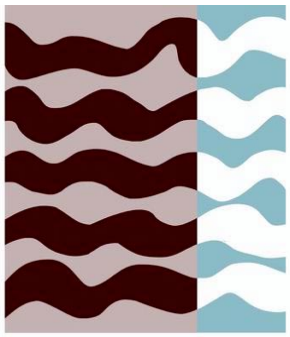
PROYECTO FIN DE CARRERA



IMPLEMENTACIÓN EN MATLAB DE TÉCNICAS MATEMÁTICAS EN EL ANÁLISIS DE EXPERIMENTOS

Alumno: Martín Abenza Jiménez
I.T.Agrícola (Industrias)

Director/es: Sergio Amat Plata
Sonia Busquier Sáez

IMPLEMENTACIÓN EN MATLAB DE TÉCNICAS MATEMÁTICAS EN EL ANÁLISIS DE EXPERIMENTOS

1. Objetivo del proyecto

El objetivo principal de este proyecto es la implementación en MATLAB de ANOVA (análisis de la varianza). Esta técnica estadística fue introducida para obtener de forma rigurosa la dependencia entre las variables de experimentos agronómicos. Posteriormente se extendió a otras ramas de la ciencia. Por su sencillez nos centraremos en el análisis de un solo factor. Confirmada dicha dependencia, el objetivo es ver como es ésta. Los ajustes lineales o cuadráticos son los más utilizados. Presentaremos la implementación de estos ajustes.

El proyecto será autocontenido, poniendo en relieve todos los aspectos teóricos necesarios para la comprensión de los métodos matemáticos así como una introducción a MATLAB. Cabría destacar que nuestra Universidad tiene licencia Campus de MATLAB.

Un experimento que puede ser estudiado con estas técnicas es el uso del purín de cerdo como abono orgánico, cabría citar el gran número de granjas en la Región de Murcia. Por otra parte, el gran contenido de nitrato (contaminante) hace necesaria la búsqueda de la cantidad óptima del purín vertido. Comentaremos los resultados obtenidos en un proyecto de investigación de nuestra Universidad en el Valle del Guadalentín.

2. INTRODUCCIÓN A MATLAB

Matlab es un programa interactivo para cálculo numérico y tratamiento de datos. Contiene muchas herramientas y utilidades que permiten a demás diversas funcionalidades, como la presentación gráfica en dos y tres dimensiones. Estos útiles están agrupados en "paquetes" *(toolboxes)*. A Matlab se le pueden añadir paquetes especializados para algunas tareas (por ejemplo, para tratamiento de imágenes). Trabajar con Matlab comporta aprender un lenguaje simple. En esta introducción se explican los elementos básicos de este lenguaje.

Matlab es un programa *command-driven*, es decir, que se introducen las órdenes escribiéndolas una a una a continuación del símbolo (*prompt)* que aparece en una interfaz de usuario (una ventana). Esta introducción contiene ejemplos que se pueden escribir directamente en la línea de comandos de Matlab. Para distinguir esos comandos, junto con la respuesta del programa, se emplean un tipo de letra diferente:

```
>> 2+2
ans =
      4
```

Una manera de seguir esta introducción consiste en abrir Matlab e ir utilizando sus comandos. Este documento contiene los siguientes apartados:

- Gráficos
- Programación
- Introducción a la programación
- Cálculo simbólico

2.1 Comandos básicos

En esta sección se explica cómo usar Matlab a modo de calculadora. Vamos a empezar con un ejemplo sencillo: las operaciones matemáticas elementales.

```
>> X =2+3
X =
     5
```

Si no se asigna el resultado a ninguna variable, Matlab lo asigna por defecto a la variable *ans (answer)*:

```
>> 2+3
ans =
     5
```

Para saber cuál es el valor asignado a una sola variable, basta introducir el nombre de la variable:

```
>> x
x =
     5
```

La notación para las **operaciones matemáticas elementales** es la siguiente:

^	exponenciación
*	multiplicación
/	división
+	suma
-	resta

El orden en que se realizan las operaciones de una línea es el siguiente: primero, la exponenciación; luego, las multiplicaciones y divisiones; y finalmente, la suma y las restas. Si se quiere forzar un determinado orden, se deben de utilizar paréntesis, que se evalúan siempre al principio. Por ejemplo, para hallar dos entre tres,

```
>> 2/2+1
ans =
              2
```

(en efecto: primero se calcula 2/2 y luego se le suma 1).

```
>> 2/ (2+1)
ans =
         0.6667
```

Primero se calcula el paréntesis (2+1) y luego se realiza la división.

Dos observaciones. El punto decimal es "." (no una coma). Y en Matlab, las mayúsculas y las minúsculas son distintas. Es decir, X es una variable diferente de x.

En Matlab están también definidas algunas funciones elementales. Las funciones, en Matlab, se escriben introduciendo el argumento

entre paréntesis a continuación del nombre de la función, sin dejar espacios. Por ejemplo:

```
>> y=exp(0)
y =
        1
```

(la función exp es la exponencial). He aquí una tabla con algunas **funciones elementales:**

sin	seno
cos	coseno
tan	tangente
sec	secante
csc	cosecante
cot	cotangente
exp	exponencial
log	logaritmo natural
sqrt	raíz cuadrada
abs	valor absoluto

Para obtener las funciones trigonométricas inversas, basta añadir una a delante del nombre. Y para las funciones hiperbólicas, una h al final. Por ejemplo, atanh(x) es el arcotangente hiperbólica de x:

```
>> z=atanh(2)
z =
        0.5493 + 1.5708i
```

(z es un número complejo).

2.2 Ayuda en línea

Este documento es tan sólo una introducción -muy resumida- del lenguaje y del manejo de Matlab. Antes de seguir, es conveniente indicar cómo puede obtenerse más información sobre cualquier detalle referente a Matlab, desde dentro del mismo, se puede obtener cualquier explicación sobre un tema en particular.

El comando *help.* Para obtener información sobre una determinada función, basta teclear desde la línea de comandos *help* seguido del nombre de la función. Por ejemplo:

>> help round

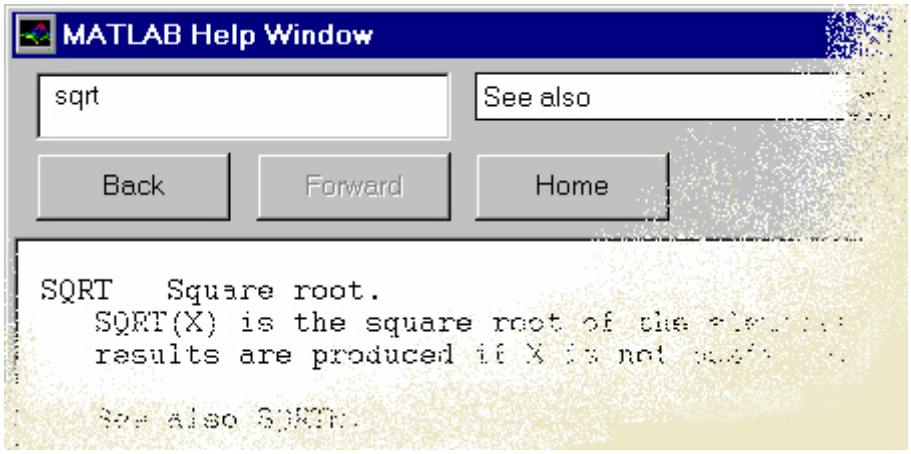
ROUND Round towards nearest integer.

ROUND(X) rounds the elements of X to the nearest integers.

See also FLOOR, CEIL, FIX.

Si se escribe sólo *help*, se obtiene un índice de temas. También puede obtenerse información sobre uno de los temas de esa lista: así, *help elfun* proporciona información sobre las funciones matemáticas elementales.

La ventana de ayuda. Puede llamarse tecleando *helpwin* o bien escogiendo del menú Help el ítem Help Window. Se obtiene una ventana nueva, y haciendo doble click con el ratón sobre un capítulo se pasa a un elenco de los ítems contenidos, que a su vez pueden escogerse para una explicación más detallada. Con los botones Back y Forward se navega hacia atrás o hacia adelante. También puede escribirse directamente en la zona superior izquierda el nombre del comando deseado: por ejemplo, para buscar información sobre *sqrt* ...



En la barra *See also* aparecen comandos relacionados. La información es la misma que la obtenida con el comando *help*, pero con la comodidad de presentarse en una ventana aparte en vez de en la línea de comandos.

La ayuda interactiva. Se obtiene escogiendo del menú Help el ítem Help Desk, o tecleando *helpdesk* en la barra de comandos. Se lanza el navegador y se obtiene un documento de inicio con un índice de temas en hipertexto donde están los manuales y otras utilidades, como un buscador. Para leer el manual, se necesita el programa Acrobat Reader.

La información que se obtiene es mucho más completa que en los otros dos casos, lo cual puede resultar inconveniente si uno desea simplemente, por poner un caso, conocer la sintaxis de una función.

2.3 El entorno Matlab

Edición de la línea de comandos. Con las flechas del teclado se pueden recuperar las órdenes anteriores, sin tener que volver a teclearlas. Así, en el caso de una equivocación en un comando complicado

```
>> d2_f=(y2-2*y1+y3)/deltax^2)

??? -2*y1+y3)/deltax^2)

Missing operator, comma, or semi-colon.
```

en vez de volver a teclear todo, puede recuperarse la instrucción pulsando la tecla "flecha hacia arriba", desplazarse hasta el error (falta un paréntesis) con la flecha hacia a la izquierda, y arreglarlo:

```
>> d2_f=(y2-2*y1+y3)/(deltax^2)
```

En ocasiones, es interesante **no presentar el resultado en la pantalla** (por ejemplo, cuando se trata de una lista de datos muy larga). Eso se consigue poniendo un punto y coma al final de la instrucción.

```
>> y=sqrt(4);
```

El resultado no aparece, pero sin embargo el cálculo se ha realizado:

```
>> y

y =

2
```

El comando *who* indica las **variables** con las que se está trabajando:

```
>> who
```

Your variables are:

Fy f indice n_punt t_m

delta_f f_max manchas t y

Comandos relacionados con el **sistema operativo**:

pwd	Present working directory (directorio de trabajo actual)
cd	cambiar de dirección
dir	listado de los ficheros del directorio actual

Estos comandos son muy similares a los análogos de MS-DOS o UNIX.

Guardar y cargar ficheros de datos. Se emplean los comandos *save y load*, respectivamente.

- para guardar datos: *save [nombre del fichero] [variable] -ascii*
- para recuperar datos: *load [nombre del fichero] [variable] -ascii*

Por ejemplo: con estas dos órdenes

>> cd a:

```
>> save toto.dat y -ascii
```

se cambia el directorio de trabajo a a:\ y se guarda allí el contenido de la variable y en el fichero *toto.dat* con formato texto (por eso se pone -*ascii*).

2.4 Vectores y matrices

Un vector se define introduciendo los componentes, separados por espacios o por comas, entre corchetes:

```
>> v=[sqrt(3) 0 -2]

v =

    1.7321   0   -2.0000
```

Para definir un vector columna, se separan las filas por puntos y comas:

```
>> w=[1;0;1/3]

w =

    1.0000

    0

    0.3333
```

La operación transponer (cambiar filas por columnas) se designa por el apóstrofe:

```
>> w'

ans =

    1.0000   0   0.3333
```

Las operaciones matemáticas elementales pueden aplicarse a los vectores:

```
>> v*w

ans =

    1.0654
```

```
>> v+w'

ans =

    2.7321   0   -1.6667
```

Para crear un vector de componentes equiespaciados se emplean los dos puntos:

```
>> x=4:2:10

x =

    4 6 8 10
```

(los componentes de x van desde 4 de 2 en 2 hasta 10).

Para introducir matrices, se separa cada fila con un punto y coma:

```
>> M = [1 2 3 ;4 5 6 ;7 8 9]

M =

    1 2 3

    4 5 6

    7 8 9
```

Para referirse a un elemento de la matriz se hace así:

```
>> M(3,1)

ans =

     7
```

Para referirse a toda una fila o a toda una columna se emplean los dos puntos:

```
>> v1=M(:,2)

v1 =

     2

     5

     8
```

(v1 es la segunda columna de M).

