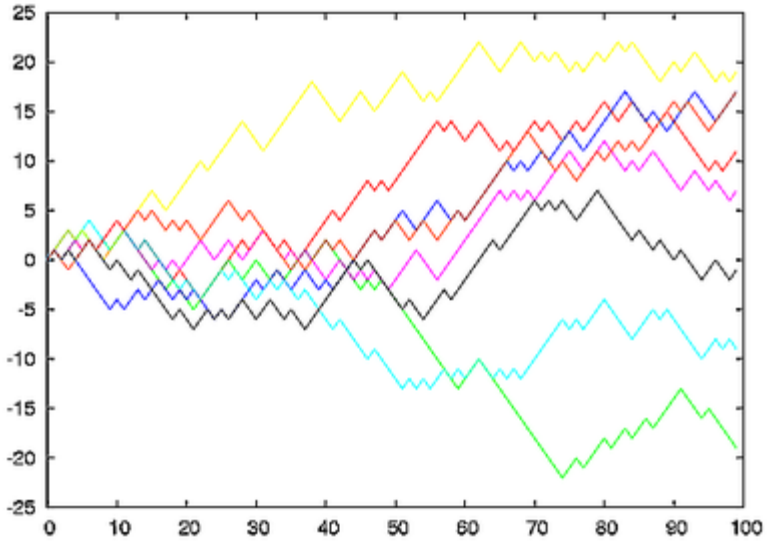
Unidad 4: Probabilidad, estadística y análisis de datos

CONTENIDO

Variables aleatorias (V.A.).
Caracterización estadística de V.A.: media, varianza, desviación estándar, moda, mediana.
Histogramas de V.A. y distribuciones de probabilidad.
Generación de números aleatorios.
Distribuciones gaussiana y de Poisson.
Integrales mediante el método de Montecarlo.
Caminos aleatorios.
Ajuste de funciones a datos experimentales.



Análisis estadístico de conjuntos de datos

Se dispone de un número *N* de medidas de una sola magnitud (experimental). Estas medidas (números reales) constituyen una variable aleatoria (V.A.) en el sentido de que su valor no se conoce a priori (antes de hacer las medidas del experimento), por la presencia de *ruido*, por ejemplo.

Este conjunto de datos puede caracterizarse estadísticamente mediante las siguientes funciones (entre otras): media, varianza, desviación estándar, moda, mediana, que se verán a continuación.

Definición de **media estadística** `(mean)` de una variable aleatoria *x*, con *N* valores x_i:

$$\overline{x} = \langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i$$

```
a=[0.7675 2.3022 1.4939 1.2179 1.4902 1.1683 1.1965 1.7614 1.7332 1.7360] ;
plot(a,'o') ;
mean(a)
sum(a)/10
```

Definición de **varianza estadística** (σ^2) de una variable aleatoria *x*, con *N* valores x_i: la media de las desviaciones cuadráticas de la media, esto es, expresa cuánto se devían los valores de la variable aleatoria de la media, sin importar el signo.

$$\sigma^2 = \langle (x_i - \overline{x})^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

```
a=[0.7675 2.3022 1.4939 1.2179 1.4902 1.1683 1.1965 1.7614 1.7332 1.7360] ;
var(a,1)
a_mean=mean(a)  ;
sum((a-a_mean).^2)/10
```

Definición de **valor esperado de la varianza estadística** ($E(\sigma^2)$) de una variable aleatoria *x*, con *N* valores x_i:

$$E(\sigma^2) = \langle (x_i - \overline{x})^2 \rangle = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

NOTA: La única diferencia (aunque importante) con la varianza es el *N-1* en el denominador. Para una aclaración de esta diferencia véase:
'Corrección de Bessel' http://en.wikipedia.org/wiki/Bessel%27s_correction

```
a=[0.7675 2.3022 1.4939 1.2179 1.4902 1.1683 1.1965 1.7614 1.7332 1.7360] ;
var(a,0)
a_mean=mean(a)  ;
sum((a-a_mean).^2)/9
```

Definición de **desviación estándar** $\sigma = (\sigma^2)^{\wedge 1/2}$ de una variable aleatoria *x*, con *N* valores x_i , es la raíz cuadrada de la varianza:

$$\sigma = +\sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

```
a=[0.7675 2.3022 1.4939 1.2179 1.4902 1.1683 1.1965 1.7614 1.7332 1.7360] ;
std(a,1)
a_mean=mean(a)  ;
sqrt(sum((a-a_mean).^2)/10)
```

Definición de **valor esperado de la desviación estándar** ($E(\sigma)=E(\sigma^2)^{\wedge 1/2}$) de una variable aleatoria *x*, con *N* valores x_i:

$$E(\sigma) = +\sqrt{E(\sigma^2)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

NOTA: La única diferencia (aunque importante) con la desviación estándar es el *N-1* en el denominador.

```
a=[0.7675 2.3022 1.4939 1.2179 1.4902 1.1683 1.1965 1.7614 1.7332 1.7360] ;
std(a,0)
a_mean=mean(a)  ;
sqrt(sum((a-a_mean).^2)/9)
```

Histogramas y distribuciones de probabilidad de V.A.

