TALLER DE ESTADÍSTICA

3. ANÁLISIS DE DATOS Y REGRESIÓN CON PROGRAMAS INFORMÁTICOS

MAURICIO CONTRERAS

DESCRIPCIÓN DE DATOS ESTADÍSTICOS CON EXCEL

Introducción

Vamos a ver cómo podemos usar la hoja de cálculo Excel para describir y analizar datos estadísticos. En los ejemplos que siguen se observará la potencia del ordenador para construir gráficos estadísticos atractivos y para realizar con facilidad cálculos. De esta forma, podremos dedicar más tiempo a la reflexión sobre los conceptos y los datos y el foco de atención dejará de ser la correcta realización de los cálculos. Excel no es un programa diseñado específicamente para hacer estadística, pero dada su facilidad de acceso en los centros, pueden aprovecharse algunas funciones especiales para el tratamiento de datos y sus posibilidades gráficas para describir datos.

1. Distribución estadística

- Abre la hoja de cálculo **Microsoft Excel**.
- Abre de tu carpeta de trabajo **Estadística1** el libro **Unidimensional**. Utiliza la **Hoja1** de dicho libro.
- Completa la siguiente distribución estadística:

	A	B	C	D	E	F	G
1		**Frecuencia**			**Frecuencia acumulada**		
2	**Modalidad**	**Absoluta**	**Relativa**	**Porcentual**	**Absoluta**	**Relativa**	**Porcentual**
3	X_i	n_i	f_i	p_i	N_i	F_i	P_i
4	3	4					
5	4	8					
6	5	18					
7	6	12					
8	7	5					
9	8	3					
10	**Total**						

Sigue los siguientes pasos:

- Haz clic en la celda **B10**, elige el botón **Σ Autosuma**, selecciona el rango **B4: B9** y haz clic en el botón 4 **Introducir**.
- En la celda **C4** introduce **=B4 / $B$10** y arrastra el controlador de relleno de la celda hasta **C9**.
- En la celda **D4** introduce **=C4*100** y arrastra el controlador de relleno de la celda hasta **D9**.
- En la celda **E4** introduce **=SUMA($B$4: B4)** y arrastra el controlador de relleno de la celda hasta **E9**.
- En la celda **F4** introduce **=SUMA($C$4: C4)** y arrastra el controlador de relleno de la celda hasta **F9**.
- En la celda **G4** introduce **=SUMA($D$4: D4)** y arrastra el controlador de relleno de la celda hasta la celda **G9**.
- Selecciona la celda **B10** y arrastra el controlador de relleno de la celda hasta **D10**.
- Ajusta todos los datos a dos decimales y comenta los resultados obtenidos.

2. Gráficos estadísticos

- Haz clic en la etiqueta **Hoja2** y comprueba que están introducidos los siguientes datos (en caso contrario, introdúcelos):

	A	B
1	**Congreso de los Diputados**	
2	**Partidos políticos**	**Escaños**
3	$\mathbf{X_i}$	$\mathbf{n_i}$
4	PP	156
5	PSOE	141
6	IU	16
7	OTROS	35
8	**Total**	348

Dibuja un diagrama de sectores tridimensional para esta distribución estadística. Para ello:

- Haz clic en el botón **Asistente para gráficos**.
- **Paso 1 de 4 – Tipo de gráfico**: Circular. **Subtipo de gráfico**: Circular con efecto 3D y haz clic en el botón **Siguiente**.
- **Paso 2 de 4 – Datos de origen**: en la ficha **Rango de Datos**, selecciona con el ratón **B4: B7**, en la ficha **Serie** en **Rótulos de las categorías** selecciona el rango **A4: A7** y haz clic en el botón **Siguiente**.
- **Paso 3 de 4 – Opciones de gráfico**: rellena la ficha **Títulos**: Congreso de los Diputados, desactiva en la ficha **Leyenda**: **Mostrar leyenda**, en la ficha **Rótulos de datos** marca **Mostrar rótulo y porcentaje** y haz clic en **Siguiente**.
- **Paso 4 de 4 – Ubicación del gráfico**: elige en la misma hoja, y haz clic en **Terminar**.
- Mejora la presentación del gráfico a través del menú contextual que se obtiene al hacer clic con el botón derecho del ratón.

3. Medidas de centralización y dispersión

- Haz clic en la etiqueta Hoja3 y comprueba que en dicha hoja están disponibles los siguientes datos:

	A	B	C	D	E
1	**Estudio de la natalidad**				
2	2	2	1	5	2
3	3	1	2	1	3
4	1	5	0	2	1
5	0	4	1	3	2
6	1	1	2	1	0
7	Nº de datos				
8	**Medidas de centra lización**				
9	Media				
10	Mediana				
11	Moda				
12	**Medidas de disper sión**				
13	Rango				
14	Varianza				
15	Desv. típica				

Calcula todas las medidas de centralización y de dispersión de esta tabla estadística. Para ello:

- En la celda **E7** escribe la fórmula **=CONTAR (A2: E6).**

Obtén las medidas de centralización:
- En la celda **E9** escribe la fórmula **=PROMEDIO (A2: E6)**.
- En la celda **E10** escribe la fórmula **=MEDIANA(A2: E6)**.
- En la celda **E11** escribe la fórmula **=MODA(A2:E6)**.

Obtén las medidas de dispersión:
- En la celda **E13** escribe la fórmula **=MAX(A2: E6) – MIN(A2: E6)**.
- En la celda **E14** escribe la fórmula **=VARP(A2: E6)**.
- En la celda **E15** escribe la fórmula **=DESVESTP(A2: E6)**.
- Comenta e interpreta los resultados obtenidos.

4. Tres ejemplos comentados

Ejemplo 1.- *Estudiando el número de hijos de 30 familias elegidas al azar en una ciudad se han obtenido los siguientes datos:*

1	*2*	*3*	*5*	*6*	*0*	*7*	*8*	*4*	*1*	*3*	*4*	*5*	*2*	*6*
5	*2*	*3*	*4*	*6*	*2*	*3*	*4*	*6*	*4*	*3*	*6*	*6*	*3*	*3*

Dibuja el diagrama de sectores.

Una vez abierta la hoja de cálculo Excel, introducimos los datos en la columna **A**, desde la fila 1 hasta la 30.

En la columna **C** introducimos los distintos valores de la variable (0, 1, 2, 3, 4, 5, 6, 7, 8) desde la fila 1 hasta la fila 9.

Para obtener la columna de las frecuencias absolutas utilizamos la función **CONTAR.SI( ).** La columna **D** contendrá dichos valores.

Para ello situaremos el cursor en la posición **D1**. Abrimos el menú **Insertar**, seleccionamos el comando **f_x Función**, y de las funciones estadísticas elegimos la función **CONTAR.SI**.

En la nueva ventana que aparece especificamos en **Rango A$1: A$30** y en **Criterio C1**, pulsando **Aceptar**.

Como resultado, en la celda **D1** aparecerá un 1, es decir, el número de veces que aparece el valor de la celda **C1**, desde la fila A1 hasta A30. Podemos repetir el procedimiento para todos los valores de la variable, cambiando el cursor de celda a la fila siguiente y en la ventana **Criterio** de la función **CONTAR. SI ( )** colocando el valor correspondiente.

No obstante, se puede realizar de una manera más automática, mediante las utilidades de **Copiar** y **Pegar**. Para ello se selecciona con el botón derecho del ratón la celda **D1**, pulsando con el botón izquierdo la opción de **Copiar**, con lo cual la celda **D1** aparecerá recortada con trazos intermitentes. Seleccionamos con el ratón las celdas donde deseamos copiar la fórmula situando el cursor en la celda **D2** y pulsando el botón izquierdo arrastramos el puntero del ratón hasta la celda **D9** (las celdas **D2** hasta **D9** deben aparecer en fondo negro), pulsando **ENTER** para terminar.

Una vez calculadas las frecuencias absolutas, para realizar el diagrama de sectores, seleccionamos en el menú **Insertar** la opción **Gráficos**. En dicha opción, **Excel** nos proporciona una gran variedad de gráficos a elegir.

Seleccionamos **Circular** y pulsamos **Siguiente** con el botón izquierdo del ratón. La siguiente ventana nos muestra el rango de valores que vamos a representar y los rótulos que deseamos que nos muestre. Para ello en la pantalla **2** (**Datos de origen**) deberemos pulsar en **Serie**. En la pantalla que aparece deberemos rellenar las casillas correspondientes a la casilla **Rótulos de las categorías =Hoja1!$C$1: $C$9** para indicar cuáles son los datos que vamos a representar y la casilla **Valores** con **=Hoja1!$D$1: $D$9** para indicar las veces que se encuentra cada dato repetido, es decir, las frecuencias absolutas de dichos valores.

Una vez completada dicha pantalla pulsamos **Siguiente**, y aparece la tercera pantalla del asistente para gráficos.

En dicha pantalla, en la ficha **Títulos** y en la casilla **Título del gráfico**, teclearemos **NÚMERO DE HIJOS**. A continuación pulsaremos con el botón izquierdo del ratón en la ficha **Rótulos de datos** y activaremos la opción **Mostrar rótulo** para que aparezca en el gráfico cada uno de los valores que estamos representado. Como se observa, **Excel** también permite mostrar el porcentaje de cada valor, ambos o nada.

Pulsamos **Siguiente** para pasar a la última pantalla del asistente en la que bastará con pulsar en **Terminar** para que aparezca el diagrama de sectores en la hoja de cálculo. Observa que si situamos el puntero del ratón sobre los distintos sectores del gráfico se muestra el número de veces que aparece cada valor, es decir, las frecuencias absolutas.

Ejemplo 2.- *Dada la siguiente información aparecida en un diario, representa los datos mediante un diagrama de barras:*

Uso de anticonceptivos	***1978***	***1997***
Ogino	*6,4*	*0,87*
Coitus interruptus	*23,6*	*1,49*
Diu	*0,5*	*5,68*
Preservativo	*5,0*	*21,0*
Píldora	*12,0*	*14,26*
Otros	*3,9*	*0,19*
No usan	*48,6*	*50,92*
Óvulos espermicidas	*—*	*0,1*
Diafragma	*—*	*0,29*
Ligadura de trompas	*—*	*5,2*

Si utilizamos el programa **Excel** para realizar el diagrama de barras, bastará con introducir:

- en la columna **A**, los distintos tipos de anticonceptivos, desde **A1** hasta **A10**;
- en la columna **B**, los porcentajes de las personas que utilizaban cada uno de ellos en 1978, y
- en la columna **C**, los porcentajes de las personas que utilizaban cada uno de ellos en 1997.

Los pasos a seguir son básicamente los mismos que en el ejemplo anterior: Seleccionar el comando **Insertar**, **gráficos**; en la primera pantalla del Asistente para gráficos, seleccionamos **Columnas**, para realizar el diagrama de barras, pulsando con el ratón en **Siguiente**.

En la pantalla 2 del Asistente nos aparece el diagrama de barras comparando los datos de las columnas **B** y **C**. Para ello, en la casilla **Rango de datos** hemos de **especificar =Hoja1!$B$1: $C$10** y hemos de seleccionar **Series en: Columnas**. Pulsando en **Siguiente** pasamos a la pantalla número 3 del Asistente.

En la pantalla 3 pulsamos **Siguiente** si no deseamos modificar nada o previamente pulsamos con el ratón en **Leyenda** para no mostrar los rótulos de series. Una vez pulsado en **Siguiente** pasamos a la pantalla 4 del Asistente, donde basta pulsar **Terminar** con el ratón.

Una vez aparezca el gráfico en la pantalla podemos modificar su tamaño pulsando en **Vista preliminar** de la barra de herramientas estándar.

Ejemplo 3.- *Al lanzar dos dados 30 veces y anotar la suma de las caras superiores hemos obtenido los datos que representamos en la siguiente tabla:*

Resultados	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Frecuencia	*1*	*2*	*4*	*3*	*2*	*1*	*4*	*3*	*5*	*4*	*1*

Halla los parámetros estadísticos de centralización, dispersión y posición.

Una vez introducidos los datos en las celdas **A1: A30** (para lo que hay que escribir un 2, dos 3, cuatro 4, tres 5, etc), seleccionamos el comando **Insertar, f_x Función**, apareciendo la ventana de diálogo donde seleccionamos **Estadísticas** y la función que queramos calcular.

- Para calcular la media seleccionamos la función **=PROMEDIO** del menú e introducimos el rango de valores **A1: A30**. Al hacer clic en **Aceptar** aparece el resultado.

- Para calcular los percentiles, hay que seleccionar la función **=PERCENTIL**. Como segundo parámetro introducimos el decimal entre 0 y 1 que indica el porcentaje. Por ejemplo, para obtener el percentil 45, en la casilla **Matriz** introducimos el rango **A1:A30** y en la casilla **K** introducimos 0,45. Haciendo clic en **Aceptar** obtendremos el valor del percentil.

- Para calcular la desviación típica, seleccionamos la función **=DESVESTP** y en la casilla **Número1** introducimos el rango **A1: A30**. Hacemos clic en **Aceptar** y se mostrará el valor de la desviación típica.

- Todas las medidas estadísticas se calculan de la misma forma: la moda (**=MODA(A1: A30)**), la mediana (**=MEDIANA(A1: A30)**), los cuartiles (**=CUARTIL(A1: A30, 1 2 ó 3)**) — 1 si es el primer cuartil, 2 si es el segundo y 3 si es el tercero—, la desviación media (**=DESVPROM(A1: A30)**), la varianza (**=VARP(A1: A30)**), el coeficiente de asimetría (**=COEFICIENTE.ASIMETRÍA(A1:A30)**) y el coeficiente de apuntamiento (**=CURTOSIS(A1: A30)**).