Con las matrices también funcionan las operaciones matemáticas elementales. Así

```
>> M^2

ans =

    30  36  42

    66  81  96

    102 126 150
```

Si se quiere operar en los elementos de la matriz, uno por uno, se pone un punto antes del operador. Si se quiere elevar al cuadrado *cada uno de los elementos de M*, entonces

```
>> M.^2
```

ans =

1 4 9

16 25 36

49 64 81

Algunas **funciones** definidas sobre matrices:

Det	determinante
Inv	matriz inversa
Poly	polinomio característico
´	transpuesta

(Para más información: *help elmat*)

2.5 Polinomios

Sea

$$P(x)=x^2-3x+2$$

Este polinomio se representa por un vector p

>> p=[1 -3 +2]

p = 1 -3 2

Para hallar las raíces del polinomio, se hace

>> roots(p)

ans =

2

1

y si se quiere hallar el valor de P(x) para un determinado valor de x (por ejemplo, para x=0)

>> polyval(p,0)

ans =

2

2.6 Gráficos

Las posibilidades de Matlab son muy grandes. Se indica a continuación cómo realizar gráficos sencillos. Para más información, o para conocer la versatilidad de Matlab: capítulo *Handle Graphics Object* del Help Desk, el manual *Using MATLAB Graphics* o la ayuda en línea help graph2d.

Veamos cómo se puede representar la función seno entre 0 y 10. Para empezar creemos una variable x que vaya de cero a 10:

>> x=0:0.1:10;

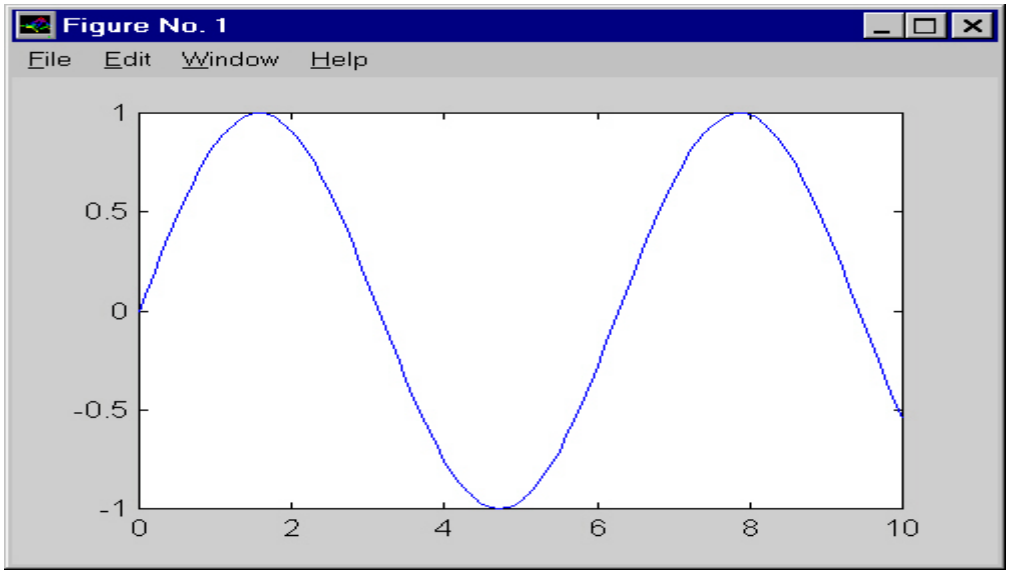
y a continuación, calculemos sin(x) almacenando el resultado en la variable y:

>> y=sin(x);

Para trazar el gráfico, se emplea la función *plot*:

>> plot(x,y)

y se obtiene en otra ventana el gráfico:



Entre los muchos comandos que se pueden utilizar para modificar los gráficos, es muy útil el empleado para cambiar la escala de los ejes. La orden es:

axis([x1 x2 y1 y2])

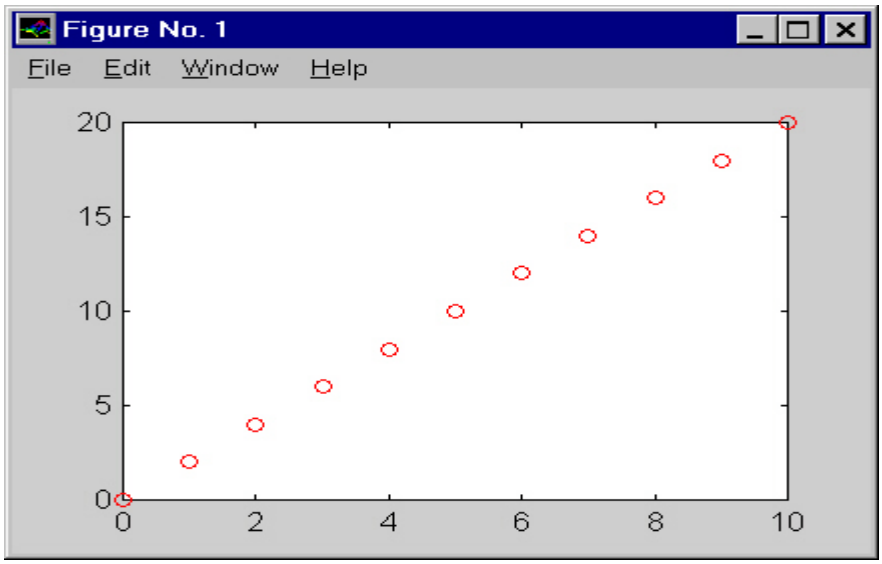
donde x1, x2 son los límites inferior y superior del eje x, e y1 e y2 los del eje y.

Para representar unos datos con símbolos de colores, se añade al comando *plot*, entre apóstrofes, la especificación. Vamos a crear una variable con dos filas que contenga los números del 1 al 10 en la primera fila, y el doble de esos números en la segunda, y dibujarlos con puntos rojos:

```
>> x(1,:)=0:10;

>> x(2,:)=2*x(1,:);

>> x

x =

     0   1   2   3   4   5   6   7   8   9   10
```

0 2 4 6 8 10 12 14 16 18 20

>> plot(x(1,:),x(2,:),'ro')



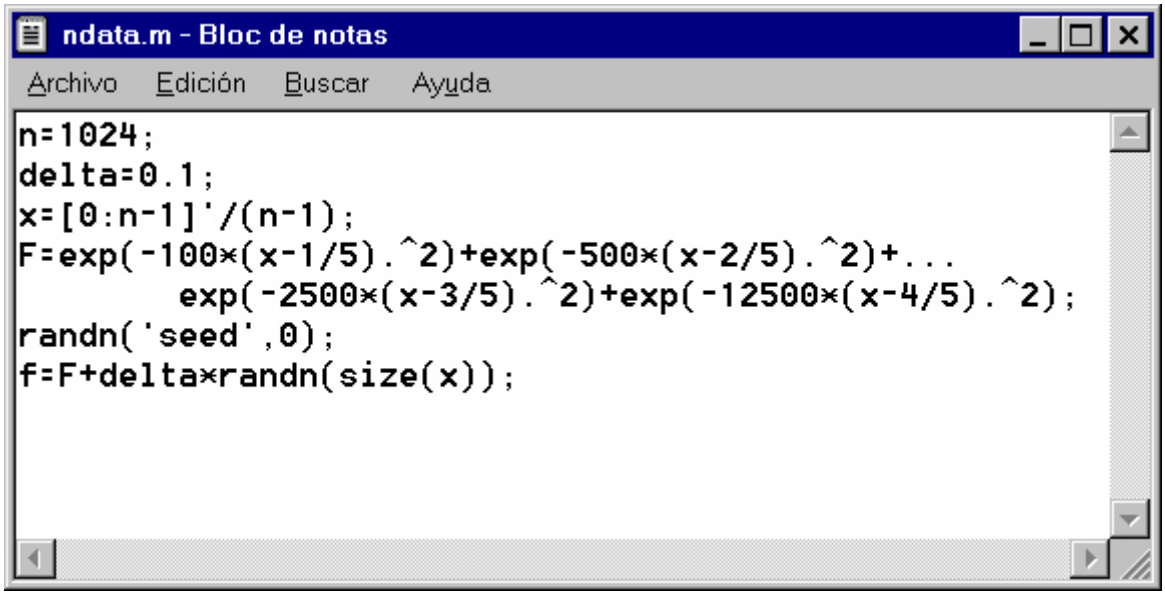
(Para ver las especificaciones posibles, teclear *help plot*. Por ejemplo, *'ro'* establece un gráfico de color rojo: r y de puntos: o.) Si no se indica nada, el gráfico se traza con una línea azul.

Otras funciones muy útiles: *grid*, que traza una cuadrícula, *xlabel('títulox')* e *ylabel('tituloy')*, que sirven para poner un título en los ejes.

Para imprimir una figura, basta seleccionar *print del menú de la figura* y para guardarla *print*ps.*

2.7 Programas

Realizar un programa en Matlab es fácil. Basta abrir un editor de texto (como el Bloc de Notas de Windows) y escribir los comandos uno a continuación de otro. Luego ese fichero de texto debe guardarse con la extensión .m, y a eso se le llama un *programa*:



Una vez guardado el fichero (en el ejemplo, ndata.m) en el directorio actual, desde la línea de comandos de Matlab basta escribir ndata para que se ejecute el programa.

2.8 Introducción a la programación en MATLAB

Las condiciones se construyen con **operadores relacionales**, como son los siguientes:

>	mayor que
<	menor que
==	igual que
~=	Diferente que
<=	menor o igual que
>=	mayor o igual que

La ramificación más simple, expresada en este diagrama de flujo, se obtiene con la siguente sintaxis:

if (*condición*)

sentencias

```
end
```

(lo que va en cursiva, hay que sustituirlo por las expresiones adecuadas; `if` y `end` son *palabras clave* del lenguaje informático, y no se pueden utilizar para otra cosa, p. ej. una variable no puede -no debería- llamarse `if`).

Un caso concreto:

```
if(length(sitios)>1)
recta=polyfit(x,y,1);
end
```

Leído en lenguaje corriente: si la longitud del vector `sitios` es mayor que 1, se realiza el ajuste lineal indicado en la instrucción `recta=polyfit(x,y,1)`. Caso contrario (si la longitud del vector `sitios` es menor o igual a 1) esa instrucción no se ejecuta (y el programa sigue en la instrucción que venga después de `end`).

Existe la posibilidad de ejecutar ciertas sentencias si la condición es verdadera, y otras diferentes si la condición es falsa:

```
if (condición)
sentencias A
else
sentencias B
end
```

dicho de otra manera: si la condición se cumple, se ejecutan las *sentencias A*; si no, se ejecutan las *sentencias B*.

Otra posibilidad de ramificación múltiple la ofrece la construcción **switch**. La sintaxis es:

switch *variable*

case *valor1*,

sentencias A

case *valor2*,

sentencias B

case ...

...

end

(Como antes, lo escrito en cursiva debe sustituirse por las expresiones adecuadas). Las palabras clave son switch, case, end.

La ramificación switch opera de la siguiente manera. Al llegar a la expresión switch *variable*, si *variable* tiene el valor *valor1* se ejecutan las *sentencias A*; si *variable* toma el valor *valor2*, las *sentencias B*; y así sucesivamente. Es importante notar que la variable sólo debe tomar unos pocos valores: *valor1*, *valor2*, etc. para que el programa se ramifique en unas pocas ramas. No tiene sentido intentar una ramificación switch con una variable que pueda tomar un número infinito de valores.

Hay ocasiones en las que es necesario repetir el mismo conjunto de instrucciones muchas veces, cambiando algunos detalles. Pongamos un caso. Sea un vector x(i) con n componentes; se quiere construir la "media móvil" de x con tres elementos, que consiste en ir tomando la media aritmética de cada tres puntos consecutivos. Es decir: desde i=2 hasta n-1, media(i-1)=(x(i)+x(i-1)+x(i+1))/3.

(Detalles: se empieza a contar en i=2 porque para el primer elemento de x no existe el elemento anterior; y se acaba en n-1 por análoga

razón; además, el primer componente de media es el correspondiente a i=2, de ahí que se asigne el resultado a media(i-1)).

Eso es lo que se consigue con un bucle **for**, cuya sintaxis es:

for *contador*=*inicio*:*paso*:*fin*,

sentencias

end

Las palabras claves son for y end. Este bucle pone en marcha una variable llamada *contador* que va desde *inicio* hasta *fin* de *paso* en *paso*. Cada vez que las *sentencias* se ejecutan, *contador* aumenta en un valor *paso* (que si se omite, se le asigna automáticamente el valor 1). Cuando *contador* llega al valor *fin*, el bucle se acaba y el programa continúa con las sentencias que haya más allá de end.

Obsérvese que un bucle como el indicado se implementa un número fijo de veces: desde inicio hasta fin de paso en paso. En ocasiones, sin embargo, no se sabe de antemano cuántas veces habrá que ejecutar las sentencias del bucle. Por ejemplo: si es necesario repetir una serie de sentencias hasta que se cumpla una determinada condición, y no se sabe a priori cuántas veces será necesario realizar esas operaciones. En ese caso se emplea un bucle **while**:

while(*condición*),

sentencias

end

Este bucle ejecuta las *sentencias* mientras la *condición* sea verdadera.

