ESTADÍSTICA

Prácticas con Microsoft Excel®

Departamento de Estadística e Investigación Operativa

Universitat de València

Índice

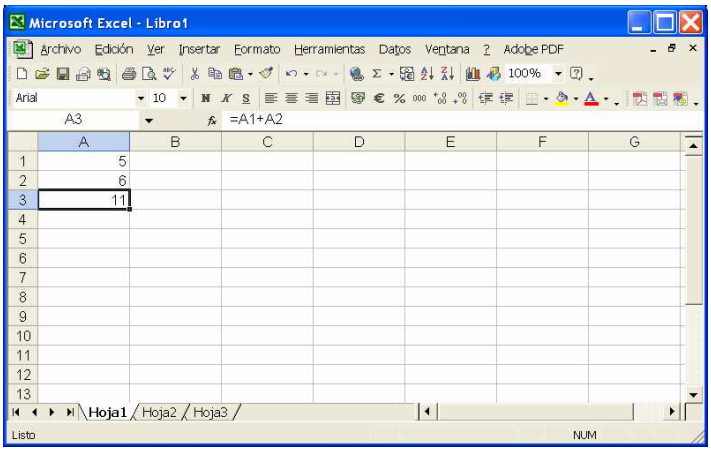
Este cuaderno describe el uso de algunas funciones del paquete de software *Microsoft Excel®*. En concreto está orientado a la resolución de problemas estadísticos en el contexto de las ciencias sociales. Hemos utilizado la versión *Excel 2003®* para su elaboración.

Francisco Montes
Rafael Martí

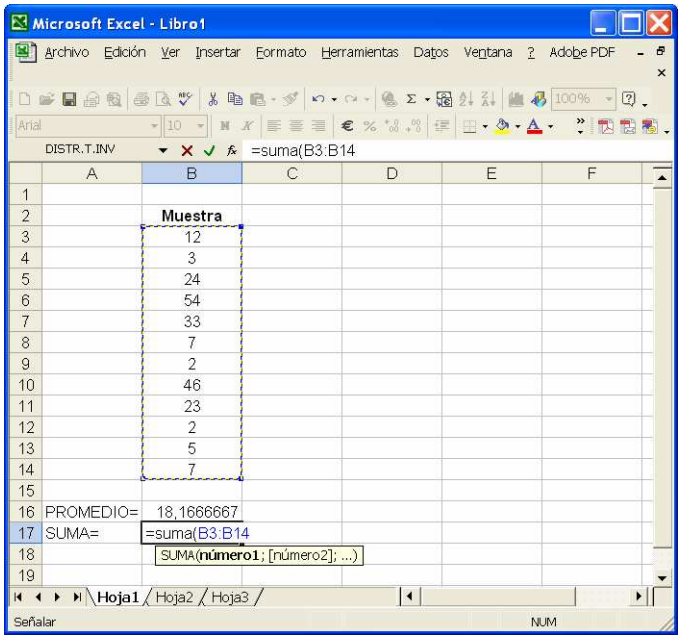
Departamento de Estadística e Investigación Operativa
Universitat de València

1 Descripción de Datos

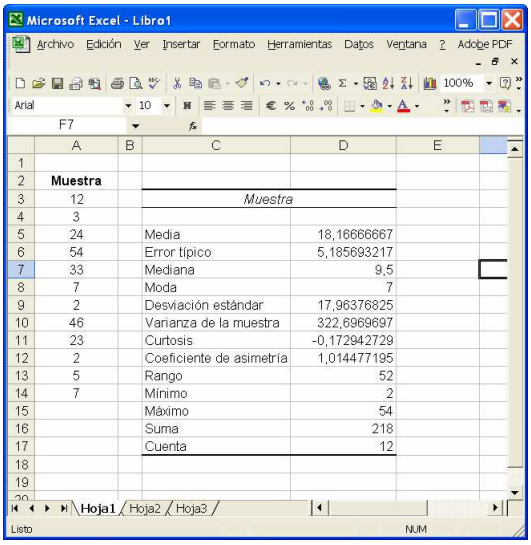
Al abrir el programa nos encontramos con una cuadrícula en la que podemos escribir tanto texto como números. Excel es una *hoja de cálculo*, lo que quiere decir que su propósito es precisamente calcular expresiones matemáticas. En las casillas o celdas de la ventana de Excel podemos introducir tanto números como expresiones. Así por ejemplo, si en la casilla A1 hemos introducido el número 5, y en la casilla A2 el número 7 y queremos calcular su suma, podemos introducir en la casilla A3 la expresión "=A1+A2" obteniendo el valor de dicha suma. Notad que el símbolo "=" indica que el programa ha de calcular la expresión que viene a continuación y no se trata de un mero texto a insertar. En ocasiones podemos ver que se intercala el símbolo "$" junto a la referencia de una celda, por ejemplo "A2". Esto indica que la referencia de la celda es absoluta y no relativa; es decir, que si copiáramos la expresión en otra celda, al ser absoluta la formula quedará tal cual está, pero si es relativa (no lleva los símbolos $) modificará la fórmula que copiamos.