En un histograma se calcula el número de resultados de una V.A. en un cierto rango de valores, que puede representarse gráficamente en un diagrama de barras.

Sintaxis MATLAB: `[Nx,x_barra_mid]=hist(x,n_barras)`

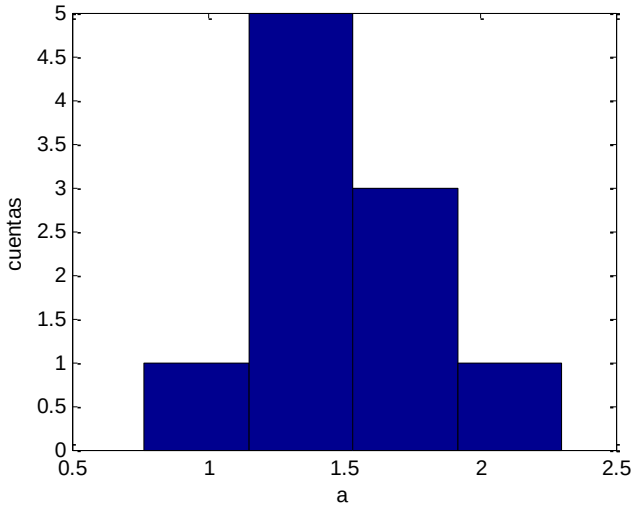
`x` son los valores de la V.A.

`n_barras` es el número de intervalos en los que se agrupan los valores de la V.A.

`Nx` es un vector que contiene el número de ocurrencias de la V.A. en cada barra.

`x_barra_mid` es el valor de la V.A. en el centro de la barra (*caja o bin*).

```
a=[0.7675 2.3022 1.4939 1.2179 1.4902 1.1683 1.1965 1.7614 1.7332 1.7360] ;
hist(a,4) ; xlabel('a') ; ylabel('cuentas') ;
mean(a) , std(a,1)
% para obtener el histograma
% sin que aparezca el gráfico
% utilizar [Na,a_mid]=hist(a,4)
% que se puede representar
% como un gráfico de barras con bar(a_mid,Na)
```



Distribución de probabilidad de una V.A.

Es una función discreta $P(x_i)$ que refleja la probabilidad de que una V.A. tenga un cierto valor x_i. Se relaciona con la definición de probabilidad de que una V.A. adquiera un valor x_i, que después de haber hecho N medidas de la magnitud se haya obtenido N_i veces el resultado x_i , esto es, la *proporción de casos favorables*.

$$P(x_i) = \lim_{N\to\infty} \frac{N_i}{N}$$

$$0 < P(x_i) < 1 \qquad \sum_{Dominio \;\; x} P(x_i) = 1$$

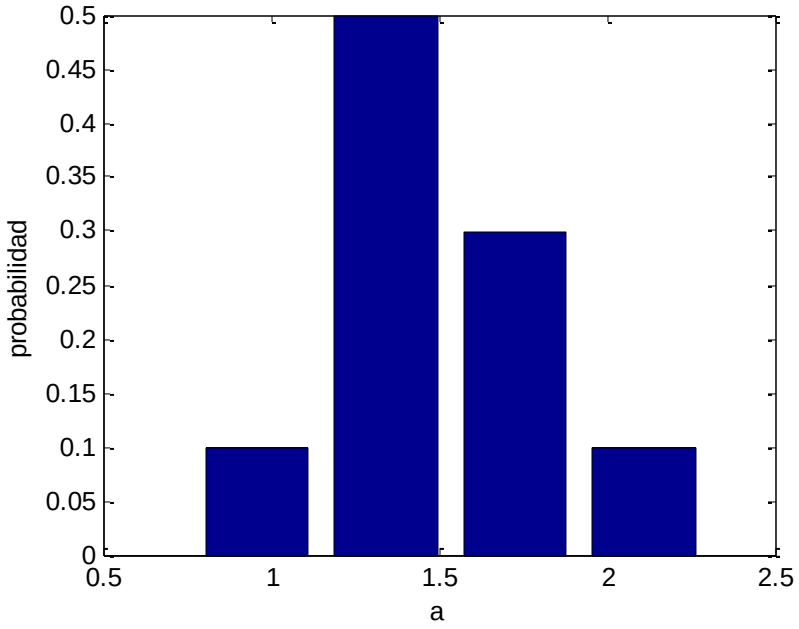
En la última igualdad (normalización) la suma debe hacerse sobre todo el Dominio de la V.A., constituido por todos los posibles valores de la V.A.

Ejemplo: Sea la V.A. que es el resultado de lanzar un dado de seis caras. El dominio es el conjunto de posibles valores {1,2,3,4,5,6} y su distribución de probabilidad es $P(1)$=1/6, $P(2)$=1/6, $P(3)$=1/6, $P(4)$=1/6, $P(5)$=1/6, $P(6)$=1/6, que está normalizada. Se trata de una distribución de probabilidad constante.

