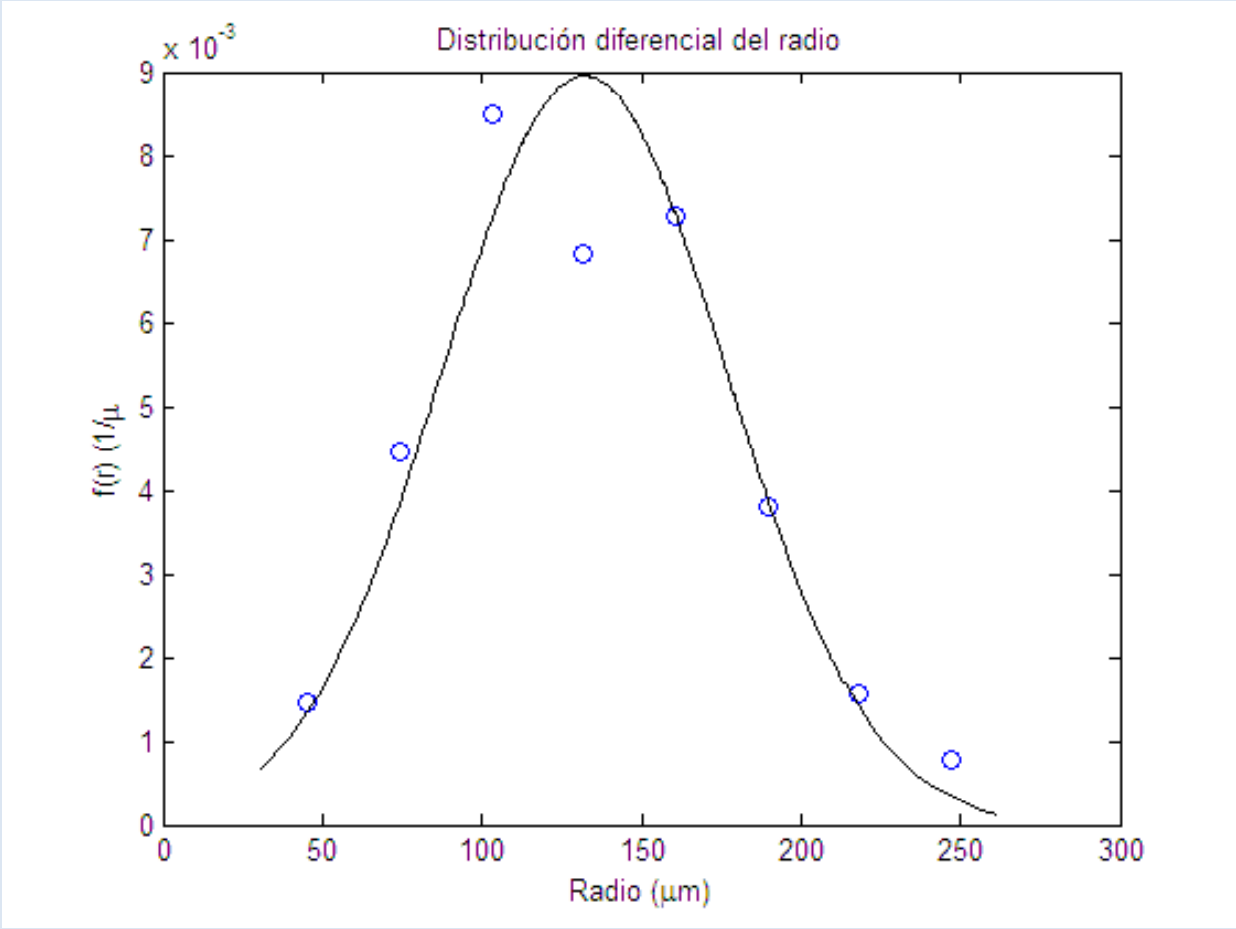
N: número de valores de la muestra.

• Si la distribución no es gaussiana, se puede usar esta regla como una primera estimación.

Ejemplo

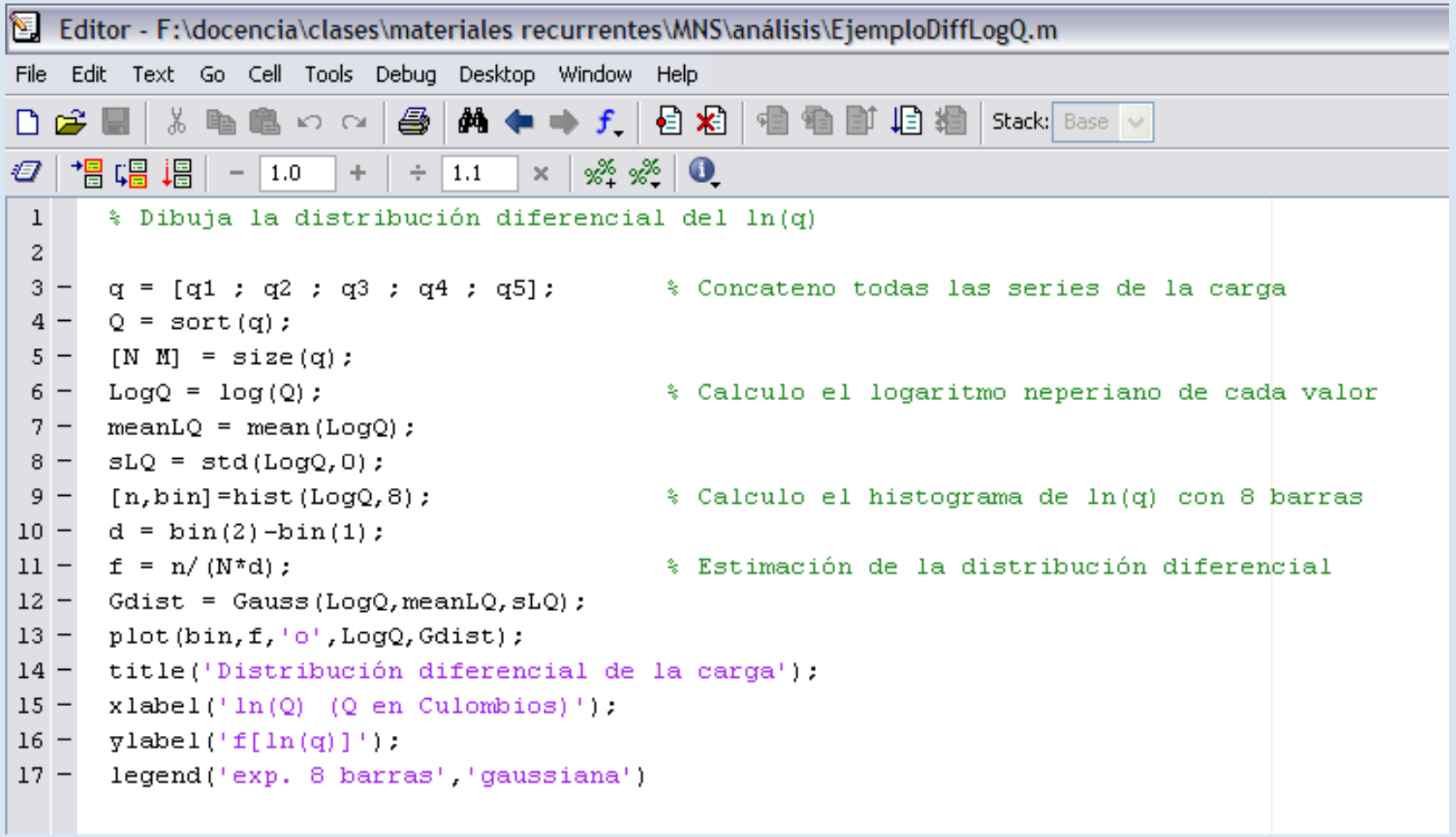
En el caso de la distribución del radio, N = 310, s(R) = 44.54 μm

$$\Delta R \approx 3.49 \frac{s(R)}{N^{1/3}} = 3.49 \frac{44.54}{310^{1/3}} = 22.97\ \mu\text{m}$$