Para comenzar a trabajar con una muestra, una vez introducidos sus datos, utilizaremos las funciones que Excel nos proporciona. Por ejemplo, la función =SUMA(*Rango*) o =PROMEDIO(*Rango*) calculan respectivamente la suma y el promedio de una muestra. Si los datos están contiguos, el *Rango* se especifica poniendo la referencia de la primera celda de la muestra, el símbolo ":", y la referencia de la última. Por ejemplo: =SUMA(B3:B14).

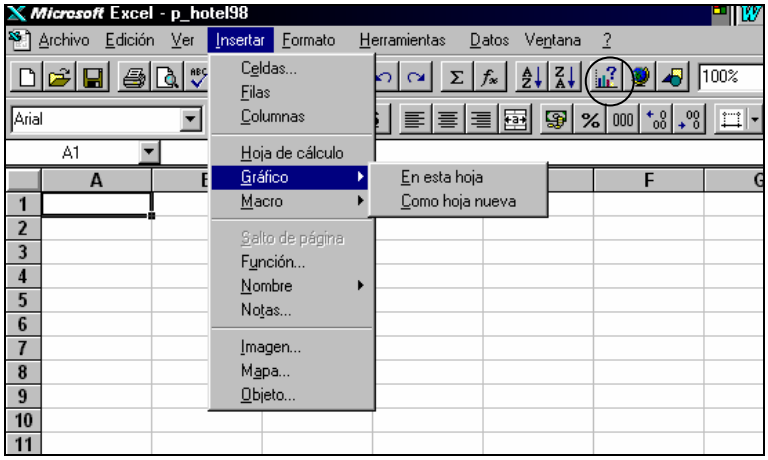


Dentro del menú Herramientas, podemos encontrar la opción *Análisis de Datos*. Si no se encuentra, la tendremos que instalar mediante la opción *Complementos*. Una vez instalada, seleccionamos la opción Estadística Descriptiva, que nos proporciona fácilmente las medidas de centrado y dispersión más habituales de la muestra.



2. Representaciones Gráficas

La hoja de cálculo EXCEL® dispone de un asistente para gráficos que, mediante un sencillo cuadro de diálogo, permite elegir fácilmente la gráfica más adecuada a los datos que queremos representar. El asistente puede ser activado, indistintamente, a través de un icono en la barra de herramientas o mediante la opción Gráfico del apartado Insertar en el menú principal, tal como se muestra en la figura. Esta segunda alternativa permite insertar la gráfica en una nueva hoja de cálculo que se abre automáticamente. En ambos casos, el cuadro de diálogo aparece de inmediato en la pantalla y va desarrollándose a medida que *dialogamos* con él.



Para explicar los distintos pasos a seguir lo más conveniente será hacerlo a partir de una tabla cuyos datos queramos representar. La tabla 1 contiene la distribución, por categorías de los establecimientos, de las plazas hoteleras del año 1998 en la Comunidad Valenciana y en todo el territorio español. Estos datos admiten dos tipos de representación gráfica, la de cada territorio por separado y una conjunta que permita comparar las distribuciones en ambos territorios. Como el orden de magnitud de los datos es distinto en cada territorio, parece aconsejable, al menos para la representación conjunta, utilizar los datos porcentuales. Veamos paso a paso cómo obtener las gráficas.

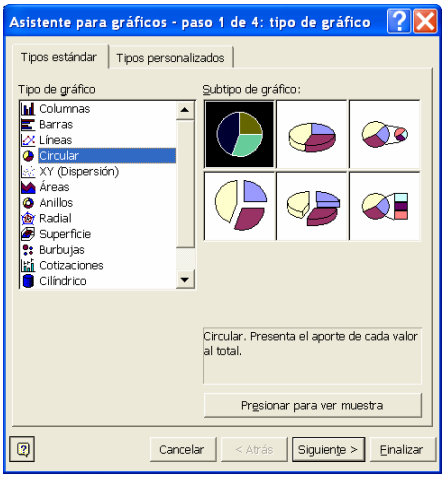
Plazas Hoteleras

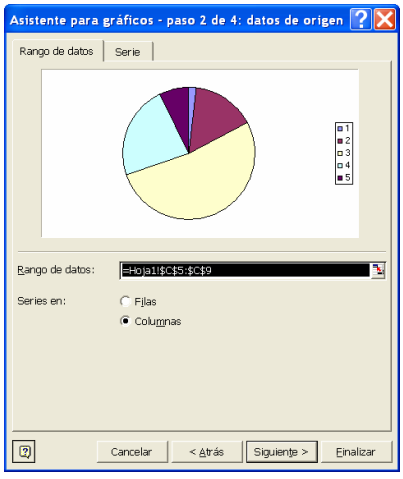
	C. Valenciana		España	
categoría	**98**	**%98**	**98**	**%98**
5 estr.	1.602	2,08	24.962	2,61
4 estr.	11.769	15,28	238.700	24,98
3 estr.	40.179	52,17	467.792	48,96
2 estr.	17.962	23,32	145.857	15,27
1 estr.	5.501	7,14	78.097	8,17
	77.013		**955.408**	