Relación entre el histograma y la distribución de probabilidad

Si el histograma de una V.A. aleatoria se **normaliza**, esto es, se divide por el número de valores analizados de la V.A., los valores de las barras reflejan la probabilidad de que se dé un resultado (en un cierto rango) de la V.A., esto es, la proporción de *casos favorables*.

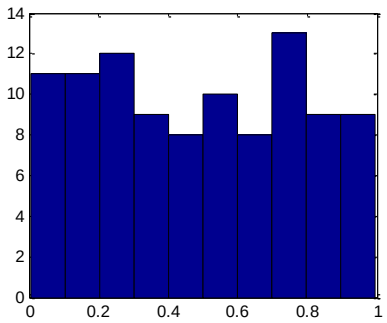
```
a=[0.7675 2.3022 1.4939 1.2179 1.4902 1.1683 1.1965 1.7614 1.7332 1.7360] ;
[Na,a_mid]=hist(a,4) ;
bar(a_mid,Na/10) ; xlabel('a') ; ylabel('probabilidad') ;
```

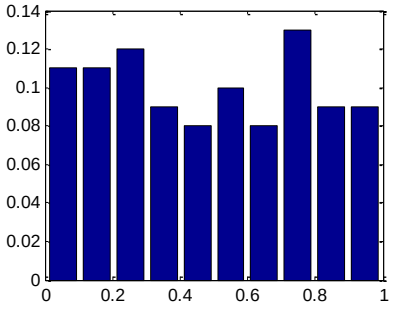


Generación de número aleatorios (valores de V.A.)

Función **rand**(n_filas,n_cols) genera una matriz de de n_filas x n_cols cuyos elementos tienen un valor entre 0 y 1, con distribución de probabilidad constante.

```
b=rand(1,100) ; hist(b,10) ; % 10 barras % sin normalizar (histograma)
[Nb,b_mid]=hist(b,10) ; bar(b_mid,Nb/100) ; % normalizado (probabilidad)
```





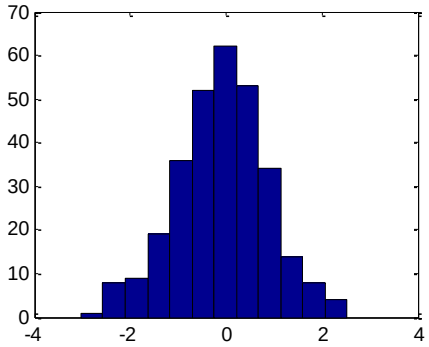
Función **randn**(n_filas,n_cols) genera una matriz de de n_filas x n_cols cuyos elementos tienen un valor entre -Inf y +Inf, con distribución de probabilidad *gaussiana* de media cero y varianza uno.

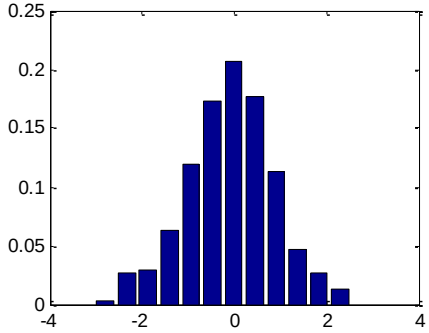
Función gaussiana $$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\bar{x})^2/2\sigma^2}$$

```
c=randn(1,300) ; hist(c,12) ; % 12 barras % sin normalizar (histograma)
[Nc,c_mid]=hist(c,12) ; bar(c_mid,Nc/300) ; % normalizado (probabilidad)
```

(SIGUE)

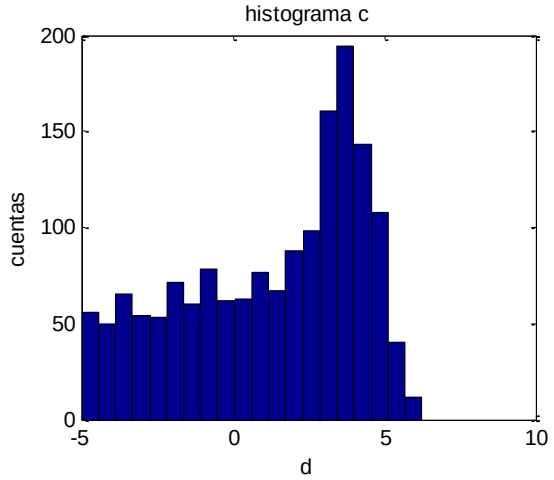
Histograma y distribución de probabilidad gaussiana del ejemplo anterior. Nótese que tiene forma de campana gaussiana, centrada en cero (la media) y con anchura igual (aproximadamente) a la desviación estándar (uno).



