



# Prácticas de Estadística con *Excel*

Cargar o importar datos

Cargar el módulo de Estadística

Generación de números (pseudo)aleatorios

## Una población o muestra

Obtener un resumen de medidas

Dibujar el histograma de una muestra

Cálculo de la media muestral

Distribución (en el muestreo) de la media muestral

## Dos poblaciones o muestras

Diagrama de dispersión entre dos variables

Recta de regresión entre dos variables

## Cargar o importar datos

Los datos podemos meterlos a mano, generarlos con el programa o cargarlos desde un archivo en el que estén. Veamos esto último. Normalmente este tipo de programas pueden abrir archivos que estén en varios formatos, además del suyo propio. En el foro de una tutoría están disponibles los datos que vamos a utilizar, tanto en formato de *Excel* como en formato de texto plano (los datos puestos en columnas separadas por tabulaciones). Estos archivos se llaman «países.xls» y «países.txt». Para abrir los datos:

Seleccionar *Archivo --> Abrir...*

	A	B	C	D	E	F	G
1	Nº	País	POB	AREA	ESPERANZ	AGRICUL	I
2	1	Mozambique	16.5	802	44	64	
3	2	Etiopía	54.8	1222	49	48	
4	3	Tanzania	25.9	945	51	61	
5	4	Sierra Leona	4.4	72	43	38	
6	5	Nepal	19.9	141	54	52	
7	6	Uganda	17.5	236	43	57	
8	7	Bhutan	1.5	47	48	42	
9	8	Burundi	5.8	28	48	54	
10	9	Malawi	9.1	118	44	28	
11	10	Bangladesh	114.4	144	55	34	
12	11	Chad	6	1284	47	44	
13	12	Guinea-Bisau	1	36	39	44	
14	13	Madagascar	12.4	587	51	33	

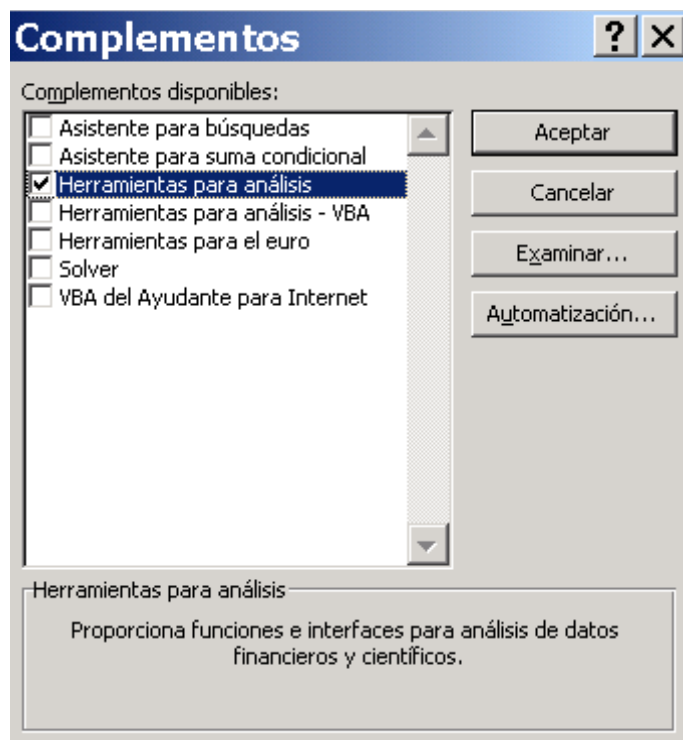
Vemos que cada país está en una fila y cada variable en una columna. Los nombres de los países están en la segunda columna (la primera es un índice para numerarlos) y los de las variables están en la primera fila. Es importante tener esto presente para no incluir esos nombre entre las celdas de la muestra a la hora de hacer cálculos con ellos.



## Cargar el módulo para Estadística

*Excel* tiene un módulo para hacer cálculos estadísticos. Como no es un módulo básico, hay que cargarlo expresamente. Para hacer que el submenú *Análisis de datos...* aparezca en el menú *Herramientas* del programa:

Seleccionar *Herramientas* --> *Complementos...* --> Marcar *Herramientas para análisis*.



## Generación de números (pseudo)aleatorios

### Objetivo

Aprender a generar muestras aleatorias simples de variables aleatorias de distintos modelos,  $x_1, \dots, x_n$ .

### Enunciado

Generar en columnas diferentes 10 muestras de tamaño 100 de una variable aleatoria que sigue una  $N(0, 4^2)$ .