Tabla 1.- Distribución de las plazas hoteleras por categorías

Los recursos de la Comunidad Valenciana vamos a representarlos mediante un diagrama de sectores. Al activar el Asistente para gráficos, pulsando sobre él con el ratón, aparece sobre la hoja de cálculo una cruz que se desplaza con el ratón y mediante la cual hemos de seleccionar el espacio (un rectángulo) sobre el que queremos que aparezca la gráfica. Las dimensiones iniciales son irrelevantes, por cuanto el rectángulo puede redimensionarse con facilidad una vez finalizado el proceso, lo que permite también una modificación del aspecto de la gráfica. Los pasos a seguir a continuación son los siguientes:

En el primer paso, el Asistente para gráficos nos muestra los distintos tipos de gráficas. Basta con pulsar con el ratón sobre cualquiera de ellas para llevar a cabo la elección. El tipo elegido aparece entonces sobre un fondo negro, tal como se aprecia en la figura. En el segundo paso, Seleccionamos aquella parte de la tabla que deseamos representar. La selección se lleva a cabo mediante arrastre sobre la tabla. La tecla *Control* permite seleccionar partes no contiguas de la tabla.





En los siguientes pasos seguimos las indicaciones del asistente hasta llegar a Terminar. La gráfica aparece en el rectángulo que hayamos elegido al principio, de donde es posible copiarla y trasladarla a cualquier otro documento, como hemos hecho nosotros para incluirla en este texto. En la figura 1 mostramos la gráfica obtenida mediante el anterior proceso y también un diagrama de barras para los mismos datos.

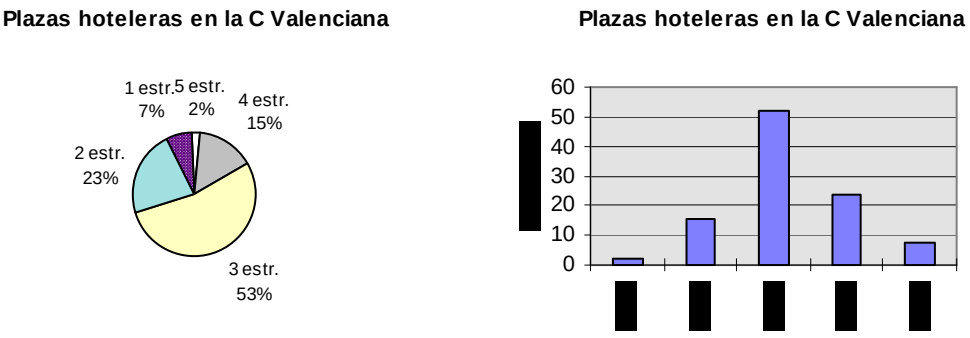


Figura 1.- Representaciones gráficas de los datos relativos a plazas hoteleras en la C. Valenciana

Si queremos comparar los datos de ambos territorios, un diagrama de dobles barras puede ser lo más apropiado. Seleccionaremos en primer lugar las columnas adecuadas en nuestra tabla y elegiremos después, paso 2, el tipo de gráfica deseado. Las dobles barras se dibujan automáticamente puesto que el Asistente para gráficos reconoce la presencia de dos series de datos. El resultado se muestra en la figura 2. Hay que advertir que los encabezamientos originales de ambas columnas eran los mismos (%98), por lo que han sido cambiados previamente para evitar confusión en el etiquetado.

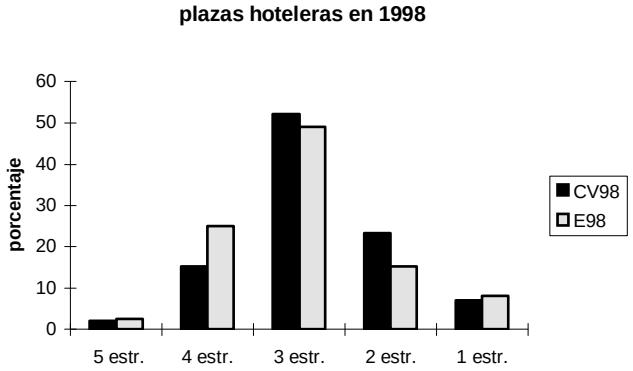


Figura 2.- Distribución de las plazas hoteleras en España y en la CV (año 1998)

Ya hemos visto que cada tipo de datos requiere un tipo de gráficas, sin que ello signifique que haya de ser necesariamente un solo tipo de gráfica el que mejor los describa visualmente. Afortunadamente el Asistente para gráficos permite llevar a cabo el proceso con rapidez y facilidad, por lo que podemos *probar* en muy poco tiempo diversas gráficas.

La tabla 2 recoge los datos de la evolución del precio medio del m^2 construido en España a lo largo de varios semestres. Este tipo de datos puede representarse utilizando diagramas de barras, pero sin duda una línea que una los distintos precios permitirá seguir su evolución mucho mejor.

evolución precio m^2 construido

dic-94	jun-95	dic-95	jun-96	dic-96	jun-97	dic-97	jun-98	dic-98	jun-99
158.700	161.500	164.500	166.800	166.700	168.100	172.700	174.100	181.200	188.700

Tabla 2.- Evolución del precio del m^2 construido

Entre los distintos tipos de gráficas podemos elegir la de Líneas, con una de cuyas variantes hemos llevado a cabo la representación de la figura 3, en la que se aprecia fácilmente la evolución alcista del precio de la vivienda a lo largo del período Dic94-Jun99.