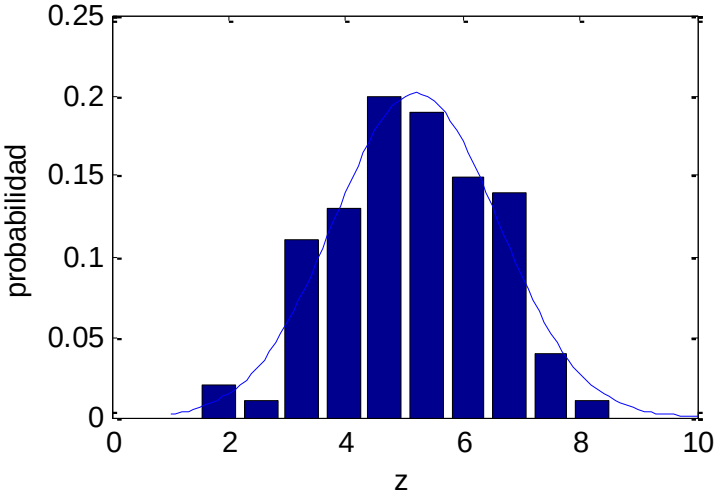


Combinación de histogramas y distribuciones de probabilidad.

```
d1=9*rand(1,1000)-5 ; d2=0.8*randn(1,600)+4 ; hist([d1 d2],20) ;
xlabel('d') ; ylabel('cuentas') ; title('histograma c') ;
```



Ejercicio: generar 100 valores aleatorios con distribución de probabilidad gaussiana de media 5.2 y desviación estándar 1.4. Representar gráficamente la distribución de probabilidad junto con la función analítica gaussiana correspondiente. Obtener la media y la desviación estándar.




```
N=100 ;
z=1.4*randn(1,N) +5.2 ;
x=[1:0.1:10] ;
xpdf=pdf('norm',x,5.2,1.4) ;
[Nz,z_mid]=hist(z) ;
bar(z_mid,Nz/N) ; hold on ;

% ATENCIÓN a la normalización de la gaussiana
bin=z_mid(2)-z_mid(1) ;
plot(x,xpdf*bin) ; hold off ; xlabel('z') ;
ylabel('probabilidad') ;

mean(z) , std(z)
```

Ejercicio: empleando los datos y la distribución de probabilidad anterior, encontrar:

a) La probabilidad de que la V.A. tenga un valor mayor que 5.2
b) La probabilidad de que la V.A. tenga un valor menor que 7.1
c) La probabilidad de que la V.A. tenga un valor entre 4.0 y 6.2

Responder a estas preguntas analizando tanto los valores de la V.A. como su histograma.

a)

```
% Atención a la utilización del vector lógico
b=z>5.2 ; sum(b)/N              % a partir del conjunto de datos
bb=z_mid>5.2 ; sum(Nz(bb)/N)    % a partir del histograma
```

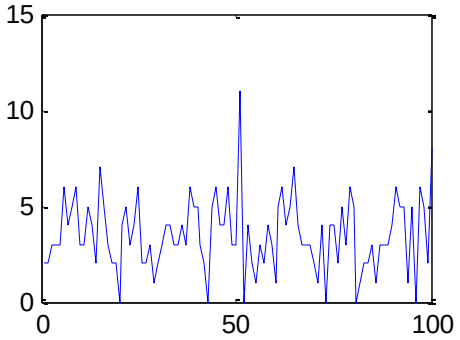
Ejercicio: Se considera el lanzamiento de un proyectil con ángulo 30 grados, con velocidad inicial aleatoria con distribución de probabilidad constante entre 30 y 50 m/s, bajo la acción de la gravedad terrestre.
a) Encontrar la distribución de probabilidad de la distancia de impacto con el suelo (alcance máximo), que es horizontal.
b) Encontrar la media de la distancia de impacto.
c) Encontrar la probabilidad de que la distancia de impacto sea mayor o igual que la media.
NOTA: efectuar al menos 1E5 lanzamientos y un histograma de 10 barras.

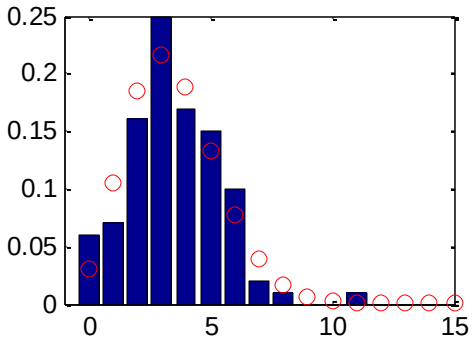
Distribución de probabilidad de Poisson (***)

La **distribución de Poisson** expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo. Concretamente, se especializa en la probabilidad de ocurrencia de sucesos con probabilidades muy pequeñas, o sucesos "raros".