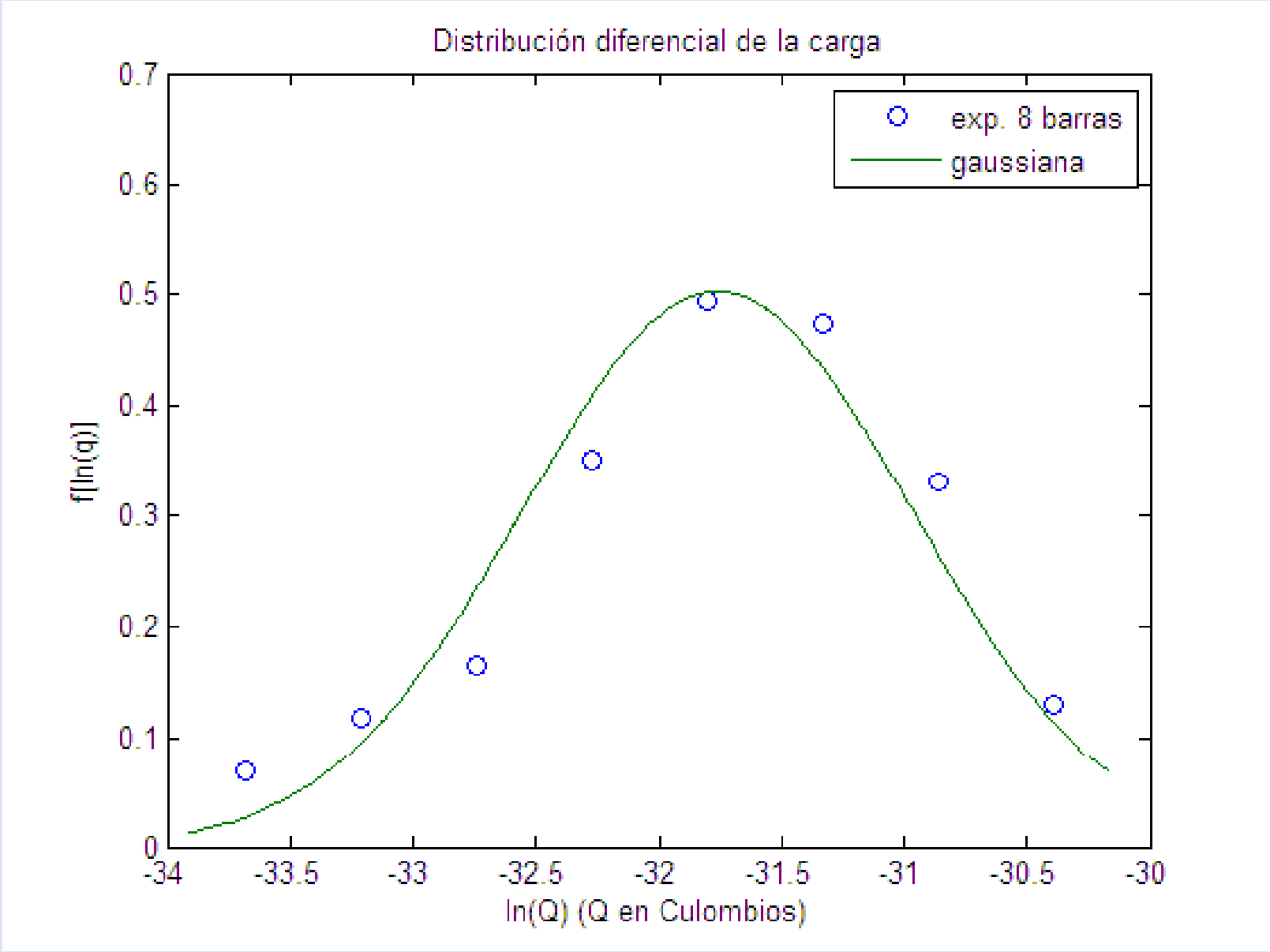
Ejemplo

Cuando calculamos la distribución diferencial para la carga, nos damos cuenta de que es el ln(Q) el que está distribuido (aproximadamente) como una distribución gausiana.




```
1   % Dibuja la distribución diferencial del ln(q)
2
3 - q = [q1 ; q2 ; q3 ; q4 ; q5];       % Concateno todas las series de la carga
4 - Q = sort(q);
5 - [N M] = size(q);
6 - LogQ = log(Q);                      % Calculo el logaritmo neperiano de cada valor
7 - meanLQ = mean(LogQ);
8 - sLQ = std(LogQ,0);
9 - [n,bin]=hist(LogQ,8);               % Calculo el histograma de ln(q) con 8 barras
10 - d = bin(2)-bin(1);
11 - f = n/(N*d);                       % Estimación de la distribución diferencial
12 - Gdist = Gauss(LogQ,meanLQ,sLQ);
13 - plot(bin,f,'o',LogQ,Gdist);
14 - title('Distribución diferencial de la carga');
15 - xlabel('ln(Q) (Q en Culombios)');
16 - ylabel('f[ln(q)]');
17 - legend('exp. 8 barras','gaussiana')
```

Ejemplo


Distribución diferencial de la carga
exp. 8 barras
gaussiana
f[ln(q)]
ln(Q) (Q en Culombios)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-34
-33.5
-33
-32.5
-32
-31.5
-31
-30.5
-30

Cambios de variable en la distribución diferencial

Supongamos que queremos cambiar de la variable x a otra variable y que nos resulte más conveniente.

Para respetar la probabilidad, ha de cumplirse:

$$f(y)dy = f(x)dx \quad \Rightarrow \quad f(y) = f(x)\frac{dx}{dy}$$

Por ejemplo, en el caso de la carga de las partículas queremos pasar de la carga q al logaritmo de la carga ln(q)

Hacemos y=ln(q), x=q

$$\frac{dy}{dx} = \frac{d(\ln q)}{dq} = \frac{1}{q} \quad \Rightarrow \quad \frac{dx}{dy} = q$$

$$f(\ln q) = f(q)q$$

La distribución log-normal

Sabemos que:

$$y = \ln(q) \qquad f_{gauss}(y; < y >, \sigma_y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\left[- \frac{(y - < y >)^2}{2\sigma_y^2} \right]$$