Es posible sustituir la condición por una variable. En efecto: una variable que toma el valor cero corresponde a una condición falsa. Si

la variable toma un valor diferente de cero, es equivalente a una condición verdadera. Así, se puede escribir

```
x=10;
while(x)
sentencias
x=x-1;
end
```

Para `x=10`, la "condición" es verdadera puesto que `x` es diferente de cero. Nótese que el contador `x` hay que modificarlo manualmente (línea `x=x-1`) puesto que, al revés que lo que ocurre con el bucle for, este no gestiona ningún contador. En cuanto `x` tome el valor cero, la "condición" es falsa y el bucle acaba.

Atención: es fácil caer en bucles infinitos. En el ejemplo anterior, si falta la línea `x=x-1` y las *sentencias* no modifican el valor de `x`, la "condición" siempre será cierta (pues `x=10`) y el programa nunca saldrá del bucle: ejecutará una y otra vez las *sentencias*. El programa se "cuelga", y hay que interrumpirlo desde el teclado apretando las teclas Ctrl+C.

Los comentarios son líneas que no se ejecutan, en las que se escriben aclaraciones explicativas. Para que una línea no se ejecute, basta escribir al principio de ella el símbolo %.

Suele ser bueno **definir las variables al principio**. Ello evita tener que buscarlas a lo largo del código para cambiar su valor cuando sea necesario. Además, si es posible, es mejor definir los vectores y matrices al principio con su dimensión adecuada. En el caso de que haya que ir rellenando los valores de un vector, el programa va más rápido si se define el vector vacío al principio (con el comando `ones` o

zeros) que ir añadiendo componentes al vector conforme se van calculando.

Por último, también pueden programarse funciones. La primera instrucción de un fichero que contenga una función de nombre fun debe ser:

function [argumentos de salida]=fun(argumentos de entrada)

Es conveniente que el fichero que contenga la función se llame como ella; así, la función anterior debería guardarse en el fichero fun.m; por ejemplo, si se desea programar una función que calcule, mediante el algoritmo de Euclides, el máximo común divisor de dos números naturales, basta escribir un fichero euclides.m cuyo contenido sea:

```
function m=euclides(a,b)
% Cálculo del máximo común divisor de dos números naturales
% mediante el algoritmo de Euclides
if a<b
c=b;
b=a;
a=c;
end
while b>0
c=rem(a,b);
a=b;
b=c;
end
m=a;
```

Si, una vez escrito el fichero anterior, en el espacio de trabajo o en un programa se escribe la instrucción

```
mcd=euclides(33,121)
```

en la variable mcd se almacenará el valor 11.

Las variables de una función son siempre locales. Por tanto, aunque en el seno de la función se modifiquen los argumentos de entrada, el valor de las variables correspondientes queda inalterado. Por ejemplo, en la función euclides.m se modifica el valor de los argumentos de entrada, pero, sin embargo:

```
>>x=15;
>>mcd=euclides(x,3);
>>x
x =
15
```

Si se pretende que las modificaciones de un argumento de entrada afecten a la variable correspondiente, deberá situarse dicho argumento, además, en la lista de argumentos de salida.

2.9 Cálculo simbólico

Hasta ahora, las operaciones que se han mostrado se han realizado con números. El *toolbox* de cálculo simbólico permite realizar **cálculos abstractos**:

```
>> diff('sin(x)')

ans =

    cos(x)
```

Las expresiones simbólicas se introducen entre apóstrofes.

A continuación se da una tabla con algunas funciones de este toolbox, junto con un ejemplo de cada una:

diff	derivada	diff('sin(x)')
int	integral	int('x^2')
solve	resolución de ecuaciones	solve('x^2-3*x+2=0')
ezplot	gráficos	ezplot('exp(x)')

Evidentemente, las expresiones pueden ser todo lo complicadas que se quiera

```
>> solve('x=cos(x)')
```

```
ans =

    .73908513321516064165531208767387

>>
int('(x^4+4*x^3+11*x^2+12*x+8)/((x^2+2*x+3)^2*(x+1))'
)

ans =

    log(x+1)+1/8*(-4*x-8)/(x^2+2*x+3)-
    1/4*2^(1/2)*atan(1/4*(2*x+2)*2^(1/2))
```

3. ANÁLISIS DE LA VARIANZA

El análisis de la varianza (o Anova: **An**alysis **o**f **va**riance) es un método para comparar dos o más medias. Cuando se quiere comparar más de dos medias es incorrecto utilizar repetidamente el contraste basado en la *t de Student* por dos motivos:

En primer lugar, y como se realizarían simultánea e independientemente varios contrastes de hipótesis, la probabilidad de encontrar alguno significativo por azar aumentaría. En cada contraste se rechaza la H_0 si la *t* supera el nivel crítico, para lo que, en la hipótesis nula, hay una probabilidad α. Si se realizan *m* contrastes independientes, la probabilidad de que, en la hipótesis nula, ningún estadístico supere el valor crítico es $(1 - \alpha)^m$, por lo tanto, la probabilidad de que alguno lo supere es $1 - (1 - \alpha)^m$, que para valores de α próximos a 0 es aproximadamente igual a α m. Una primera solución, denominada *método de Bonferroni*, consiste en bajar el valor de α, usando en su lugar α/m, aunque resulta un método muy conservador.

Por otro lado, en cada comparación la hipótesis nula es que las dos muestras provienen de la misma población, por lo tanto, cuando se hayan realizado todas las comparaciones, la hipótesis nula es que todas las muestras provienen de la misma población y, sin embargo, para cada comparación, la estimación de la varianza necesaria para el contraste es distinta, pues se ha hecho en base a muestras distintas.

El método que resuelve ambos problemas es el *anova*, aunque es algo más que esto: es un método que permite comparar varias medias en diversas situaciones; muy ligado, por tanto, al diseño de experimentos y, de alguna manera, es la base del análisis multivariante.

3.1 Bases del análisis de la varianza

Supónganse *k* muestras aleatorias independientes, de tamaño *n*, extraídas de una única población normal. A partir de ellas existen dos maneras independientes de estimar la varianza de la población σ^2:

1) Una llamada *varianza dentro de los grupos* (ya que sólo contribuye a ella la varianza dentro de las muestras), o *varianza de error*, o *cuadrados medios del error*, y habitualmente representada por *MSE* (*M*ean *S*quare *E*rror) o *MSW* (*M*ean *S*quare *W*ithin) que se calcula como la media de las *k* varianzas muestrales (cada varianza muestral es un estimador centrado de σ^2 y la media de *k* estimadores centrados es también un estimador centrado y más eficiente que todos ellos). *MSE* es un cociente: al numerador se le llama *suma de cuadrados del error* y se representa por *SSE* y al denominador *grados de libertad* por ser los términos independientes de la suma de cuadrados.

2) Otra llamada *varianza entre grupos* (sólo contribuye a ella la varianza entre las distintas muestras), o *varianza de los tratamientos*, o *cuadrados medios de los tratamientos* y representada por *MSA* o *MSB* (*M*ean *S*quare *B*etween). Se calcula a partir de la varianza de las medias muestrales y es también un cociente; al numerador se le llama *suma de cuadrados de los tratamientos* (se le representa por *SSA)* y al denominador (*k*-1) grados de libertad.

MSA y *MSE*, estiman la varianza poblacional en la hipótesis de que las *k* muestras provengan de la misma población. La distribución muestral del cociente de dos estimaciones independientes de la varianza de una población **normal** es una *F* con los grados de libertad correspondientes al numerador y denominador respectivamente, por lo tanto se puede contrastar dicha hipótesis usando esa distribución.

Si en base a este contraste se rechaza la hipótesis de que *MSE* y *MSA* estimen la misma varianza, se puede rechazar la hipótesis de que las *k* medias provengan de una misma población.

Aceptando que las muestras provengan de poblaciones con la misma varianza, este rechazo implica que las medias poblacionales son distintas, de modo que con un único contraste se contrasta la igualdad de *k* medias.

Existe una tercera manera de estimar la varianza de la población, aunque no es independiente de las anteriores. Si se consideran las *kn* observaciones como una única muestra, su varianza muestral también es un estimador centrado de σ^2:

Se suele representar por *MST*, se le denomina *varianza total* o *cuadrados medios totales*, es también un cociente y al numerador se

le llama *suma de cuadrados total* y se representa por *SST*, y el denominador (*kn* -1) grados de libertad.

Los resultados de un anova se suelen representar en una tabla como la siguiente:

Fuente de variación	G.L.	*SS*	*MS*	*F*
Entre grupos Tratamientos	*k*-1	*SSA*	*SSA* /(*k*-1)	*MSA* /*MSE*
Dentro Error	(*n*-1)*k*	*SSE*	*SSE* /*k*(*n*-1)	
Total	*kn*-1	*SST*		

F se usa para realizar el contraste de la hipótesis de medias iguales. La *región crítica* para dicho contraste es $F > F_{\alpha(k-1,(n-1)k)}$

3.2 Algunas propiedades

Es fácil ver en la tabla anterior que

$$GL_{error} + GL_{trata} = (n - 1)\,k + k - 1 = k + k - 1 = nk - 1 = GL_{total}$$

No es tan inmediato, pero las sumas de cuadrados cumplen la misma propiedad, llamada *identidad* o *propiedad aditiva* de la suma de cuadrados:

SST = SSA + SSE

El análisis de la varianza se puede realizar con tamaños muestrales iguales o distintos, sin embargo es recomendable iguales tamaños por dos motivos:

1) La *F* es insensible a pequeñas variaciones en la suposición de igual varianza, si el tamaño es igual.

2) Igual tamaño minimiza la probabilidad de error tipo II.

3.3 Pruebas para la homocedasticidad

Para que este contraste de hipótesis, basado en la *F*, lo sea de la igualdad de medias es necesario que todas las muestras provengan de una población con la misma varianza σ^2, de la que *MSE* y *MSA* son estimadores. Por lo tanto es necesario comprobarlo antes de realizar el contraste. Del mismo modo que no se puede usar repetidamente la prueba basada en la *t* para comparar más de dos medias, tampoco se puede usar la prueba basada en la *F* para comparar más de dos varianzas. La prueba más usada para contrastar si varias muestras son homocedásticas (tiene la misma varianza) es la prueba de Bartlett.

La prueba se basa en que, en **la hipótesis nula de igualdad de varianzas** y poblaciones normales, un estadístico calculado a partir de las varianzas muestrales y *MSE* sigue una distribución χ^2_{k-1} .

Otras pruebas para contrastar la homocedasticidad de varias muestras son la de Cochran y la de la *F* del cociente máximo, ambas similares y de cálculo más sencillo pero restringidas al caso de iguales tamaños muestrales. La de Cochran es particularmente útil para detectar si una varianza es mucho mayor que las otras.

En el caso de que las muestras no sean homocedásticas, no se puede, en principio, realizar el análisis de la varianza.

Existen, sin embargo, soluciones alternativas: Sokal y Rohlf describen una prueba aproximada, basada en unas modificaciones de las fórmulas originales.

Hay situaciones en que la heterocedasticidad es debida a falta de normalidad. En estos casos existen transformaciones de los datos que estabilizan la varianza: la raíz cuadrada en el caso de Poisson, el arco seno de la raíz cuadrada de *p* para la binomial, el logaritmo cuando la desviación estándar es proporcional a la media.

En la práctica, si las pruebas de homocedasticidad obligan a rechazar la hipótesis nula, se prueba si con alguna de estas transformaciones los datos son homocedásticos, en cuyo caso se realiza el anova con los datos transformados.

Hay que tener en cuenta que estas pruebas van "*al revés*" de lo habitual. La hipótesis nula es lo que se quiere probar, en consecuencia hay que usarlas con precaución.

3.4 Modelos de análisis de la varianza

El anova permite distinguir dos modelos para la hipótesis alternativa:

modelo I o de *efectos fijos* en el que la H_1 supone que las *k* muestras son muestras de *k* poblaciones distintas y fijas.

modelo II o de *efectos aleatorios* en el que se supone que las *k* muestras, se han seleccionado aleatoriamente de un conjunto de $m>k$ poblaciones.

La manera más sencilla de distinguir entre ambos modelos es pensar que, si se repitiera el estudio un tiempo después, en un modelo I las muestras serían iguales (no los individuos que las forman) es decir corresponderían a la misma situación, mientras que en un modelo II las muestras serían distintas.

Aunque las suposiciones iniciales y los propósitos de ambos modelos son diferentes, los cálculos y las pruebas de significación son los mismos y sólo difieren en la interpretación y en algunas pruebas de hipótesis suplementarias.