Ejemplo: Consideremos una cuadrícula de *M=10 x 10* celdas. Se lanzan *N=350* partículas aleatoriamente sobre la cuadrícula. Evidentemente al final del proceso el número medio de partículas en cada celda es *N/M=3.5*. Aunque no todas las celdas tendrán el mismo número de partículas. Consideremos la V.A. *z*=número de partículas en cada celda. Queremos obtener la distribución de probabilidad de *z*.

```
M=10*10 ; % celdas
p=zeros(1,M) ; % n partículas en cada celda
N=350 ; % n partículas lanzadas
% LANZAMIENTO
for i=1:N
    n_celda=randi(M,1) ; % ATENCIÓN a randi
    p(n_celda)=p(n_celda)+1 ;
end
x=[0:15] ; % bins del histograma: números enteros
[Nz,z_mid]=hist(p,x) ;
subplot(2,1,1) ; plot(p)
subplot(2,1,2) ;
bar(z_mid,Nz/sum(Nz)) ; hold on ;
plot(x,pdf('pois',x,N/M),'or') ; hold off ;
xlim([-1 15]) ; % N/M es la media
```



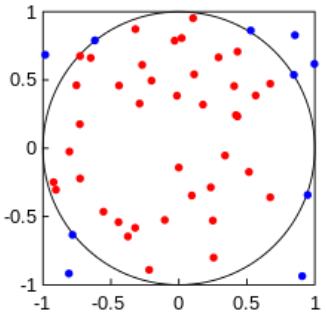


(***) 'Distribución de Poisson' http://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_Poisson

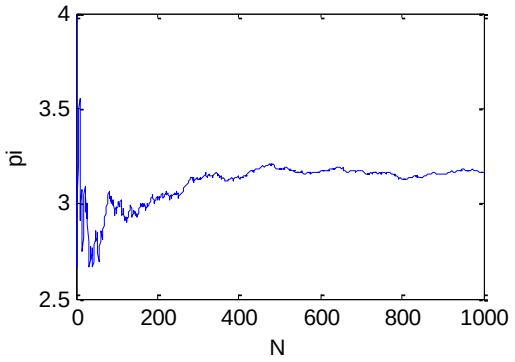
Integrales mediante el método de Montecarlo (***)

Se aprovecha el generador de números aleatorios para calcular una integral definida. Veamos un ejemplo, que obtiene el número *pi* calculando la integral de un cuadrante de círculo de radio uno:

$$\pi = 4 \int_0^1 \sqrt{1 - x^2}\, dx$$



El método consiste en generar aleatoriamente coordenadas dentro de un cuadrado y determinar qué proporción de ellas se encuentran dentro del cuadrandte de círculo, que es igual al cociente del área del cuadrante del círculo y el área del cuadrado.

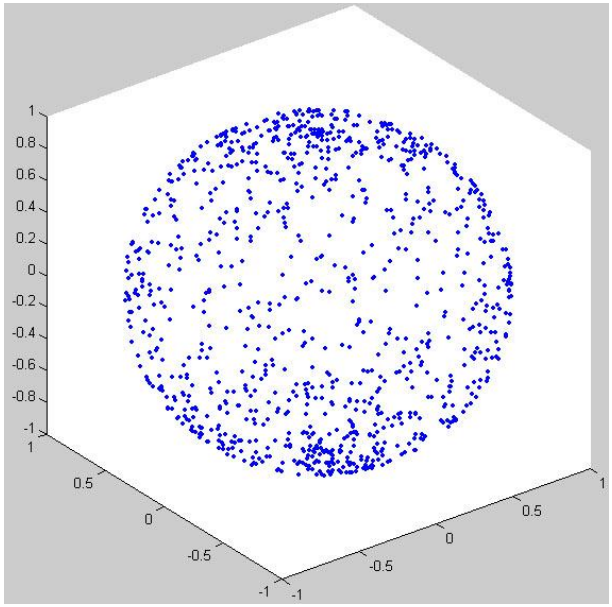


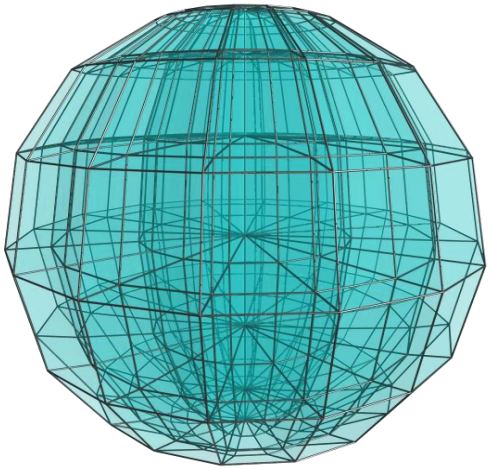

```
% cálculo de pi mediante Montecarlo
N=1E3 ; % número de puntos aleatorios generados
x=rand(1,N) ; y=rand(1,N) ; % genera puntos x,y en un cuadrado de área uno.
b=(x.^2+y.^2)<1 ; % encontrar elementos que están dentro del círculo
mi_pi=4*sum(b)/N   % encontrar la proporción dentro del círculo: sum(b)/N
```