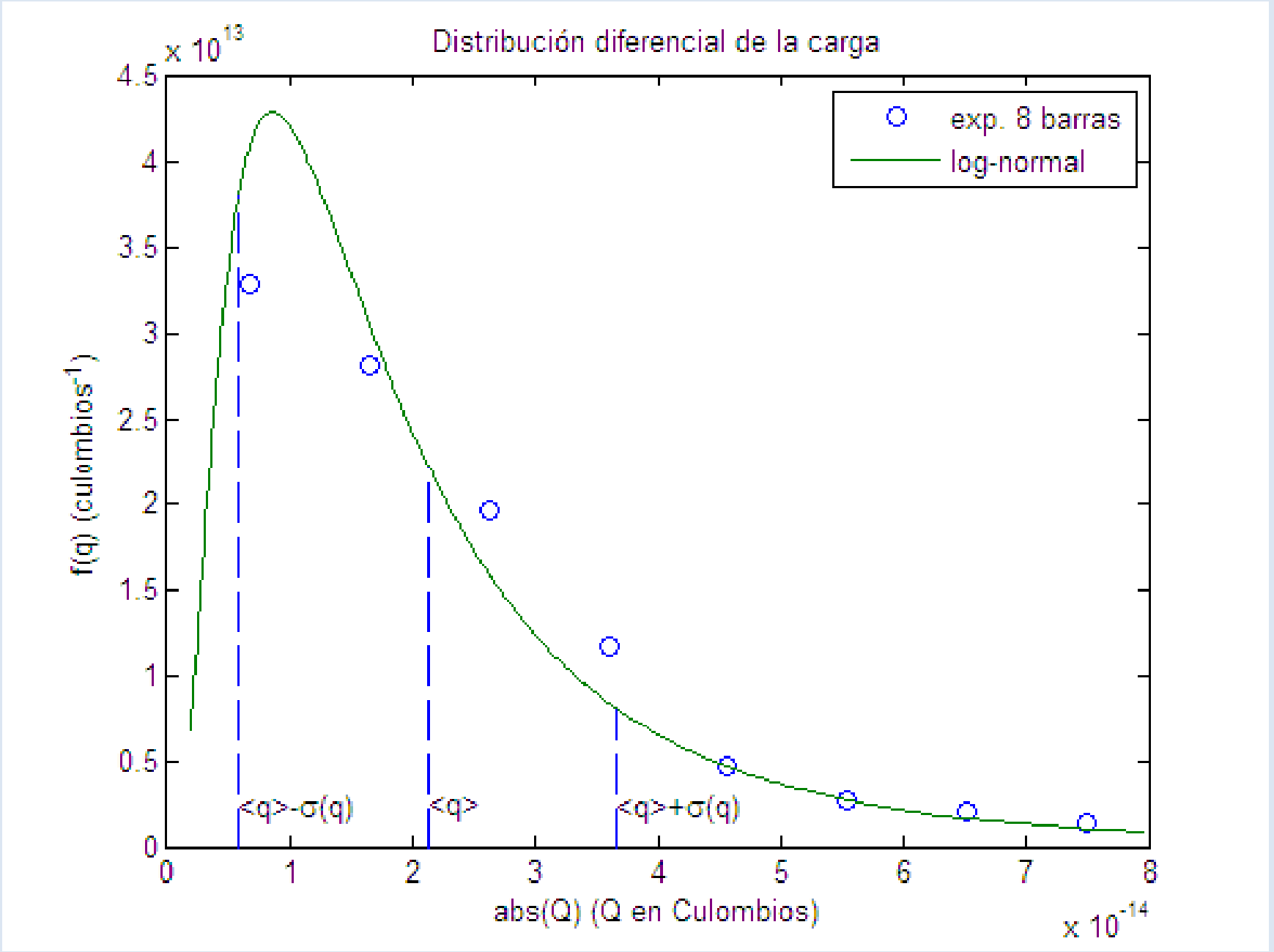
De lo que hemos dicho:

$$f(\ln q) = f(q) q \quad \Rightarrow \quad f(q) = \frac{1}{q} f(\ln(q))$$

$$f_{\log - norm}(q) = \frac{1}{q} \frac{1}{\sigma_{\ln(q)} \sqrt{2\pi}} \exp\left[- \frac{(\ln q - < \ln q >)^2}{2\sigma_{\ln q}^2} \right]$$

A esta distribución se la llama distribución log-normal

Ejemplo (distribución log-normal)



Otras distribuciones

Otras distribuciones que trae Matlab y que nos resultarán útiles son:

- La distribución beta-incompleta.
- La distribución gamma-incompleta

Matlab trae sólo las distribuciones cumulativas.

Las distribuciones diferenciales y la inversas de estas funciones de las distribuciones cumulativas están programadas en un paquete extra llamado *Statistics toolbox*.

La función beta

- La función beta se define como:

  En matlab: ***y = beta(a,b)***.

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$$

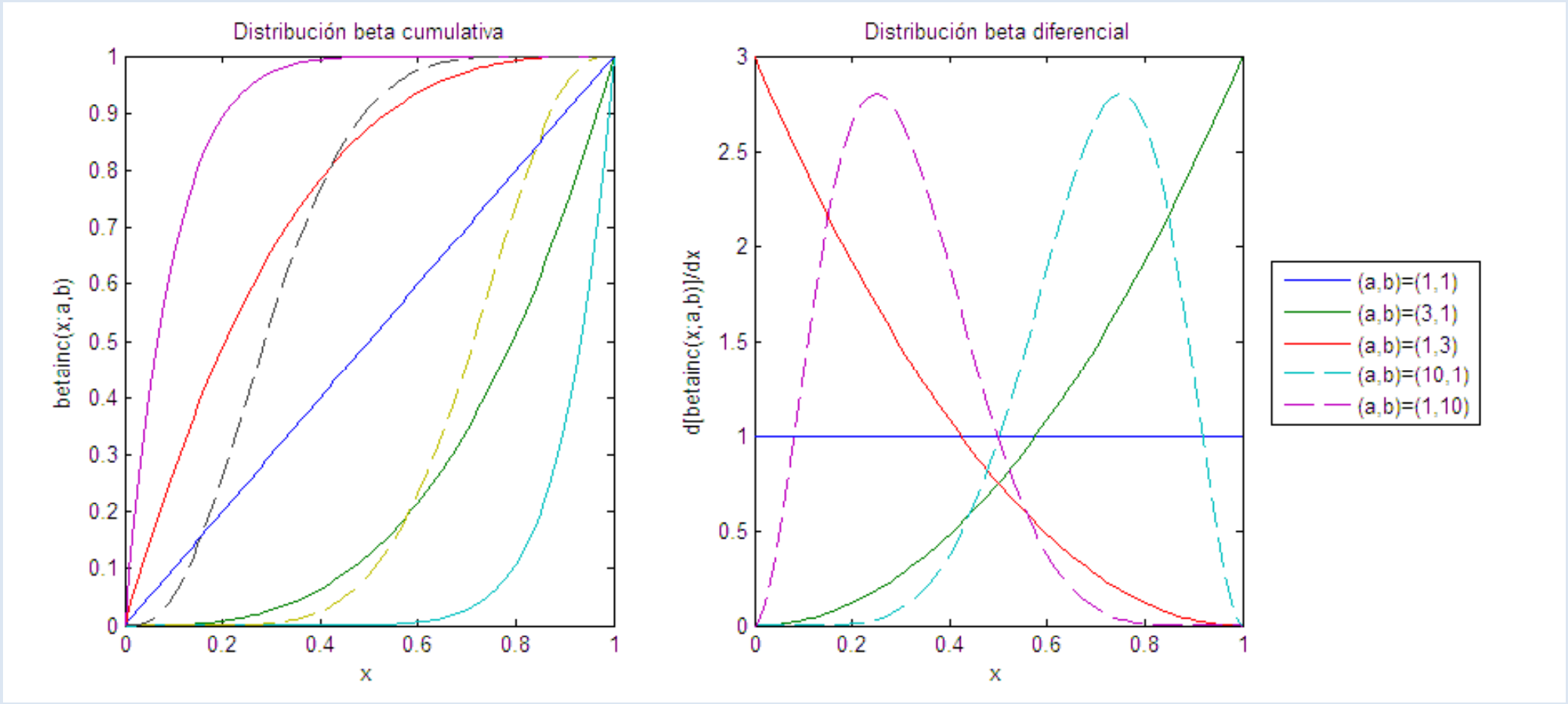
- La función beta incompleta es:

$$I_x(a,b) = \frac{1}{B(a,b)} \int_o^x t^{a-1}(1-t)^{b-1}\,dt$$

En matlab: ***y = betainc(x,a,b)***

Con a , b >0 y x<1.

Forma de la distribución beta



La función gamma

- Se define la función gamma como:

  En matlab: ***y = gamma(a)***.