3.4.1 Modelo I o de efectos fijos

Un valor individual se puede escribir en este modelo como

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \ldots, k \quad y \quad j = 1, \ldots, n$$

μ es la media global, α_i es la constante del efecto, o efecto fijo, que diferencia a las *k* poblaciones. También se puede escribir:

$$\mu_i = \mu + \alpha_i \quad \varepsilon_{ij}$$

representa la desviación de la observación *j*-ésima de la muestra *i*-ésima, con respecto a su media. A este término se le suele llamar *error aleatorio* y, teniendo en cuenta las suposiciones iniciales del análisis de la varianza son *k* variables (una para cada muestra), todas con una distribución normal de media 0 y varianza σ^2.

La hipótesis nula en este análisis es que todas las medias son iguales

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$
H_1: al menos una es diferente

que puede escribirse en términos del modelo como:

$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
H_1: al menos dos α_i distintas de 0

Como en H_0 se cumplen las condiciones del apartado anterior se tratará de ver como se modifican las estimaciones de la varianza en H_1.

En H_0 *MSA* y *MSE* son estimadores centrados de σ^2, es decir y usando el superíndice 0 para indicar el valor de las variables en H_0

$E[MSA^0] = \sigma^2$

$E[MSE^0] = \sigma^2$

Se puede ver que *MSE* es igual en la hipótesis nula que en la alternativa. Por lo tanto:

$E[MSE] = E[MSE^0] = \sigma^2$

Sin embargo al valor esperado de *MSA* en la hipótesis alternativa se le añade un término con respecto a su valor en la hipótesis nula

$$E[MSA] = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k} \alpha_i^2$$

Al segundo sumando dividido por *n* se le llama *componente de la varianza añadida por el tratamiento*, ya que tiene forma de varianza, aunque estrictamente no lo sea pues α_i no es una variable aleatoria.

La situación, por lo tanto, es la siguiente: en H_0, *MSA* y *MSE* estiman σ^2; en H_1, *MSE* estima σ^2 pero *MSA* estima $\sigma^2 + n\sigma_a^2$. Contrastar la H_0 es equivalente a contrastar la existencia de la componente añadida o, lo que es lo mismo, que *MSE* y *MSA* estimen, o no, la misma varianza.

El estadístico de contraste es *F=MSA/MSE* que, en la hipótesis nula, se distribuye según una *F* con *k* - 1 y (*n* - 1)*k* grados de libertad. En caso de rechazar la H_0, *MSA* - *MSE* estima $n\sigma_a^2$.

3.4.2 Modelo II o de efectos aleatorios

En este modelo se asume que las *k* muestras son muestras aleatorias de *k* situaciones distintas y aleatorias. De modo que un valor aislado Y_{ij} se puede escribir como:

$$Y_{ij} = \mu + A_i + \varepsilon_{ij} \quad i = 1, \ldots k \quad \text{y} \quad j = 1, \ldots n$$

donde μ es la media global, ε_{ij} son variables (una para cada muestra) distribuidas normalmente, con media 0 y varianza σ^2 (como en el modelo I) y A_i es una variable distribuida normalmente, independiente de las ε_{ij}, con media 0 y varianza σ_a^2.

La diferencia con respecto al modelo I es que en lugar de los efectos fijos α_i ahora se consideran efectos aleatorios A_i.

Igual que en el modelo I se encuentra que *MSE* no se modifica en la H_1 y que al valor esperado de *MSA* se le añade el término de *componente añadida* (que aquí es una verdadera varianza ya que A_i es una variable aleatoria):

$$\frac{n}{k-1} \sum_{i=1}^{k} A_i^2 = n\sigma_a^2$$

Para llegar a este resultado se utiliza la asunción de independencia entre A_i y ε_{ij} y es, por tanto, muy importante en el modelo y conviene verificar si es correcta en cada caso.

Por tanto, en H_0 tanto *MSA* como *MSE* estiman σ^2, mientras que en H_1, *MSE* sigue estimando σ^2 y *MSA* estima $\sigma^2 + n\sigma_a^2$. La existencia de esta componente añadida se contrasta con *F= MSA*/*MSE* y en caso afirmativo, la varianza de A_i se estima como:

$$\sigma_a^2 = \frac{1}{n}(MSA - MSE)$$

3.5 Pruebas "*a posteriori*"

En general, en un modelo II el interés del investigador es averiguar si existe componente añadida y en su caso estimarla.

Sin embargo, en un modelo I, lo que tiene interés son las diferencias entre los distintos grupos.

Las pruebas "a posteriori" son un conjunto de pruebas para probar todas las posibles hipótesis del tipo $\mu_i - \mu_j = 0$.

Existen varias, (Duncan, Newman-Keuls, LSD): todas ellas muy parecidas. Usan el rango (diferencia entre medias) de todos los pares de muestras como estadístico y dicho rango debe superar un cierto valor llamado *mínimo rango significativo* para considerar la diferencia significativa.

La principal diferencia con respecto a la *t de Student* radica en que usan *MSE* como estimador de la varianza, es decir un estimador basado en todas las muestras.

Una manera semigráfica habitual de representar los resultados es dibujar una línea que una cada subconjunto de medias adyacentes entre las que no haya diferencias significativas.

3.6 Análisis de la varianza de dos factores

Es un diseño de *anova* que permite estudiar simultáneamente los efectos de dos fuentes de variación.

Una observación individual se representa como:

$Y_{ijk} \quad i = 1, \ldots, a \quad j = 1, \ldots, b \quad k = 1, \ldots, n$

El primer subíndice indica el nivel del primer factor, el segundo el nivel del segundo factor y el tercero la observación dentro de la muestra. Los factores pueden ser ambos de efectos fijos (se habla entonces de *modelo I*), de efectos aleatorios *(modelo II)* o uno de efectos fijos y el otro de efectos aleatorios *(modelo mixto)*. El modelo matemático de este análisis es:

$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ modelo I

$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk}$ modelo II

$Y_{ijk} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \varepsilon_{ijk}$ modelo mixto

donde μ es la media global, α_i o A_i el efecto del nivel *i* del 1º factor, β_j o B_j el efecto del nivel *j* del 2º factor y ε_{ijk} las desviaciones aleatorias alrededor de las medias, que también se asume que están normalmente distribuidas, son independientes y tienen media 0 y varianza σ^2.

A las condiciones de muestreo aleatorio, normalidad e independencia, este modelo añade la de aditividad de los efectos de los factores.

A los términos $(\alpha\beta)_{ij}$, $(AB)_{ij}$, $(\alpha B)_{ij}$, se les denomina *interacción* entre ambos factores y representan el hecho de que el efecto de un determinado nivel de un factor sea diferente para cada nivel del otro factor.

3.7 Identidad de la suma de cuadrados

La suma de cuadrados total en un *anova de 2 vías*, es:

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (Y_{ijk} - \bar{Y}_{...})^2$$

(donde para representar las medias se ha usado la convención habitual de poner un punto (.) en el lugar del subíndice con respecto al que se ha sumado) que dividida por sus grados de libertad, *abn* - 1, estima la varianza σ^2 en el supuesto de que las *ab* muestras provengan de una única población.

Se puede demostrar que

$$SST = SSA + SSB + SSAB + SSE$$

que es la llamada *identidad de la suma de cuadrados* en un anova de dos factores. Los sucesivos sumandos reciben respectivamente el nombre de suma de cuadrados del 1º factor (tiene *a* -1 grados de libertad y recoge la variabilidad de los datos debida exclusivamente al

1º factor), del 2º factor (con *b* -1 grados de libertad y recoge la variabilidad de los datos debida exclusivamente al 2º factor), de la interacción (con (*a* - 1)(*b* - 1) grados de libertad, recoge la variabilidad debida a la interacción) y del error (con *ab*(*n* - 1) grados de libertad, recoge la variabilidad de los datos alrededor de las medias de cada muestra).

Los resultados de un análisis de la varianza de dos factores se suelen representar en una tabla como la siguiente:

Fuente de variación	**GL**	**SS**	**MS**
1º factor	a - 1	SSA	SSA/(a - 1)
2º factor	b - 1	SSB	SSB/(b - 1)
Interacción	(a - 1)(b - 1)	SSAB	SSAB/[(a - 1)(b - 1)]
Error	ab(n - 1)	SSE	SSE/[ab(n - 1)]
Total	abn - 1	SST	

Los grados de libertad también son aditivos.

En ocasiones se añade una primera línea llamada *de tratamiento* o *de subgrupos* cuyos grados de libertad y suma de cuadrados son las sumas de los del primer, segundo factor y la interacción, que corresponderían a la suma de cuadrados y grados de libertad del tratamiento de un análisis de una vía en que las *ab* muestras se considerarán como muestras de una clasificación única.

Para plantear los contrastes de hipótesis hay que calcular los valores esperados de los distintos cuadrados medios.

3.8 Contrastes de hipótesis en un análisis de la varianza de dos factores

Del mismo modo que se hizo en el anova de una vía, para plantear los contrastes de hipótesis habrá que calcular los valores esperados de los distintos cuadrados medios. Los resultados son:

3.8.1 Modelo I

MS	**Valor esperado**
MSA	$\sigma^2 + \frac{n\,b}{a-1}\sum_{i=1}^{a}\alpha_i^2$
MSB	$\sigma^2 + \frac{n\,a}{b-1}\sum_{j=1}^{b}\beta_j^2$
MSAB	$\sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha_i\beta_j)^2$
MSE	σ^2

Por lo tanto, los estadísticos *MSAB*/*MSE*, *MSA*/*MSE* y *MSB*/*MSE* se distribuyen como una *F* con los grados de libertad correspondientes y permiten contrastar, respectivamente, las hipótesis:

i) no existe interacción (*MSAB*/*MSE*)

$H_0: (\alpha\beta)_{ij} = 0 \quad i = 1,\ldots,\ a \quad j = 1,\ldots,\ b$

ii) no existe efecto del primer factor, es decir, diferencias entre niveles del primer factor (*MSA*/*MSE*)

$H_0: \mu_{1.} = \ldots = \mu_{a.}$

iii) no existe efecto del segundo factor (*MSB*/*MSE*)

$H_0: \mu_{.1} = \ldots = \mu_{.b}$

Si se rechaza la primera hipótesis de no interacción, no tiene sentido contrastar las siguientes. En este caso lo que está indicado es realizar un análisis de una vía entre las *ab* combinaciones de tratamientos para encontrar la mejor combinación de los mismos.

3.8.2 Modelo II

	MS		**Valor esperado**
	MSA		$\sigma^2 + n\,\sigma^2_{ab} + nb\,\sigma^2_a$
	MSB		$\sigma^2 + n\,\sigma^2_{ab} + na\,\sigma^2_b$
	MSAB		$\sigma^2 + n\,\sigma^2_{ab}$
	MSE		σ^2

donde σ^2_a, σ^2_b y σ^2_{ab} son, respectivamente las componentes añadidas por el primer factor, por el segundo y por la interacción, que tienen la misma forma que los del modelo I, sin más que cambiar α_i y β_j por A_i y B_j, respectivamente.

La interacción se contrasta, como en el modelo I, con *MSAB*/*MSE*, si se rechaza la hipótesis nula se contrastarían cada uno de los factores con *MSA*/*MSAB* y *MSB*/*MSAB*.

En un modelo II, como no se está interesado en estimar los efectos de los factores sino sólo la existencia de la componente añadida, **sí** tiene sentido contrastar la existencia de la misma para cada factor incluso aunque exista interacción.

Aquí el problema se plantea cuando no se puede rechazar la hipótesis nula y se concluye que no existe interacción: entonces tanto *MSE* como *MSAB* estiman σ^2, entonces ¿cuál se elige para contrastar la componente añadida de los factores?

En principio, parece razonable escoger su media (la media de varios estimadores centrados es también un estimador centrado y más eficiente), sin embargo si se elige *MSAB* se independiza el contraste para los factores de un posible error tipo II en el contraste para la interacción. Hay autores que por ello opinan que es mejor usar *MSAB,* pero otros proponen promediar si se puede asegurar baja la probabilidad para el error tipo II. La media de los cuadrados medios se calcula dividiendo la suma de las sumas de cuadrados por la suma de los grados de libertad.

Ejemplo

A partir de la siguiente tabla de un *anova* de 2 factores modelo II, realizar los contrastes adecuados.

Fuente de variación	**G.L.**	**SS**	**MS**
1º factor	4	315,8	78,95
2º factor	3	823,5	274,5
Interacción	12	328,9	27,41
Error	100	2308,0	23,08
Total	119	3776,2	

Se empezaría contrastando la existencia de interacción: f = 27,41/23,08 = 1,188 como $F_{0,05(12,100)}$ = 1,849 no se puede, al nivel de significación del 95%, rechazar la hipótesis nula y se concluye que no existe interacción.