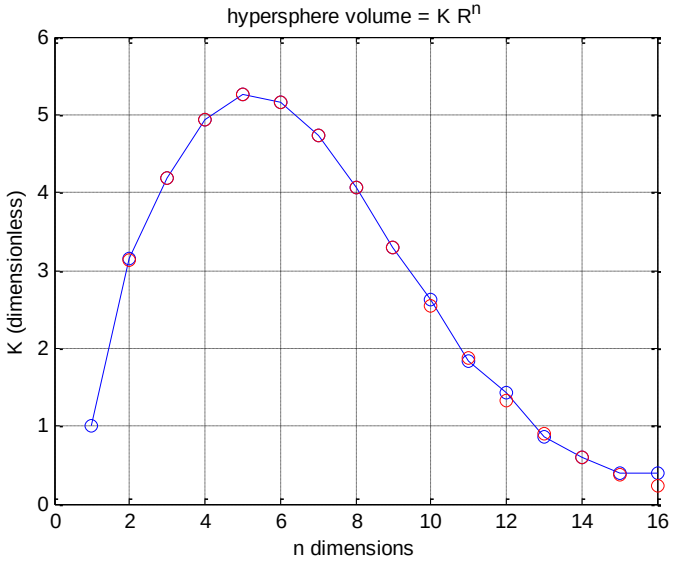
Ejercicio: determinar cómo converge la solución hacia *pi* según el número de puntos aleatorios (*N*).

(***) 'Método de Montecarlo' http://es.wikipedia.org/wiki/M%C3%A9todo_de_Montecarlo

Ejercicio: Calcular mediante el método de Montecarlo, el hiper-volumen de una hiper-esfera de 11 dimensiones.
Solución: V=1.8841*R^11

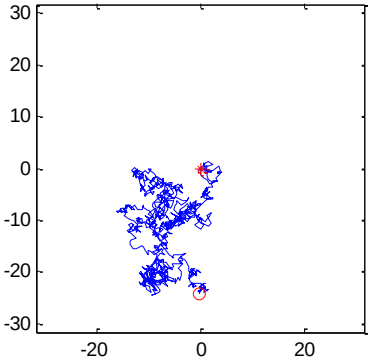






Caminos aleatorios (***)

Es una formalización matemática de la trayectoria que resulta de hacer sucesivos pasos aleatorios. Por ejemplo, la ruta trazada por una molécula mientras viaja por un líquido o un gas (movimiento browniano), el camino que sigue un animal en su búsqueda de comida, el precio de una acción fluctuante y la situación financiera de un jugador pueden tratarse como un camino aleatorio.




```
N=1000 ; % número de pasos
% tamaño máximo 1, a izquierda/derecha o arriba/abajo
x=2*rand(1,N)-1 ; y=2*rand(1,N) -1 ;
xx=cumsum(x) ; yy=cumsum(y) ; % coordenada acumulada
R=sqrt(N) ; % estimación de posición radial final
plot(xx,yy,0,0,'r*',xx(end),yy(end),'ro') ;
xlim([-R,R]) ; ylim([-R,R]) ; axis square ;
```

Ejercicio: determinar la distribución de probabilidad de V.A. distancia final al origen (*R*) para una secuencia de *M* caminos aleatorios. Encontrar la relación entre la media y la desviación estándar de *R* y el número de pasos (*N*) y el número de caminos efectuados (*M*). ¿Se obtiene una distribución gaussiana?

(***) 'Camino aleatorio' http://es.wikipedia.org/wiki/Camino_aleatorio

Generar distribuciones de probabilidad de usuario

Puede hacerse a partir de la distribución de probabilidad constante (`rand`) a la que se aplican comandos condicionales u operadores de comparación.

Ejemplo: generar números aleatorios de valores 2,3,7,8, cada uno de los cuales con probabilidad *P(2)=0.2, P(4)=0.1, P(7)=0.5, P(8)=0.2*. NOTA: la suma de probabilidades debe sumar 1. Obtener el histograma para comprobar el resultado.

```
N=1000 ; % generar 1000 valores
r=rand(1,N) ;
ran=zeros(1,N) ; % crear V.A.
for i=1:N
    if r(i)<0.2 ran(i)=2 ;          % P(2)
    elseif r(i)<0.3 ran(i)=4 ;      % P(2)+P(4)=0.3
    elseif r(i)<0.8 ran(i)=7 ;      % P(2)+P(4)+P(7)=0.8
    else ran(i)=8 ;           % P(2)+P(4)+P(7)+P(8)=1.0
    end
end
subplot(2,1,1) ; hist(ran,[2,4,7,8]) ;
subplot(2,1,2) ; plot(ran,'o') ;          % etiquetar ejes
```