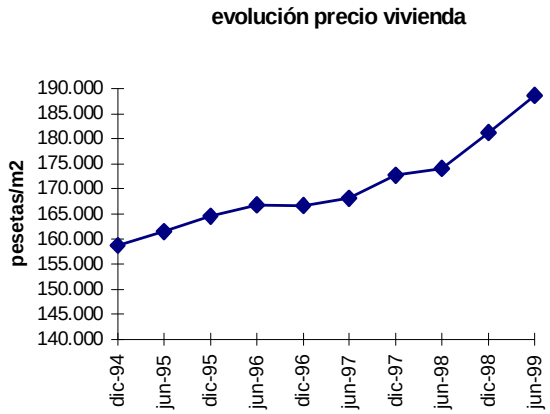


Figura 3.- Evolución del precio medio del m^2 construido

3 Histogramas y tablas de frecuencias

La hoja de cálculo EXCEL® dispone también de utilidades que permiten obtener tablas de frecuencias e histogramas para variables estadísticas a partir de datos desagregados. Como hemos hecho anteriormente, vamos a explicar el procedimiento mediante un ejemplo concreto.

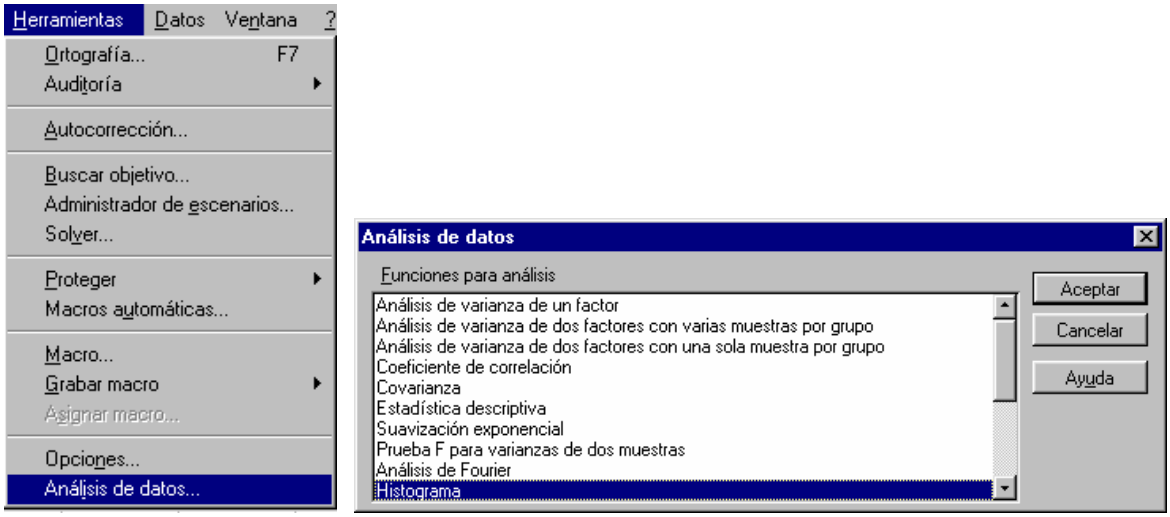
La tabla 3 recoge, parcialmente, el resultado de una encuesta a la que fueron sometidas 250 personas con edad igual o superior a 15 años (archivo encuesta.xls). Las variables y el significado de sus códigos asociados son los siguientes:

- La primera columna indica el número de caso
- **Sexo:** indica el sexo del entrevistado, **v** = varón, **m** = mujer
- **E_civil:** indica el estado civil, **1** = casado/a, **2** = soltero/a, **3** = viudo/a, **4** = div/sep
- **Edad:** edad expresada en años
- **Niv_ins:** nivel de instrucción, **1** = analfabeto/a, **2** = sin estudios, **3** = est. primarios, **4** = BUP o similares, **5** = est. universitarios
- **Peso:** peso expresado en kilogramos
- **Altura:** altura expresada en centímetros
- **E_penal:** opinión sobre el adelanto de la edad penal, **1** = a favor, **2** = en contra, **3** = ns/nc
- **35horas:** opinión sobre la semana laboral de 35 horas **1** = a favor, **2** = en contra, **3** = ns/nc
- **C_alcohol:** consumo medio diario de alcohol medido en el equivalente a vasos de vino de 200cc, la escala va de 1 a 5, indicando esta última cifra 5 o más vasos diarios