En la hoja de cálculo colocamos el puntero del ratón en las celdas donde queremos escribir las distintas medidas estadísticas:

Media	7,43333333
Moda	10
Mediana	8
Tercer cuartil (Q3)	10
Primer cuartil (Q1)	5
Percentil 45	8
Desviación media	2,57555556
Varianza	8,37888889
Desviación típica	2,89463105
Coeficiente de asimetría	–0,24933238
Coeficiente de apuntamiento	–1,28708607

ACTIVIDADES

- **CONTINENTES**

Representa en un diagrama de sectores la superficie de los continentes, utilizando los datos de la siguiente tabla (la superficie está expresada en millones de km^2:

Continente	**X**	Europa	Asia	África	América	Oceanía	Antártida
Superficie	**S**	10'05	44'15	30'50	41'98	8'97	13'18

- **GASTOS SANITARIOS**

Representa en el tipo de gráficos que quieras los Gastos de Sanidad en % del PIB de la Unión Europea. Utiliza para ello los datos de la siguiente tabla:

País	**Ale**	**Aus**	**Bél**	**Din**	**Esp**	**Fin**	**Fra**	**Gre**	**Hol**	**Irl**	**Ita**	**Lux**	**Por**	**Re.U**	**Sue**
1990	8'9	6'5	7'6	6'5	6'9	7'1	8'9	4'2	8'3	6'6	8'1	6'6	6'0	8'0	8'0
1992	9'3	7'4	8'1	6'6	7'2	7'5	9'4	4'4	8'8	7'1	8'5	6'6	6'9	8'2	9'3
1994	7'9	7'8	8'1	6'6	7'3	7'8	9'7	5'5	8'8	7'6	8'4	6'5	6'9	8'1	7'9

- **AGRUPACIÓN Y RECUENTO DE DATOS**

a) Contar datos cualitativos

Abre el libro **Recuento** situado en la carpeta **Estadística1**. Haz clic, si es preciso, en la etiqueta **Hoja1** y comprueba que contiene los siguientes datos:

	A	B	C	D	E
1	SF	SF	SF	Modalidad, X_i	Recuento, n_i
2	IS	SF	IS	IS	
3	BI	IS	SF	SF	
4	NT	BI	BI	BI	
5	SF	NT	NT	NT	
6	SB	SB		SB	
7	SF	NT			
8	IS	IS			
9	SF	SF			
10	BI	SF			

Selecciona el rango **E2: E6**. A continuación introduce la fórmula **=CONTAR. SI (A1: C10; D2: D6)** y pulsa al mismo tiempo la combinación de teclado **[CTRL] + [MAYÚS] + [ENTER]**. Observa que el resultado es la tabla de frecuencias absolutas correspondientes a las distintas calificaciones obtenidas.

b) Agrupar datos cuantitativos discretos

Haz clic en la etiqueta **Hoja2** y comprueba que contiene los siguientes datos:

	A	B	C	D	E	F
1	1	2	2	2	Modalidad, X_i	Recuento, n_i
2	2	3	3	3	1	
3	3	3	4	1	2	
4	4	2	2	4	3	
5	5	3	3	3	4	
6	4	2	4	2	5	
7	3	1	1	3		
8	2	4	5	4		
9	3	3	2	5		
10	3	2	3	1		

Selecciona el rango **F2: F6**. A continuación introduce la fórmula **=FRECUENCIA(A1:D10; E2: E6)** y pulsa la combinación de teclado **[CTRL] + [MAYÚS] + [ENTER]**. Comprueba que obtienes la tabla de frecuencias de las distintas modalidades 1, 2, 3, 4 y 5.

c) Agrupar datos cuantitativos continuos

Haz clic en la etiqueta Hoja3 del libro Recuento y comprueba que tiene los siguientes datos:

	A	B	C	D	E	F
1	70	61	65	Intervalos	Modalidad, X_i	Recuento, n_i
2	68	79	67	64	62	
3	60	70	70	68	66	
4	65	71	67	72	70	
5	80	72	77	76	74	
6	72	75		80	78	
7	71	72				
8	68	66				
9	65	67				
10	76	79				

Los datos de la columna D son los extremos superiores de los intervalos.

Selecciona el rango **F2: F6**. Introduce la fórmula **=FRECUENCIA(A1: C10; D2: D6)** y pulsa la combinación de teclado **[CTRL] + [MAYÚS] + [ENTER]**. Comprueba que obtienes la tabla de frecuencias correspondiente a cada uno de los intervalos.

5. Procedimientos rápidos con Excel

También con Excel podemos utilizar algoritmos rápidos que permiten describir los datos con suma facilidad. Veamos un ejemplo.

Diòxid de sofre i fum, per mesos, segons zones. 1998

La siguiente tabla muestra las concentraciones de dióxido de azufre y humo en una zona de tráfico denso de la ciudad de Valencia, durante cada mes de 1998.

	Zona Trànsit dens	
	Diòxid de sofre	**Fum**
Gener	26	44
Febrer	19	58
Març	22	49
Abril	20	31
Maig	22	33
Juny	24	28
Juliol	28	36
Agost	19	38
Setembre	24	50
Octubre	20	60
Novembre	23	55
Desembre	21	79
Mitjana Anual	22	47

Nota: Concentracions en micrograms / m3
Font: Laboratori Municipal. Ajuntament de València

a) Obtén los parámetros estadísticos correspondientes a las concentraciones de azufre y humo.
b) Describe gráficamente los datos, mediante histogramas.
c) Halla los percentiles de las dos distribuciones e interpreta los resultados.

- **PARÁMETROS ESTADÍSTICOS**

- Tras iniciar Excel, haz clic en el botón **Nuevo** para crear un nuevo libro de trabajo.

- Introduce los datos a partir de la celda **A1**, de forma que el primer dato numérico (correspondiente al dióxido de azufre del mes de Enero) aparezca en la celda **B3**.

- A continuación elige el comando **Herramientas / Análisis de datos** y en la ventana **Funciones para ánalisis** selecciona **Estadística descriptiva** y haz clic en **Aceptar**.

- Con el cursor en la caja **Rango de entrada**, selecciona el rango de celdas **$B$2: $C$14**. Comprueba que está activada la opción **Agrupado por Columnas** (ya que los datos están dispuestos en columnas). Haz clic en la casilla **Rótulos en la primera fila** para indicar que la primera fila del rango seleccionado contiene rótulos y no datos numéricos.

- Haz clic en **Rango de salida** e introduce en dicha caja la referencia de una celda vacía de la misma hoja, por ejemplo la **$E$1**. De esta forma visualizarás los resultados en la misma hoja, a partir de la celda indicada. También puedes hacer que se muestren los resultados en una hoja nueva o en un nuevo libro de trabajo.

- Haz clic en la casilla **Resumen de estadísticas** y haz clic en **Aceptar**.

Observa que como resultado se muestran, para cada variable, los siguientes parámetros estadísticos: **Media**, **Error típico**, **Mediana**, **Moda**, **Desviación estándar**, **Varianza de la muestra**, **Curtosis**, **Coeficiente de asimetría**, **Rango**, **Mínimo**, **Máximo**, **Suma** y **Cuenta** (número de datos). Por tanto, mediante este procedimiento podemos obtener simultáneamente los estadísticos más importantes de un conjunto de variables.

- **HISTOGRAMAS**

Vamos a construir un histograma para el dióxido de azufre, agrupando los datos en los intervalos [19, 21], [22, 23], [24, 26] y [27, +∞]. Para ello, en el rango **A17: A19**, teclea los extremos superiores de los tres primeros intervalos, es decir, **21**, **23** y **26**. A continuación sigue los siguientes pasos:

- Elige el comando **Herramientas / Análisis de datos** y en la ventana **Funciones para análisis** selecciona **Histograma** y haz clic en **Aceptar**.
- En la casilla **Rango de datos** introduce el rango de celdas **$B$3: $B$14**, correspondiente al dióxido de azufre.
- Haz clic en la casilla **Rango de clases** y, con el ratón, selecciona el rango de celdas **$A$16: $A$19**, procurando elegir la celda **A16**, aunque esté en blanco.
- Haz clic en la casilla **Rango de salida** y selecciona la celda **$C$19** para que Excel muestre los resultados a partir de dicha celda. También puedes mostrar los resultados en una hoja nueva o en un libro nuevo.
- Haz clic en la casilla **Crear gráfico** y haz clic en **Aceptar**.

Observa que, a partir de la celda **C19**, aparece una tabla con los límites superiores de cada clase y la frecuencia correspondiente a cada clase y se muestra también el histograma correspondiente. Observa que la última clase se denomina **y mayor…** para indicar los valores de la última clase.

ACTIVIDAD

Utiliza el mismo procedimiento para dibujar el histograma correspondiente a la concentración de humo. Elige de manera adecuada los límites superiores de cada clase. Una forma apropiada es procurar que la amplitud de cada clase sea igual a (Máximo−Mínimo) / Número de clases. Teniendo en cuenta que sólo disponemos de 12 datos, conviene agrupar los datos en tres o cuatro clases.

- **PERCENTILES**

Vamos a averiguar qué percentiles representan cada uno de los datos de las concentraciones de dióxido de azufre.

- Elige el comando **Herramientas / Análisis de datos** y en la ventana **Funciones para análisis** selecciona **Jerarquía y percentil** y haz clic en **Aceptar**.

- En la casilla **Rango de entrada** introduce el rango de celdas **$B$2: $B$14** correspondiente al dióxido de azufre.

- Haz clic en la casilla **Rótulos en la primera fila** para indicar a Excel que la primera fila del rango contiene rótulos y no datos numéricos.

- Haz clic en la casilla **Rango de salida** y selecciona una celda vacía, por ejemplo **$I$2**, a partir de la cual se mostrarán los resultados. Finalmente, haz clic en **Aceptar**.

Observa que a partir de la celda **$I$2** aparece una tabla que contiene las siguientes columnas:

- **Posición**: indica la posición en la tabla inicial de cada uno de los valores que aparecen en la segunda columna.

- **Dióxido de azufre**: muestra las concentraciones de dióxido de azufre ordenados de mayor a menor.

- **Jerarquía**: indica la ordenación de los datos respecto de los percentiles. Por ejemplo, el primero es el dato que deja por debajo de él a todos los demás. En cambio, el último es el dato al que superan todos los demás.

- **Porcentaje**: indica el percentil correspondiente a cada valor de la variable. En esta columna observamos que una concentración de dióxido de azufre de 24 microgramos / m^3 se sitúa en el percentil 75%, es decir, supera al 75% de los datos. En cambio, una concentración de 21 microgramos / m^3 se sitúa en el percentil 33,3%, y por tanto, solamente supera a la tercera parte de los datos.

ACTIVIDAD

Utiliza el mismo procedimiento para obtener los percentiles que corresponden a cada uno de los datos sobre concentraciones de humo. Interpreta los resultados obtenidos.

DESCRIPCIÓN DE DATOS ESTADÍSTICOS CON STATGRAPHICS PARA WINDOWS

1. Introducción

Statgraphics es uno de los programas existentes en el mercado para su utilización en Estadística, en sus técnicas más comunes: análisis descriptivo de datos, creación de gráficos, contrastes de hipótesis, análisis de la varianza, análisis multivariante, etc. Aunque hay otros programas que realizan similares tareas, éste cubre la mayoría de las necesidades de cualquier usuario, siendo uno de los de manejo más sencillo. En estas actividades veremos las características fundamentales del programa, y lo usaremos para realizar distintos análisis estadísticos.

2. Operaciones básicas

- **INICIAR STATGRAPHICS**

- Haz doble clic sobre el icono **Sgwin** de Statgraphics en el Escritorio. Cierra la ventana StatWizard haciendo clic en ×.

- Abre el menú **Inicio** y selecciona el comando **Programas** / **Statgraphics Plus 5.1** / **sgwin**.

- **TIPOS DE VENTANAS**

Statgraphics ofrece distintos tipos de ventanas:

Ventana de aplicación	Contiene todos los elementos para trabajar
Hoja de cálculo (datos)	Contiene los datos
Ventana de comentarios	Permite añadir comentarios a los análisis efectuados
Ventana de análisis	Muestra un gráfico o el resultado de un análisis estadístico
StatAdvisor	Muestra una interpretación estadística de los resultados
StatGallery	Muestra gráficos y textos tal como los desees imprimir

- Restaura la ventana de datos. Maximiza la ventana. Minimiza la ventana.

- Haz lo mismo con las ventanas **Comentarios**, **StatAdvisor, StatReporter** y **StatGallery**.

- **LOS MENÚS DE STATGRAPHICS**

- Efectúa un paseo por los distintos menús de Statgraphics. Abre cada uno de los menús y observa las opciones disponibles.

La barra de menús presenta las diez opciones siguientes:

Archivo	Permite abrir, cerrar, imprimir y almacenar ficheros de datos y statfolios
Edición	Realiza operaciones de copiar, cortar y pegar, y otras operaciones de edición
Gráficos	Permite hacer distintas representaciones gráficas de los datos
Descripción	Analiza datos, ajusta distribuciones a conjuntos de datos, etc
Comparación	Compara datos de dos o más muestras, hace análisis de la varianza
Dependencia	Permite realizar análisis de regresión: simple, polinomial y múltiple
Avanzado	Accede a control de calidad, diseño experimental, series temporales, etc
SnapStats!	Realiza análisis estadísticos sobre una y dos muestras y efectúa predicciones.
Ver	Muestra u oculta las barras de herramientas, de estado y StatAdvisor
Ventana	Realiza operaciones con ventanas: disponerlas en mosaico, cascada, etc
Ayuda	Accede a la ayuda del programa

- **LA BARRA DE HERRAMIENTAS**

- Sitúa el puntero del ratón sobre cada uno de los botones de la barra de herramientas y observa la pista y el comentario que aparece en la barra de estado.

La barra de herramientas contiene 21 botones que permiten hacer las operaciones habituales:

❑ Abrir un Statfolio	□ Guardar un Statfolio
❑ Abrir un archivo de datos	□ Guardar un archivo de datos
❑ Cortar	□ Copiar
❑ Pegar	□ Deshacer
❑ Imprimir	□ Vista preliminar
❑ Gráfico de dispersión.	□ Gráfico de cajas.
❑ Histograma	□ Resumen estadístico
❑ Regresión múltiple.	□ Gráficos X–bar y R
❑ Análisis de capacidad.	□ Predicción
❑ Abrir archivo de diseño.	□ Análisis Cluster.
❑ Modelos generales lineales.	□ StatAdvisor
❑ StatWizard	□ Ayuda
❑ Evaluar.	

- **LA BARRA DE ESTADO**

Muestra informaciones sobre las tareas que se están realizando en cada momento y el estado activado o desactivado de las teclas **<BloqMayús>** y **<BloqNum>**.

- **DATOS Y VARIABLES**

Para manejar datos, Statgraphics utiliza distintos tipos de variables: *Numérica, Caracter, Entera, Fecha, Mes, Trimestre, Hora (HH:MM), Fecha–Hora, Hora (HH:MM:SS)* y *Decimal Fijo*. También se puede usar la opción *Fórmula*, que calcula una variable en función de otras conocidas.

Cada variable se introduce en una columna de la hoja de datos y tiene un nombre que puede contener letras y números, pero no ciertos caracteres especiales (como las vocales acentuadas). El primer carácter debe ser una letra. El programa muestra un mensaje de error cuando se introduce un carácter no válido.