### Menús

i) Para hacer que en el menú *Análisis de datos...* aparezca en el menú *Herramientas* del programa:

Seleccionar *Herramientas --> Complementos... --> Marcar Herramientas para análisis.*

## ii) Para **generar números aleatorios**:

Entrar en *Herramientas --> Análisis de datos... --> Seleccionar Generación de números aleatorios -->* Elegir el número de variables (muestras), la cantidad de números aleatorios (longitud de las muestras), la distribución deseada y los valores de sus parámetros, e indicar que muestre los resultados en una nueva hoja de cálculo



## Obtener un resumen de medidas de una muestra

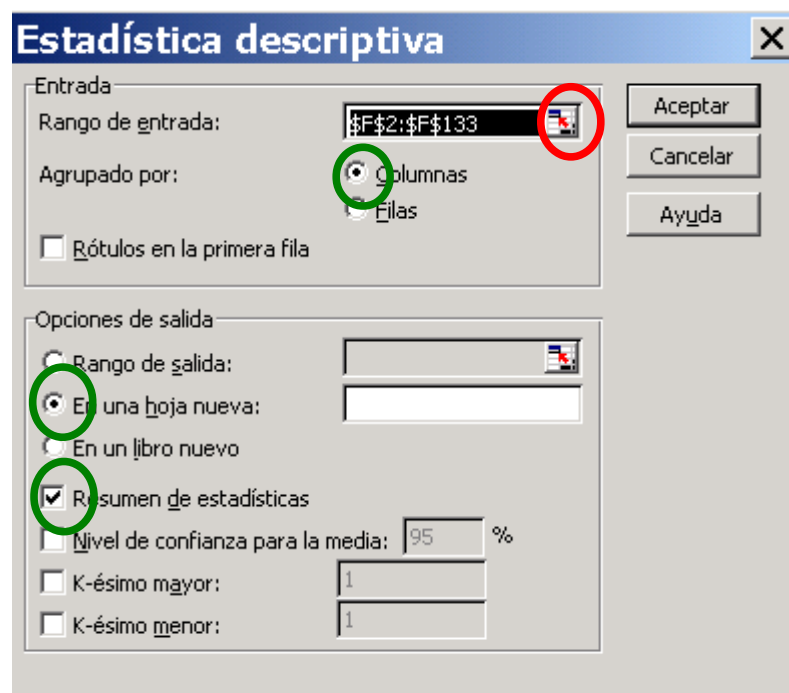
Ahora vamos a obtener un resumen de estadísticos de una muestra de datos. Para ello vamos a

*Herramientas --> Análisis de datos... --> Estadística descriptiva*

Le indicamos con el ratón en qué celdas están los datos, si están por columnas o por filas, que muestre los resultados en una nueva hoja de cálculo y que muestre el resumen de medidas. Para seleccionar los datos que queremos analizar, pulsamos el botón que están en la circunferencia roja del dibujo de más abajo (no de éste primero). Aparece el siguiente cuadro



No le hacemos caso, sencillamente seleccionamos con el ratón la segunda celda de la columna AGRICUL y extendemos la selección con el ratón hasta el final de esa columna. Pulsamos el botón que está en el círculo azul (figura de arriba) y automáticamente se introduce la selección en el campo adecuado.



Y se obtiene

	A	B	C
2			
3	Media	19,373	
4	Error típico	1,3896	
5	Mediana	16	
6	Moda	2	
7	Desviación estándar	15,966	
8	Varianza de la muestra	254,91	
9	Curtosis	0,1111	
10	Coefficiente de asimetría	0,8809	
11	Rango	65	
12	Mínimo	0	
13	Máximo	65	
14	Suma	2557,2	
15	Cuenta	132	
16			



## Dibujar el histograma de una muestra

Hay que entrar en el menú:

*Herramientas --> Análisis de datos... --> Histograma*

La información que necesita Excel para hacer un histograma es la muestra y los intervalos. Si no se le indican los intervalos, él los construye. Veamos primero este caso.

### Si no indicamos las clases

Le indicamos dónde está la muestra como se ha hecho antes (de hecho la mantiene si no se ha cerrado el programa) y si queremos que muestre la frecuencias acumuladas porcentuales o alguno de los gráficos que ofrece.