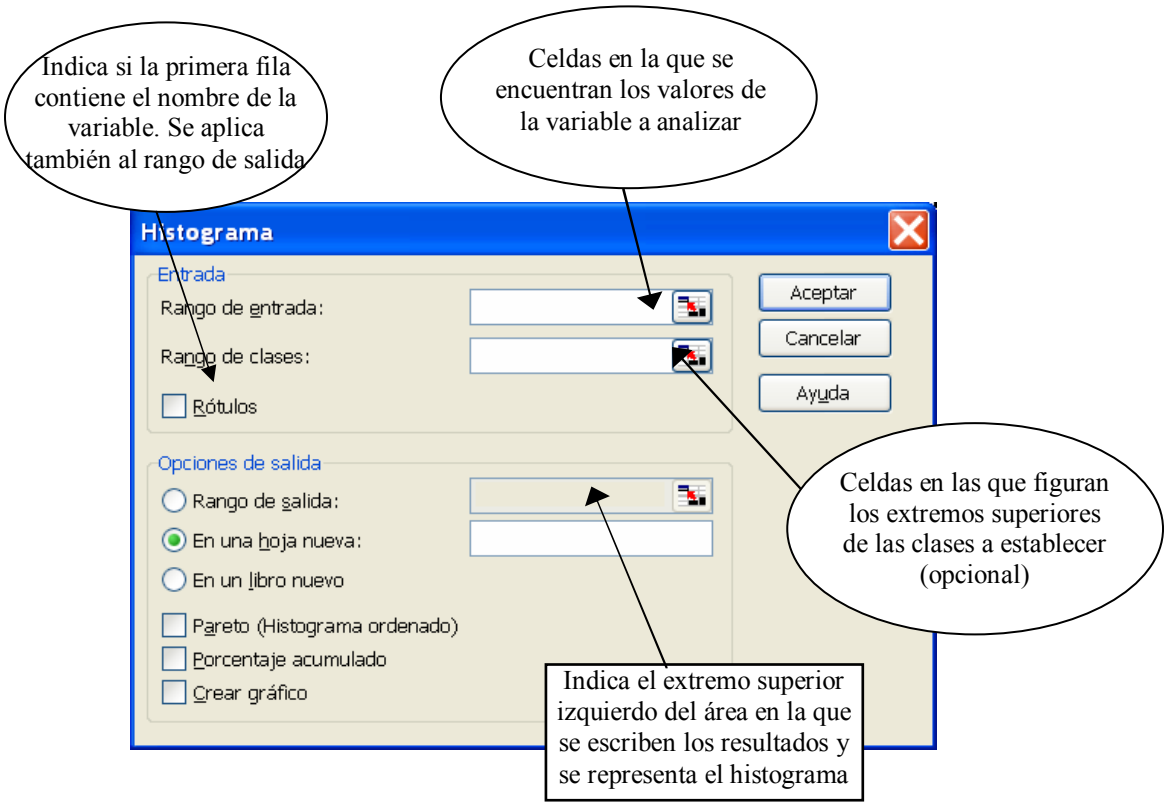
	SEXO	E_CIVIL	EDAD	NIV_INS	PESO	ALTURA	E_PENAL	35HORAS	C_ALCOHOL
01	v	1	63	3	80,30	190	1	1	3
02	v	1	79	4	56,16	155	3	1	2
03	m	1	52	3	64,37	151	3	1	2
04	m	3	41	3	63,02	146	2	2	2
05	v	2	18	4	75,50	164	2	1	3
06	m	2	68	3	35,00	136	3	2	2
07	v	2	35	2	62,79	145	1	1	2
08	m	2	46	2	78,92	190	2	1	3
09	m	2	20	3	58,27	171	2	1	0
10	v	1	61	4	52,17	159	2	2	2
11	m	1	69	3	70,82	169	1	2	2
12	m	2	50	3	41,10	167	2	1	3
13	m	1	67	2	49,46	171	2	3	1

Tabla 3.- Reproducción parcial de las 250 observaciones contenidas en el archivo encuesta.xls

La distribución de frecuencias y el histograma de la variable *C_alcohol* lo obtendremos mediante la macro Análisis de Datos que es una de las opciones del menú Herramientas (si no aparece, seleccionar Complementos para instalarlo). Al activar la macro Análisis de Datos se despliega un segundo cuadro entre cuyas opciones figura la de Histograma, tal y como mostramos a continuación.

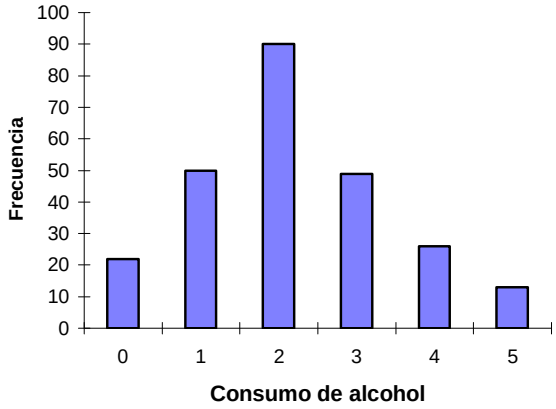


El cuadro de diálogo correspondiente al Histograma nos solicita la información necesaria que le permita identificar la variable cuyo histograma queremos obtener, las clases a establecer y el lugar donde aparecerán los resultados, que puede ser en la misma hoja o en una hoja distinta. El cuadro correspondiente al histograma de la variable *C_alcohol* aparece completo a continuación.



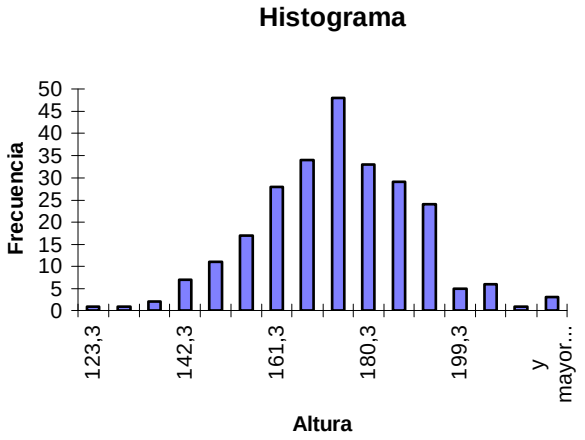
El resultado final es la tabla de frecuencias que figura abajo y el diagrama de barras (histograma) que aparece al lado. Hay que señalar que el aspecto del histograma puede ser modificado, para ello basta hacer doble *clic* sobre la gráfica para que aparezca un cuadro de diálogo que permite acceder a sus características (forma, diseño, fuentes en los ejes, colores, etc.).