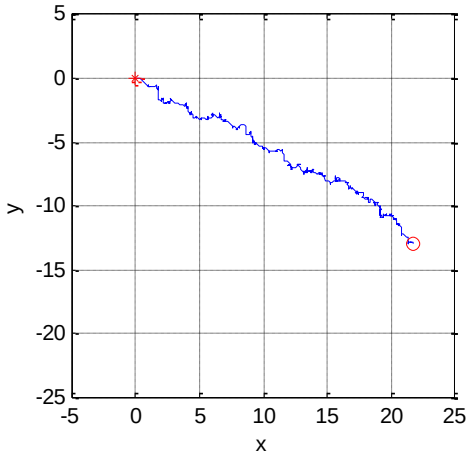
Ejercicio:

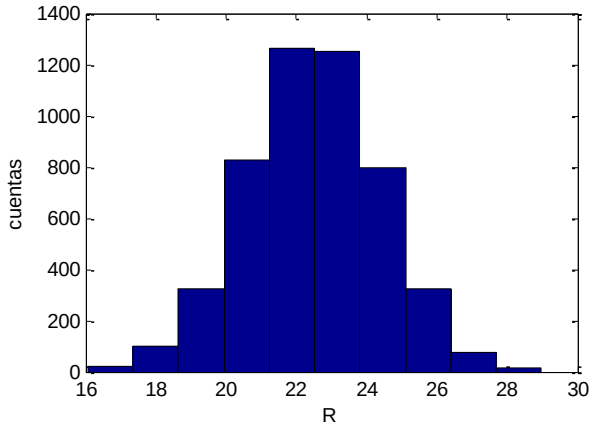
Generar un camino aleatorio en el plano *XY* en el que la partícula puede moverse cada vez un incremento fijo *d=0.1* o quedarse en el sitio (no moverse) con las siguientes probabilidades:

P(no se mueve)=0.1 ;
P(+d en dirección X)=0.3 ;
P(-d en dirección X)=0.1 ;
P(+d en dirección Y)=0.2 ;
P(-d en dirección Y)=0.3 ;

Representar gráficamente la trayectoria para 1000 pasos.

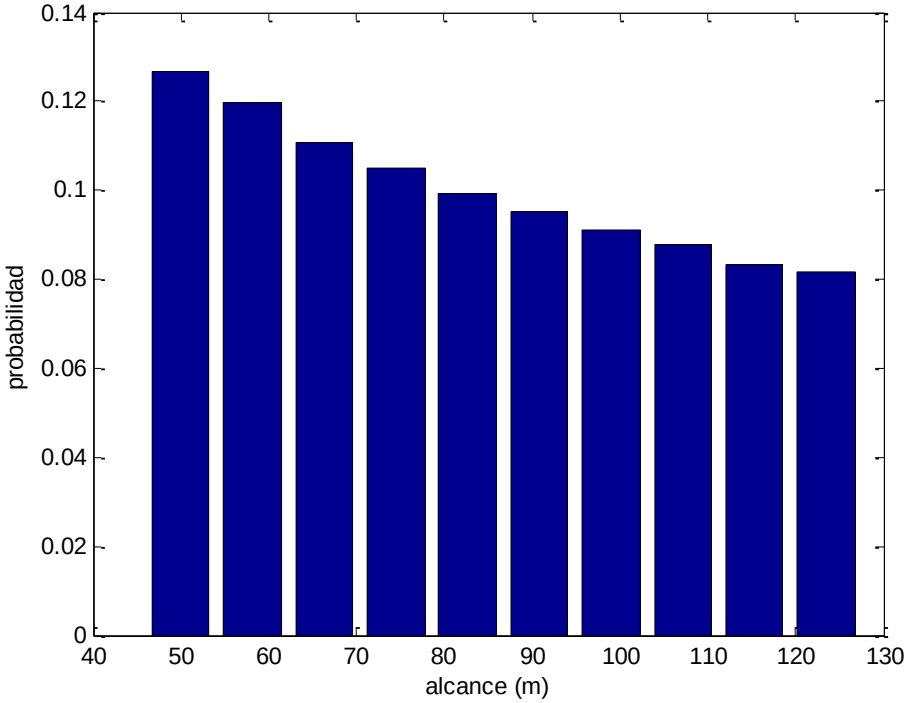
Para 5000 caminos de este tipo, representar un histograma de la distancia de la posición final al origen.





Ejercicio:

Se lanza un proyectil sobre la superficie plana de la Tierra, con ángulo de 45 grados sobre la horizontal y velocidades aleatorias (uniformes) entre 30 y 50 m/s. Obtener la distribución de probabilidad del alcance del proyectil.



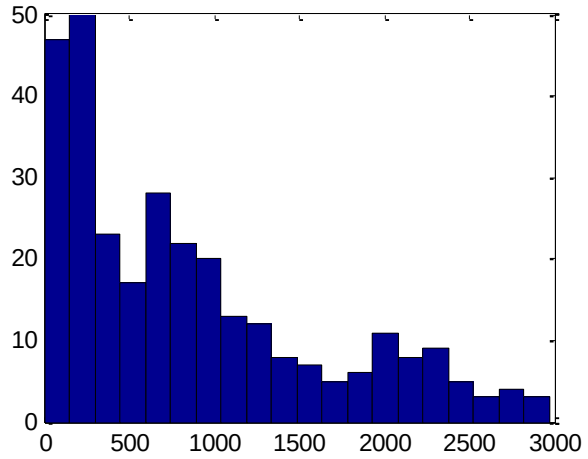
Otras caracterizaciones estadísticas de una V.A. o conjunto de medidas.