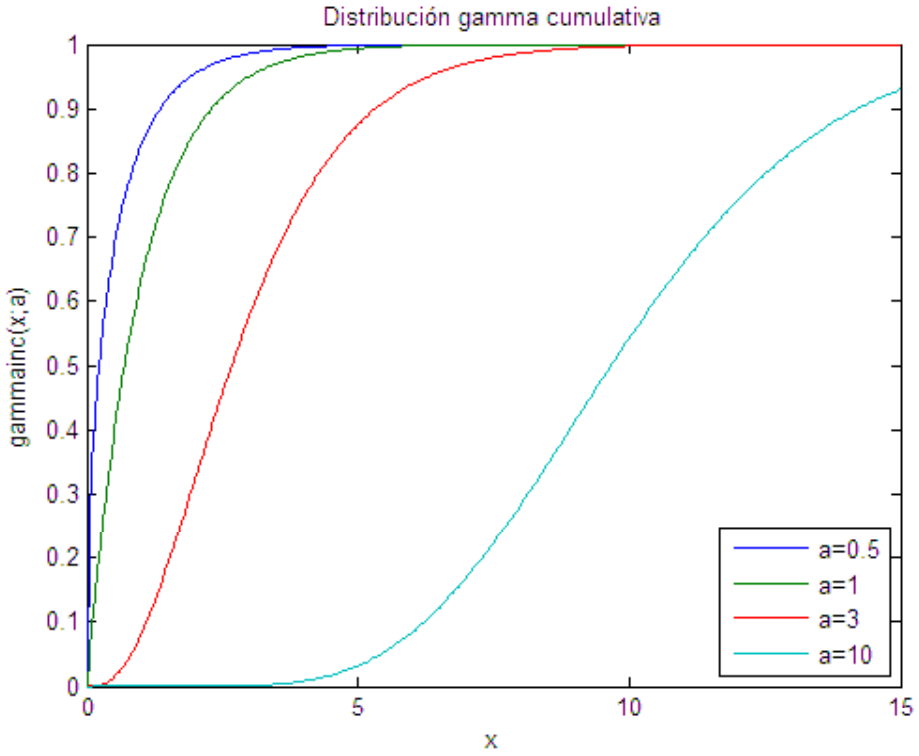
$$\Gamma(a) = \int_0^{\infty} e^{-t} t^{a-1} dt$$

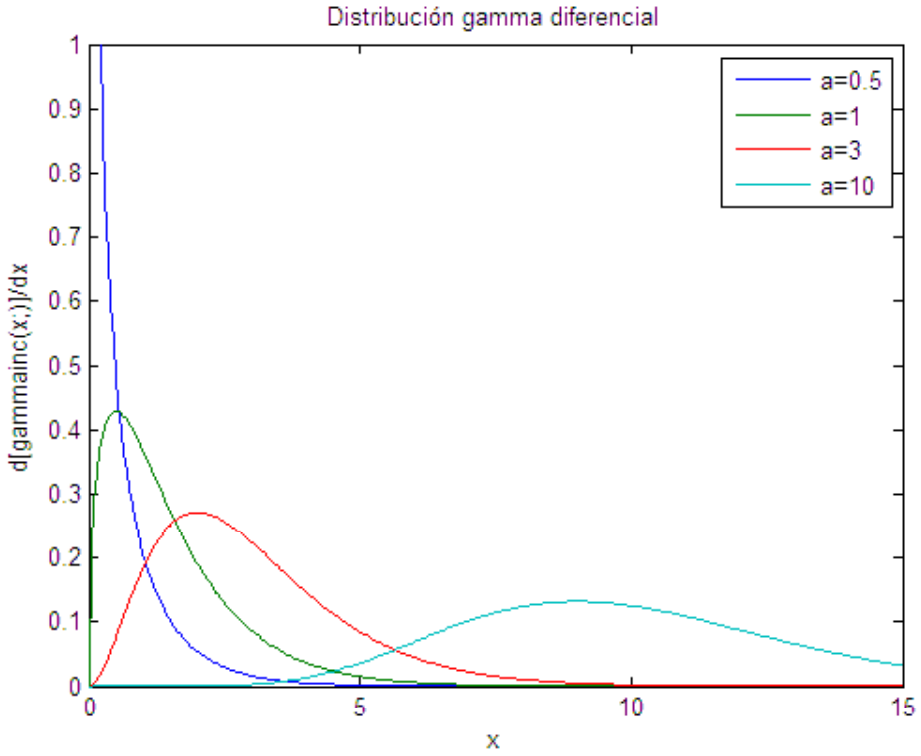
- Se define la función gamma incompleta como:

  En matlab: ***y = gammainc(x,a)***.

$$P(x,a) = \frac{1}{\Gamma(a)} \int_0^{x} e^{-t} t^{a-1} dt$$

Forma de la distribución gamma





Indeterminación en la media

- Supongamos que hemos tomado una muestra de N valores de x y que hemos encontrado que el valor medio de la muestra es $\langle x \rangle$.

Por ejemplo en nuestro caso, para el radio $\langle R \rangle = 132.16\ \mu m$ para una muestra con N=310.

- El valor medio de x en la población, que llamaremos $\langle x \rangle_{true}$ nos es desconocido.

En nuestro caso, medir $\langle R \rangle_{true}$ supone medir el radio de todas las partículas del bote y hacer la media.

Se nos plantea la pregunta:

¿Está $\langle R \rangle$ suficientemente cerca de $\langle R \rangle_{true}$ o tenemos que medir más partículas?

Probabilidad de obtener un valor de la media

- Si cogemos M muestras, con el mismo número de valores N, la media $<x>_k$ de cada una de ellas es una variable aleatoria. Como:

$$< x >= \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Si P(x) es la probabilidad de obtener el valor x, entonces, la probabilidad de obtener la media <x> es:

$$P(< x >) = P(x_1) \cdot P(x_2) \cdot \ldots \cdot P(x_N)$$

condicionada a $x_1 + x_2 + x_3 + .. + x_N = N<x>$

Teorema central del límite

El Teorema Central del Límite nos dice que sea cual sea la forma de P(x), cuando N es muy grande la distribución P(<x>) tiende a una distribución gausiana con

$$\sigma(<x>) = \frac{\sigma(x)}{\sqrt{N}}$$

$\sigma(x)$ es la desviación estándar de la población de x

Normalmente no conocemos la desviación estándar $\sigma(x)$ de la población, sino sólo la de una muestra, así que estimamos

$$\sigma(<x>) \approx \frac{s(x)}{\sqrt{N}}$$

s(x): desviación estándar de la muestra.

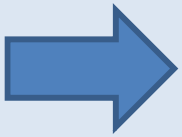
Podemos usar esta información para responder a la pregunta que nos hemos hecho.

Intervalo de confianza para la media

Como conocemos la forma de la distribución de <x>, podemos dar la probabilidad de que la diferencia entre <x> y $<x>_{true}$ sea mayor de un determinado valor.

Dado un número k >0, la probabilidad P de que...

$$\left| < R >_{true} - < R > \right| < k\sigma(< R >)$$



$$< R >_{true} - k\sigma \; << R >< < R >_{true} + k\sigma$$

...viene dada por:

$$P = \int_{<R>_{true}-k\sigma}^{<R>_{true}+k\sigma} f_{gauss}(t;< R >_{true},\sigma)dt = F_{gauss}(< R >_{true} + k\sigma;< R >_{true},\sigma)$$

$$- F_{gauss}(< R >_{true} - k\sigma;< R >_{true},\sigma)$$

Cálculo del intervalo de confianza

Puede parecer que no podemos calcular P, porque no conocemos <R>true. Sin embargo, de las forma de F_{gauss}.

$$F_{gauss}(x;<x>,\sigma) = 1 - 0.5\,\mathrm{erfc}\left(\frac{z}{\sqrt{2}}\right) \qquad z = \frac{x - <x>}{\sigma}$$

$$F_{gauss}(<R>_{true} + k\sigma;<R>_{true},\sigma) = 1 - 0.5\mathrm{erfc}(\frac{k}{\sqrt{2}})$$

$$F_{gauss}(<R>_{true} - k\sigma;<R>_{true},\sigma) = 1 - 0.5\,\mathrm{erfc}(-\frac{k}{\sqrt{2}})$$

$$P = 0.5\left[\mathrm{erfc}\left(-\frac{k}{\sqrt{2}}\right) + \mathrm{erfc}\left(\frac{k}{\sqrt{2}}\right)\right]$$

Ejemplo de cálculo (I)

Para la muestra del radio:

N = 310 <R> = 132.16 μm s(R) = 44.54 μm

Sabemos que <R> tiene una distribución gausiana centrada en $<R>_{true}$ micras y con

$$\sigma(<R>) \approx \frac{44.54}{\sqrt{310}} = 2.53 \ \mu\text{m}$$

¿Cuál es la probabilidad de que $|<R>_{true} - <R>| < 2$ μm?