Vamos a definir tres variables con los valores que se indican en la siguiente tabla:

Nombre	*Edad*	*Sexo*
Pepe	32	H
Juan	35	H
María	40	M
Luis	25	H
Teresa	28	M
Jaime	35	H

En primer lugar, hay que activar la ventana de Hoja de datos, si no lo estuviera ya:

- Haz clic sobre el botón Maximizar o sobre el botón Restaurar de la hoja de datos.
- Selecciona la primera columna, haciendo clic sobre la etiqueta de columna, rotulada inicialmente como **Col_1**.
- Pulsa el botón derecho del ratón sobre la columna seleccionada. Aparecerá su menú contextual.

- Selecciona **Modificar Columna** del menú de contexto.

- Escribe el nombre de la variable, el comentario (si procede y con no más de 32 caracteres) y selecciona el tipo de variable, en este caso, Carácter.

- Introduce también un ancho apropiado para los datos que va a contener, en el campo **Anchura**.

- Finalmente, haz clic en el botón **OK**.

- Repite el procedimiento con el resto de variables, teniendo en cuenta que la variable edad es del tipo numérico y la variable sexo del tipo carácter.

- **INTRODUCCIÓN DE DATOS**

Una vez creadas las variables, hay que activar cada celda (haciendo clic sobre ella o utilizando las teclas de cursor y/o el tabulador) e introducir en ella el dato correspondiente. Si el dato no corresponde al tipo de variable (lo que ocurre al teclear caracteres en una variable numérica y viceversa), aparecerá un mensaje de error.

Cuando se introducen datos, es posible seleccionar algunos de ellos para copiarlos en el Portapapeles y pegarlos posteriormente en otra zona de la hoja de cálculo.

- Introduce los datos de la tabla anterior en las tres primeras columnas de la hoja de datos.

- **INSERTAR UNA FILA**

Podemos intercalar datos entre otros ya introducidos, insertando una nueva fila:

- Selecciona la fila, encima de la cual se desea insertar otra, haciendo clic sobre su etiqueta (por ejemplo, selecciona la fila 3).

- Pulsa el botón derecho del ratón sobre la fila.

- Selecciona la opción **Insertar** del menú de contexto.

- Introduce los nuevos datos: Conchita 33 M.

Los mismos pasos (pero seleccionando la opción **Borrar**) sirven para borrar una fila. Lo mismo puede hacerse para insertar o eliminar un columna, pero las columnas se insertan a la izquierda de la columna seleccionada.

- **INSERTAR UNA VARIABLE (COLUMNA) CALCULADA**

Si las variables a crear son función del resto de variables ya insertadas en la hoja, podemos usar la opción Generar columna del menú de contexto. Vamos a crear una nueva variable que indique la edad dentro de 20 años:

- Haz clic sobre la etiqueta correspondiente a una columna vacía.

- Sitúa el puntero sobre dicha columna y pulsa el botón derecho del ratón.

- Selecciona la opción **Generar Datos** del menú de contexto.

- En la siguiente ventana introduce la expresión que generará los nuevos datos. Para ello haz doble clic sobre las variables existentes para seleccionarlas y clic sobre los diferentes operadores y números que aparecen. También puedes escribir la fórmula directamente en la caja de texto de la ventana.

- Haz clic sobre el botón **Aceptar**.

- Cambia el nombre de la columna seleccionada. Para ello, pulsa el botón derecho del ratón sobre la columna y selecciona la opción **Modificar Columna**, introduce un nombre apropiado (por ejemplo, *edad20*) y confirma haciendo clic sobre el botón **Aceptar**.

En caso de error puedes deshacer la última acción seleccionando **Edición / Deshacer…**

- **ORDENAR UN ARCHIVO**

El orden en que se introducen los datos es también el que se sigue en el análisis. Pero puede mostrarse en otro orden distinto.

- Selecciona **Edición / Ordenar datos**.

- Selecciona el tipo de ordenación (ascendente o descendente), si se va a aplicar a todo el archivo, y el nombre de la variable por la que se ordenará (en este caso el nombre).

- Haz clic en el botón **Aceptar**.

- **GUARDAR UN ARCHIVO**

Completada la introducción de datos, y hechos los cambios, podemos guardarlos en un fichero:

- Selecciona **Archivo / Guardar como / Guardar Datos como**. También puedes pulsar la tecla **<F12>** o hacer clic sobre el botón **Guardar Archivo de Datos**.

- Escribe el nombre del archivo (EDADES) y selecciona la carpeta en la que se guardará. Haz clic el botón **Guardar**.

Los archivos de datos se almacenan con la extensión SF3.

- **CERRAR UN ARCHIVO**

Para cerrar un archivo de datos y que aparezca de nuevo la hoja de cálculo vacía:

- Selecciona **Archivo / Cerrar / Cerrar Datos**.

- **ABRIR UN ARCHIVO**

Para abrir un archivo de datos, previamente guardado, sigue los siguientes pasos:

- Selecciona **Archivo / Abrir / Abrir Datos**. También puedes pulsar la combinación **<Control+F12>** o hacer clic en el botón **Abrir archivo de datos** de la barra de herramientas.

- Selecciona la carpeta en la que se encuentra y haz doble clic sobre el nombre del archivo (en este caso, EDADES).

- Cierra de nuevo el archivo mediante **Archivo / Cerrar / Cerrar Datos**.

- **SALIR DE STATGRAPHICS**

Para salir del programa:

- Selecciona **Archivo / Salir de STATGRAPHICS** o pulsa **<ALT+F4>**. También puedes hacer clic sobre el botón **Cerrar**, de la esquina superior derecha.

- Si se ha llevado a cabo algún tipo de tarea, el programa ofrece la posibilidad de almacenar el statfolio. En nuestro caso no queremos almacenarlo. Haz clic en el botón **No**.

ACTIVIDADES

- **NOTAS**

Las notas de un grupo de alumnos en Biología y Química se muestran en la siguiente tabla. Crea un archivo de datos a partir de ellas:

B	5	6	6	7	5	7	8	3	5	4	8	5	5	8	8	8	5
Q	5	5	8	7	7	9	10	4	7	4	10	5	7	9	10	5	7

a) Utilizando dos variables: Nota y Carrera y guardando los datos con el archivo de nombre Notas1.

b) Utilizando dos variables: Nota_Bio y Nota_Qui y guardando los datos en el fichero de nombre Notas2.

- **FÓSILES**

Se considera el siguiente banco de datos sobre algunas características de una población fósil. De los 10 fósiles encontrados se ha medido la altura total en cm. (Altu), la altura de la última vuelta (Altu_v), la altura de la boca (Altu_b) y la anchura máxima (Anchu). Además, se ha anotado su color (Col=blanco, gris o marrón) y la altitud (Alti) en la que han sido encontrados, en metros.

Introduce el siguiente banco de datos y guárdalo en el archivo de nombre FOSILES.

Ind	*Altu*	*Altu_v*	*Altu_b*	*Anchu*	*Col*	*Alti*
1	2,400	1,886	1,482	1,585	blanco	220
2	2,569	1,994	1,555	1,573	marrón	750
3	2,606	2,037	1,655	1,781	blanco	310
4	3,271	2,570	1,925	2,048	gris	120
5	2,528	2,021	1,555	1,590	marrón	1330
6	2,730	2,085	1,579	1,693	blanco	350
7	2,605	2,037	1,519	179	gris	400
8	2,400	1,913	1,493	1,575	gris	670
9	2,630	2,192	1,675	1,862	marrón	1100

3. Estadística descriptiva con Statgraphics para Windows

Vamos a realizar un análisis estadístico con datos contenidos en una única variable. Calcularemos las medidas estadísticas más representativas, así como diversas representaciones gráficas de los datos. Comentaremos también el tratamiento de los denominados “statfolios”.

- **CONCEPTOS ESTADÍSTICOS**

POBLACIÓN.- Es el conjunto de todos los elementos (individuos) objeto de estudio.

MUESTRA.- Es una parte o subconjunto de la población.

1) ¿Cómo hay que seleccionar la muestra para que sea representativa de la población?.
2) ¿Qué tamaño debe tener la muestra?.
3) ¿Hasta qué punto es fiable la información contenida en la muestra?.

Todas estas cuestiones se abordan en la **INFERENCIA ESTADÍSTICA**.

VARIABLES ESTADÍSTICAS.- Los datos dan lugar a variables estadísticas que pueden ser:

* **cualitativas o categóricas** (si recogen características no numéricas).

* **cuantitativas** (si contienen datos numéricos). Se pueden clasificar, a su vez, en:

  - ✓ **continuas**: pueden tomar infinitos valores en un intervalo dado (estatura,…)
  - ✓ **discretas**: solamente toman un conjunto aislado de valores (nº de hermanos, etc).

MEDIDAS DE CENTRALIZACIÓN.– Pretenden resumir el conjunto de datos en valores lo más representativos posible:

MEDIA.– $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

MODA.– Es el valor que se repite más veces.

MEDIANA.- Una vez ordenados los datos de forma creciente, la mediana deja a la izquierda el 50% de los datos , y a la derecha el otro 50%. Cuando hay valores extremos, la mediana es más representativa que la media.

MEDIDAS DE DISPERSIÓN.– Indican la dispersión de los datos, es decir, la distancia de los datos al centro de la distribución.

RANGO.– Es la diferencia entre el valor mayor y el menor del conjunto de datos.

VARIANZA.– Es la media de los cuadrados de las desviaciones de cada dato a la media.

$$S_x^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n}$$

En STATGRAPHICS, se utiliza la denominada cuasivarianza muestral, que se obtiene dividiendo entre n–1, en lugar de n:

$$S_x^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}$$

Si el nº de datos es muy grande, no hay diferencia entre varianza y cuasivarianza.

DESVIACIÓN TÍPICA.– Es la raíz cuadrada de la varianza. Se mide en las mismas unidades que los datos. STATGRAPHICS usa siempre la cuasidesviación típica muestral.

COEFICIENTE DE VARIACIÓN.– Compara conjuntos de datos que miden cosas distintas.

$$CV = \frac{Sx}{\bar{x}}$$

ERROR ESTÁNDAR.– Es la cuasi–desviación típica muestral dividida por el nº total de datos.

$$e = \frac{S_x}{n}$$

RANGO INTERCUARTÍLICO.– Es la diferencia entre el cuartil superior y el cuartil inferior.

MEDIDAS DE POSICIÓN.– Miden la posición de los datos dentro de la distribución y si están muy juntos, o muy alejados.

CUARTILES.– Dividen al conjunto de datos (previamente ordenados) en cuatro partes iguales. El primer cuartil, Q_1, deja a la izquierda la cuarta parte (25%) de los datos. El segundo cuartil, Q_2, deja las dos cuartas partes (coincide con la mediana), y el tercero, Q_3, deja a la izquierda las tres cuartas partes (75%) de los datos.

PERCENTILES.– Dividen al conjunto ordenado de datos en 100 partes iguales. Por ejemplo, el percentil 32 es el que deja a la izquierda el 32% de los datos.

MEDIDAS DE FORMA.– Indican la forma que tiene la distribución asociada, si es aplanada o puntiaguda.

ASIMETRÍA.–Si es 0 indica que la distribución es simétrica, si es positiva indica que hay asimetría a la derecha (valores extremos mayores que la media); y si es negativa, hay asimetría a la izquierda (valores extremos menores que la media).

ASIMETRÍA ESTÁNDAR.– Coeficiente de asimetría corregido, proporciona un criterio de normalidad de la distribución. Si está comprendido entre –2 y 2 puede suponerse que la distribución de partida es normal.

CURTOSIS.– Indica si la distribución en cuestión está más o menos apuntada que la normal. Si es positiva, está más apuntada (puntiaguda), si es igual a cero tiene la misma forma que la normal, si es negativa, es menos apuntada (forma aplanada).

CURTOSIS ESTÁNDAR.- Coeficiente de curtosis corregido, proporciona el mismo criterio de normalidad de la distribución: si está comprendido entre –2 y 2 puede considerarse que la distribución de partida es normal.

- **UN PRIMER ANÁLISIS DE LOS DATOS**

Para proceder al análisis de los datos, en primer lugar debes abrir el archivo de datos. Para ello:

- Haz clic sobre el botón **Abrir archivo de datos** de la barra de herramientas y en la siguiente ventana selecciona el archivo EDADES y haz clic en **Abrir**.
- Selecciona **Descripción / Datos Numéricos / Análisis unidimensional**. Aparece un cuadro de diálogo mostrando las variables contenidas en el archivo. Por defecto, dichas variables aparecerán ordenadas, pero se puede cambiar esto y dejar el orden natural haciendo clic sobre la casilla **Ordenar** para desactivarla.
- Haz clic sobre la variable numérica a analizar, en este caso Edad, y sobre el botón **Datos**.
- Haz clic sobre el botón **Aceptar**.

En la siguiente ventana se muestran cuatro zonas en las que se indica un resumen del procedimiento, un resumen de los parámetros estadísticos más significativos, un diagrama de dispersión y un diagrama de caja. En esta ventana aparece una barra de herramientas que permite acceder a distintas opciones. Si los comentarios del StatAdvisor no se visualizan, puedes seleccionar **Ver / StatAdvisor** para activar la opción.

La barra de herramientas de la ventana de análisis consta, entre otros, de los siguientes botones:

Introducir Texto	Permite cambiar la variable objeto de estudio
Opciones tabulares	Permite elegir los paneles de texto que se visualizarán
Opciones gráficas	Permite elegir los gráficos que se visualizarán
Guardar resultados	Permite guardar los datos generados

Haz clic en cada uno de los botones de la barra de herramientas de la ventana de análisis y observa las opciones disponibles. En cada caso haz clic en botón **Cancelar**.

- **OBTENCIÓN DE MEDIDAS ESTADÍSTICAS**

- Haz clic en el botón **Opciones Tabulares** de la barra de herramientas.

- Haz clic sobre la casilla **Resumen de Procedimiento** para desactivarla y sobre el botón **Aceptar**. De esta forma desaparece el panel de resumen del procedimiento y se muestran los tres restantes.

- Haz clic en el botón **Opciones Tabulares** de la barra de herramientas y haz clic en la casilla **Resumen Estadístico** para desactivarlo y haz clic en **Aceptar**. Desaparece el panel de parámetros estadísticos y se muestran solamente los gráficos.

- Haz clic en el botón **Opciones Tabulares** de la barra de herramientas y selecciona la casilla **Resumen Estadístico** para activarlo y haz clic en **Aceptar**. Aparece de nuevo el panel con las medidas seleccionadas por defecto, ya descritas anteriormente. Para verlo maximizado, haz doble clic sobre él.