Clase	Frecuencia
0	22
1	50
2	90
3	49
4	26
5	13



La variable *C_alcohol* es una variable cuantitativa discreta y ha resultado sencillo establecer el número de clases para el diagrama de barras (histograma). En el caso de variables continuas el proceso puede resultar más complicado, aunque siempre existe la posibilidad de dejar en blanco el rango de clases, recordemos que es un campo opcional, y que el procedimiento determine el número de clases que considere más conveniente. Así lo hemos hecho con la variable altura, obteniendo la distribución de frecuencias y el histograma que mostramos a continuación.

Clase	Frecuencia
123,3	1
129,6	1
135,9	2
142,3	7
148,6	11
154,9	17
161,3	28
167,6	34
173,9	48
180,3	33
186,6	29
192,9	24
199,3	5
205,6	6
211,9	1
y mayor...	3



Siempre podemos definir nosotros las clases fijando los extremos superiores de las mismas y escribiéndolos en las celdas del rango de clases. Conviene tener presente que el procedimiento añade siempre una última clase, la de aquellos valores de la variable que exceden la última de las clases fijadas por nosotros. Esta clase viene siempre etiquetada como *y mayor...* En el caso de la variable *C_alcohol* hemos proporcionado solamente cuatro clases (0,1,2,3,4) y la quinta ha resultado ser la de aquellos valores mayores que 4, es decir, 5.

3 Tablas dinámicas

Si queremos obtener la distribución de frecuencias de la variable *Sexo* nos encontraremos con el inconveniente de que el procedimiento Histograma admite solamente valores numéricos a la hora de definir las clases, y nuestra variable es cualitativa y toma los valores *h*-hombre, *m*-mujer. Existe una solución que consiste en definir una nueva variable, numérica en este caso, que tome, por ejemplo, el valor 1 para los hombres y el 2 para las mujeres. La definición es sencilla utilizando una función condicional. A partir de aquí podemos proceder como anteriormente.

Pero nuestro interés es introducir otra utilidad de la hoja de cálculo, la conocida como Tablas dinámicas, que permite obtener nuevas tablas que resumen los datos provenientes de otras preexistentes. En nuestro caso se trata de contar las apariciones de hombres (*h*) y mujeres (*m*) entre los 250 encuestados. Para ello añadiremos una nueva variable, *V_muda*, que vale 1 en todos los casos y mediante una tabla dinámica la sumaremos a lo largo de cada uno de los grupos que establece la variable *Sexo*. Al asistente para Informe de Tablas y gráficos dinámicos se accede a través del menú Datos, obteniendo la ventana siguiente:

Los restantes pasos del procedimiento permiten, siempre mediante los cuadros de diálogo habituales, introducir la información necesaria para elaborar la tabla deseada. Comenzamos por introducir el rango o parte de la tabla que contiene las variables a utilizar en la elaboración de la tabla dinámica. Para ello señalaremos con el ratón el extremo superior izquierdo del área con los datos, y el extremo inferior derecho.

Tras especificar el rango y pulsar Siguiente, aparece una nueva ventana y en la casilla Opciones es donde definimos la estructura de la tabla que deseamos tal y como muestra la figura siguiente. En el ejemplo le hemos pedido que sume los valores de *V_muda* para los grupos que establece la variable *Sexo*, y que disponga el resultado por filas. El procedimiento consiste en arrastrar las variables al campo deseado. En el campo de Datos las opciones son varias (suma, media, varianzas, max, min, ...) y se accede a ellas haciendo doble *clic* sobre la función que, por defecto, haya aparecido al arrastrar sobre el campo.

Finalmente indicamos si la nueva tabla la ha de crear en otra hoja o en la actual, en cuyo caso nos solicita donde queremos que escriba la nueva tabla (solo la celda inicial = superior izquierda) y un nombre para la misma (opcional). Al pulsar Terminar el resultado final es la tabla siguiente:

SEXO	Total
m	124
v	126
Total	**250**

Es posible construir tablas cruzadas mediante esta función de Tablas dinámicas sin más que arrastrar a los campos Fila y Columna las variables que deseemos cruzar. En el ejemplo que sigue hemos cruzado las variables *Sexo* y *C_alcohol*.