Respuesta:

$$k = \frac{2}{2.53} = 0.79$$

$$P = 0.5\left[\text{erfc}\left(-\frac{0.79}{\sqrt{2}}\right) + \text{erfc}\left(\frac{0.79}{\sqrt{2}}\right)\right] = 0.57$$

Ejemplo de cálculo (II)

¿Cuántas partículas tengo que medir para que $|<R>_{true} - <R>| < 2\ \mu m$ con probabilidad P=0.95?

Respuesta:

Necesito resolver la ecuación:

$$P = 0.5\left[\operatorname{erfc}\left(-\frac{k}{\sqrt{2}}\right) + \operatorname{erfc}\left(\frac{k}{\sqrt{2}}\right)\right] = 0.95$$

$$k = 1.96 = \frac{2}{\sigma(< R >)}$$

Pero:

$$\sigma(< R >) \approx \frac{44.54}{\sqrt{N'}}$$

$$N' = \left(\frac{1.96}{2} \times 44.54\right)^2 = 1905$$

Y estimo que tengo que agrandar mi muestra a N' = 1905 partículas.

¿Tienen dos series la misma media?

Supongamos que en un experimento he tomado una muestra 1 de una población y he hallado que la media es $<x>_1$.

Si tomo otra muestra 2, en general, no obtendremos el mismo valor para la media, es decir:

$$< x >_1 \neq < x >_2$$

Se nos plantea la siguiente pregunta:

¿Es la diferencia fruto del azar, o ha cambiado la población entre las dos medidas?

La población puede haber cambiado porque:

- He alterado algún parámetro de mi experimento.
- Ha evolucionado en el tiempo transcurrido entre las dos medidas.

Ejemplo

Cada serie de datos de la carga y del radios se ha hecho con distintos parámetros experimentales. Por ejemplo, tenemos:

Serie 4	**Serie 5**
N = 72	N = 17
<q>=2.04e-14 culombios	<q>=1.74e-14 culombios
s(q) = 1.24e-14 culombios	s(q) = 1.22e-14 culombios
<q> = 0.146e-14 culombios	σ(<q>) = 0.296e-14

La desviación estándar es parecida, pero la media es muy diferente. ¿Se debe la diferencia al azar? (p. ej. porque hemos tomado muy pocos datos en la serie 5). ¿Han afectado los cambios de los parámetros a los resultados del experimento?

El test de Student o t-test

• Este test determina si dos muestras A y B **con la misma desviación estándar** difieren de forma significativa en su media.

1) Calcular la desviación estándar conjunta s_D como:

$$s_D = \sqrt{\frac{(N_A - 1)s_A(x)^2 + (N_B - 1)s_B(x)^2}{N_A + N_B - 2}\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}$$

2) Calcular el valor de:

$$t = \frac{< x >_A - < x >_B}{s_D}$$

De ahí el nombre de t-test.

• Llamemos P(t|ν) representa la probabilidad de obtener por azar un valor de |t| mayor o igual que $|<x>_A - <x>_B|/s_D$

$\nu = N_A + N_B - 2$, recibe el nombre de número de grados de libertad.

El t-test (II)

Puede demostrarse que la probabilidad P(t|ν) viene dada por:

$$P(t\,|\,\nu)=\frac{1}{\nu^{2}B\left(\frac{1}{2},\frac{\nu}{2}\right)}\int_{-t}^{t}\left(1+\frac{z^{2}}{\nu}\right)^{-\frac{\nu+1}{2}}dz$$

P(t|ν) puede escribirse usando la función beta incompleta.

$$P(t\,|\,\nu)=\text{betainc}\left(x,\frac{\nu}{2},\frac{1}{2}\right)\qquad x=\frac{\nu}{\nu+t^{2}}$$

A esta distribución también se la llama distribución de Student cumulativa.

Si P(t|ν) es bajo, entonces la diferencia entre las medias de las muestras A y B es poco probable que se deba al azar.

Se concluye entonces que las medias son diferentes.

Ejemplo (test-t)

Las cuentas de nuestro ejemplo son:

$$s_D = \sqrt{\frac{71\times(1.24e-14)^2 + 16\times(1.22e-14)^2}{87}\left(\frac{1}{72}+\frac{1}{17}\right)}$$

$$s_D = 3.33e-15$$

$$t = \frac{(2.04-1.74)e-14}{3.33e-15} = 0.901 \qquad x = \frac{87}{87+0.901^2} = 0.991$$

$$P(0.901|87) = \text{betainc}\left(0.991, \frac{87}{2}, \frac{1}{2}\right) = 0.376$$

La probabilidad de que la diferencia se deba al azar es de un 37.6%. Se concluye que es muy probable que ambas medias sean diferentes por haber cambiado los parámetros experimentales

Indeterminación en la desv. estándar

• Supongamos que hemos tomado una muestra de N valores de x y que la desviación estándar de la muestra es s(x).

Por ejemplo en nuestro caso, para el radio s(R)=44.54 μm para una muestra con N=310.

• La desviación estándar de la población, que llamaremos σ(x) nos es desconocida.

En nuestro caso, medir σ(x) supone medir el radio de todas las partículas del bote y hallar su desviación estándar.

Se nos plantea la pregunta:

¿Está s(R) suficientemente cerca de σ(R) o tenemos que medir más partículas?

NOTA: Este problema no es tan común como el de la media porque muchas veces no conocemos la forma de la distribución y/o no nos interesa saber con mucha precisión su desviación estándar.

Probabilidad de obtener una variancia

• Si tomamos M muestras, la variancia $Var_k(x)$ de cada una de ellas es una variable aleatoria. Como:

$$\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - < x >)^2 = < x^2 > - < x >^2$$

• Si P(x) es la probabilidad de obtener el valor x, entonces, la probabilidad de obtener Var(x) es:

$$P\big(Var(x)\big) = P(x_1)\cdot P(x_2)\cdot \ldots \cdot P(x_N)$$

condicionada a:

$$x_1^2 + x_2^2 + \ldots + x_N^2 - N < x >^2 = N\,\mathrm{Var}(x)$$

$$x_1 + x_2 + \ldots + x_3 = < x >$$

La distribución chi-cuadrado

Puede demostrarse teóricamente que si xj es una variable aleatoria **con distribución gausiana** con desviación estándar σ, la variable

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(x_i - < x >\right)^2 = \frac{(N-1)s(x)}{\sigma(x)}$$

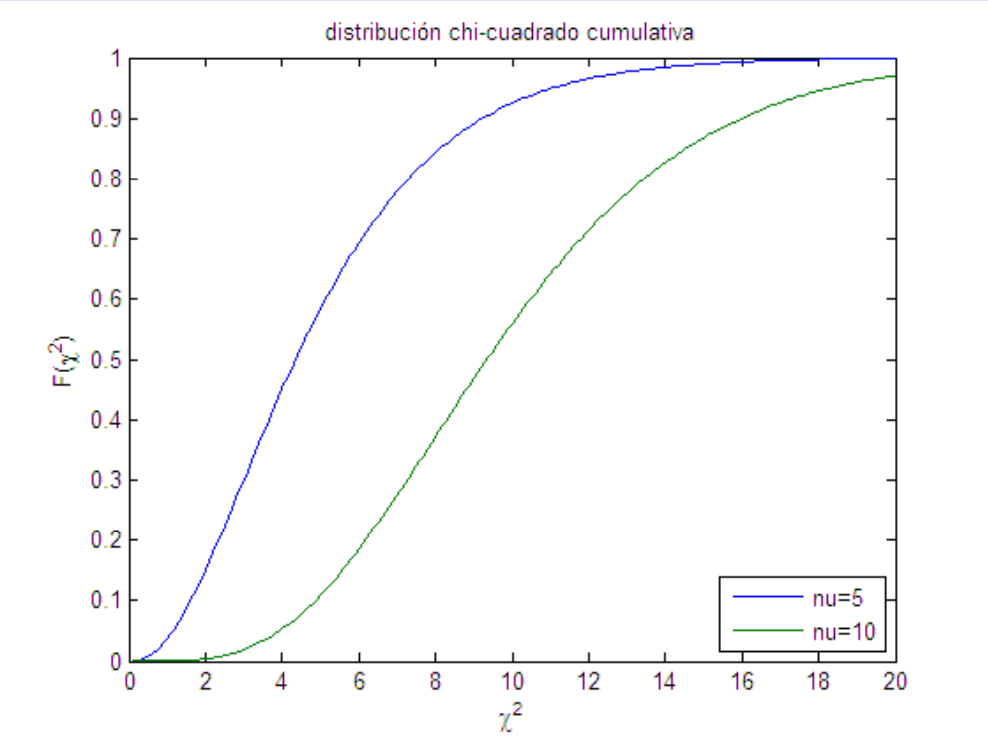
sigue la siguiente distribución cumulativa:

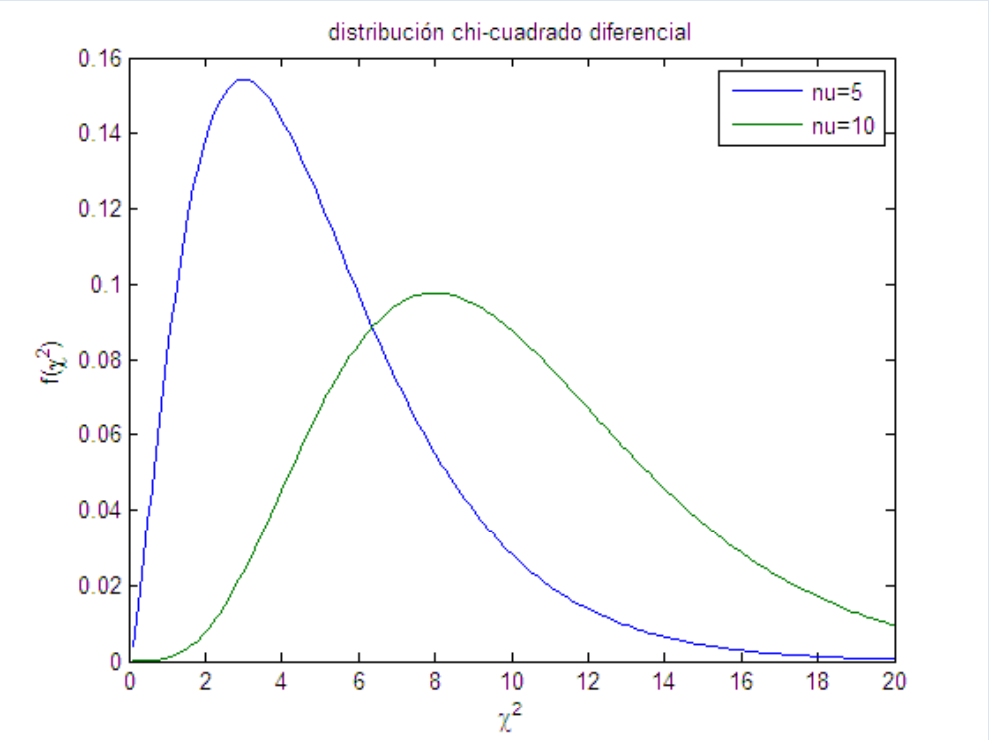
$$F_{chi-cuad}\left(\chi^2;\nu\right) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}\int_o^{\chi^2} t^{\frac{\nu}{2}-1}e^{-\frac{t}{2}}dt = \text{gammainc}(\frac{\chi^2}{2};\frac{\nu}{2})$$

que recibe el nombre de distribución chi-cuadrado

ν=N-1 es el número de grados de libertad.

Representación gráfica (chi-cuadrado)





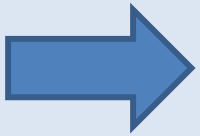
$<\chi^2> = \nu$, $\mathrm{Var}(\chi^2) = 2\nu$

Intervalos de confianza para la varianza

Si **estamos seguros de que x sigue una distribución gausiana**, podemos dar la probabilidad de que la diferencia entre s(x) y σ(x) sea mayor que un determinado valor.

Dado un número k>0 queremos saber cuál es la probabilidad P de que:

$$\left|\sigma(R) - s(R)\right| < k\sigma(R)$$



$$\sigma(R) - k\sigma(R) < s(R) < \sigma(R) + k\sigma(R)$$

$$(1-k)\sigma(R) < s(R) < (1+k)\sigma(R)$$

Multiplicando por (N-1) y dividiendo por σ(x)

$$(1-k)(N-1) < \frac{(N-1)s(R)}{\sigma(R)} < (1+k)(N-1)$$

Cálculo del intervalo de confianza

Y usando la definición de χ^2

$$(1-k)(N-1) < \chi^2 < (1+k)(N-1)$$

Luego la probabilidad P se encuentra partir de la distribución chi-cuadrado con ν=N-1 grados de libertad haciendo:

$$P = \int_{(1-k)(N-1)}^{(1+k)(N-1)} f_{chi-cuad}(t;\nu)dt$$

$$P = F_{chi-cuad}\left((1+k)(N-1);\nu\right) - F_{chi-cuad}\left((1-k)(N-1);\nu\right)$$

Y usando la definición de la distribución chi-cuadrado:

$$P = \text{gammainc}\left(\frac{(1+k)\nu}{2};\frac{\nu}{2}\right) - \text{gammainc}\left(\frac{(1-k)\nu}{2};\frac{\nu}{2}\right)$$

ν=N-1 es el número de grados de libertad.

Cálculo del intervalo de confianza

Sabemos que la distribución de R en la muestra de los datos del radio **se aproxima bien por una forma gausiana**.

N = 310 <R> = 132.16 μm s(R) = 44.54 μm

Var(R) = 1984.44 micras^2

¿Cuál es la probabilidad de que |σ - s(R)| sea menor que un 1% de σ(R)?

Respuesta: $k = 0.01$

$$P = \text{gammainc}\left(\frac{1.01\times 309}{2}; \frac{309}{2}\right) - \text{gammainc}\left(\frac{0.99\times 309}{2}; \frac{309}{2}\right)$$

$$P = 0.0976$$

¿Tienen dos series la misma variancia?

Supongamos que hemos tomado una muestra 1 de una población y hemos hallado que su variancia es $Var_1(x)=s_1(x)^2$.

Si sacamos otra muestra 2, en general, no obtendremos el mismo valor para la variancia, es decir:

$$s_1(x) \neq s_2(x)$$

Se nos plantea la siguiente pregunta:

¿Es la diferencia fruto del azar, o es que ha cambiado la población entre las dos medidas?

La población puede haber cambiado porque:

- He alterado algún parámetro de mi experimento.
- Ha evolucionado en el tiempo transcurrido entre las dos medidas.

¿Tienen dos series la misma variancia?

Ejemplo: si consideramos separadamente los valores del radio para la serie 4 y 5:

Serie 4:

N = 72

<R>=133.41 μm

s(R) = 40.64 μm

σ(<R>) = 14.79 μm

Serie 5

N = 17

<R>=133.88 μm

s(R) = 34.29 μm

σ(<R>) = 8.32 μm

La media es aproximadamente la misma, pero la desviación estándar es diferente. ¿Puede deberse la diferencia a que hemos tomado muy pocos datos en la serie 5?

Nota: Ambas series no tienen por qué tener la misma media para poder hacer el F-test

El test-F

1) Calcula el cociente

Si $Var_A(x) > Var_B(x)$ $\quad F = Var_A(x)/Var_B(x) \quad \nu_1 = N_A - 1 \quad \nu_2 = N_B - 1$

Si $Var_B(x) > Var_A(x)$ $\quad F = Var_B(x)/Var_A(x) \quad \nu_1 = N_B - 1 \quad \nu_2 = N_A - 1$

de modo que el valor de F sea siempre mayor que uno.