Puedes ocultar alguna de las medidas visualizadas en el panel, o añadir otras no seleccionadas previamente. Para ello:

- Pulsa el botón derecho del ratón sobre el panel para abrir su menú contextual.

- Selecciona **Opciones de Ventana**.

- Selecciona las medidas **Mediana**, **Moda**, **Primer Cuartil**, **Tercer Cuartil**, **Rango Intercuartílico** y haz clic en el botón **Aceptar**. Observa el panel resultante. De esta forma se visualizan la mediana, moda, cuartil inferior, cuartil superior y rango intercuartílico, además de las activadas por defecto.

- Pulsa el botón derecho del ratón sobre el panel y selecciona **Opciones de Ventana** en el menú de contexto.

- Haz clic en el botón **Todos** para seleccionar todas las medidas estadísticas y haz clic en el botón **Aceptar**. Observa el resultado.

- Haz doble clic sobre el panel para restaurar la ventana a su tamaño original.

- **MODIFICACIÓN DEL TEXTO**

Podemos cambiar el tipo y tamaño de letra que aparece en los paneles. Para ello:

- Selecciona **Edición / Cambiar Fuente**, o bien pulsa **<F2>**.
- Selecciona, en el cuadro de diálogo Fuente el tipo de letra, el tamaño y el estilo (negrita, cursiva,…). Por ejemplo, selecciona el tipo de fuente *Lucida Console*, *negrita cursiva*, tamaño *14 puntos*.
- Haz clic en el botón **Aceptar** y observa el nuevo aspecto de la ventana.
- Repite la operación para dejar el tipo de letra *Courier New*, *normal*, de tamaño *10 puntos*.

- **CAMBIO DE LA VARIABLE DE ANÁLISIS**

Una vez activado un panel, podemos actualizarlo para otra variable numérica contenida en el archivo de datos. Para ello:

- Haz clic sobre el botón **Introducir Texto**.
- Haz doble clic sobre la nueva variable a analizar, en este caso, *edad20*.
- Haz clic sobre el botón **Aceptar**. Los paneles abiertos se adaptarán a la nueva variable.

- **CERRAR LA VENTANA DE ANÁLISIS**

- Haz clic sobre el botón **Cerrar** × de la ventana.
- Haz clic sobre el botón **Sí** para confirmar el cierre, en el cuadro de diálogo que aparece y selecciona la opción de no almacenar los cambios.
- Cierra el archivo de datos seleccionando **Archivo / Cerrar / Cerrar Datos**.

- **PERCENTILES**

- Crea un nuevo archivo a partir de los datos de la siguiente tabla, que corresponden a los niveles de precipitaciones anuales en mm. para un conjunto de 40 zonas forestales.

640	356	786	567	445	677	564	781
543	384	941	769	590	831	564	612
735	497	742	681	992	821	589	635
437	726	492	751	964	527	883	399
734	165	528	467	628	679	820	775

- El archivo a crear tendrá una única variable, PRECIPIT, y puede guardarse con el nombre PREC_ANUAL.
- Selecciona **Descripción / Datos Numéricos / Análisis unidimensional**.
- Haz clic sobre la variable numérica PRECIPIT y sobre el botón **Datos**.
- Haz clic sobre el botón **Opciones Tabulares**. Desactiva la casilla Resumen de Procedimiento. Si eliges de nuevo la opción de cálculo de las medidas estadísticas, **Resumen Estadístico**, y seleccionas el total de las disponibles (clic en el botón **Todos**), podrás comprobar que sus valores son los de la siguiente tabla:

Media	642,9	Cuartil 1º	527,5
Mediana	637,5	Cuartil 3º	772
Moda	564	Rango IC	244,5
Media Geom.	613,6	Coef. asimetría	–0,239
Varianza	31987	Coef. asimetría (estándar)	–0,618
Desv. típica	178,8	Coef. curtosis	0,10
Error estándar	28,2	Coef. curtosis	0,13
Mínimo	165	Coef. variación	27,8
Máximo	992	Suma	25717
Rango	827	nº datos	40

Ello indica que, entre otras cosas, la distribución asociada es ligeramente asimétrica a la izquierda y que es ligeramente más apuntada que la normal.

Queremos calcular el nivel de precipitación por debajo del cual quedan el 10% de los datos observados, es decir, el percentil 10. Este y otros se pueden obtener a partir del panel correspondiente:

- Haz clic sobre el botón **Opciones Tabulares**.
- Haz clic sobre la opción **Percentiles**.
- Haz clic sobre la casilla **Aceptar**.
- Haz doble clic sobre el nuevo panel para maximizarlo. Observa que aparecen por defecto los percentiles 1, 5, 10, 25, 50, 75, 90, 95 y 99.

Como se trataba de hallar el percentil 10, resulta que dicho percentil es 418. Es decir, el 10% de los datos son inferiores a 418 mm, mientras que el 90% superan a 418 mm.

- Pulsa el botón derecho del ratón sobre el panel.
- En el cuadro de diálogo **Opciones de Ventana** introduce los valores: 32, 45, 56, 78, 83.
- Haz clic en el botón **Aceptar**. Los percentiles introducidos se visualizarán en el panel.
- Haz doble clic sobre el panel para recuperar el tamaño original.

- **TABLA DE FRECUENCIAS**

Hay ocasiones en que el número elevado de datos hace aconsejable agruparlos en clases. Esto facilita la obtención de las medidas estadísticas habituales. A la hora de los cálculos, el intervalo se representaba por su “marca de clase” (punto medio del mismo). Actualmente esta clasificación no es necesaria, porque el programa efectúa cualquier cálculo sin problemas. No obstante, puede ser interesante obtener una tabulación de los datos de forma automática. En ese caso, ¿cuántas clases debemos elegir?. No hay una regla universalmente válida. Un criterio que se suele utilizar es tomar el número de clases como la raíz cuadrada del número total de datos.

- Haz clic en el botón **Opciones Tabulares**.
- Haz clic sobre la opción **Tabla de Frecuencias**.
- Haz clic en la casilla **Aceptar**.
- Haz doble clic sobre el nuevo panel para ampliarlo. Observa que aparecen intervalos definidos por su límite inferior y su límite superior, con las frecuencias correspondientes, las marcas de clase de cada intervalo, las frecuencias relativas y las frecuencias acumuladas.

Para cambiar el nº de clases y situarlo, por ejemplo, en 4 intervalos, sigue los siguientes pasos:

- Pulsa el botón derecho sobre el panel para abrir el cuadro de opciones del panel.
- Introduce el número de clases (4).
- Haz clic en el botón **Aceptar**. Observa el panel resultante.

También puedes modificar en el cuadro de diálogo *Opciones de Tabulación de Frecuencias* el límite inferior y el límite superior de los datos. Los valores indicados aquí influirán en otros análisis que se pueden hacer y que dependen de ellos (por ejemplo, el histograma, el polígono de frecuencias, etc.).

- Haz doble clic sobre el panel para recuperar el tamaño original.

- **DIAGRAMA DE TALLO Y HOJAS**

Dentro del análisis exploratorio de datos, uno de los diagramas que se ha impuesto en el mundo estadístico es el denominado de tallo y hojas (Stem−and−Leaf). Se trata de representar los datos de forma que el resultado dé una idea de la forma y características de la distribución.

- Haz clic sobre el botón **Opciones Tabulares**.
- Haz clic sobre la opción **Diagrama de Tallo y Hojas**.
- Haz clic sobre la casilla **Aceptar**.
- Haz doble clic sobre el nuevo panel para maximizarlo.

Observa que aparecen nueve tallos: los números del 1 al 9, y varias hojas, que se interpretan de la forma siguiente: 1 | 2, en este caso, significa 120. Así, los valores del diagrama hay que leerlos, de arriba abajo, como 160, 350, 380, etc. También aparecen las frecuencias acumuladas de arriba abajo, y de abajo arriba, hasta confluir en el intervalo que contiene a la mediana, señalado con unos paréntesis que contienen su frecuencia.

Es fácil contar hasta llegar a la mitad del conjunto de datos de la distribución, y obtener así gráficamente su mediana. Por ejemplo, y ya que el número de datos es 40 y hay que considerar los que ocupan los lugares 20 y 21 una vez ordenados, habría que comenzar por el tallo 6 y contar de izquierda a derecha a partir del número 17 (valor de la frecuencia acumulada del tallo anterior), hasta llegar al lugar 20 (que está ocupado por 630) y al lugar 21 (ocupado por 640), concluyendo que la mediana es el promedio de los valores, es decir 635. La diferencia con su verdadero valor, 637,5 se debe al redondeo que se lleva a cabo en el diagrama.

Cuando el número de hojas es excesivo, el programa divide automáticamente un mismo tallo en dos (con dos caracteres de inicio distintos).

- **GRÁFICO DE CAJAS**

STATGRAPHICS permite representar distintos gráficos relacionados con una variable estadística. Una vez abierto el archivo de datos y la ventana de análisis, hay que utilizar el botón correspondiente de la barra de herramientas.

- Haz clic sobre el botón **Opciones Gráficas**.
- En el siguiente cuadro de diálogo, activa solamente la casilla **Gráfico de Dispersión**.
- Haz clic en el botón **Aceptar**.
- Haz doble clic sobre la ventana gráfica para maximizarla. El resultado es un diagrama de puntos de la variable PRECIPIT.

Uno de los gráficos estadísticos más importantes es el denominado “diagrama de cajas”. Este gráfico da una impresión visual de varias características del conjunto de datos: la simetría o asimetría, dispersión, existencia de valores atípicos.

- Haz clic en el botón **Opciones**.
- En el siguiente cuadro de diálogo, selecciona solamente la casilla **Gráfico de Caja y Bigotes**.
- Haz clic sobre el botón **Aceptar**. Obtendrás el gráfico en un nuevo panel.
- Haz doble clic sobre el panel para maximizarlo.

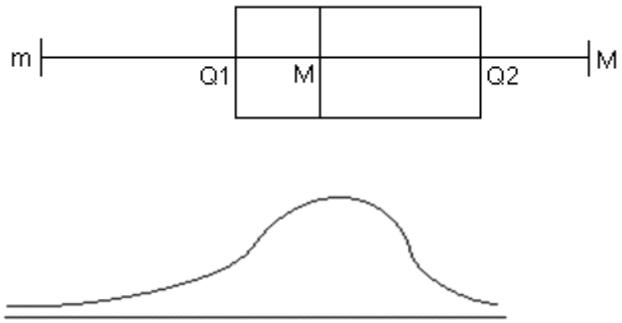
El gráfico está construido de la siguiente forma:

El trazo vertical más a la izquierda representa el valor más pequeño del conjunto de datos, y el trazo vertical de la derecha el valor mayor: esos son los límites del gráfico. En el caso de que haya valores excepcionalmente grandes o pequeños que impidan una correcta representación del gráfico, éstos aparecerán como puntos aislados: son los valores atípicos o “outliers”.

La caja central está construida de forma que el trazo vertical izquierdo corresponde al primer cuartil, el trazo central a la mediana, y el trazo de la derecha al tercer cuartil.

La mayor o menor longitud de las líneas horizontales dan una idea de mayor o menor dispersión (se observa que es más disperso el primer 25% de los datos que el resto).

También se puede comentar la asimetría de la distribución en función de dicha longitud (en este caso, ligeramente asimétrica a la izquierda).



Podemos cambiar la orientación del diagrama. Para ello:

- Pulsa el botón derecho sobre cualquier punto del panel gráfico.
- Haz clic sobre la opción **Opciones de Ventana** del menú de contexto.
- Haz clic sobre la opción **Vertical**.
- Haz clic sobre el botón **Aceptar**. Comprueba el resultado.
- Pulsa el botón derecho del ratón sobre cualquier punto del panel y haz clic sobre la opción **Opciones de Ventana** del menú contextual.
- Desactiva la opción **Vertical** activando **Horizontal** y haz clic sobre la casilla **Muesca de Mediana**.
- Haz clic sobre el botón **Aceptar**. Observa el panel resultante.

El programa permite personalizar el gráfico creado en función de lo que se desee. Se puede, por ejemplo, cambiar el texto que aparece (y el tipo de letra, el tamaño,…), los colores, las tramas, el sombreado,…

- Haz clic sobre el título para seleccionar el objeto. Aparecerán unos marcadores bordeándolo.
- Pulsa el botón derecho sobre el objeto.
- Haz clic sobre la opción **Opciones Gráficas** del menú de contexto.
- En el siguiente cuadro de diálogo escribe el nuevo título: DIAGRAMA DE CAJAS.
- Haz clic en el botón **Línea 1 Fuente**.
- Elige el tipo de letra *Arial, negrita cursiva*, tamaño *16 puntos*.
- Haz clic en el botón **Aceptar** y observa el resultado.
- Haz clic sobre uno de los números que aparecen en el eje horizontal. De esta forma queda seleccionado el eje de abcisas.
- Pulsa el botón derecho del ratón sobre dicho eje.
- Selecciona la opción **Opciones Gráficas** en el menú de contexto.
- En el siguiente cuadro de diálogo puedes modificar la escala de dicho eje, indicando el valor inicial (**Desde:**) y el final (**Hasta:**). En la caja de texto **Por** introduce 100.
- Haz clic en el botón **Marca Fuente**, selecciona el color rojo y haz clic en **Aceptar**.
- Haz clic en **Aceptar** y observa el resultado.
- Haz clic sobre uno de los “bigotes” de la caja para seleccionar la línea.
- Pulsa el botón derecho del ratón sobre dicha línea del diagrama.
- Selecciona **Opciones Gráficas** del menú contextual. Haz clic sobre la solapa **Líneas**.
- En el siguiente cuadro de diálogo, cambia el tipo de trazado, haciendo clic en el segundo botón. Cambia el grosor de la línea arrastrando el botón de tamaño.
- Haz clic en el botón **Colores** y selecciona uno de los colores del panel.
- Haz clic en el botón **Aceptar** y observa el resultado.
- Vuelve a seleccionar la línea y pulsa el botón derecho del ratón para abrir el menú de contexto.
- Selecciona **Opciones Gráficas**. Selecciona la solapa **Líneas** y, en el cuadro de diálogo haz clic sobre el botón de traza continua.
- Arrastra el botón **Grosor de línea** hasta llegar al ancho de línea que desees.
- Haz clic en el botón **Aceptar** y observa el resultado.
- Haz clic sobre la trama de la caja para seleccionarla.
- Pulsa el botón derecho del ratón sobre dicha trama.
- Selecciona **Opciones Gráficas** en el menú de contexto.