**Histograma**

Entrada

Rango de entrada:

Rango de clases:

Rótulos

Opciones de salida

Rango de salida:

En una hoja nueva:

En un libro nuevo

Pareto (Histograma ordenado)

Porcentaje acumulado

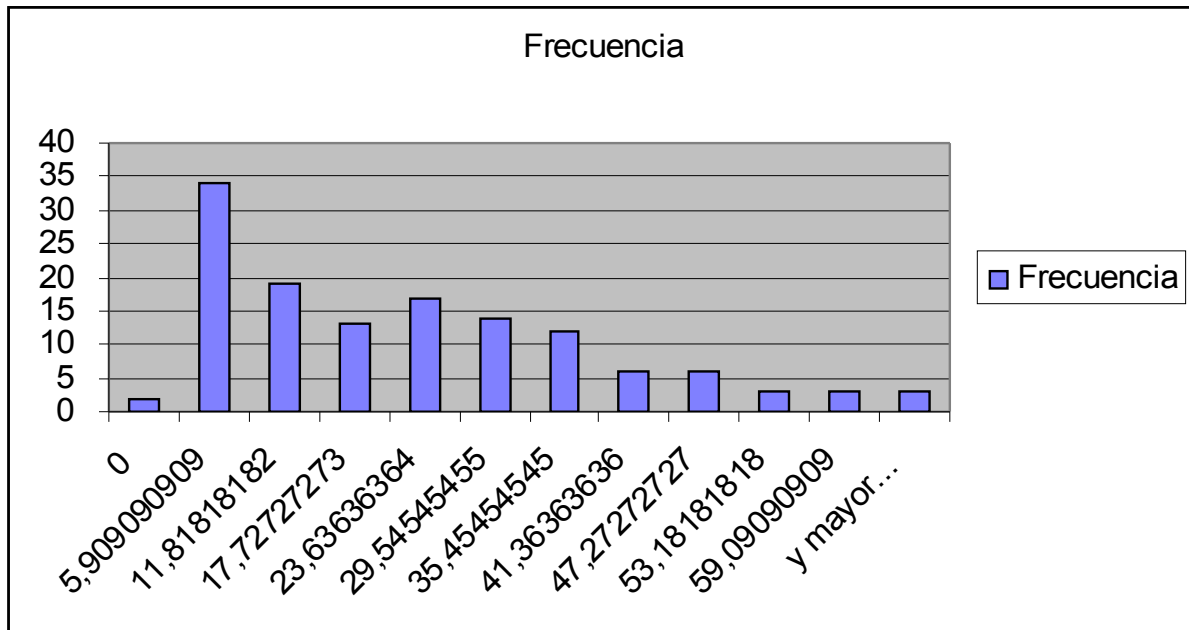
Crear gráfico

Botones: Aceptar, Cancelar, Ayuda

	A	B	C
1	Clase	Frecuencia	
2	0	2	
3	5,9090909	34	
4	11,818182	19	
5	17,727273	13	
6	23,636364	17	
7	29,545455	14	
8	35,454545	12	
9	41,363636	6	
10	47,272727	6	
11	53,181818	3	
12	59,090909	3	
13	y mayor...	3	

En una hoja nueva (o en un libro nuevo, según se le haya indicado) proporciona la tabla con las clases (indica el extremo inferior) y con las frecuencias absolutas (imagen superior derecha)

Si no le hemos indicado que haga el gráfico, podemos hacerlo ahora. Por último, para que muestre la gráfica del histograma, pulsamos el botón del asistente de gráficos que *Excel* tiene en la barra de herramientas y vamos siguiendo los pasos:



### Si queremos indicar las clases

Para indicárselo tenemos que introducir los extremos inferiores de los intervalos en una columna. En este caso, del resumen anterior vemos que:

- La muestra tiene 132 datos
- El valor máximo es 65
- El valor mínimo es 0

Como  $\sqrt{132} = 11,48913$  podemos considerar 11 ó 12 intervalos que cubran todo el rango. Como  $65/11 = 5,90$  y  $65/12 = 5,42$  parece que es más cómodo **tomar 11 intervalos de longitud 6**, ¿no? Ahora,  $11 \cdot 6 = 66$ , así que podemos centrar los intervalos y empezar desde el -0,5 al 65,5. Si tuviésemos que hacer los cálculos a mano, no pasaría nada grave por empezar el primer intervalo en 0 y terminar en 66. Vemos en el apartado anterior que el programa ha tomado precisamente 11 intervalos de longitud 5,90 empezando en el 0, es decir, para calcular el número de clases *Excel* aplica la regla de la raíz del número de datos y redondea.

Vamos a elegir una columna que esté en blanco y vamos a escribir en una celda el número -0,5. Seleccionamos la celda de debajo y vamos a la barra de funciones. Escribimos lo siguiente:

$f_x = L2 + 6$

porque en mi caso la celda en que he puesto la primer cantidad es la L2. Al pulsar el símbolo verde vemos que aplica la fórmula. Ahora queremos que *Excel* entienda que queremos hacer

esa misma operación en las celdas de debajo, es decir, que la generalice y haga que cada celda sea la anterior más seis. Esto es fácil, basta seleccionar esta segunda celda y arrastrar con el ratón de la esquina inferior hasta la fila que queramos. En nuestro caso, como hemos empezado en la segunda, queremos que lo aplique hasta la decimotercera. Esta selección se muestra a la izquierda en las siguientes imágenes, y el resultado a la derecha.