2) Llamo $P(F|\nu_1;\nu_2)$ a la probabilidad de obtener al azar un valor de F igual o mayor que el que he obtenido. Puede demostrarse que esa probabilidad vale:

$$P(F\,|\,\nu_1,\nu_2) = \int_0^F \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{z^{(\nu_1-2)/2}}{\left(1+\frac{\nu_1}{\nu_2}z\right)^{\frac{\nu_1+\nu_2}{2}}} dz$$

El test-F

3) Puede demostrarse que

$$I = \text{betainc}\left(\frac{\nu_2}{\nu_2 + \nu_1 F}; \frac{\nu_2}{2}, \frac{\nu_1}{2}\right)$$

$$P(F \mid \nu_1, \nu_2) = 2\min\left[I, 1 - I\right]$$

Si P(t|ν) es bajo, entonces la diferencia entre las medias de las muestras A y B es poco probable que se deba al azar.

Se concluye entonces que las medias son diferentes.

Ejemplo (test-F)

Para la serie 4: $\mathrm{Var(R)} = 1651.5$

Para la serie 5: $\mathrm{Var(R)} = 1176.1$

$$F = \frac{1651.5}{1176.1} = 1.404 \qquad N_1 = 72 \qquad N_2 = 17$$

$$\frac{\nu_2}{\nu_2 + \nu_1 F} = \frac{16}{16 + 71 \times 1.404} = 0.1383$$

$$I = \mathrm{betainc}(0.1383; \frac{16}{2}, \frac{71}{2}) = 0.2275$$

$$P(F \mid \nu_1, \nu_2) = 2 \min\left[I, 1 - I\right] = 0.455$$

Se concluye entonces que hay un 45.5% de posibilidades de que la desviación estándar de ambas muestras sea igual.

¿Se ajusta una distribución a una fórmula teórica?

Supongamos que hemos tomado una muestra de N valores de x y que nos han dicho que la expresión de la distribución de x en la población es F(x).

Representamos $F_{exp}(x)$ y no coincide con F(x)

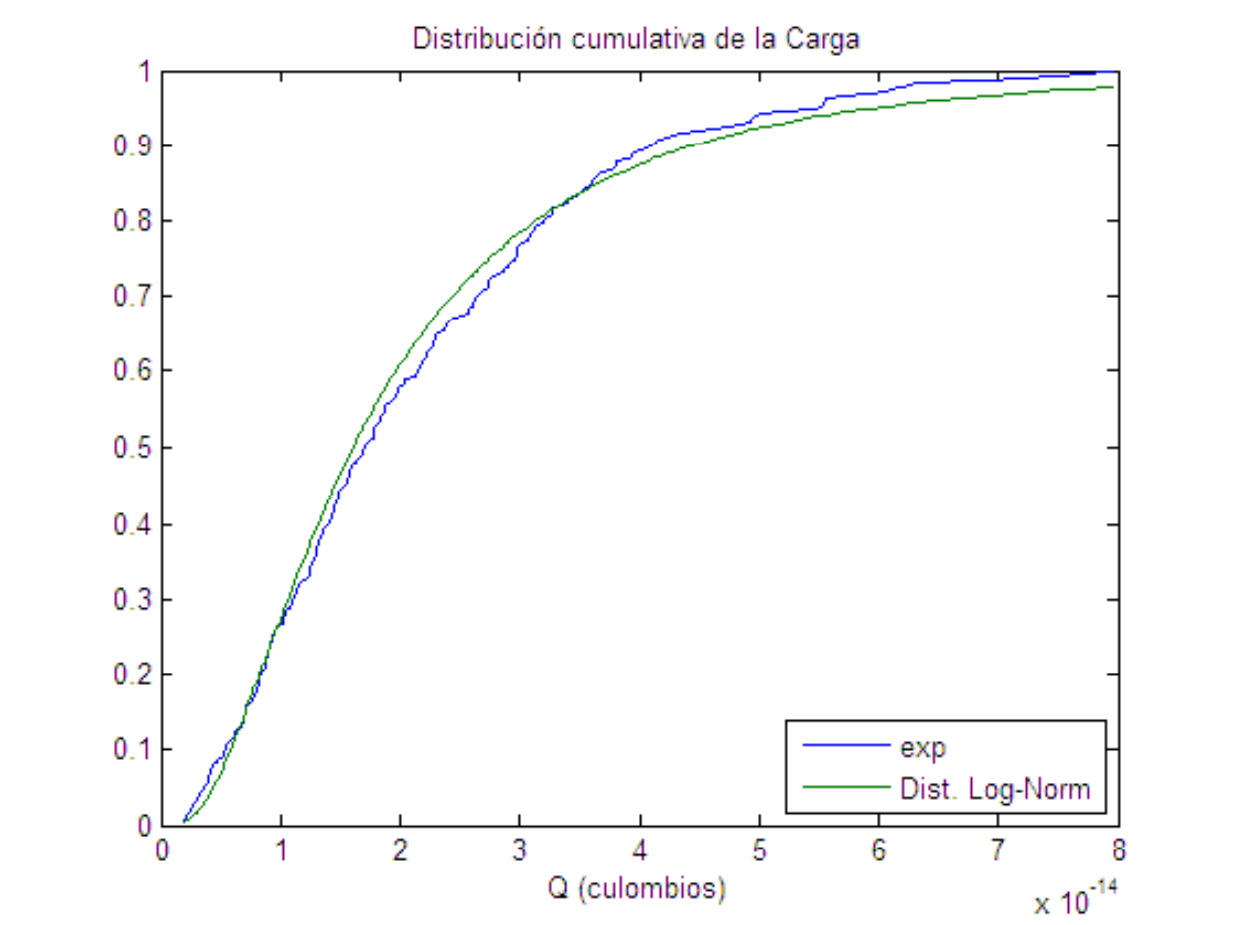
Se nos plantea la pregunta

¿Es esto fruto del azar porque hemos tomado muy pocos valores de N? O por el contrario ¿es falso que la distribución de x en la población es F(x)?

Ejemplo

Por ejemplo, parece que los datos de la carga se ajustan a una distribución log-normal:

¿Cuál es la probabilidad de que las diferencias se deban al azar?