- En el cuadro de diálogo que aparece elige la solapa Relleno, selecciona el estilo de relleno y color deseado.
- Haz clic en el botón **Aceptar** y observa el resultado.
- Haz clic sobre el fondo del diagrama, fuera del mismo, para seleccionarlo.
- Pulsa el botón derecho del ratón sobre el área seleccionada.
- En el menú de contexto elige **Opciones Gráficas**.
- En el cuadro de diálogo que aparece puedes elegir, entre otras cosas, un color para el fondo del gráfico (opción **Fondo**), distinto del seleccionado por defecto y otro para el fondo del marco (opción **Borde**). Para ello debes activar cada una de estas opciones y a continuación el botón **Colores**, seleccionando el color deseado.
- Haz clic en el botón **Aceptar** para comprobar los cambios efectuados.
- Haz doble clic sobre el panel gráfico para recuperar el tamaño original.

De la misma forma se puede modificar cualquiera de los gráficos que pueden crearse con el programa.

- **HISTOGRAMA Y POLÍGONO DE FRECUENCIAS**

Otro gráfico muy utilizado es el histograma que consiste en un diagrama de rectángulos, donde las alturas son proporcionales a las frecuencias.

- Haz clic sobre el botón **Opciones Gráficas**.
- Haz clic sobre la casilla **Histograma**. Haz clic en el botón **Aceptar**. Obtendrás el gráfico en un nuevo panel.
- Haz doble clic sobre el panel para maximizarlo.

Podemos utilizar el histograma para analizar si los datos se ajustan a una determinada distribución teórica, lo que queda determinado por la forma del histograma.

Observa que los intervalos del histograma coinciden con los reflejados en el panel **Tabla de Frecuencias**. Podemos cambiar el número de intervalos desde el panel del histograma.

- Pulsa el botón derecho del ratón sobre el panel.
- Selecciona la opción **Opciones de Ventana** del menú de contexto.
- Escribe un nuevo número de clases, por ejemplo, 5, y haz clic en el botón **Aceptar**. Observa el resultado.
- Pulsa de nuevo el botón derecho del ratón sobre el panel para abrir el menú contextual y selecciona el comando **Opciones de Ventana**.
- En la caja de texto **Limite Inferior** introduce 500 y haz clic en el botón **Aceptar**. ¿Qué ocurre?.
- Vuelve a abrir el menú de contexto y selecciona **Opciones de Ventana**. Vuelve a la situación anterior, es decir, introduce 0 en la caja **Limite Inferior**.
- Activa la opción **Relativa** para que se dibuje un histograma de frecuencias relativas.

- Haz clic en el botón **Aceptar** y observa el resultado. ¿Qué diferencias hay entre un histograma de frecuencias absolutas y uno de frecuencias relativas?.

- Abre de nuevo el menú de contexto y selecciona **Opciones de Ventana**. Activa la opción **Acumulada** y haz clic en el botón **Aceptar**. El resultado es un histograma de frecuencias relativas acumuladas.

- Utiliza un procedimiento parecido para dibujar un histograma de frecuencias absolutas acumuladas.

- Abre de nuevo el menú contextual y selecciona **Opciones de Ventana**. En la caja **Tipo de Gráfico**, selecciona **Polígono**. Desactiva las opciones **Relativa** y **Acumulada**.

- Haz clic en el botón **Aceptar**. El resultado es un polígono de frecuencias absolutas. Este diagrama se obtiene uniendo con segmentos de recta los puntos medios de los lados superiores de los rectángulos del histograma.

- Utiliza un procedimiento similar para dibujar el polígono de frecuencias relativas. ¿Qué diferencias presenta con un polígono de frecuencias absolutas?.

- Repite el procedimiento para dibujar un polígono de frecuencias relativas acumuladas. Dibuja también el polígono de frecuencias absolutas acumuladas y comenta las diferencias entre los dos.

- Haz doble clic sobre el panel para restaurar su tamaño original.

- **IMPRIMIR LOS ANÁLISIS ESTADÍSTICOS**

En cualquier momento puedes imprimir el panel activo.

- Pulsa el botón derecho del ratón sobre el panel.

- Selecciona la opción **Imprimir** del menú de contexto.

- Haz clic en el botón **Aceptar**.

En caso de que quieras imprimir todos los paneles que constituyen la ventana de análisis, puedes proceder de la siguiente forma:

- Selecciona **Archivo / Vista Previa**. Haz clic en el botón **Imprimir**.

En el siguiente cuadro de diálogo puedes elegir la opción que más te interese: imprimir todos los paneles, imprimir sólo la selección, imprimir sólo las filas de los paneles de texto seleccionados o imprimir solo los paneles gráficos. También puedes configurar la impresora, haciendo clic en el botón **Configurar**.

- En nuestro caso, selecciona **Todo** y haz clic en el botón **Aceptar**. Observa el resultado.

- Haz clic en el botón Opciones Gráficas y selecciona solamente la opción Histograma. Haz clic en el botón Aceptar.

- Haz doble clic sobre el panel Histograma para maximizarlo. Haz clic en el botón Añadir texto. En la siguiente ventana introduce el texto “Precipitación anual”. Arrastra el texto hasta una zona libre del panel.

- Haz clic en el botón **Vista Previa**. Haz clic en el botón **Acercar**. Haz clic en el botón **Alejar**.

- Haz clic en el botón **Imprimir**. Haz clic en el botón **Aceptar**. Observa el resultado.

- **GUARDAR UN STATFOLIO**

Todo el entorno construido después de un análisis estadístico, formado por los distintos paneles de texto y de gráficos, constituyen un "statfolio" y se puede guardar como tal. Ello permite trabajar posteriormente con él. Los paneles obtenidos pueden ser utilizados en adelante para otras variables del mismo archivo o de otro distinto. Para guardar todo el entorno en archivo de "statfolio" sigue los siguientes pasos:

- Selecciona **Archivo / Guardar como / Guardar StatFolio como**. También puedes pulsar **<F11>** o hacer clic sobre el botón **Guardar StatFolio** de la barra de herramientas.
- Selecciona la carpeta donde quieres guardar el archivo e introduce su nombre en la caja **Nombre**, por ejemplo PREC_ANUAL. El programa asigna al archivo la extensión SGP.
- Haz clic sobre el botón **Guardar**.

- **CERRAR UN STATFOLIO**

- Selecciona **Archivo / Cerrar / Cerrar StatFolio**. El archivo de datos no se guardará. Para hacerlo:
- Selecciona **Archivo / Guardar / Guardar Datos**.

- **ABRIR UN STATFOLIO**

Para abrir un Statfolio:

- Selecciona **Archivo / Abrir / Abrir StatFolio**, También puedes pulsar **<Control + F11>** o hacer clic sobre el botón **Abrir StatFolio**.
- Selecciona la carpeta donde está almacenado el statfolio y haz doble clic sobre el nombre del archivo. Todos los análisis hechos y el fichero de datos fuente de los mismos se mostrarán automáticamente.
- Cierra el StatFolio seleccionando **Archivo / Cerrar / Cerrar Statfolio**.

- **EL ARCHIVO CARDATA**

El programa STATGRAPHICS contiene varios archivos de ejemplo que pueden utilizarse para practicar las diferentes técnicas. Un archivo clásico en este y otros paquetes estadísticos es CARDATA, que contiene datos relativos a 155 coches de distintas procedencia (Estados Unidos, Europa y Japón) del periodo 1978–82. Algunas de las variables que contiene son: mpg (millas por galón), cylinders (cilindros), horsepower (caballos de potencia), price (precio), etc.

Como ya tenemos abierto el statfolio correpondiente a los análisis anteriormente realizados, al abrir el archivo de datos CARDATA este statfolio se actualizará a los nuevos datos.

- Haz clic sobre el botón **Abrir archivo de datos** de la barra de herramientas.
- Localiza el archivo CARDATA.SF y haz doble clic sobre él.
- Selecciona el comando Descripción / Datos numéricos / Análisis unidimensional.
- En la siguiente ventana selecciona la variable Price y haz clic en Datos. Haz clic en Aceptar. Los paneles se adaptarán a la nueva situación. Observa que en el panel correspondiente al gráfico de cajas aparecen valores atípicos.

Teniendo en cuenta las equivalencias aproximadas galón–litro (1 galón = 4,5 litros) y milla–kilómetro (1 milla = 1,6 km), y a partir de la variable mpg, se puede crear una nueva variable, consumo, que calcule el consumo en litros cada 100 kilómetros. Para ello:

- Minimiza la ventana Análisis unidimensional–price para visualizar la ventana de hoja de cálculo.
- Selecciona la columna de la variable mpg.
- Pulsa el botón derecho sobre dicha columna y selecciona la opción **Insertar** del menú contextual para insertar una nueva variable.
- Selecciona la nueva columna, pulsa el botón derecho del ratón sobre ella, selecciona la opción **Modificar Columna** del menú de contexto, escribe el nombre consumo y haz clic sobre el botón **Aceptar**.
- Selecciona la nueva columna, pulsa el botón derecho sobre ella y selecciona **Generar Datos**.
- Escribe la fórmula que genera la columna: (100*4.5) / (mpg*1.6).
- Haz clic sobre el botón **Aceptar**. Comprueba los valores obtenidos en la columna: 6,52; 7,79; 8,57;…

Podemos realizar un análisis exploratorio de los datos, mediante un gráfico de cajas múltiple, de la variable consumo con respecto a la variable Origin (origen del vehículo) y a la variable year (año de fabricación). Para ello:

- Selecciona **Gráficos / Gráficos Exploratorios / Gráfico de caja y bigotes múltiple**.
- Introduce la variable de datos, consumo, en el campo **Datos** y la variable de "agrupación de datos" (origin) en el campo **Códigos de nivel**.
- Comprueba que está activada la opción **Ordenar** y haz clic en el botón **Aceptar**.

Observa que el gráfico resultante permite comparar el consumo en los tres ámbitos geográficos estudiados, concluyendo que en EEUU, en la fecha del estudio, los coches consumían más que en Europa y bastante más que en Japón.

- Haz clic en el botón **Introducir Texto** de la ventana de análisis.
- En la caja de texto **Códigos de nivel** introduce la variable year y haz clic en **Aceptar**. Observa que la ventana gráfica se actualiza para la variable seleccionada.

Observa que, a partir del año 80, hay un descenso significativo del consumo, seguramente como consecuencia de la crisis del petróleo del año 79 (que obligó a los fabricantes de coches a tener en cuenta este dato).

- Cierra la ventana de análisis, pero no el archivo de datos CARDATA.

- **ANALIZANDO SIMULTÁNEAMENTE VARIAS VARIABLES**

Supongamos que queremos hallar las medidas estadísticas de varias de las variables contenidas en el archivo CARDATA, por ejemplo price, consumo y weight. En lugar de utilizar la opción **Descripción / Datos Numéricos / Análisis unidimensional**, usaremos directamente la barra de herramientas de la ventana de STATGRAPHICS. Para ello:

- Haz clic sobre el botón **Resumen Estadístico** de la barra de herramientas. Aparecerá la ventana de selección de variables.

- Haz doble clic sobre el nombre de las variables de las que se desea calcular sus medidas, price, consumo y weight.
- Haz clic sobre el botón **Aceptar**.
- Haz clic sobre el botón **Opciones Tabulares**.
- Selecciona solamente la opción **Resumen Estadístico**.
- Haz clic sobre el botón **Aceptar**. En la pantalla aparecerán las medidas de las tres variables mencionadas.

	price	consumo	weight
Frecuencia	154	154	154
Media	4609,74	10,4814	2672,19
Varianza	4,05568E6	8,42663	363630,0
Desviación típica	2013,87	2,90287	603,017
Minimo	1900,0	6,03541	1755,0
Máximo	15475,0	18,1452	4360,0
Asmetría tipificada	10,7032	3,51397	2,77899
Curtosis tipificada	19,7655	–1,3462	–1,34697
Suma	709900,0	1614,13	411518,0

También se indican otros parámetros, como mediana, moda, media geométrica, error estándar, rango, primer cuartil, segundo cuartil, rango intercuartílico, asimetría, curtosis y coeficiente de variación.

Vamos a comparar las variables price y consumo, utilizando para ello un diagrama de puntos.

- Haz clic en el botón **Gráfico de dispersión** de la barra de herramientas.
- En el siguiente cuadro de diálogo introduce en la caja **Y** la variable consumo y en la caja **X** la variable price. Haz clic en el botón **Aceptar**. Observa el resultado.

- **USANDO LA BARRA DE HERRAMIENTAS DE STATGRAPHICS**

Vamos a dibujar un diagrama de cajas de la variable price.

- Haz clic en el botón **Gráfico de caja** de la barra de herramientas.
- En el siguiente cuadro de diálogo introduce, en el campo **Datos** la variable price.
- Haz clic en el botón **Aceptar**. Observa que hay valores atípicos.

Vamos a dibujar un histograma de la variable weight.

- Haz clic en el botón **Histograma** de la barra de herramientas.
- En la siguiente ventana, introduce en el campo **Datos** la variable weight.
- Haz clic en el botón **Aceptar**. Observa el resultado.
- Haz clic con el botón derecho del ratón sobre el panel gráfico y selecciona la opción **Opciones de Ventana** del menú de contexto.
- Introduce en la siguiente ventana 12 clases y haz clic en **Aceptar**.

- Repite el proceso para ver cómo se transforma el histograma al aumentar el número de clases. Introduce 16, 20, 25, 30 clases sucesivamente. ¿Qué pasaría si clasificamos los datos en 50 clases?. ¿Qué conclusiones puedes extraer?.

- **DIAGRAMAS DE BARRAS Y DIAGRAMAS DE SECTORES**

Vamos a representar un diagrama de barras que represente la edad de los distintos individuos que aparecen en el archivo EDADES.

- Selecciona **Archivo / Abrir / Abrir Datos**.
- Selecciona el archivo EDADES en la carpeta donde esté situado y haz clic en **Abrir**.
- Selecciona **Gráficos / Diagramas de Presentación / Diagrama de Barras**.
- Introduce en el campo **Recuentos** la variable a representar, en este caso, Edad, y en el campo **Etiquetas** la variable que contiene los rótulos de los datos (nombre).
- Haz clic en el botón **Aceptar**. Observa el resultado.
- Haz clic con el botón derecho del ratón sobre el panel gráfico. Elige **Opciones de Ventana** en el menú de contexto.
- En el campo **Dirección**, selecciona **Vertical** y haz clic en **Aceptar**. ¿Qué ocurre?.