Si usamos *MSAB* para contrastar los factores:

1º factor: f = 78,95/27,41 = 2,880 como $F_{0,05(4,12)}$ = 3,26 no se rechaza la hipótesis nula y se concluye la no existencia de componente añadida por este factor.

2º factor: f = 274,5/27,41 = 10,015 como $F_{0,05(3,12)}$ = 3,49 se rechaza la hipótesis nula y se acepta la existencia de componente añadida por este factor.

El resultado del análisis es: no existe componente añadida por la interacción, tampoco por el 1º factor y sí existe componente añadida por el 2º.

La estimación de esta componente es: como a partir de los grados de libertad de la tabla podemos calcular a = 5, b = 4 y n = 6 resulta que la estimación de $n\,a\,\sigma_b^2$ es 274,5 - 27,41 = 247,09; por lo tanto $\sigma_b^2 = 247{,}09 / 30 = 8{,}24$ que representa un 35,7% de componente añadida por el segundo factor.

Si se hubiera optado por promediar, los cuadrados medios promediados son (328,9+2308,0)/(12+100)=23,54 con 112 grados de libertad y hubiera resultado significativo también el 1º factor.

La salida de un paquete estadístico, p.e. el Statgraphics, para un anova de 2 factores modelo II

Analysis of Variance for Hum.Relativa - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:altura	2381,75	3	793,917	6,14(1)	0,0851
B:bosque	2145,13	1	2145,13	16,60(1)	0,0267
INTERACTIONS					
AB	387,625	3	129,208	7,30(0)	0,0012
RESIDUAL	425,0	24	17,7083		
TOTAL (CORRECTED)	5339,5	31			

F-ratios are based on the following mean squares:
(0) Residual
(1) AB

3.8.3 Modelo mixto

Supóngase el primer factor de efectos fijos y el segundo de efectos aleatorios, lo que no supone ninguna perdida de generalidad, ya que el orden de los factores es arbitrario.

MS	Valor esperado
MSA	$\sigma^2 + n\,\sigma^2_{ab} + \frac{n\,b}{a-1}\sum_{i=1}^{a}\alpha i^2$
MSB	$\sigma^2 + n\,a\,\sigma^2_b$
MSAB	$\sigma^2 + n\sigma^2_{ab}$
MSE	σ^2

Se contrastan la interacción y el factor aleatorio con el término de error, si la interacción fuera significativa no tiene sentido contrastar el efecto fijo y si no lo fuera, el efecto fijo se contrasta con el término de interacción o con el promedio de interacción y error.

3.9 Tamaños muestrales desiguales en un anova de dos factores

Aunque los paquetes estadísticos suelen hacer el anova de dos factores, tanto en el caso de tamaños muestrales iguales como desiguales, conviene resaltar que el análisis es bastante más

complicado en el caso de tamaños desiguales. La complicación se debe a que con tamaños desiguales hay que ponderar las sumas de cuadrados de los factores con los tamaños muestrales y no resultan ortogonales (su suma no es la suma de cuadrados total) lo que complica no sólo los cálculos sino también los contrastes de hipótesis.

Por esto, cuando se diseña un análisis factorial de la varianza se recomienda diseñarlo con tamaños iguales. Hay ocasiones en que, sin embargo, por la dificultad de obtener los datos o por pérdida de alguno de ellos es inevitable recurrir al análisis con tamaños desiguales. Algunos autores recomiendan, incluso, renunciar a alguno de los datos para conseguir que todas las muestras tengan el mismo tamaño. Evidentemente esta solución es delicada pues podría afectar a la aleatoriedad de las muestras.

3.10 Casos particulares: Anova de dos factores sin repetición

En ciertos estudios en que los datos son difíciles de obtener o presentan muy poca variabilidad dentro de cada subgrupo es posible plantearse un anova sin repetición, es decir, en el que en cada muestra sólo hay una observación (n=1). Hay que tener en cuenta que, como era de esperar con este diseño, no se puede calcular *SSE*. El término de interacción recibe el nombre de *residuo* y que, como no se puede calcular *MSE*, no se puede contrastar la hipótesis de existencia de interacción.

Esto último implica también que:

a) en un modelo I, para poder contrastar las hipótesis de existencia de efectos de los factores no debe haber interacción (si hubiera interacción no tenemos término adecuado para realizar el contraste).

b) en un modelo mixto existe el mismo problema para el factor fijo.

Bloques completos aleatorios

Otro diseño muy frecuente de *anova* es el denominado de *bloques completos aleatorios* diseñado inicialmente para experimentos agrícolas pero actualmente muy extendido en otros campos. Puede considerarse como un caso particular de un *anova* de dos factores sin repetición o como una extensión al caso de *k* muestras de la comparación de medias de dos muestras emparejadas. Se trata de comparar *k* muestras emparejadas con respecto a otra variable cuyos efectos se quieren eliminar.

En este diseño a los datos de cada individuo se les denomina *bloque* y los datos se representan en una tabla de doble entrada análoga a la del *anova* de clasificación única en la que las *a* columnas son los tratamientos y las *b* filas los bloques, el elemento Y_{ij} de la tabla corresponde al tratamiento *i* y al bloque *j*. Las hipótesis que se pueden plantear son:

$H_0: \mu_{1.} = \mu_{2.} = \dots = \mu_{a.}$ (igualdad de medias de tratamientos)

y también, aunque generalmente tiene menos interés:

$H_0: \mu_{.1} = \mu_{.2} = \dots = \mu_{.b}$ (igualdad de medias de bloques)

A pesar del parecido con la clasificación única, el diseño es diferente: allí las columnas eran muestras independientes y aquí no. Realmente es un diseño de dos factores, uno de efectos fijos: los tratamientos, y el otro de efectos aleatorios: los bloques, y sin repetición: para cada bloque y tratamiento sólo hay una muestra.

El modelo aquí es:

$Y_{ij} = \mu + \alpha_i + B_j + \varepsilon_{ij}$ donde α_i es el efecto del tratamiento *i* y B_j el del bloque *j*. No hay término de interacción ya que, al no poder contrastar su existencia no tiene interés. Al ser un modelo mixto

exige la asunción de no existencia de interacción y los contrastes se hacen usando el término *MSE* como divisor.

3.11 Análisis de la varianza de más de dos factores

Es una generalización del de dos factores. El procedimiento, por lo tanto, será:

1) encontrar el modelo, teniendo en cuenta si los factores son fijos o aleatorios y todos los términos de interacción.

2) subdividir la suma de cuadrados total en tantos términos ortogonales como tenga el modelo y estudiar los valores esperados de los cuadrados medios para encontrar los estadísticos que permitan realizar los contrastes de hipótesis.

Un modelo de tres factores fijos, por ejemplo, será:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \delta_k + (\alpha\beta)_{ij} + (\alpha\delta)_{ik} + (\beta\delta)_{jk} + (\alpha\beta\delta)_{ijk} + \varepsilon_{ijkl}$$

Los tres primeros subíndices para los factores y el cuarto para las repeticiones, nótese que aparecen términos de interacción de segundo y tercer orden, en general en un modelo de *k* factores aparecen términos de interacción de orden 2, 3,... hasta *k* y el número de términos de interacción de orden *n* será el número combinatorio $C_{k;n}$. Este gran número de términos de interacción dificulta el análisis de más de dos factores, ya que son difíciles de interpretar y complican los valores esperados de los cuadrados medios por lo que también resulta difícil encontrar los estadísticos para los contrastes. Por estas razones no se suele emplear este tipo de análisis y cuando interesa estudiar varios factores a la vez se recurre a otros métodos de análisis multivariante.

ANEXO 1

Contrastes de hipótesis

Una *hipótesis estadística* es una suposición relativa a una o varias poblaciones, que puede ser cierta o no. Las hipótesis estadísticas se pueden contrastar con la información extraída de las muestras y tanto si se aceptan como si se rechazan se puede cometer un error.

La hipótesis formulada con intención de rechazarla se llama *hipótesis nula* y se representa por H_0. Rechazar H_0 implica aceptar una *hipótesis alternativa* (H_1).

La situación se puede esquematizar:

	H_0 cierta	H_0 falsa H_1 cierta
H_0 rechazada	Error tipo I (α)	Decisión correcta (*)
H_0 no rechazada	Decisión correcta	Error tipo II (β)

(*) Decisión correcta que se busca

α = p(rechazar H_0|H_0 cierta)
β = p(aceptar H_0|H_0 falsa)
Potencia =1-β = p(rechazar H_0|H_0 falsa)

Detalles a tener en cuenta

1 α y β están inversamente relacionadas.
2 Sólo pueden disminuirse las dos, aumentando *n*.

Los pasos necesarios para realizar un contraste relativo a un parámetro θ son:

1. Establecer la hipótesis nula en términos de igualdad

$H_0 : \theta = \theta_0$

2. Establecer la hipótesis alternativa, que puede hacerse de tres maneras, dependiendo del interés del investigador

$H_1 : \theta \neq \theta_0 \qquad \theta > \theta_0 \qquad \theta < \theta_0$

en el primer caso se habla de contraste *bilateral* o de *dos colas*, y en los otros dos de *lateral* (*derecho* en el 2º caso, o *izquierdo* en el 3º) o *una cola*.

3. Elegir un *nivel de significación*: nivel crítico para α

4. Elegir un *estadístico de contraste*: estadístico cuya distribución muestral se conozca en H_0 y que esté relacionado con θ y establecer, en base a dicha distribución, la *región crítica*: región en la que el estadístico tiene una probabilidad menor que α si H_0 fuera cierta y, en consecuencia, si el estadístico cayera en la misma, se rechazaría H_0.

Obsérvese que, de esta manera, se está más seguro cuando se rechaza una hipótesis que cuando no. Por eso se fija como H_0 lo que se quiere rechazar. Cuando no se rechaza, no se ha demostrado nada, simplemente no se ha podido rechazar. Por otro lado, la decisión se toma en base a la distribución muestral en H_0, por eso es necesario que tenga la igualdad.

5. Calcular el estadístico para una muestra aleatoria y compararlo con la región crítica, o equivalentemente, calcular el "valor p" del estadístico (probabilidad de obtener ese valor, u otro más alejado de la H_0, si H_0 fuera cierta) y compararlo con α.

Ejemplo:

Estamos estudiando el efecto del estrés sobre la presión arterial. Nuestra hipótesis es que la presión sistólica media en varones jóvenes estresados es mayor que 18 cm de Hg.

Estudiamos una muestra de 36 sujetos y encontramos

$$\overline{X} = 18{,}5 \qquad S = 3{,}6$$

1. Se trata de un contraste sobre medias. La hipótesis nula (lo que queremos rechazar) es:

$$H_0 : \mu = 18$$

2. La hipótesis alternativa

$$H_1 : \mu > 18$$

es un contraste lateral derecho.

3. Fijamos "a priori" el nivel de significación en 0,05 (el habitual en Biología).

4. El estadístico para el contraste es

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

y la región crítica $T>t_{\alpha}$

Si el contraste hubiera sido lateral izquierdo, la región crítica sería $T<t_{1-\alpha}$ y si hubiera sido bilateral $T<t_{1-\alpha/2}$ o $T>t_{\alpha/2}$.

En este ejemplo $t_{(35)0,05}=1,69$.

5. Calculamos el valor de t en la muestra

$$T = \frac{18{,}5 - 18}{3{,}6/\sqrt{36}} = 0{,}833$$

no está en la región crítica (no es mayor que 1,69), por tanto no rechazamos H_0.

Otra manera equivalente de hacer lo mismo (lo que hacen los paquetes estadísticos) es buscar en las tablas el "valor p" que

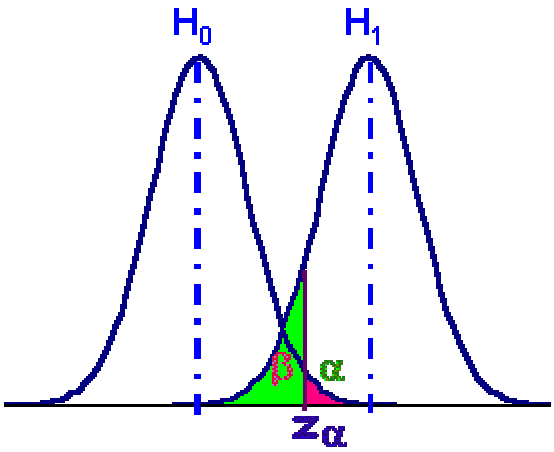
corresponde a T=0,833, que para 35 g.l. es aproximadamente 0,20. Es decir, si H_0 fuera cierta, la probabilidad de encontrar un valor de T como el que hemos encontrado *o mayor* (¿por qué mayor? Porque la H_1 es que μ es mayor , lo que produciría una media muestral mayor y por tanto mayor valor de t) es 0,20, dicho de otra manera la probabilidad de equivocarnos si rechazamos H_0 es 0,20, como la frontera se establece en 0,05 no la rechazamos.