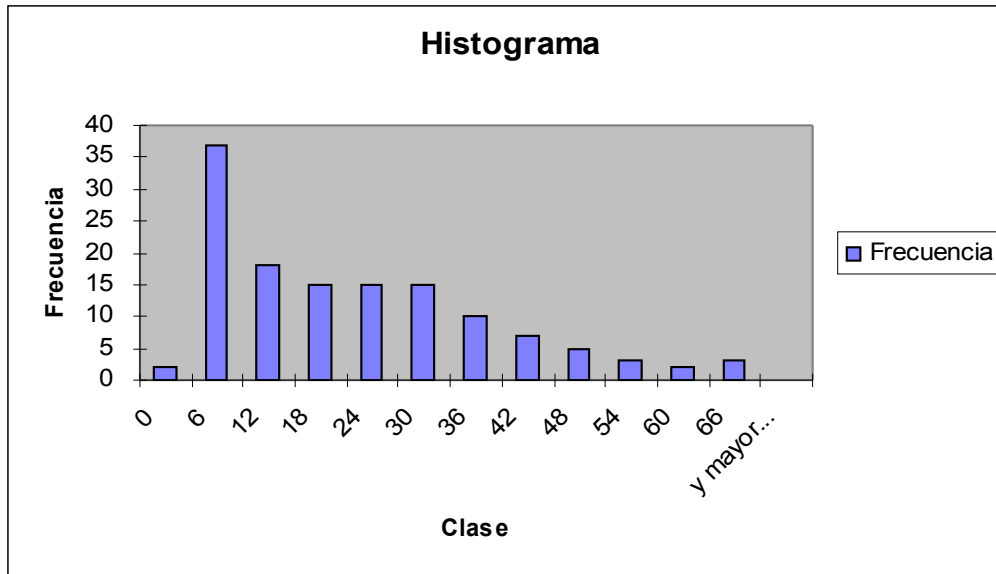
The image shows two parts of an Excel spreadsheet. On the left, a column of cells in column L is selected, containing the following values: -0,5, 5,5, 11,5, 17,5, 23,5, 29,5, 35,5, 41,5, 47,5, 53,5, 59,5, and 65,5. A small icon with a plus sign is visible at the bottom right of the selection. On the right, a table with two columns, A and B, is shown. Column A is labeled 'Clase' and column B is labeled 'Frecuencia'. The data in the table is as follows:

	A	B
1	<i>Clase</i>	<i>Frecuencia</i>
2	-0,5	0
3	5,5	36
4	11,5	19
5	17,5	13
6	23,5	17
7	29,5	14
8	35,5	12
9	41,5	6
10	47,5	6
11	53,5	3
12	59,5	3
13	65,5	3
14	y mayor...	0

Fácil, ¿verdad? Sí, como los cestos... cuando uno hace el primero. Vemos que ha metido los dos datos que valían 0 en la clase del 5,5, no en la del -0,5. Esto se debe a que el programa mete en cada clase los valores que hay entre su límite inferior y el del intervalo siguiente, pero para los intervalos de los extremos computa respectivamente el número de datos que hay menores y mayores que sus límites. Es decir, para la clase del -0,5 ve que no hay ningún valor menor que él y pone frecuencia 0. Para arreglar esto, podemos empezar en 0.

The image shows two parts of an Excel spreadsheet. On the left, a column of cells in column L is selected, containing the following values: 0, 6, 12, 18, 24, 30, 36, 42, 48, 54, 60, and 66. On the right, a table with two columns, A and B, is shown. Column A is labeled 'Clase' and column B is labeled 'Frecuencia'. The data in the table is as follows:

	A	B
1	<i>Clase</i>	<i>Frecuencia</i>
2	0	2
3	6	37
4	12	18
5	18	15
6	24	15
7	30	15
8	36	10
9	42	7
10	48	5
11	54	3
12	60	2
13	66	3
14	y mayor...	0



## Cálculo de la media muestral

### Objetivo

Aprender a calcular la media muestral,  $\bar{x}$ , de cada muestra  $x_1, \dots, x_n$ .

### Enunciado

*Dada una muestra (en una columna), calcular la media muestral de cada columna (si hay varias muestras).*

### Menús

Para **calcular la media** de cada una de las columnas elegimos primero una casilla debajo de la primera columna, y la «programamos» para que *Excel* calcule la media de la muestra y en ella.

Seleccionamos la celda --> Pulsamos el símbolo de función que aparece en la barra de herramientas --> Elegimos *PROMEDIO* --> Con el ratón le indicamos a *Excel* cuál es el rango de las celdas donde está esa muestra --> No hay que hacer eso con cada columna, basta seleccionar esa primera y arrastrar con el ratón para seleccionar el resto de celdas de debajo de cada columna; *Excel* suele entender bien que se quiere hacer la misma operación para cada columna.