Cuando los datos son categóricos (cualitativos) hay que tabularlos previamente. Por ejemplo, para crear un gráfico de barras de la variable sexo del archivo EDADES, hay que seguir los siguientes pasos:

- Selecciona **Descripción / Datos Cualitativos / Tabulación**.
- Selecciona la variable a tabular, en este caso, sexo.
- Haz clic sobre el botón **Aceptar**.
- En la siguiente ventana, haz clic sobre el botón **Opciones Gráficas** y selecciona solamente **Diagrama de Barras**.
- Haz clic en **Aceptar**. Aparece el gráfico con barras horizontales
- Haz clic con el botón derecho del ratón sobre el panel. En el menú de contexto selecciona **Opciones de Ventana**.
- En el campo **Dirección**, selecciona **Vertical** y haz clic en **Aceptar**. Observa el resultado.
- Pulsa el botón derecho del ratón sobre el panel gráfico y selecciona **Opciones de Ventana** del menú contextual.
- En el campo **Escala**, selecciona **Porcentajes** y haz clic en **Aceptar**. ¿En qué se diferencia este diagrama del anterior?.

Vamos a crear un gráfico de sectores de la variable sexo. Para ello:

- Haz clic sobre el botón **Opciones Gráficas** de la barra de herramientas de la ventana de análisis.

- En la siguiente ventana, selecciona **Diagrama de Sectores** y haz clic en **Aceptar**.
- Haz doble clic sobre el nuevo panel para maximizarlo.
- Selecciona el título del diagrama y haz clic con el botón derecho del ratón. En el menú de contexto, elige **Opciones Gráficas**.
- En la siguiente ventana introduce como título del gráfico "DIAGRAMA DE LA VARIABLE SEXO". Haz clic en el botón **Línea 1 Fuente** y selecciona el tipo de letra Arial, color rojo, negrita cursiva de tamaño 16 puntos. Haz clic en **Aceptar** y haz clic en **Aceptar**.
- Haz clic sobre uno de los sectores y vuelve a hacer clic con el botón derecho del ratón para que aparezca el menú contextual.
- Elige el comando **Opciones Gráficas**. En la siguiente ventana, en la solapa **Relleno**, haz clic sobre el botón **Colores** y selecciona el color verde. Haz clic en **Aceptar** y haz clic en **Aceptar**.

Vamos a crear un diagrama de sectores de la variable Edad y, posteriormente, lo editaremos.

- Cierra la ventana de análisis.
- Selecciona **Gráficos / Diagramas de Presentación / Diagrama de Sectores**.
- En el campo **Frecuencias** introduce la variable Edad y en el campo **Etiquetas** la variable Nombre.
- Haz clic en **Aceptar**.
- Haz doble clic en el panel gráfico, para maximizarlo.
- Sustituye el título del gráfico por este otro: "DIAGRAMA DE EDADES".
- Modifica los colores de cada uno de los sectores.
- Haz que en cada sector se muestren frecuencias absolutas y no porcentajes.
- Cierra la ventana de análisis.
- Cierra el archivo de datos EDADES.
- Cierra el programa STATGRAPHICS.

ACTIVIDADES

- **HERMANOS**

El número de hermanos de 40 alumnos es:

3	4	2	3	4	3	4	4	4	2	3	4	4	3	4	1	2	3	5	4
2	2	2	5	3	4	4	6	2	6	4	3	2	1	2	3	2	4	3	1

¿Cuántos alumnos tienen 5 o más hermanos?. ¿Cuántos 3 o menos?. Dibuja el histograma y el polígono de frecuencias.

- **SOCIOLOGÍA**

En un grupo de Sociología se han obtenido estas puntuaciones en un test de habilidad mental:

50	23	45	36	56	34	56	49	53	23	66	31	45	22
67	45	34	23	45	23	67	33	44	48	53	57	77	31
54	21	34	43	12	78	36	23	47	52	33	37	64	21

a) Construye el histograma, calcula la media $\bar{x}$ y la desviación típica σ y comprueba si en el intervalo $(\bar{x}-\sigma, \bar{x}+\sigma)$ se encuentra aproximadamente el 68% de los datos.

b) Agrupa los datos en intervalos de amplitud 10 y construye la tabla de frecuencias. Obtén los parámetros estadísticos de forma y relaciónalos con la forma del histograma.

- **ÁCIDO ÚRICO**

Los valores del nivel de ácido úrico medidos en miligramos por decilitro correspondiente a 50 individuos son los siguientes:

1,9	3,0	2,9	4,1	3,7	1,8	4,3	2,4	5,4	4,6
2,6	3,1	3,0	3,4	2,8	4,3	5,8	3,3	7,4	3,1
3,6	5,2	6,7	3,2	5,5	5,1	7,0	2,7	3,6	2,5
3,6	4,8	3,8	2,1	4,1	3,4	4,6	4,4	4,3	3,9
5,6	3,6	8,9	2,7	3,2	3,8	4,2	3,6	6,0	2,3

a) Calcula las medidas estadísticas de la siguiente tabla:

Media	Desv. típica	Cuartil 1º	Coef. curtosis
Mediana	Error estándar	Curartil 3º	Coef. curtosis (estánd.)
Moda	Mínimo	Rango IC	Coef. variación
Media geom.	Máximo	Coef. asimetría	Suma
Varianza	Rango	Coef. asimetría (estánd.)	nº datos

b) Utilizando 1,0 mg / dl como límite inferior del primer intervalo de clase, construye una tabla de distribución de frecuencias donde la amplitud de las clases sea 1,0 mg / dl.

c) ¿Qué se puede afirmar sobre la asimetría y forma de la distribución?.

d) Determina los percentiles 10, 25, 75 y 90.

e) Construye un diagrama de tallo y hojas, e identifica en él la mediana y la moda.

f) Obtén el polígono de frecuencias absolutas de la distribución.

g) Construye el polígono de frecuencias relativas acumuladas.

h) Obtén el histograma de frecuencias absolutas.

i) Obtén el histograma de frecuencias relativas acumuladas.

- **TEMPERATURAS**

Las temperaturas máximas, en ºC, medidas en Valencia durante un mes fueron las que se muestran en la siguiente tabla:

26	24	28	30	33	35	25	36	34	26	32	31	37	25	29	38
30	34	38	36	27	29	39	30	25	31	32	40	35	32	24	

a) Construye la tabla de frecuencias y representa gráficamente la distribución mediante un diagrama de caja. ¿Hay valores atípicos?.

b) Utilizando el diagrama de caja, comenta la simetría de la distribución.

c) Obtén los siguientes parámetros estadísticos: moda, media, mediana, cuartiles, varianza y desviación típica. Interpreta el significado de cada uno.

d) Calcula los percentiles 30, 40, 50, 60 y 70 de la distribución, interpretando su significado.

- **SUPERFICIE AGRÍCOLA**

Las dimensiones (en hectáreas)de 50 explotaciones agrícolas de una comarca son las que se muestran en la siguiente tabla:

10	32	27	19	24	21	13	16	17	16
16	10	15	16	15	20	15	14	21	11
12	14	12	17	14	18	17	15	25	13
19	17	18	16	16	17	16	18	20	11
26	11	13	12	15	12	14	28	17	17

a) Construye la tabla de frecuencias, tomando clases de amplitud 4.

b) Obtén el histograma de frecuencias absolutas y el histograma de frecuencias relativas.

c) Calcula la media, mediana, varianza, desviación típica.

d) Calcula los coeficientes de asimetría y apuntamiento y comenta sus valores comparándolos con la forma del histograma.

- **TITULACIONES**

En la siguiente tabla se muestra el número de colocaciones según titulación durante 1999 en la ciudad de Valencia (fuente revista DADES):

Analfabets i sense estudis	1104
Certificat d'escolaritat	28918
EGB	32151
FP1–FP2–Cicles formatius	9095
BUP–COU–Batxillerat	15196
Titulació grau mitjá	4692
Titulació grau superior	4978
Total	96134

Representa gráficamente los datos mediante un diagrama de barras y un diagrama de sectores.

- **PUNTUACIONES**

Un profesor es calificado de 1 a 5 por sus alumnos obteniendo los siguientes resultados:

2	5	4	4	1	2	2	3	4	4	5	3	3	1	2
5	4	4	3	3	4	2	3	3	2	3	4	3	4	4

a) Dibuja un gráfico de tronco y hojas.

b) Dibuja un gráfico de caja. ¿Es simétrico?.

c) ¿Sale bien parado el profesor con la calificación de sus alumnos?.

- **TEST DE INTELIGENCIA**

Se aplica un test de inteligencia general (cociente intelectual) a 40 alumnos de primero de Bachillerato) de un centro, obteniendo los siguientes resultados:

106	136	81	110	95	92	99	106	81	95
110	103	88	81	81	99	110	114	128	103
103	74	95	136	95	88	106	121	106	114
117	92	85	125	95	110	132	95	103	81

Construye una tabla tomando intervalos de amplitud 10 comenzando por 80, y determina:

a) Frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

b) ¿Cuántos alumnos tienen un cociente intelectual por debajo de 100?.

c) Si se consideran superdotados a partir de 130, ¿hay alguno en clase?.

d) ¿Qué porcentaje de alumnos tiene un cociente intelectual 110 o más?.

e) Representa mediante un histograma y un polígono de frecuencias los datos de la tabla.

- **HORAS DE SOL**

El número de horas de sol registradas en el mes de enero de 2000 en 50 estaciones metereológicas es:

83	82	78	72	107	107	93	72	85	98
71	76	75	83	72	126	102	67	112	99
155	118	150	129	119	148	181	151	167	156
180	173	149	80	131	121	110	200	162	214
176	186	187	186	141	212	186	199	198	219

Forma una tabla de intervalos de amplitud 20 tomando 70 como extremo inferior del primer intervalo y determina media, mediana y moda. ¿Cuál de ellas crees que resulta más representativa?.

- **ELECCIONES AUTONÓMICAS**

Las elecciones autonómicas de Andalucía del 12 de junio de 1994 arrojaron los siguientes resultados:

	Censo	Votantes
Almería	358834	243636
Cádiz	815074	490488
Córdoba	588278	429121
Granada	634462	436173
Huelva	342787	214608
Jaén	498134	374801
Málaga	888952	570837
Sevilla	1263031	866381
Andalucía	4389552	3626045

a) Calcula el porcentaje de votantes.

b) Dibuja el diagrama de barras para cada provincia del censo y votantes.

c) Dibuja el gráfico de sectores de los votantes por provincia.

- **ZAPATERÍA**

Una zapatería de caballeros vende en un día 40 pares de zapatos de los siguientes números (tallas):

39	39	40	43	39	38	41	40	41	39
40	41	41	37	42	40	41	42	42	43
42	40	41	43	38	41	42	41	42	42
44	41	42	39	41	40	44	43	40	40

a) Construye una tabla de frecuencias absolutas de tipo discreto y dibuja un polígono de frecuencias. ¿Qué porcentaje de zapatos es de cada talla?.

b) ¿Cuántos pares hay en el intervalo $(\bar{x}-\sigma, \bar{x}+\sigma)$?.

c) ¿Qué talla de zapatos se vende más?.

d) ¿Qué talla de zapatos es la mediana?.

- **CALIFICACIONES**

Las calificaciones en Matemáticas de 25 alumnos del grupo A son:

6	6	7	6	7	5	5	6	7	5	4	5	4
9	3	3	5	5	5	9	5	4	5	4	8	

mientras que las de los 20 alumnos del grupo B fueron:

6	6	7	3	10	3	5	5	2	5
4	3	9	4	9	5	6	6	6	7

a) ¿En qué grupo los alumnos obtuvieron mejor nota media?.

b) ¿En qué grupo las notas están más dispersas?.

- **TEMPERATURAS**

Las temperaturas en Valencia a lo largo del año 2001 fueron:

	E	F	Mz	Ab	My	Jn	Jl	Ag	S	O	N	D
Máxima	17,0	15,5	19,5	23,6	25,7	31,0	32,5	33,9	29,8	23,7	18,9	18,8
Mínima	2,8	6,8	7,0	9,2	12,9	17,1	19,9	20,7	16,1	12,1	9,0	5,5

a) ¿Cuál fue la temperatura media (máxima y mínima) del año?.

b) ¿Cuál fue la mediana de las temperaturas máximas?. ¿Y de las mínimas?.

c) Si se seleccionan como meses más estables aquellos cuya temperatura se encuentra en el intervalo $(\bar{x}-\sigma, \bar{x}+\sigma)$, ¿qué meses son estables en sus temperaturas máximas?. ¿Y en sus mínimas?. ¿Qué meses son estables en ambas?.

- **COCHES**

En una población de 25 familias se ha observado la variable X=número de coches que tiene la familia y se han obtenido los siguientes datos:

0	1	2	3	1	0	1	1	1	4	3	2	2	1	1	2	2	1	1	1	2	1	3	2	1

a) Construye la tabla de frecuencias de la distribución de X.

b) Construye el diagrama de barras y explica si es simétrica la distribución.

c) Calcula la moda, la media y la mediana.

- **EDADES**

Dos equipos de baloncesto tienen cada uno en su plantilla 15 jugadores con las siguientes edades:

Equipo A	18	20	22	23	25	25	26	26	26	26	26	26	27	27	27
Equipo B	20	20	21	21	21	21	22	23	23	23	24	24	25	25	27

Representa gráficamente los datos y calcula, para cada equipo:

a) Edad media.

b) Varianza.

c) Desviación típica.

- **VOCABULARIO**

Se ha realizado un test de vocabulario a un grupo de 100 alumnos obteniéndose los siguientes resultados:

46	58	52	49	42	45	56	38	57	42	50	60	17	54	41	51	27	62	51	64
57	63	31	41	47	30	54	67	50	47	48	53	50	39	64	49	60	37	51	63
62	28	44	31	56	16	68	39	46	70	39	59	30	57	40	51	53	55	50	42
24	51	47	57	34	58	40	47	63	46	57	20	70	52	38	59	46	51	57	26
58	52	61	53	44	62	50	62	48	55	33	68	32	61	38	49	25	45	56	48

a) Realiza la tabla de frecuencias agrupando los datos en intervalos.

b) Representa gráficamente la distribución.

- **DIANAS**

Las dianas logradas por 26 jugadores en un campeonato fueron:

8	10	12	12	10	10	11	11	10	13	9	11	10
9	9	11	12	12	9	10	9	10	9	10	8	10

a) Resume los datos en una tabla de frecuencias.

b) Dibuja el diagrama de barras y el polígono de frecuencias correspondiente.

c) Halla los parámetros estadísticos de centralización y dispersión e interprétalos.