Obtención de valor mínimo y máximo: funciones `max` y `min`.

Definición de **moda** (`mode`): el valor que se encuentra más frecuentemente, esto es, el valor en el que el histograma presenta un máximo. Si hay varios iguales, devuelve el menor de ellos.

Definición de **mediana** (`median`): es aquel valor x_i que, ordenados de menor a mayor los N valores de x en una lista:
a) si N es impar el valor x_i en la posición de la mitad de la lista
b) si N es par, el valor promedio de x_i en las posiciones alrededor de la mitad de la lista.

```
% Atención a la función randi
z=randi(10,1,301) .*[1:301] ;
hist(z,20) ; min(z) , max(z) , mean(z)
mode(z) , median(z)
```



Ajuste de funciones a datos experimentales.

En el capítulo de sistemas de ecuaciones lineales se estudió cómo utilizar el comando **** para ajustar a datos por minimización.

Además MATLAB dispone de funciones (`polyfit`) para ajustar datos a funciones polinómicas de grado *n* (minimizando la diferencia cuadrática media) y generar el resultado del ajuste (`polyval`).

$$p(x)=p_1x^n+p_2x^{n-1}+...+p_nx+p_n$$

Sintaxis:

`p = polyfit(x,y,n)`

`p` es un vector con los coeficientes del polinomio (atención al orden)

`x, y` son los datos a los que se pretende ajustar

`n` es el grado máximo del polinomio

`y = polyval(p,x)`

`p` es el resultado de un `polyfit` previo

`y` es el valor correspondiente al polinomio evaluado en `x`, que puede ser un vector.

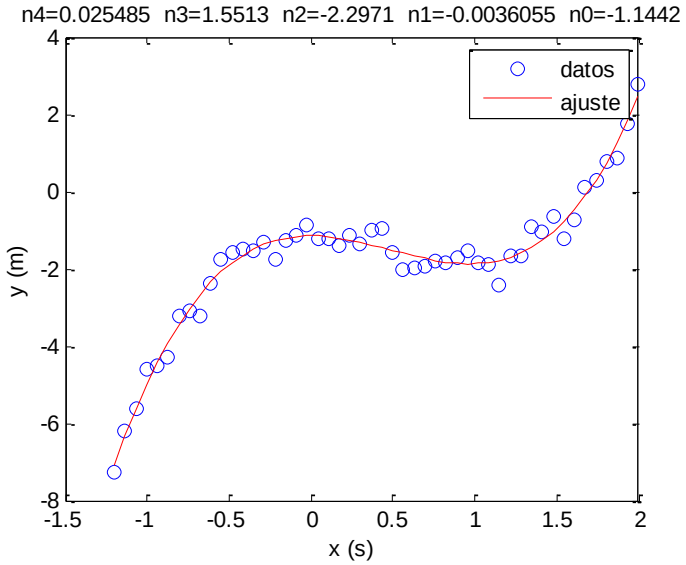
Ejemplo:

```
% ajuste a polinomio de cuarto orden
% simula datos con ruido aleatorio
N=50 ;
x=linspace(-1.2,2.0,N) ;
y=1.6*x.^3-2.3*x.^2-1.2+0.3*randn(1,N) ;

% ajustar
p=polyfit(x,y,4) ;
% p(1) debería ser próximo a cero (potencia 4)
y_fit=polyval(p,x) ;

% representación gráfica
plot(x,y,'ob',x,y_fit,'-r') ;
xlabel('x (s)') ; ylabel('y (m)') ;
legend('datos','ajuste') ;

% mostrar resultados en el título
title(['n4=' num2str(p(1))...
    '  n3=' num2str(p(2)) '  n2=' num2str(p(3))...
    '  n1=' num2str(p(4)) '  n0=' num2str(p(5))]) ;
```



Unidad 4. Resumen.

Funciones MATLAB

```
hist      bar
rand    randn   randi
mean    var     std      mode      median     max    min
pdf 'norm'   'pois'
polyfit    polyval
```

Destrezas

- Generar números aleatorios (`rand randn randi`).
- Generar números aleatorios con distrib. de probabilidad de *usuario*.
- Utilizar distribuciones de probabilidad (`pdf`).
- Distribuciones de probabilidad constante, gaussiana y Poisson.
- Obtener histogramas y normalizarlos (`hist  bar`).
- Caracterización estadística (`mean var std mode median min max`).
- Generar caminos aleatorios.
- Integrales mediante Montecarlo.
- Ajuste de datos a funciones polinómicas (`polyfit  polyval`)