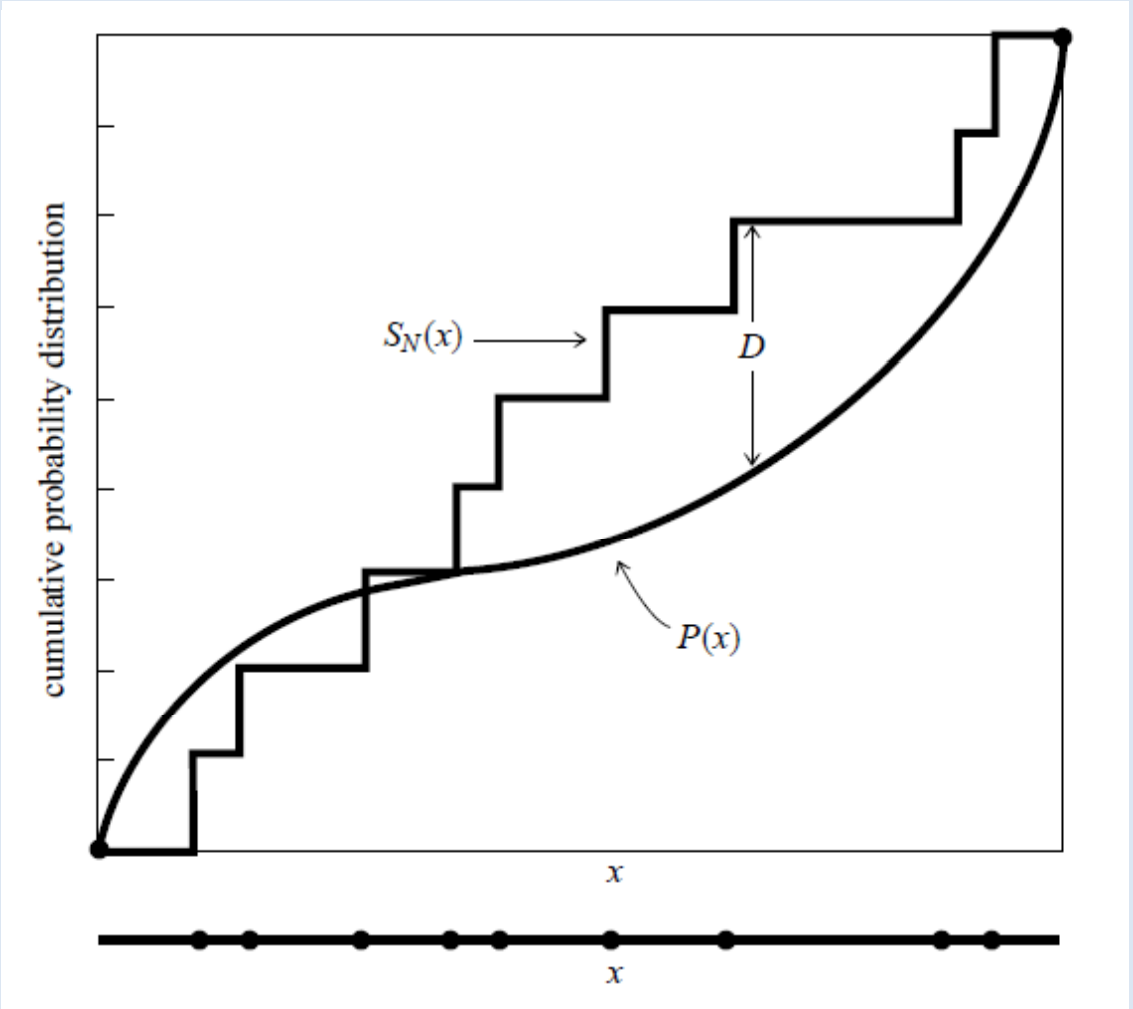
El test de Kolmogorov-Smirnoff

1) Calculamos:

$$D_o = \max \left| F_{\exp}(x) - F_{teo}(x) \right|$$

$F_{teo}(x)$: distribución cumulativa teórica que queremos comprobar

D: máxima distancia en vertical entre la distribución cumulativa experimental y la teórica que proponemos.



El test de Kolmogorv-Smirnoff

Puede demostrarse que la probabilidad de haber obtenido un valor de la distancia D mayor o igual que D_o por azar viene dada por:

$$P(D > D_o) = 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp\left(-2 j^2 z^2\right)$$

$$z \approx \left[\sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}}\right] D_o$$

Si $P(D>D_o)$ es bajo, es muy probable que $F_{exp}(x)$ no se corresponda con $F_{teo}(x)$.

Código

Editor - F:\docencia\clases\materiales recurrentes\MNS\analisis\KSone.m

File Edit Text Go Cell Tools Debug Desktop Window Help

Stack: Base

Save

```
1   function [prob,d] = KSone(data,func)
2   % PARÁMETROS DE ENTRADA: un vector data y una función anónima func de una única variable que
3   % representa la distribución cumulativa teórica que suponemos tiene la población
4   % de la que se han tomado los datos.
5   % PARÁMETROS DE SALIDA: la máxima separación d entre los datos contenidos en el
6   % vector data y la probabilidad prob de que el valor de d se seba al azar.
7   % Un valor pequeño de prob significa que la distribución cumulativa de los
8   % datos en data es muy probablemente differente de func (es decir, que las diferencias
9   % observadas no se deben al azar)
10  % Esta función está tomada del capítulo 14.3 del Numerical Recipes en C
11  data=data(:);                          % Convierte a data en un vector columna sea cual sea su forma inicial
12  n = length(data);
13  fo=0.0;
14  data=sort(data);
15  en=n;
16  d=0.0;
17  for j=1:1:n                            % Loop over the sorted data points.
18      fn=j/en;                           % Data's c.d.f. after this step.
19      ff=func(data(j));                  % Compare to the user-supplied function.
20      dt=max(abs(fo-ff),abs(fn-ff));     % Maximum distance.
21      if (dt > d)
22          d=dt;
23      end
24      fo=fn;
25  end
26  en=sqrt(en);
27  prob=probks((en+0.12+0.11/en)*d); % Compute significance.
28  end
29
30  function [sum] = probks(alam)
31  % Kolmogorov-Smirnov probability function.
32  EPS1=0.001;
33  EPS2=1e-8;
34  fac=2.0;
35  sum=0.0;
```

KStwo.m × KSone.m ×

¿Tienen dos muestras la misma distribución?

Supongamos que hemos tomado una muestra de N_1 valores de x y hemos representado su distribución cumulativa $F_{exp1}(x)$.

Ahora tomamos otra muestra de N_2 valores de x y representamos su distribución cumulativa $F_{exp2}(x)$

En general $F_{exp1}(1) \neq F_{exp2}(x)$

Se nos plantea la pregunta:

¿Es la diferencia fruto del azar, o es que ha cambiado la población entre las dos medidas?

KS-test para dos muestras

1) Calculamos:

$$D_o = \max \left| F_{\exp 1}(x) - F_{\exp 2}(x) \right|$$

Puede demostrarse que la probabilidad de haber obtenido un valor de la distancia D > D_o viene dada por:

$$P(D > D_o) = 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp\left(-2 j^2 z^2\right)$$

$$z \approx \left[\sqrt{N_{eff}} + 0.12 + \frac{0.11}{\sqrt{N_{eff}}} \right] D_o \qquad N_{eff} = \frac{N_1 N_2}{N_1 + N_2}$$

Código

Editor - F:\docencia\clases\materiales recurrentes\MNS\analisis\KStwo.m

File Edit Text Go Cell Tools Debug Desktop Window Help

```
function [prob,d] = KStwo(data1,data2)
% PARÁMETROS DE ENTRADA: dos vectores data1 y data2
% con los valores de dos muestras diferentes que queremos comparar para ver
% si tienen la misma distribución cumulativa
% PARÁMETROS DE SALIDA: la máxima separación d entre los datos contenidos en el
% vector data y la probabilidad prob de que el valor de d se seba al azar.
% Un valor pequeño de prob significa que la distribución cumulativa de los
% datos en data es muy probablemente differente de func (es decir, que las diferencias
% observadas no se deben al azar)
% Esta función está tomada del capítulo 14.3 del Numerical Recipes en C
data1=data1(:); % Convierte a data1 en un vector columna sea cual sea su forma inicial
data2=data2(:);
n1=length(data1);
n2=length(data2);
data1=sort(data1);
data2=sort(data2);
d=0.0;
j1=1;
j2=1;
fn1=0.0;
fn2=0.0;
while (j1 <= n1 && j2 <= n2)
    d1=data1(j1);
    d2=data2(j2);
    if (d1 <= d2)
        j1=j1+1;
        fn1=j1/n1;                      % Next step is in data1.
    end
    if (d2 <= d1)
        j2=j2+1;
        fn2=j2/n2;                       % Next step is in data2.
    end
    dt = abs(fn2-fn1);
    if (dt > d)
        d=dt;
```

KStwo.m × KSone.m ×

Tests usando la *Statistics Toolbox*

Si tenemos una instalación de Matlab con la Statistics Toolbox, ya tenemos programados los tests que hemos explicado en este tema:

Test de student: ***ttest(X), ttest2(X,Y)***

Test F: ***vartest2(X,Y)***

Test de Kolmogorov-Smirnoff: ***kstest(X), kstest2(X1,X2)***

Además de toda una serie de distribuciones cumulativas, diferenciales y sus inversas.

Bibliografía del tema:

• W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery. Numerical Recipes in C / Fortran 77 / Fortran 90, Cambridge University Press (1997). Disponible de forma gratuita en: http://www.nr.com/oldverswitcher.html

• Milton Abramowitz and Irene A. Stegun . Handbook of mathematical functions with formlas, graphs and numerical tables. Dover Books, 1972. Disponible de forma gratuita en: http://www.nrbook.com/abramowitz_and_stegun/

• Joaquim P. Marqués de Sá, Applied statistics using SPSS, STATISTICA, MATLAB and R , Springer Verlag, (2007