Este valor crítico de 0,05 es *arbitrario* pero es la convención habitual. ¿Cuán razonable es?

Problema al respecto: en la hipótesis de que un mazo de cartas esté bien barajado, la probabilidad de que al sacar dos cartas sean, p.e.:1 el as de oros y 2 el rey de bastos es 1/40 x 1/39=0,000833.

Si hacemos la experiencia y obtenemos ese resultado ¿rechazaríamos la hipótesis de que el mazo está bien barajado? ¿Cuánto se parece esto a la lógica del contraste de hipótesis?

Volvamos al problema del estrés. Como no se rechaza H_0, se puede cometer un error tipo II. ¿Cuál es β ?. De hecho, sería la información relevante a comunicar en este estudio (la probabilidad del error que se pude cometer en él). Habitualmente, sin embargo, no se da porque los paquetes estadísticos no la calculan.
Para calcularla se debe concretar H_1, p.e. $\mu = 20$ (el criterio para este valor **no** es estadístico)



β =p(aceptar H_0|H_1 cierta)

Supongamos que el tamaño muestral sea suficientemente grande para poder aproximar *t* a *z*.

¿Cuándo se acepta H_0? si $z \leq 1{,}69$

$$\frac{\bar{X}-18}{3{,}6/\sqrt{36}} \leq 1{,}69 \Rightarrow \bar{X}-18 \leq 1{,}01 \Rightarrow \bar{X} \leq 19{,}01$$

es decir, se acepta H_0 si $\bar{X} \leq 19{,}01$

¿Qué probabilidad hay de encontrar $\bar{X} \leq 19{,}01$ si μ = 20 (zona verde del gráfico)? En esta hipótesis lo que se distribuye como una *z* es

$$\frac{\bar{X}-20}{3{,}6/\sqrt{36}} \Rightarrow z = \frac{19{,}01-20}{3{,}6/\sqrt{36}} = 1{,}65 \Rightarrow \beta = 0{,}05$$

Cálculo del tamaño muestral para contrastes sobre medias

Sea el contraste (bilateral)

H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

Para calcular el tamaño muestral debemos, además de fijar α y β, concretar H_1

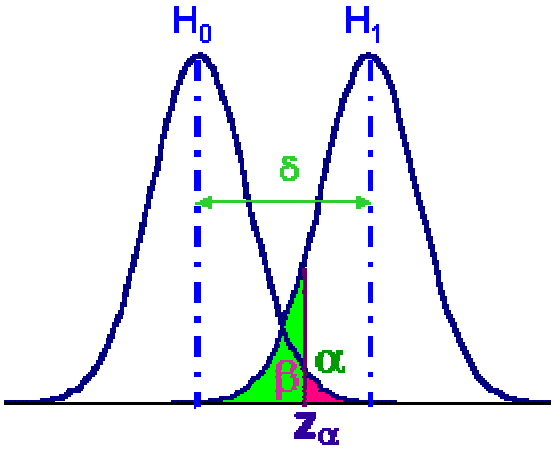
Concretando H_1: $\mu = \mu_0 + \delta$.

Si *n* suficientemente grande para poder usar la normal, es decir

$$z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \propto N(0,1)$$

resulta que

$$n = \frac{(z_\alpha + z_\beta)^2 s^2}{\delta^2}$$



Si el contraste fuera a dos colas habría que cambiar z_α por $z_{\alpha/2}$

ANEXO 2

Sucesos independientes

Dos sucesos son independientes si y sólo si $p(A \cap B) = p(A)\ p(B)$. Si dos sucesos son independientes

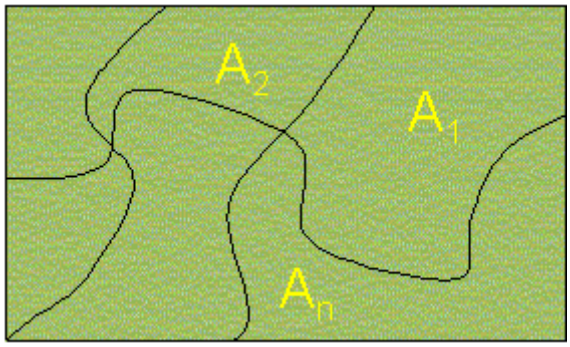
$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(A) \cdot p(B)}{p(B)} = p(A)$$

y del mismo modo p(B|A) = p(B).

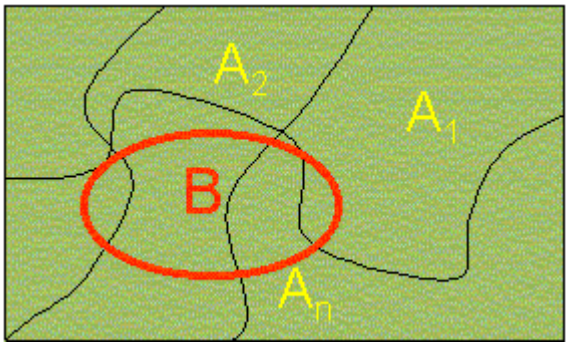
Esta propiedad coincide más con la idea intuitiva de independencia y algunos textos la dan como definición. Hay que notar, sin embargo, que ambas definiciones no son estrictamente equivalentes.

Regla de la probabilidad total

Se llama *partición* a conjunto de sucesos A_i tales que $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$ y $A_i \cap A_j = \varnothing\ \forall\ i \neq j$ es decir un conjunto de sucesos mutuamente excluyentes y que cubren todo el espacio muestral



Regla de la probabilidad total: Si un conjunto de sucesos A_i forman una partición del espacio muestral y $p(A_i) \neq 0\ \forall\ A_i$, para cualquier otro suceso B se cumple



$$p(B) = p(B|A_1)p(A_1) + p(B|A_2)p(A_2) + \ldots + p(B|A_n)p(A_n) = \sum_{i=1}^{n} p(B|A_i)p(A_i)$$

Teorema de Bayes

Si los sucesos A_i son una partición y B un suceso tal que $p(B) \neq 0$

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_{j=1}^{n} p(B|A_j)p(A_j)} \quad \text{para } i = 1,\ldots n$$

ANEXO 3

Comparación de medias

La hipótesis nula

H_0: $\mu_1 - \mu_2 = d_0$

Generalmente $d_0=0$

Hay 3 situaciones distintas:

1º σ_1^2 y σ_2^2 conocidos (poco frecuente).

2º σ_1^2 y σ_2^2 desconocidos pero iguales.

3º σ_1^2 y σ_2^2 desconocidos pero distintos.

Los estadísticos son distintos (*z* en 1 y *t* en 2 y 3) pero el procedimiento es el mismo. En los 3 casos se supone que las muestras son independientes.

Todos asumen **normalidad**. Si no se cumpliera hay que usar los llamados *test no paramétricos*.

Contrastes sobre independencia de v.a. cualitativas

Se quiere estudiar un posible factor pronóstico del éxito de una terapia, p.e. cierto grado de albuminuria como mal pronóstico en la diálisis. Los resultados de un estudio de este tipo se pueden comprimir en una tabla 2x2 del tipo

	F	**nF**	
E	a	b	m = a+b
nE	c	d	n = c+d
	e = a+c	f = b+d	T

Se estudian *T* individuos, *a* tienen al factor (**F**) y tiene éxito la terapia (**E**), *b* no tienen al factor (**nF**) y tiene éxito la terapia, ...

¡Ojo! A pesar de la aparente "inocencia" de esta tabla, puede significar cosas distintas según el diseño del estudio. **No** todas las probabilidades de las que se habla más abajo se pueden estimar siempre.

H_0 es que el factor **F** y el éxito **E** son independientes (**F** no es factor pronóstico) y H_1 que están asociados (sí es factor pronóstico). Si son independientes $p(E \cap F) = p(E)p(F)$. A partir de los datos de la tabla las mejores estimaciones de estas probabilidades son $\hat{p}(E) = m/T$ $\hat{p}(F) = e/T$, por lo tanto en H_0 $\hat{p}(E \cap F) = \frac{m \times e}{T^2}$, en consecuencia el valor esperado para esa celda en H_0 es $T \times \hat{p}(E \cap F) = \frac{m \times e}{T}$ (cociente entre el producto de los totales marginales y el gran total), del mismo modo se calculan los demás valores esperados y se construye el estadístico

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{g.l.}$$

que se distribuye según una distribución conocida denominada χ^2, que depende de un parámetro llamado "grados de libertad" (g.l.) Los g.l. en esta tabla son 1. Esto se puede generalizar a tablas CxF y los grados de libertad son (C-1)x(F-1).

Tabla de contingencia EJERC * SUPER

Recuento

		SUPER		Total
		0	1	
EJERC	0	15	25	40
	1	10	50	60
Total		25	75	100

Pruebas de χ^2

	Valor	gl	Sig. asint. (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	5,556	1	,018		
Corrección de continuidad	4,500	1	,034		
Razón de verosimilitud	5,475	1	,019		
Estadístico exacto de Fisher				,033	,017
Asociación lineal por lineal	5,500	1	,019		
N de casos válidos	100				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 10,00.

Estadísticos de *fuerza* de la asociación

¿Cuál es la *fuerza* de la asociación? Ni el estadístico χ^2 ni su valor *p* asociado miden esa fuerza, es decir se puede encontrar un alto valor de χ^2 (pequeño valor de *p*) con una asociación débil si el tamaño muestral fuera grande. Hay varios estadísticos propuestos para medir esta fuerza:

1º *Diferencia de riesgo o Reducción absoluta del riesgo (RAR)*: A partir de la tabla del ejemplo anterior podemos estimar la probabilidad (*riesgo* en la terminología epidemiológica) de que un individuo que haga ejercicio tenga éxito: $\hat{p}_1 = \frac{50}{60} = 0{,}83$ y también la probabilidad de que lo tenga uno que no lo haga: $\hat{p}_1 = \frac{25}{40} = 0{,}63$. Se llama *Diferencia de riesgo* o *Reducción absoluta del riesgo* a esta diferencia: 0,20 que puede oscilar entre -1 y 1; 0 indica no asociación.

2º *Reducción relativa del riesgo (RRR)*: La magnitud de la diferencia de riesgo es difícil de interpretar: una diferencia de 0,001 puede ser mucho o poco dependiendo del riesgo basal. Para superar esta dificultad se define la *RRR* como la reducción absoluta del riesgo dividida por el riesgo basal o riesgo del grupo de referencia. En el ejemplo, si consideramos como referencia el no hacer ejercicio, el *RRR* sería 0,20/0,63 = 0,32.

3º *Riesgo relativo (RR)*: Otro índice relativo es el riesgo relativo definido como el cociente entre los riesgos. En el ejemplo anterior RR=0,83/0,63=1,32. Los individuos que hacen ejercicio tienen una

probabilidad de éxito 1,32 veces mayor que los que no. El *RR* puede oscilar entre 0 y ∞; 1 indica no asociación. Es el estadístico preferido.

4º *Odds ratio (OR)*: Es un estadístico menos intuitivo que el *RR*. Para caracterizar un proceso binomial se puede usar su probabilidad (*p*) o el cociente *p/q* llamado *odds*. En el ejemplo anterior, para el ejercicio *p* = 0,83 y el odds = 0,83/0,17=4,88, es decir es 4,88 veces más probable tener éxito que no tenerlo si se hace ejercicio y para el no ejercicio *p* = 0,63 y el odds = 0,63/0,37=1,70. Para comparar ambos procesos podemos usar su cociente u *odds ratio* OR = 4,88/1,70 = 2,87. El odds para el ejercicio es 2,87 veces mayor que para el no ejercicio. El OR también puede oscilar entre 0 y ∞; 1 indica no asociación. Queda como ejercicio para el lector comprobar que el *OR* se puede estimar como el cociente de los productos cruzados de los elementos de la tabla, OR=(50x15)/(10x25)=3. La diferencia con el anterior es debida a errores de redondeo.

¿Qué ventajas tiene el *OR* frente al *RR*?. En principio parece menos intuitivo aunque un jugador no opinaría lo mismo. De hecho el *OR* proviene del mundo de las apuestas. Si queremos comparar dos juegos ¿qué da más información el *OR* o el *RR*? ... y ¿si queremos comparar dos estrategias terapéuticas?

Por otro lado si el estudio del ejemplo anterior se hubiera hecho de otra forma: muestreando por un lado individuos con éxito y por otro sin éxito (*diseño caso-control*) el *RR* no se podría estimar y sin embargo el *OR* sí **y de la misma forma** (se puede demostrar usando el teorema de Bayes).