	C_ALCOHOL						
SEXO	**0**	**1**	**2**	**3**	**4**	**5**	**TOTAL**
m	20	26	41	25	9	3	**124**
v	2	24	49	24	17	10	**126**
TOTAL	**22**	**50**	**90**	**49**	**26**	**13**	**250**

4. Regresión

En esta sección vamos a analizar la relación lineal entre dos variables, y en su caso obtener la recta de regresión. Consideremos las dos variables de la tabla siguiente:

ingresos (*x*)	ahorros (*y*)
1,9	0,19
2,4	0,30
2,1	0,43
3,3	0,52
4,6	0,35
3,0	0,41
2,4	0,19
2,8	0,37
1,2	0,15

Considerando que la variable *x* está situada en la columna B de la tabla y sus valores van de la fila 4 a la 12, los valores necesarios para realizar el análisis sobre la relación lineal de ambas variables se pueden obtener mediante las siguientes expresiones:

Promedio de *x*	=Promedio(B4:B12)	2,633
Promedio de y	=Promedio(C4:C12)	0,323
Varianza de x	=Var(B4:B12)	0,932
Covarianza	=Covar(B4:B12;C4:C12)	0,061
Correlación	=Coef.de.correl(B4:B12;C4:C12)	0,565

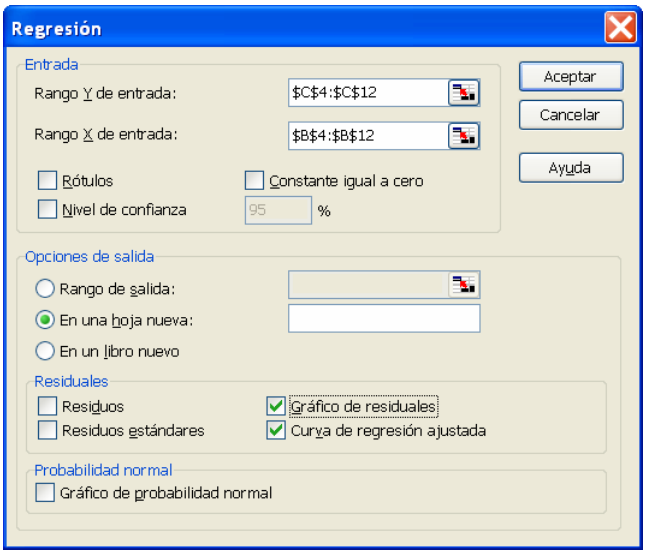
Hemos de notar que Excel calcula la covarianza dividiendo entre *n* y no entre *n*-1, por lo que para obtener el valor que necesitamos en este caso, basta con hacer 0,061*9/8 = 0,068.

Así pues los coeficientes de la recta de regresión los obtendremos mediante las expresiones:

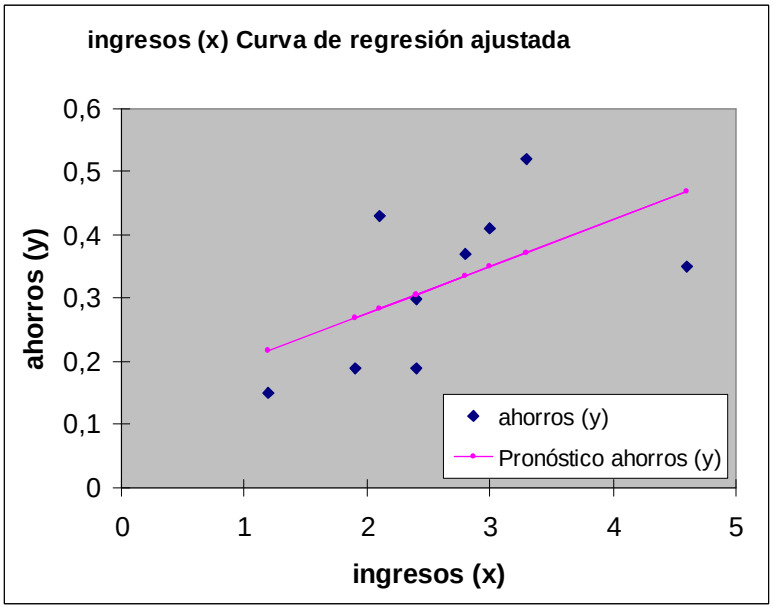
Pendiente	a=Covar / Var(x)	0,0736
Ordenada en origen	b=promedio(y) - a promedio(x)	0,1295

Alternativamente, podemos obtener directamente todos estos cálculos mediante la opción Regresión que se encuentra dentro de la ventana de Análisis de Datos en el menú Herramientas. Aparece la siguiente

ventana en la que podemos especificar el rango de los datos y si queremos el gráfico con los puntos y la recta ajustada



Además de obtener unas tablas con la información sobre los estadísticos mencionados, recomendamos unir mediante una línea los puntos ajustados para ver la recta de regresión en el gráfico.



5. Análisis de muestras

La herramienta Análisis de datos cuenta con otras utilidades estadísticas de interés. Entre ellas, las que nos permiten obtener los estadísticos descriptivos de una o varias variables, comparar las medias de varias poblaciones, extraer muestras aleatorias de un conjunto de datos, ... A través de un nuevo ejemplo veremos el funcionamiento de algunas de ellas.