## Distribución (en el muestreo) de la media muestral

### Objetivo

Estudiar la distribución (en el muestreo) de la media muestral.

### Enunciado

*Dada una distribución de probabilidad  $X$ , se pueden generar valores numéricos  $x_1, \dots, x_n$ . Se pueden considerar tanto la medida numérica  $\bar{x}$ , formada a partir de los números  $x_1, \dots, x_n$ , como, antes de conocer los valores de estos números,  $\bar{X}$ , formada a partir de una muestra aleatoria simple (ideal) de variables*

aleatorias  $X_1, \dots, X_n$ . La cantidad  $\bar{X}$  es ella misma una variable aleatoria, y a su distribución, que hay que distinguir de la de  $X$ , se le llama distribución (en el muestreo). Para estudiar esta distribución empíricamente, hay que generar una muestra aleatoria simple  $\bar{x}_1, \dots, \bar{x}_n$  (nótese que para generar cada uno de estos valores hay que generar a su vez una muestra  $x_1, \dots, x_n$ ).

## Teoría

La media muestral de una muestra aleatoria simple de normales sigue una distribución cuya media es la misma que la de las variables, y su varianza es la misma dividida por el tamaño de la muestra:

Si  $X$  sigue una distribución  $N(\mu, \sigma^2)$ , entonces  $\bar{X} = (X_1 + X_2 + \dots + X_n)/n$  sigue una  $N(\mu, \sigma^2/n)$

## Menús

Para **estudiar la media muestral**, es necesario generar una muestra aleatoria simple  $\bar{x}_1, \dots, \bar{x}_n$ . Lo hacemos como se indica en :

Entramos en *Herramientas --> Análisis de datos... --> Estadística descriptiva -->* Le indicamos en qué celdas están los datos (son las que acabamos de calcular), que los datos están por filas, que muestre los resultados en una nueva hoja de cálculo y que muestre el resumen de estadísticas.

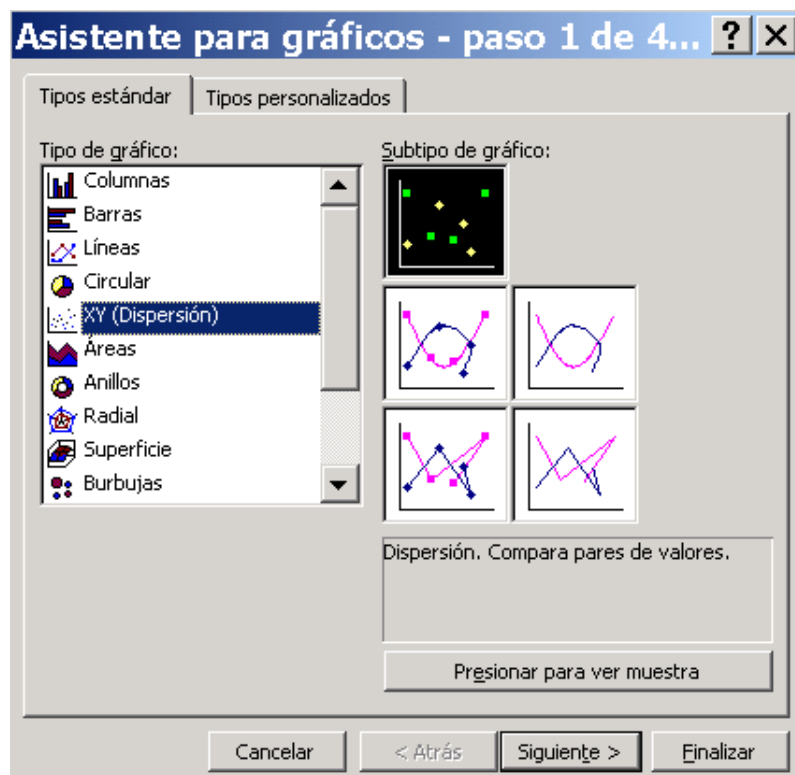
Ya tenemos un resumen de estadísticos de la muestra de la media muestral.