Además, cuando se estudian fenómenos con probabilidades bajas (típicamente enfermedades) el *OR* tiende al *RR*.

Sean dos fenómenos con probabilidades p_1 y p_2 próximas a cero, enconsecuencia q_1 y q_2 estarán próximos a 1 y su cociente también, por lo tanto

$$OR = \frac{\frac{p_1}{q_1}}{\frac{p_2}{q_2}} = \frac{p_1 q_2}{p_2 q_1} \approx \frac{p_1}{p_2}$$

Resumiendo, el *OR* se puede estimar en diseños como el caso-control en los que el *RR* no se puede y si se estudian fenómenos con baja *prevalencia* el *OR* estima el *RR*. Además el *OR* es un buen indicador en sí mismo.

5º *Número necesario a tratar (NNT)*: En el contexto de la evaluación de tratamientos (ensayos clínicos) se suele usar este índice definido como el número de personas que se necesitaría tratar con un tratamiento para producir, o evitar, una ocurrencia adicional del evento. Del mismo modo se define *número necesario para perjudicar (NNP)* para evaluar efectos indeseables. Se calcula como el inverso del RAR. En el ejemplo NNT = 1/0,20 = 5 que se interpreta como por cada 5 pacientes que hagan ejercicio se consigue que uno tenga éxito.

4. Ajuste por mínimos cuadrados

Muy a menudo se encuentra en la práctica que existe una relación entre dos (o más) variables. Por ejemplo: los pesos de los hombres adultos dependen en cierto modo de sus alturas; las longitudes de las circunferencias y las áreas de los círculos dependen del radio, y la presión de una masa de gas depende de su temperatura y de su volumen.
Si todos los valores de las variables cumplen exactamente una relación exacta, entonces se dice que las variables están perfectamente correlacionadas o que hay una correlación perfecta entre ellas o, mas sencillamente, que existe una función o una fórmula que las relaciona.

Estamos interesados en determinar la relación funcional entre dos magnitudes x e y como resultado de experimentos.
Supongamos que intuimos que entre x e y existe la relación lineal $y=ax+b$ y deseamos determinar los parámetros a y b a partir de n medidas de x e y. a es la pendiente de la recta, es decir, la tangente del ángulo que forma con el eje de abscisas, y b la ordenada en el origen, es decir la altura a la que corta la recta al eje de ordenadas.

Ante un problema de este tipo, lo primero que conviene hacer es representar gráficamente los resultados para observar si los valores medidos se aproximan a una recta o no.

Para determinar la recta que mejor se adapta a los puntos se emplea el llamado método de los mínimos cuadrados. Para un valor de x determinado, la recta de ajuste proporciona un valor diferente de y del medido en el experimento. Esta diferencia será positiva para algunos puntos y negativa para otros, puesto que los puntos se

disponen alrededor de la recta. Por este motivo, la suma de estas diferencias para todos los puntos es poco significativa (las diferencias negativas se compensan con las positivas).

Por ello, para medir la discrepancia entre la recta y los puntos, se emplea la suma de los *cuadrados* de las diferencias, con los que nos aseguramos de que todos los términos son positivos.

De todas las posibles rectas que podemos trazar, caracterizadas por los parámetros *a* y *b*, la recta que mejor se ajusta a los puntos es la que hace mínima la suma.

Las condiciones de mínimo (primeras derivadas nulas) conducen a las ecuaciones normales que permiten determinar los parámetros.

Algunas consideraciones importantes sobre el coeficiente de correlación lineal:

- Es una cantidad sin dimensiones, es decir no depende de las unidades empleadas. Por ejemplo, si se está buscando hallar el coeficiente de correlación entre el peso y la altura de los niños en determinada ciudad, entonces el resultado será el mismo independientemente de si el peso de *todos* los niños se mide en Kilogramos o en gramos e independientemente de si la altura de *todos* los niños se mide en metros o centímetros.
- Se verifica siempre que:
- Si el coeficiente de correlación es igual a 1, entonces hay una correlación lineal positiva perfecta, es decir que los datos se ajustan perfectamente a una recta de pendiente positiva, es decir una recta *que crece*, o sea que *cuando x aumenta, entonces también lo hace y.*
- Si el coeficiente de correlación es igual a -1, entonces hay una correlación lineal negativa perfecta, es decir que los datos se ajustan perfectamente a una recta de pendiente negativa, es

decir una recta *que decrece*, o sea que *cuando x aumenta, entonces y disminuye.*

- En cualquier otro caso, para aceptar si hay una correlación lineal aceptable, no hay ninguna regla estricta. Normalmente, para aceptar la existencia de dicha correlación, el coeficiente debe ser mayor que 0,7 o menor que -0,7. En caso contrario, se suele rechazar la existencia de correlación lineal.

¿Que puede deducirse si se rechaza la existencia de correlación lineal si, por ejemplo, se encuentra un coeficiente de correlación lineal de 0,3 entre dos variables?

- Lo único que puede deducirse es que los datos no se ajustan a una recta.
- Pero esto no significa que no haya relación entre ellos dado que podrían ajustarse a una parábola o a cualquier otra curva. Sólo se deduce que *no hay correlación lineal aunque pudiera haber una correlación no lineal.*
- Este es el gran inconveniente del coeficiente de correlación lineal: no sirve para decidir si hay o no una posible *relación* entre dos variables, sólo sirve para decidir si hay o no una posible *relación lineal* entre dos variables.
- Ello hace que, definitivamente, *la única manera de decidir inicialmente si debe sospecharse o no la existencia de relación entre dos variables es estudiar detenidamente el diagrama de dispersión correspondiente, o sea la nube de puntos.*
- Y, en su caso, sólo después habrá que decidir con que curva se intentan ajustar los datos.

De forma análoga se pueden buscar ajustes polinómicos de grado mayor a uno (y por lo tanto nolineales). En este trabajo usaremos el parabólico.

5. aplicación: PURÍN DE CERDO EN EL VALLE DEL GUADALENTíN PARA BRÓCOLI Y MELÓN DE AGUA. ***(1 factor, Modelo 1).***

En resumen, el artículo describe el experimento realizado en la comarca del Valle del Guadalentín en los cultivos de brócoli y melón de agua.

En estos cultivos se va aplicar purín de cerdo como abono orgánico, para ver como influye en las características físico-químicas del suelo *(Cu, Zn, pH, CEC, nitrógeno total, conductividad electrónica, hongos, bacterias, coniformes…)* antes y después de haberlo aplicado. Se analizan los resultados con un programa original hecho en MATLAB donde se aplica ANOVA (análisis de la varianza).

Esta técnica estadística fue inventada para obtener de forma rigurosa la dependencia entre las variables de experimentos agronómicos. Posteriormente se extendió a otras ramas de la Ciencia.

En conclusión, los resultados del primer año de experimentación con cultivos de brócoli indicaron importantes cambios en las características químicas del suelo, especialmente el contenido de Cu y Zn. La conductividad eléctrica decrece debido principalmente al riego de la parcela. Hay un ligero aumento en el pH y CEC. El N_{Total} y el carbono orgánico no varían.

En cultivos de melón de agua se observa que decrece la conductividad eléctrica y se produce un ligero incremento del pH.

El Zn y Cu tienen niveles bastante altos al principio, pero luego decrecen, el carbono orgánico y el N no cambian con la aplicación de estos purines.

Los resultados de esta técnica en producción de melones de agua en el primer año de experimentación son mas pequeños que en el brócoli.

Los nitratos contenidos en la parcela que no estaba abonada con purines de cerdo eran superiores a las que si la llevaban.

En general, la microbiología que había en el suelo era más elevada en la superficie que en el subsuelo (30-60 cm) cuando se aplicaban los purines de cerdo.

Bacterias y coliformes aumentan con la aplicación de estos purines, por en cambio los actinomicetos y hongos no tenían un cambio significativo.

En este experimento, como se puede observar, se incrementa en el suelo el número de bacterias y coliformes, pero sin afectar a la población de actinomicetos y hongos.

Por más que continuáramos utilizando purines de cerdo en estos cultivos, las características microbiológicas después de un año de experimentación, bacterias, hongos, actinomicetos y coliformes retornan a los niveles del fondo del suelo.

Por último incluimos tres programas en MATLAB donde programamos ANOVA de un factor, recta y parábola de regresión. También mostramos dos gráficas correspondientes al nitrógeno y al Ph asociados a las distintas dosis de purín en el cultivo de la sandía. Los datos corresponden a la superficie. Se puede observar una influencia escasa, siendo lineal en el nitrógeno y parabólica en el Ph.

Programa 1

```
function fo=analvar1(Y,n,a,falfa)
%Calcula el fo del analisis de varianza de un solo factor
%grados de libertad
g1=a-1;
g2=0;
for i=1:a
  g2=g2+n(i);
end
N=g2; %despues se usa
g2=g2-a;

%calculo de yi.
for i=1:a
  my(i)=0;
  for j=1:n(i)
    my(i)=my(i)+Y(i,j);
  end
end

%calculo de y.. y de sum_1:a yi.^2/ni
My=0;
Cmy=0;
for i=1:a
  My=My+my(i);
  Cmy=Cmy+my(i)*my(i)/n(i);
end

%calculo de sum_1:a  sum_1:ni yij^2
Cy=0;
for i=1:a
  for j=1:n(i)
    Cy=Cy+Y(i,j)*Y(i,j);
  end
end

%calculo de fo
SSt=Cmy-My*My/N;
SST=Cy-My*My/N;
SSE=SST-SSt;
fo=(SSt/g1)/(SSE/g2);

if(fo>falfa)
  input('Las medias no son de la misma muestra, influye el factor')
else
   input('Las medias son de la misma muestra, no influye el factor')
end
```

Programa 2

```
% Programa de REGRESIÓN LINEAL // RECTA DE MÍNIMOS CUADRADOS
% Problema a estudiar:
  %  Dado un conjunto finito y discreto de puntos (x(i),y(i)), i=1:n, se trata de encontrar
  % una curva que sin pasar necesariamente por dichos puntos se ajuste en el sentido que se
  % precise a dichos datos, (depende de si usamos mínimos cuadrados, pol.de Chebychev,...).
  % RECTA DE MÍNIMOS CUADRADOS:
    % La recta y=p(x)=a+bx, donde a y b se determinan de manera que la expresión
    % Sumatorio desde 1:n de a+b(x(i)-f(i))^2 sea mínimo, la llamaremos recta de
    % mínimos cuadrados.
    % Imponiendo las condiciones de mínimo, resultan las ecuaciones:ç
    % a*n+b*(Sumatorio i=1:n de los x(i))=(Sumatorio i=1:n de los f(i))
    % a*(Sumatorio i=1:n de los x(i))+b*(Sumatorio i=1:n de los x(i)^2)=
    % =(Sumatorio i=1:n de los x(i)*f(i))
    % siendo n el número total de datos, a y b se obtienen al resolverlas.
    % La bondad del ajuste,es decir, la aproximación de la recta a los datos, se
    % halla mediante la siguiente expresión, que cuanto más próximo a cero sea,
    % mejor será la aproximación.
    % Sumatorio i=1:n de (p(x)-f(i))^2
% Programa:
  % Datos de entrada
    % x ==> vector de datos de las abcisas
    % y=f(x)==> vector de datos de las ordenadas
    % dato ==> valor de la abcisa que se quiere calcular su imagen
  % Datos de salida
    % fdato ==> valor de la imagen
    % bondad ==> bondad del ajuste
    % einf ==> error máximo
    % e1 ==> error medio
    % e2 ==> error cuadrático medio
    % los coeficientes a y b
function [fdato,bondad,einf,e1,e2,a,b]=Mregre(x,y,dato)
% n ==> número total de datos
n=length(x);
%Inicialización de las variables
  % xx ==> sumatorio de los x(i)
  xx=0;
  % yy ==> sumatorio de los y(i)
  yy=0;
  % x2 ==> sumatorio de los x(i)^2
  x2=0;
  % xy ==> sumatorio del producto de x(i)*y(i)
  xy=0;
  % bondad ==> sumatorio para la bondad del ajuste
  bondad=0;
```

```
for i=1:n
  xx=x(i)+xx;
  yy=y(i)+yy;
  x2=x(i)^2+x2;
  xy=x(i)*y(i)+xy;
end
% resolvemos el sistema
A=[n xx;xx x2];
B=[yy xy];
v=A\B';
a=v(1);
b=v(2);
% cálculo de la bondad del ajuste
for i=1:n
   bondad=(a+b*x(i)-y(i))^2+bondad;
end
% cáculo de la imagen del dato
fdato=a+b*dato;
% cálculo de errores
   % error máximo: el máximo del conjunto formado por abs((a+b*x(i))-y(i)) con i=1:n
      for i=1:n
        e(i)=abs(a+b*x(i)-y(i));
      end
        einf=max(e);
   % error medio: (1/n)*Sumatorio de abs((a+b*x(i))-y(i)) desde i=1:n
      ee1=0;
      for i=1:n
        ee1=e(i)+ee1;
      end
      e1=(1/n)*ee1;
   % error cuadrático medio: [(1/n)*Sumatorio de abs((a+b*x(i))-y(i))^2 desde i=1:n ]^0.5
      ee2=0;
      for i=1:n
        ee2=e(i)^2+ee2;
      end
      e2=((1/n)*ee2)^(1/2);
```