ANÁLISIS DE DATOS Y ANÁLISIS DE REGRESIÓN CON EXCEL

Introducción

Uno de los problemas de la Estadística consiste en determinar si existe relación o no entre dos variables y encontrar un modelo adecuado que se ajuste a los datos. Estas cuestiones son estudiadas en Correlación y Regresión. El uso de una hoja de cálculo como Excel permite realizar muchas de las tareas relativas a variables bidimensionales.

A continuación veremos como se crean y modifican distintos tipos de hojas de cálculo asociadas a problemas típicos de Correlación y Regresión. Analizaremos el uso de las posibilidades gráficas de Excel para describir datos estadísticos bidimensionales.

1. Incidencia de la TV en el rendimiento escolar

Queremos estudiar la incidencia que tiene el número de horas que dedican a ver la televisión los alumnos de un grupo de Bachillerato en su rendimiento académico. Para ello, hemos tomado una muestra de 7 alumnos, obteniendo los resultados que se muestran en la **Hoja1** del libro **Bidimensional**. Calcula:

a) El centro de gravedad.

b) Las desviaciones típicas marginales.

c) La covarianza.

d) El coeficiente de correlación.

e) Representa la nube de puntos.

f) Halla y representa la recta de regresión.

g) ¿Cuántas asignaturas cabe esperar que aprobará un alumno que ve la TV 3 horas?.

Para resolver el problema, sigue los siguientes pasos:

- Abre el libro **Bidimensional** situado en la carpeta **Estadistica1**. Comprueba que en la **Hoja1** de dicho libro están disponibles los siguientes datos:

	A	B	C
1	**Incidencia de la TV en el rendimiento escolar**	**Horas TV**	**Aprobados**
2		1	7
3		5	2
4		2	4
5		4	3
6		2	5
7		0	6
8		4	4
9	**Centro de gravedad**		
10	**DT marginales**		
11	**Covarianza**		
12	**Coeficiente de correlación, r**		
13	**Aplicar la fórmula**		

En las celdas siguientes escribe la fórmula que se indica:

- En la celda **B9** escribe la fórmula **=PROMEDIO(B2: B8)** y arrastra el controlador de relleno de la celda **B9** hasta **C9**.
- En la celda **B10** escribe la fórmula **=DESVESTP(B2: B8)** y arrastra el controlador de relleno de la celda **B10** hasta **C10**.
- En la celda **B11** escribe la fórmula **=COVAR(B2: B8; C2: C8)**.
- En la celda **B12** escribe la fórmula **=COEF.DE.CORREL(B2: B8; C2: C8)**.
- Selecciona los datos del rango **B2: C8**, haz clic en el botón **Asistente para gráficos**.
- **Paso 1 de 4 – Tipo de gráfico**: XY (Dispersión). **Subtipo de gráfico**: Compara pares de valores.
- **Paso 2 de 4 – Datos de origen**: haz clic en **Siguiente** (los datos ya están elegidos).
- **Paso 3 de 4 – Opciones de gráfico**: rellena la ficha **Títulos**, escribiendo como título del gráfico **"TV y rendimiento escolar"**, como título del eje X, **Horas TV** y como título del eje Y, **Nº de aprobados**. Desactiva la ficha **Leyenda**.
- **Paso 4 de 4 – Ubicación del gráfico**: elige en la misma hoja.
- En el gráfico selecciona los puntos haciendo clic en uno de ellos, activa el menú contextual haciendo clic con el botón derecho, elige **Agregar línea de tendencia…** en la ficha **Tipo: Lineal**, en la ficha **Opciones** activa la casilla **Presentar ecuación en el gráfico** y haz clic en **Aceptar**.
- Luego retoca el gráfico para que mejore la apariencia.
- En la celda **B13** escribe **3** y en la **C13** la fórmula **= – 0'8478 * B13 + 6'6087**.

2. Relación talla – peso

Hemos realizado una encuesta a un grupo de 8 personas que tuvieran una buena relación entre la estatura y el peso, obteniendo los resultados que se indican en la **Hoja2** del libro **Bidimensional**. Calcula:

a) El centro de gravedad.
b) Las desviaciones típicas marginales.
c) La covarianza.
d) El coeficiente de correlación.
e) Representa la nube de puntos.
f) Halla y representa la recta de regresión.
g) ¿Cuánto cabe esperar que pese una persona que mide 162 cm?.
h) ¿Cuánto cabe esperar que mida una persona que pesa 67 kg?.
i) Interpreta los resultados obtenidos en la covarianza y en el coeficiente de correlación.

Para resolver el problema sigue los siguientes pasos:

- Abre la **Hoja2** del libro **Bidimensional** y comprueba que contiene los siguientes datos:

	A	B	C
1	**Relación estatura peso**	**Talla (cm)**	**Peso (kg)**
2		156	56
3		165	65
4		170	70
5		175	75
6		165	65
7		172	72
8		178	78
9		160	60
10	**Centro de gravedad**		
11	**DT marginales**		
12	**Covarianza, Sxy**		
13	**Coeficiente de correlación, r**		
14	**Aplicar la fórmula**		
15	**Buscar objetivo…**		

- Resuelve los apartados (a), (b), (c), (d), (e) y (f), como en el ejercicio anterior, escribiendo la fórmula adecuada en las celdas correspondientes.

- Cambia la escala en el eje OY. Para ello selecciónalo haciendo clic en él con botón derecho del ratón, y en el menú contextual correspondiente, elige **Formato de ejes**. En la ficha **Escala**, en **Mínimo**, escribe **50** y en **Máximo 80**.

- En la celda **B14** escribe **162** y en la **C14** la fórmula **= B14 – 100**.

- En la celda **C15** escribe la fórmula **=B15 – 100**, elige en el menú **Herramientas / Buscar objetivos**. En la ventana que aparece, escribe en **Definir celda**: **C15**, **con el valor**: **67**, **para cambiar la celda**: **B15** y haz clic en **Aceptar**.

3. Dos ejemplos comentados

Ejemplo 1.– Se observaron las edades de cinco niños y sus pesos respectivos, y se obtuvieron los resultados siguientes:

Edad (años)	*2*	*4,5*	*6*	*7,2*	*8*
Peso (kg)	*15*	*19*	*25*	*33*	*34*

a) Halla las medias y desviaciones marginales.

b) Calcula el coeficiente de correlación lineal y la recta de regresión del peso sobre la edad.

Introducimos los datos en las celdas elegidas: en la columna **A**, la edad, y en la **B**, el peso. A continuación seleccionamos la opción **Insertar / Función**, apareciendo la ventana de diálogo **Pegar función**, donde debemos seleccionar funciones **Estadísticas** y la función que queremos calcular; o bien directamente, si conocemos la sintaxis de las funciones estadísticas, editamos dichas funciones. Situamos el puntero en la columna **D** y vamos tecleando cada una de las funciones estadísticas en la barra de fórmulas, situando el puntero cada vez en una celda distinta para ir conservando los datos.

=PROMEDIO(A1: A5) obtenemos la media de la edad, en la celda **D2**.
=PROMEDIO(B1: B5) media del peso, en la celda **D3**.
=DESVESTP(A1: A5) desviación típica de la edad, en la celda **D4**.
=DESVESTP(B1: B5) la desviación típica del peso, en la celda **D5**.

	A	B	C	D
1	2	15		
2	4,5	19	**Media edad:**	
3	6	25	**Media peso:**	
4	7,2	33	**Desv. típica edad:**	
5	8	34	**Desv. típica peso:**	
6	**Edad**	**Peso**		
7	**coeficiente de**	**correlación:**		
8	**pendiente recta**	**regresión Y sobre X:**		

Para calcular el coeficiente de correlación, tecleamos en la barra de fórmulas: **=COEF.DE.CORREL(A1: A5, B1: B5)**, una vez situados en la celda **C7**.

=PENDIENTE(B1: B5, A1: A5) nos informa sobre la pendiente de la recta de regresión del peso sobre la edad, en la celda **D8**.

Por tanto, la recta de regresión es: **Y–25,2 = 3,4049 · (X–5,54)**.

También podemos calcular la recta de regresión seleccionando dos celdas vacías contiguas, activando el comando **Insertar / Función** y eligiendo la función **ESTIMACION.LINEAL** de las funciones estadísticas. En el cuadro de diálogo correspondiente introducimos el rango **B1: B5** en la caja **Conocido_y**, e introducimos el rango **A1: A5** en la caja **Conocido_x**. Al pulsar **CTRL + ↑ + ENTER**, en la pantalla aparecen los valores **a** y **b**, siendo **Y = a ·X + b** la recta de regresión de **Y** sobre **X**. Por tanto, la recta de regresión es: **Y = 3,4049·X + 6,33678**.

Excel incorpora dos funciones que nos permiten predecir el valor de una variable, conocido el valor de la otra, por ejemplo, tecleando **=TENDENCIA(B1:B5, A1: A5, 5)** obtenemos el peso esperado para una edad de cinco años.

Y la función que nos mide el error estimado de una variable al ser estimado su valor por la recta de regresión, **=ERROR.TÍPICO.XY.(B1: B5, A1: A5)** devuelve el error típico del valor de **Y** previsto para cada **X**.

También podemos hacer el diagrama de dispersión. Marcamos los datos introducidos, pulsamos el botón de gráficos, seleccionamos diagrama de dispersión y a través de ventanas de diálogo damos nombre a los ejes y hacemos la división de los mismos.

Ejemplo 2.– La siguiente tabla representa los pesos y las alturas de 20 jóvenes:

Peso (X)	*73*	*76*	*73*	*78*	*80*	*82*
Altura (Y)	*1,65*	*1,68*	*1,70*	*1,72*	*1,76*	*1,80*
Nº de jóvenes	*4*	*3*	*2*	*5*	*4*	*2*

a) ¿Qué tipo de correlación existe entre las variables X e Y?.

b) ¿Cuál es la altura estimada para un joven que pese 75 kg?.

c) ¿Cuál será el peso estimado para un joven de altura 1,73 m?.

- En las celdas **A1** y **B1** teclea **PESO (X)** y **ALTURA (Y)** respectivamente. A continuación introduce los datos del problema en el rango de celdas **A2: B21**, de forma que cada pareja de valores (peso, altura) aparezca repetida tantas veces como indica su frecuencia.

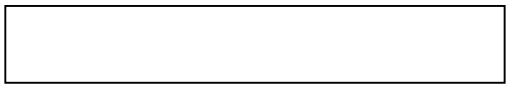
- Para ver el tipo de correlación existente entre las variables X e Y, selecciona el rango de celdas **A2: B21** y haz clic en el botón **Asistente para gráficos**. Selecciona el diagrama **(XY)Dispersión** y elige el primer subtipo de gráfico. Haz clic en **Siguiente**. Introduce como título del gráfico **PESOS Y ALTURAS**, como título del eje de valores **(X) PESOS** y como título del eje de valores **(Y) ALTURAS**. En la etiqueta **Leyenda** desactiva la opción **Mostrar leyenda**. Haz clic en **Terminar**. De esta forma hemos obtenido el diagrama de dispersión. A la vista del gráfico, ¿qué tipo de correlación crees que existe entre las variables PESO y ALTURA?.

- En la celda **C7** teclea COEFICIENTE DE CORRELACIÓN. En la celda **E7** introduce la fórmula **COEF.DE.CORREL(A2: A21; B2: B21).** El resultado obtenido es el coeficiente de correlación lineal entre las variables EDAD y PESO.

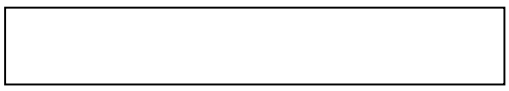
- Para estimar la altura de un joven que pese 75 kg, utilizaremos el siguiente procedimiento. En la celda **C9** teclea ALTURA ESTIMADA PARA 75 KG. En la celda **E9** introduce la fórmula **=TENDENCIA(B2: B21; A2: A21;75)** y pulsa ENTER. El resultado es la estimación de la altura de un joven de 75 kg de peso.

- Para estimar el peso de un joven de 1’73 m de altura, en la celda **C11** teclea PESO ESTIMADO PARA 1’73 M. En la celda **E11** introduce la fórmula **=TENDENCIA(A2:A21;B2:B21;1’73)** y pulsa ENTER. El resultado es la estimación del peso de un joven de 1’73 m de altura.

- Para completar el problema, vamos a calcular las rectas de regresión de pesos sobre alturas y de alturas sobre pesos. En la celda **C13** teclea RECTA DE ALTURAS SOBRE PESOS. A continuación, selecciona el rango de celdas **C14: D14**. Haz clic en el botón **Asistente para funciones** de la barra de herramientas estándar y en la categoría **Estadísticas** selecciona **ESTIMACION.LINEAL**. Haz clic en **Aceptar**. En el campo **Conocido_y** introduce el rango **B2: B21**. En el campo **Conocido_x** introduce el rango **A2: A21**. En lugar de hacer clic en **Aceptar**, pulsa la combinación de teclado **CTRL+MAYÚS+ENTER**. De esta forma aparecerá en el rango **C14: D14** la ecuación de la recta de regresión de Y sobre X: en la celda **C14** aparece la pendiente y en la celda **D14** la ordenada en el origen. Por tanto, la ecuación de la recta de regresión de alturas sobre pesos es:



- En la celda **C16** teclea RECTA DE PESOS SOBRE ALTURAS. A continuación, selecciona el rango de celdas **C17: D17**. Haz clic en el botón **Asistente para funciones** de la barra de herramientas estándar y en la categoría **Estadísticas** selecciona **ESTIMACION.LINEAL**. Haz clic en **Aceptar**. En el campo **Conocido_y** introduce el rango **A2: A21**. En el campo **Conocido_x** introduce el rango **B2: B21**. En lugar de hacer clic en **Aceptar**, pulsa la combinación de teclado **CTRL+MAYÚS+ENTER**. De esta forma aparecerá en el rango **C17: D17** la ecuación de la recta de regresión de X sobre Y: en la celda **C17** aparece la pendiente y en la celda **D17** la ordenada en el origen. Por tanto, la ecuación de la recta de regresión de pesos sobre alturas es:



- Vamos a dibujar ahora sobre el diagrama de dispersión las dos rectas de regresión. Para ello haz clic sobre el área del gráfico y elige la opción **Gráfico**, **Agregar línea de tendencia**. En la etiqueta **Tipo** elige **Lineal**. En la etiqueta **Opciones** activa las casillas **Presentar ecuación en el gráfico** y **Presentar el valor R cuadrado en el gráfico**. Haz clic en **Aceptar** y observa el resultado. La recta obtenida es la recta de regresión de Y sobre X (alturas sobre pesos). ¿Cómo puedes dibujar la recta de regresión de X sobre Y?.