## Diagrama de dispersión entre dos variables

Para hacer cualquier gráfico, llamamos al asistente pulsando el botón de la barra de herramientas o entrando en el menú

*Insertar --> Gráfico*



Seleccionamos la opción que se muestra en la imagen. En la siguiente ventana nos ofrece información de dónde cree el programa que están los datos, para nuestra comodidad. En este caso lo vamos a eliminar, para entender mejor qué estamos representando. Es decir, donde pone

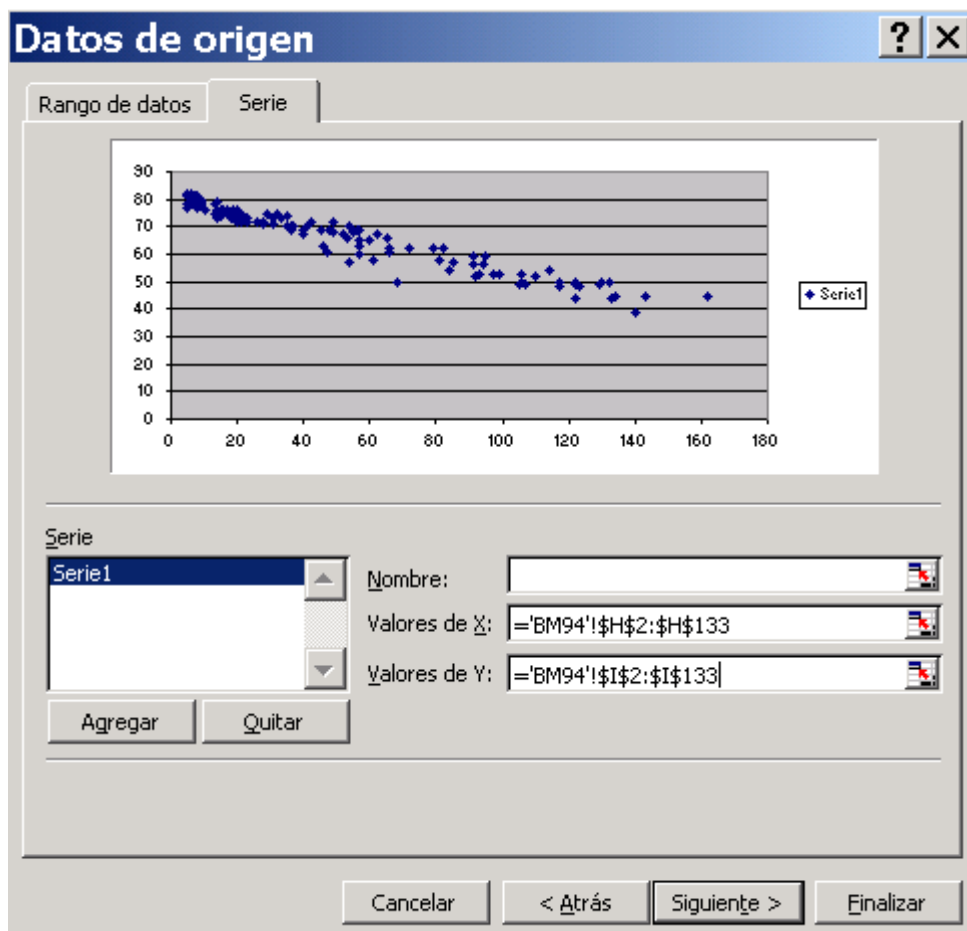


hacemos que ponga



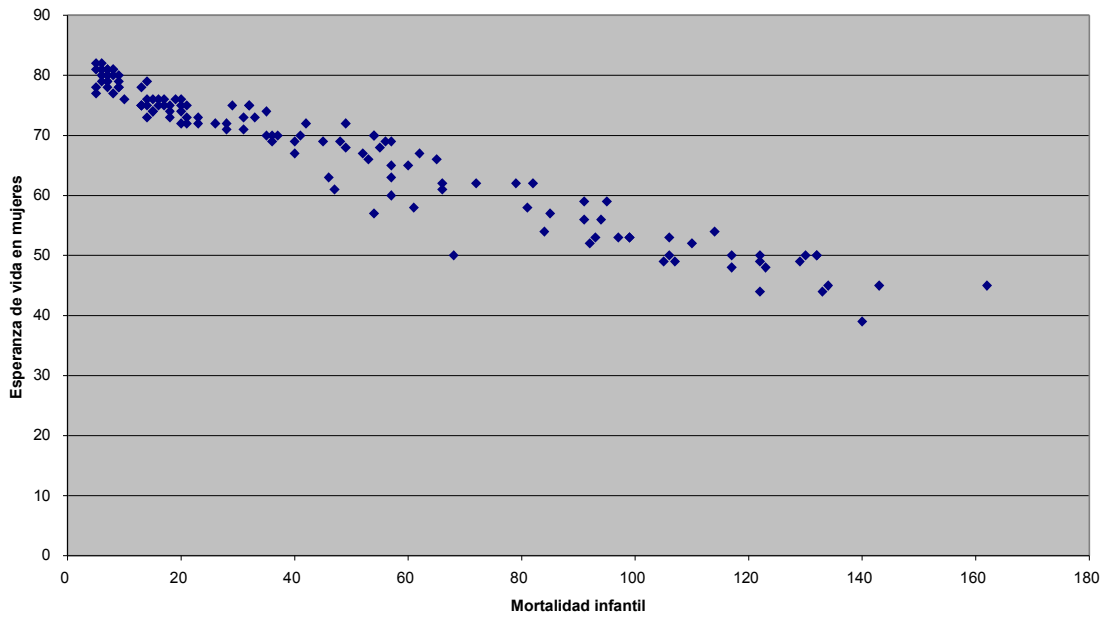
y ahora manualmente vamos a indicarle que nos represente, para cada país, el par de variables que queremos representar. Es decir, cada punto del gráfico será un país. Otra cosa distinta es representar por cada lado, pero en el mismo gráfico, las dos series. En este segundo caso se puede intuir la relación entre las dos variables, pero no es un diagrama de dispersión.