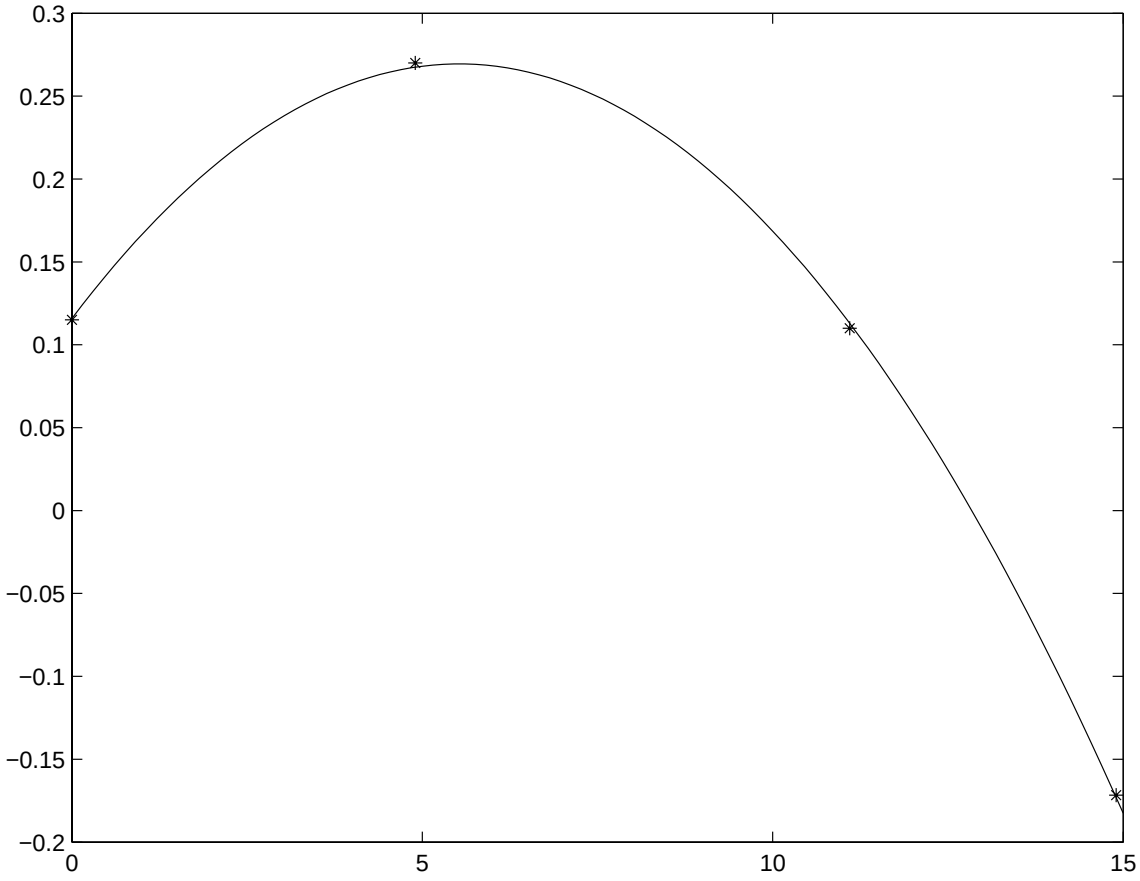
Programa 3

```
% Programa de REGRESIÓN no LINEAL // PARABOLA DE MÍNIMOS CUADRADOS
% Problema a estudiar:
  %  Dado un conjunto finito y discreto de puntos (x(i),y(i)), i=1:n, se trata de encontrar
  % una curva que sin pasar necesariamente por dichos puntos se ajuste en el sentido que se
  % precise a dichos datos, (depende de si usamos mínimos cuadrados, pol.de Chebychev,...).
  % PARABOLA DE MÍNIMOS CUADRADOS:
    % La recta y=p(x)=a+bx+cx, donde a, b y c se determinan de manera que la expresión
    % Sumatorio desde 1:n de (a+b x(i)+ c x(i)^2-f(i))^2 sea mínimo, la llamaremos parabola de
    % mínimos cuadrados.
    % Imponiendo las condiciones de mínimo (gradiente =0), resultan las ecuaciones:
    % a*n+b*(Sumatorio i=1:n de los x(i))+b*(Sumatorio i=1:n de los x(i))+
    % c*(Sumatorio i=1:n de los x(i)^2)
    %=(Sumatorio i=1:n de los f(i))
    % a*(Sumatorio i=1:n de los x(i))+b*(Sumatorio i=1:n de los x(i)^2)+
    % c*(Sumatorio i=1:n de los x(i)^3)
    % =(Sumatorio i=1:n de los x(i)*f(i))
    % a*(Sumatorio i=1:n de los x(i)^2)+b*(Sumatorio i=1:n de los x(i)^3)+
    % c*(Sumatorio i=1:n de los x(i)^4)
    % =(Sumatorio i=1:n de los x(i)^2*f(i))
    % siendo n el número total de datos, a, b y c se obtienen al resolverlas.
    % La bondad del ajuste,es decir, la aproximación de la recta a los datos, se
    % halla mediante la siguiente expresión, que cuanto más próximo a cero sea,
    % mejor será la aproximación.
    % Sumatorio i=1:n de (p(x)-f(i))^2
% Programa:
  % Datos de entrada
    % x ==> vector de datos de las abcisas
    % y=f(x)==> vector de datos de las ordenadas
    % dato ==> valor de la abcisa que se quiere calcular su imagen
  % Datos de salida
    % fdato ==> valor de la imagen
    % bondad ==> bondad del ajuste
    % einf ==> error máximo
    % e1 ==> error medio
    % e2 ==> error cuadrático medio
    % los coeficientes a, b y c
function [fdato,bondad,einf,e1,e2,a,b,c]=Mregrenoli(x,y,dato)
% n ==> número total de datos
n=length(x);
%Inicialización de las variables
 % xx ==> sumatorio de los x(i)
 xx=0;
 % yy ==> sumatorio de los y(i)
 yy=0;
```

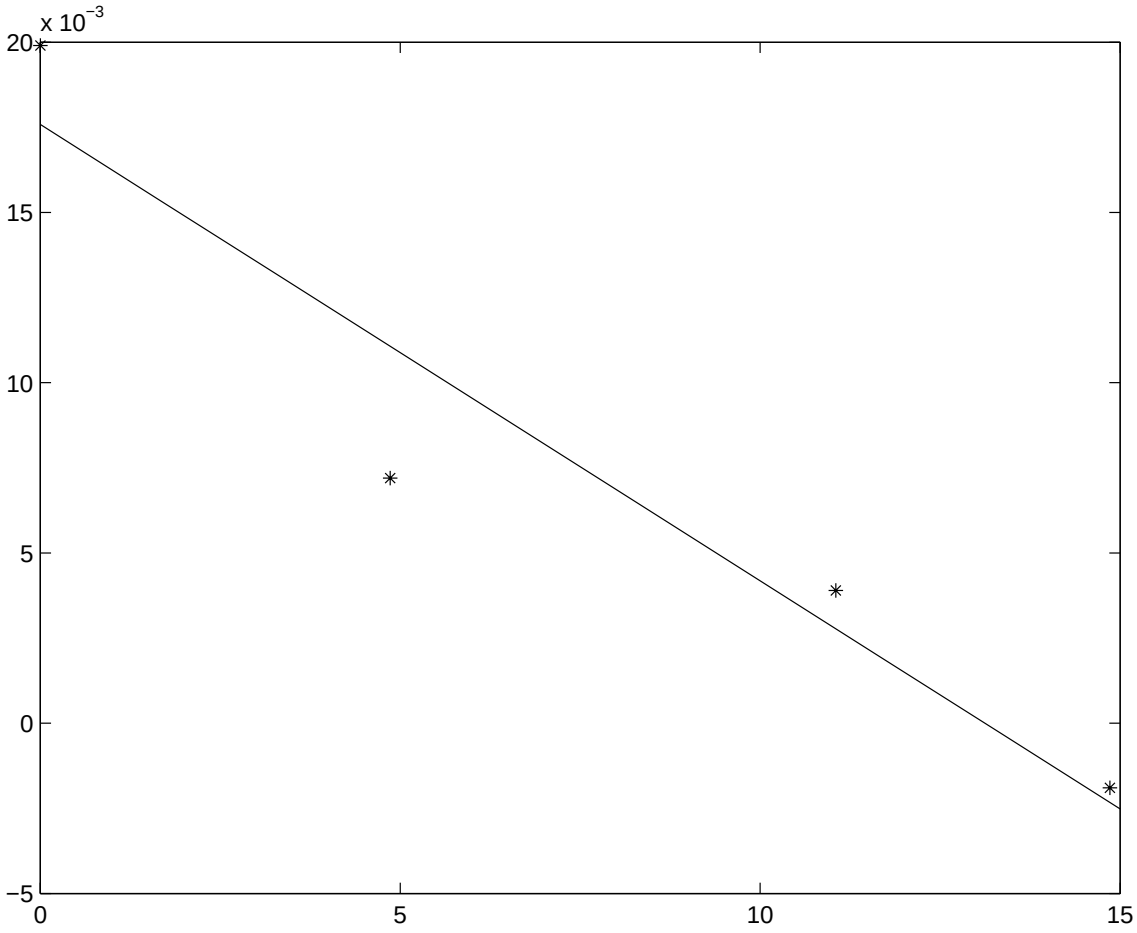
```
  % x2 ==> sumatorio de los x(i)^2
  x2=0;
  % x3 ==> sumatorio de los x(i)^3
  x3=0;
  % x4 ==> sumatorio de los x(i)^4
  x4=0;
  % xy ==> sumatorio del producto de x(i)*y(i)
  xy=0;
   % xxy ==> sumatorio del producto de x(i)*x(i)*y(i)
  xxy=0;
  % bondad ==> sumatorio para la bondad del ajuste
  bondad=0;
for i=1:n
  xx=x(i)+xx;
  yy=y(i)+yy;
  x2=x(i)^2+x2;
  x3=x(i)^3+x3;
  x4=x(i)^4+x4;
  xy=x(i)*y(i)+xy;
  xxy=x(i)*x(i)*y(i)+xxy;
end
% resolvemos el sistema
A=[n xx x2;xx x2 x3;x2 x3 x4];
B=[yy xy xxy];
v=A\B';
a=v(1);
b=v(2);
c=v(3);
% cálculo de la bondad del ajuste
for i=1:n
   bondad=(a+b*x(i)+c*x(i)*x(i)-y(i))^2+bondad;
end
% cáculo de la imagen del dato
fdato=a+b*dato;
% cálculo de errores
   % error máximo: el máximo del conjunto formado por abs((a+b*x(i))-y(i)) con i=1:n
      for i=1:n
        e(i)=abs(a+b*x(i)+c*x(i)*x(i)-y(i));
      end
        einf=max(e);
   % error medio: (1/n)*Sumatorio de abs((a+b*x(i))-y(i)) desde i=1:n
      ee1=0;
      for i=1:n
        ee1=e(i)+ee1;
      end
      e1=(1/n)*ee1;
```

```
% error cuadrático medio: [(1/n)*Sumatorio de abs((a+b*x(i))-y(i))^2 desde i=1:n ]^0.5
  ee2=0;
  for i=1:n
    ee2=e(i)^2+ee2;
  end
  e2=((1/n)*ee2)^(1/2);
```

Gráfica Nitrógeno



Gráfica Ph



6. Bibliografía

http://fisica.unav.es/angel/matlab (introducción a MATLAB)

http://www.hrc.es (análisis de la varianza)

http://www.hrc.es (anexo 1)

http://www.hrc.es (anexo 2)

http://www.hrc.es (anexo 3)

Artículo: PURÍN DE CERDO EN EL VALLE DEL GUADALENTíN PARA BRÓCOLI Y MELÓN DE AGUA. (Artículo realizado por: Ángel Faz Cano, José Luis Tortosa, Manuel Andujar, Miriam Llona, Juan B. Lobera, Alfredo Palop y Sergio Amat; profesores de la U.P.C.T), CATENA 2004.