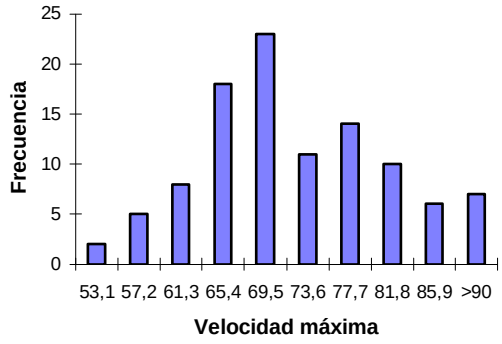
La tabla 4 contiene las velocidades máximas anuales de las rachas de viento en dos ciudades escocesas, Edimburgo y Paisley (archivo viento.xls).

edimburgo										
1918-1967	78	66	76	75	73	81	79	74	85	81
	81	87	66	65	74	86	76	69	85	87
	66	82	78	75	76	74	73	69	68	68
	76	75	79	85	69	81	75	74	87	67
	88	71	78	79	71	73	68	83	88	73
paisley	66	64	63	59	64	63	69	63	59	63
1914-1967	67	67	63	90	68	55	68	64	64	63
	67	73	83	49	70	59	73	62	58	63
	62	55	66	59	69	69	55	72	63	55
	68	66	67	75	49	63	60	70	77	56
	62	62	59	59						

Tabla 4.- Series de velocidades máximas anuales de las rachas de viento en mph en Edimburgo y Paisley

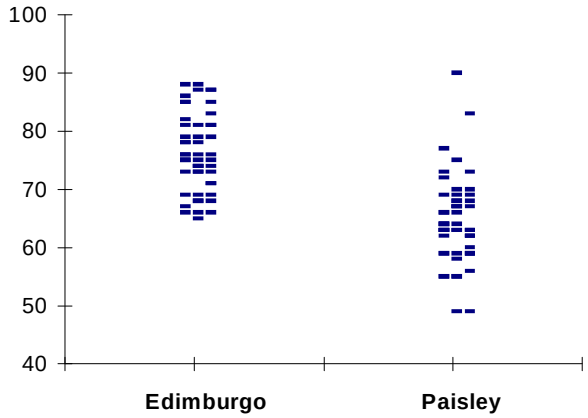
El carácter anual de cada observación y la distancia entre ambas ciudades, aproximadamente 100 kms., permiten mantener razonablemente la hipótesis de que tanto las observaciones de cada muestra como las muestras entre sí son independientes. El objetivo del estudio de ambas series de datos es obtener alguna conclusión acerca de cual de las dos ciudades soporta rachas de viento más fuerte.

La observación de una adecuada representación gráfica de los datos puede permitirnos aventurar alguna hipótesis inicial. La más inmediata de las representaciones gráficas es un histograma conjunto de todos los datos. El resultado es el siguiente:



Clase	Frecuencia
53,1	2
57,2	5
61,3	8
65,4	18
69,5	23
73,6	11
77,7	14
81,8	10
85,9	6
>90	7
Total	**104**

Un gráfico de dispersión, también disponible en el Asistente para gráficos, en el que eje de las X sean las ciudades y el eje de las Y las velocidades, puede también aportar información interesante. Dicho gráfico se muestra a continuación.



El histograma es bimodal, presenta dos clases con máxima frecuencia relativa (la 69.5 y la 77.7), lo que parece indicar que las muestras proceden de dos poblaciones distintas. Esta hipótesis es corroborada por el gráfico de dispersión, en el que se aprecia que las rachas máximas son, en general, menores en Paisley que en Edimburgo, si bien en el primer caso los valores son mucho más dispersos (mayor varianza).

El siguiente paso será obtener las características numéricas de ambas muestras. Para ello elegiremos la opción Estadística descriptiva de la herramienta Análisis de datos, y mediante un sencillo diálogo con el cuadro que despliega obtendremos el siguiente resultado

	edimburgo	paisley
Media	76,260	64,389
Error típico	0,939	0,998
Mediana	75,5	63
Moda	81	63
Desviación estándar	6,642	7,337
Varianza de la muestra	44,115	53,827
Curtosis	-0,931	2,380
Coeficiente de asimetría	0,152	0,818
Rango	23	41
Mínimo	65	49
Máximo	88	90
Suma	3813	3477
Cuenta	50	54
Nivel de confianza(95,0%)	1,888	2,003

Estos valores parecen confirmar todo cuanto las gráficas nos permitían adivinar. Pero afortunadamente la Inferencia Estadística nos proporciona una herramienta para no tener que basar nuestra decisión, acerca del distinto comportamiento del viento en una y otra ciudad, en aspectos subjetivos*: el contraste de hipótesis sobre la media de dos poblaciones*. El contraste es accesible, nuevamente, a través de la herramienta Análisis de datos, en concreto la opción Prueba t para dos muestras suponiendo varianzas iguales nos permite llevarlo a cabo. El resultado se recoge en la tabla siguiente y conduce a *rechazar* la hipótesis de igualdad de medias (Diferencia hipotética entre ambas medias igual a 0) concluyendo que las observaciones no aportan suficiente evidencia en contra de que las rachas máximas son más fuertes en Edimburgo.

Prueba t para dos muestras suponiendo varianzas iguales

	Edimburgo	Paisley
Media	76,26	64,39
Varianza	44,11	53,83
Observaciones	50	54
Varianza agrupada	49,16	
Diferencia hipotética	0	
Grados de libertad	102	
Estadístico t	8,63	
P(T<=t) una cola	4,38625E-14	
Valor crítico de t (una cola)	1,66	
P(T<=t) dos colas	8,7725E-14	
Valor crítico de t (dos colas)	1,98	