Después de borrar el rango de valores, vamos a ir a la pestaña *Serie* y le vamos a indicar quiénes queremos que sean los valores del eje X y del Y. Le damos a agregar y ponemos en la X, con el ratón, los valores de la variable MORT INF y en el eje Y la variable ESP M.



En la siguiente pantalla les damos nombre al gráfico, los ejes, etcétera.:

Diagrama de dispersión



Vemos que claramente hay una relación lineal entre las dos variables. Lo que sucede además es que cuanto mayor es la mortalidad infantil, menor es la esperanza de vida de las mujeres, y viceversa. Esto es lógico, porque ambas cosas están relacionadas con el nivel de vida del país.



## Recta de regresión entre dos variables

Queremos ahora expresar con una fórmula la relación entre las dos variables; a esto se le llama «modelar» o «modelizar». Para eso entramos en el menú

*Herramientas --> Análisis de datos...*

Elegimos *Regresión* y nos sale un cuadro de diálogo, que después de relleno queda

The screenshot shows the 'Regresión' dialog box in Excel. The 'Entrada' section has 'Rango Y de entrada' set to '\$I\$2:\$I\$133' and 'Rango X de entrada' set to '\$H\$2:\$H\$133'. There are checkboxes for 'Rótulos', 'Constante igual a cero', and 'Nivel de confianza' (set to 95%). The 'Opciones de salida' section has 'En una hoja nueva' selected. The 'Residuales' section has checkboxes for 'Residuos', 'Residuos estándares', 'Gráfico de residuales', and 'Curva de regresión ajustada'. The 'Probabilidad normal' section has a checkbox for 'Gráfico de probabilidad normal'. Buttons for 'Aceptar', 'Cancelar', and 'Ayuda' are on the right.

Notad que pide antes la variable Y, que era ESP M, y luego la X, que es MORT INF. El programa devuelve, entre otros que ahora no nos interesan, los resultados numéricos siguientes

	A	B	C
1	Resumen		
2			
3	<i>Estadísticas de la regresión</i>		
4	Coefficiente de correlación múltiple	0,968427147	
5	Coefficiente de determinación R^2	0,937851138	
6	R^2 ajustado	0,93737307	
7	Error típico	2,811841159	
8	Observaciones	132	

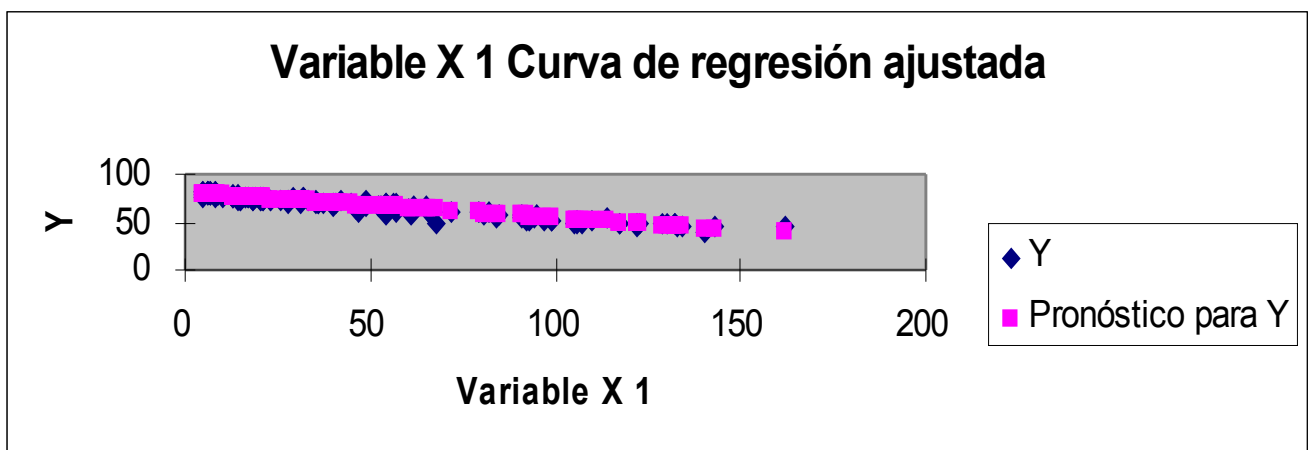
y

16		
17	Intercepción	80,10760518
18	Variable X 1	-0,26063532

que nos dice que el modelo ajustado es

$$Y = -0,26063532 * X + 80,10760518$$

Vemos que la pendiente es negativa, como esperábamos. Si además hemos indicado que muestre la curva de regresión ajustada, proporcionará

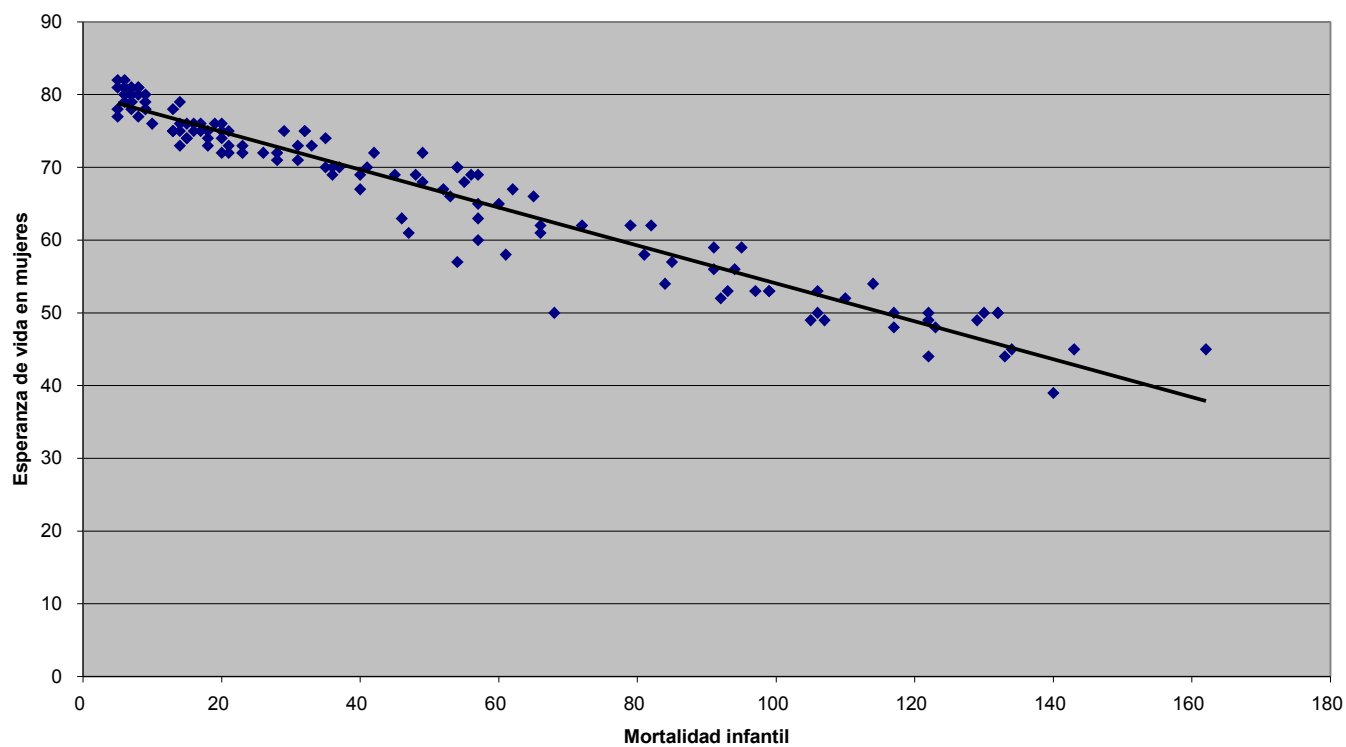


Los puntos rosas deben estar alineados. Pero para agregar a un gráfico una línea de tendencia (que coincide con la recta de regresión cuando suponemos que la tendencia es lineal), hay que entrar en el menú

*Gráfico --> Agregar línea de tendencia*

Seleccionando el tipo lineal, resulta

Diagrama de dispersión



Universidad Complutense de Madrid

└ Facultad de Ciencias Económicas y Empresariales

└ Departamento de Estadística e Investigación Operativa II

└ David Casado de Lucas

2 de marzo del 2012