4. Nube de puntos

- Abre la **Hoja3** del libro **Bidimensional** y comprueba que contiene los siguientes datos, relativos al proceso de producción de una fábrica:

Horas	40	41	42	39	40	38	42	43	38	39	40	41	40	42	41	42	39
Piezas	50	52	51	52	53	49	55	54	50	50	52	54	51	53	53	54	51

- Construye la nube de puntos, junto con la recta de regresión. Introduce como título del gráfico **"Producción de piezas"**, como título del eje OX, **"Número de horas"** y como título del eje OY, **"Número de piezas"**.

ACTIVIDADES

- **NOVIOS**

Las edades de los novios en las bodas celebradas durante una semana en una ciudad quedan reflejadas en la siguiente tabla:

Edad del novio	25	27	31	34	36	40	45
Edad de la novia	18	29	25	27	27	30	36

a) Determina el coeficiente de correlación lineal entre ambas edades.

b) Predice la edad del novio que se casa con una mujer de 20 años.

- **BIBLIOTECAS**

En las bibliotecas de seis poblaciones se han analizado conjuntamente la afluencia de lectores (X en miles de personas) y el número de libros prestados (Y), obteniéndose los datos de la tabla:

X	0'5	1	1'3	1'7	2	2'5
Y	180	240	250	300	340	400

a) ¿Cuál es el número medio de libros prestados en el conjunto de todas las bibliotecas?.

b) Ajusta una recta para explicar el nº de libros prestados a partir de la afluencia de lectores.

c) Si acudiesen 1500 lectores a una biblioteca, ¿cuántos libros se prestarían?.

- **CONTAMINACIÓN SONORA**

La siguiente tabla indica el número de turismos matriculados y el nivel de ruido por trimestre en la ciudad de Valencia durante el periodo 1995 – 1998.

		Turismos matriculados	**Nivel medio sonoro en dB**
1995	Primer trimestre	4225	69’92
	Segundo trimestre	5687	69’97
	Tercer trimestre	3873	67’91
	Cuarto trimestre	5188	70
1996	Primer trimestre	4396	70’25
	Segundo trimestre	5099	70’52
	Tercer trimestre	4951	70’11
	Cuarto trimestre	5447	70’67
1997	Primer trimestre	4994	70’45
	Segundo trimestre	5955	70’82
	Tercer trimestre	5638	70’89
	Cuarto trimestre	5776	70’98
1998	Primer trimestre	5622	70’67
	Segundo trimestre	6937	71’34
	Tercer trimestre	6747	71’23
	Cuarto trimestre	7247	71’79

a) Dibuja el diagrama de dispersión e indica qué tipo de correlación existe entre el número de turismos matriculados y el nivel sonoro.

b) Halla el coeficiente de correlación lineal entre dichas variables.

c) Dibuja, sobre el diagrama de dispersión la recta de regresión del nivel de ruido sobre el número de turismos matriculados.

d) ¿Qué ocurrirá si continua aumentando el número de turismos matriculados?. ¿Qué nivel medio de ruido cabe esperar en el trimestre en que se matriculen 8000 turismos?. ¿Y cuando se matriculen 9000?. Si el ritmo de crecimiento se mantiene, ¿ha de pasar mucho tiempo para que se alcancen 10000 vehículos matriculados en un trimestre?.

e) ¿Qué grado de seguridad te merecen las estimaciones que has hecho anteriormente?.

- **PUBLICIDAD**

La siguiente tabla muestra los gastos (en decenas de miles de €) de cinco campañas publicitarias junto con los consiguientes volúmenes de ventas (en decenas de miles de €) obtenidos de cierto artículo.

Gastos de publicidad	2	3	5	6	10
Volumen de ventas	50	60	120	150	180

Calcula el coeficiente de correlación y la recta de regresión de la variable “volumen de ventas” sobre la variable “gastos de publicidad”. Utiliza dicha recta para predecir el volumen de ventas que podría esperarse con unos gastos publicitarios de ocho millones. Valora la bondad de predicción del coeficiente de correlación obtenido.

5. Procedimientos rápidos con Excel

También con Excel podemos utilizar algoritmos rápidos que permiten describir datos bidimensionales y obtener la recta de regresión con suma facilidad. Veamos un ejemplo.

Diòxid de sofre i fum, per mesos, segons zones. 1998

La siguiente tabla muestra las concentraciones de dióxido de azufre y humo en una zona de tráfico denso de la ciudad de Valencia, durante cada mes de 1998.

	Zona Trànsit dens	
	Diòxid de sofre	Fum
Gener	26	44
Febrer	19	58
Març	22	49
Abril	20	31
Maig	22	33
Juny	24	28
Juliol	28	36
Agost	19	38
Setembre	24	50
Octubre	20	60
Novembre	23	55
Desembre	21	79
Mitjana Anual	22	47

Nota: Concentracions en micrograms / m3
Font: Laboratori Municipal. Ajuntament de València

d) Halla la covarianza y el coeficiente de correlación.

e) Obtén la ecuación de la recta de regresión que explique las concentraciones de dióxido de azufre en función de la cantidad de humo.

f) ¿Es razonable utilizar dicha recta para predecir la cantidad de azufre cuando se conoce la concentración de humo?. ¿Por qué?.

- **COVARIANZA Y COEFICIENTE DE CORRELACIÓN**

- Una vez iniciado el programa Excel, haz clic en el botón **Nuevo** para crear un nuevo libro de trabajo.

- Introduce los datos a partir de la celda **A1**, de forma que el primer dato numérico (correspondiente al dióxido de azufre del mes de Enero) aparezca en la celda **B3**.

- A continuación elige el comando **Herramientas / Análisis de datos** y en la ventana **Funciones para ánalisis** selecciona **Covarianza** y haz clic en **Aceptar**.

- Haz clic en la casilla **Rango de entrada** y, con el ratón, selecciona el rango de celdas **$B$2: $C$14**, que contiene los datos de las dos variables, incluyendo sus nombres. Procura que esté seleccionada la opción **Agrupado por Columnas**, dado que las variables se corresponden con dos columnas diferentes.

- Haz clic en la casilla **Rótulos en la primera fila** para indicar a Excel que la primera fila del rango seleccionado contiene los nombres de las variables.

- Haz clic en la casilla **Rango de salida** y, con el ratón, selecciona una celda vacía, por ejemplo, **$D$1**, a partir de la cual se mostrarán los resultados.

El resultado es la denominada matriz de covarianzas, en la que cada celda indica los valores de las varianzas y de la covarianza:

	Diòxid de sofre	Fum
Diòxid de sofre	7,878787879	
Fum	-12,09090909	219,477273

que se corresponde con la matriz:

	X	**Y**
X	Var(X)	Cov(X, Y)
Y	Cov(X, Y)	Var(X)

Por tanto, la covarianza entre las dos variables es Cov(X, Y) = –12,09.

- Elige el comando **Herramientas / Análisis de datos** y en la ventana **Funciones para ánalisis** selecciona **Coeficiente de correlación** y haz clic en **Aceptar**.

- Haz clic en la casilla **Rango de entrada** y, con el ratón, selecciona el rango de celdas **$B$2: $C$14**, que contiene los datos de las dos variables, incluyendo sus nombres. Procura que esté seleccionada la opción **Agrupado por Columnas**, dado que las variables se corresponden con dos columnas diferentes.

- Haz clic en la casilla **Rótulos en la primera fila** para indicar a Excel que la primera fila del rango seleccionado contiene los nombres de las variables.

- Haz clic en la casilla **Rango de salida** y, con el ratón, selecciona una celda vacía, por ejemplo, **$D$5**, a partir de la cual se mostrarán los resultados.

El resultado es la denominada matriz de correlacioones, en la que cada celda indica los valores de los coeficientes de correlación entre las variables.

	Diòxid de sofre	Fum
Diòxid de sofre	1	
Fum	-0,290759876	1

que se corresponde con la matriz:

	X	**Y**
X	1	Corr(X, Y)
Y	Corr(X, Y)	1

Por tanto, el coeficiente de correlación entre las dos variables es Corr(X, Y) = –0,29. Lo que parece indicar que no hay correlación entre las variables.

- **RECTA DE REGRESIÓN**

- Elige el comando **Herramientas / Análisis de datos** y en la ventana **Funciones para ánalisis** selecciona **Regresión** y haz clic en **Aceptar**.

- En la casilla **Rango Y de entrada** introduce el rango de celdas **$B$2: $B$14**, que contiene los datos del dióxido de azufre. En la casilla **Rango X de entrada** introduce el rango de celdas **$C$2: $C$14**, que contiene las concentraciones de humo.

- Haz clic en la casilla **Rótulos** para indicar a Excel que la primera fila de los rangos anteriores contiene los nombres de las variables.

- Haz clic en la casilla **Rango de salida** y selecciona una celda vacía, por ejemplo, **$D$9**, a partir de la cual se mostrarán los resultados. También puedes mostrar los resultados en una hoja nueva o en un nuevo libro de trabajo.

- Haz clic en la casilla **Curva de regresión ajustada**. Opcionalmente, puedes activar cualquiera de las casillas referidas a los residuos o al gráfico de probabilidad normal. Finalmente, haz clic en **Aceptar**.

A partir de la celda **D9** se muestra un resumen de los estadísticos, con el valor del coeficiente de correlación y del coeficiente de determinación. La recta de ajuste es **Y=Intercepción+Fum*X**, es decir, la pendiente es **Fum** y la ordenada en el origen es **Intercepción**. En este caso: **Y=24,9–0,055*X**.

Como el coeficiente de determinación, 0.08, es próximo a 0, no debe usarse la recta de regresión obtenida para hacer estimaciones. Observamos que, junto a los cálculos, Excel da la curva de regresión ajustada, junto con el diagrama de dispersión. En esta gráfica vemos que el ajuste es realmente malo. Por lo tanto, el modelo lineal no es explicativo de los datos.

ACTIVIDADES

Utiliza los procedimientos rápidos de Excel para resolver los siguientes problemas:

- **PRODUCCIÓN MUNDIAL**

La producción mundial de cereales en los años indicados es la siguiente:

AÑO	1950	1960	1970	1980	1990	2000
PRODUCCIÓN (millones de toneladas)	70	110	150	160	180	200

a) Representa gráficamente estos datos e indica si hay correlación (usa sólo las 2 últimas cifras del año). Escribe la ecuación de la recta de regresión.
b) Si se mantiene el ritmo de crecimiento, ¿cuál será la producción en año 2020?.

- **DENSIDAD**

Para calcular la densidad de cierto líquido, se pesan distintos volúmenes obteniéndose:

MASA (g)	193	68	359	511	734	947
VOLUMEN (cm^3)	250	100	500	750	1000	1350

Sabiendo que $\rho = m/v$, donde ρ es la densidad, m la masa y v el volumen, como m=ρv, esperamos que la gráfica de m en función de v sea una recta de pendiente ρ y ordenada en el origen 0. Calcula la densidad del líquido, el coeficiente de correlación y la ecuación de la recta.

ANÁLISIS DE DATOS Y ANÁLISIS DE REGRESIÓN CON STATGRAPHICS PARA WINDOWS

1. Introducción

En este capítulo intentaremos ajustar modelos funcionales entre dos variables con ayuda de STATGRAPHICS. Veremos en primer lugar un resumen de los conceptos básicos y después estudiaremos cómo hacer un análisis de regresión y medir la correlación correspondiente.

2. Operaciones básicas

- **CONCEPTOS BÁSICOS**

Cuando se realiza la observación simultánea de dos características (edad y peso, tensión arterial y número de hermanos,…) se obtiene un conjunto de pares que dan lugar a una variable estadística bidimensional. Se puede representar por el par (X, Y) y está formada por un conjunto de valores de la forma (x_1, y_1), (x_2, y_2), (x_3, y_3),...(x_n, y_n). La representación gráfica de estos puntos es un diagrama de dispersión. Los valores x_i corresponden a una variable unidimensional, lo mismo que los valores y_i. Cada una de las variables X e Y tiene sus propios parámetros estadísticos, pero pretendemos analizar la variación conjunta de las dos variables.

- **COVARIANZA**

La covarianza es un indicador de cuál es la variación conjunta de X e Y. La covarianza se calcula a partir del producto de las dos diferencias con respecto a sus medias para cada una de las variables, es decir:

$$S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n}$$

que en el programa STATGRAPHICS se convierte en: $S_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$.

La covarianza es positiva cuando la relación entre las variables es directa (cuando aumenta o disminuye una, aumenta o disminuye la otra), negativa cuando es inversa y cero cuando no hay relación. Cuanto mayor sea la relación mayor será el valor de la covarianza.

- **COEFICIENTE DE CORRELACIÓN LINEAL**

Si se divide la covarianza por el producto de las cuasi–desviaciones típicas de las dos variables, se consigue un valor adimensional al que llamamos coeficiente de correlación lineal de Pearson, r, que toma valores entre –1 y +1. Cuanto más próximo a estos dos valores, mayor será la relación entre las variables.

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

Al valor r^2 se le denomina coeficiente de determinación, e indica el porcentaje de la variabilidad de una variable explicable por la otra. Varía entre 0 (no existe relación) y 1 (relación perfecta tanto directa como inversa).

- **REGRESIÓN LINEAL**

La regresión permite encontrar una función matemática que ajuste de la mejor manera posible los valores de las dos variables X e Y. En el caso particular de la regresión lineal, se trata de obtener una recta que ajuste la nube de puntos. La ecuación de la recta, que permitirá pronosticar los valores de Y conocidos los de X, será por tanto de la forma: $y = a + bx$.

Para una observación concreta (x_i, y_i) habrá una diferencia entre el valor pronosticado a través de la recta para el valor x_i, llamémosle y_i^*, y el valor cierto real, y_i, y se produce por tanto un error que será la diferencia entre ambos valores, al que se denomina **residuo**. La suma de todos los residuos es cero (como puedes comprobar), pero lo que pretendemos es que la suma de los errores al cuadrado sea mínima. Si imponemos esta condición, obtenemos los valores de a y b de la recta de regresión (de Y sobre X) que sería:

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} \cdot \left(x - \bar{x} \right)$$

Tanto el coeficiente de correlación lineal, r, como el coeficiente de determinación, r^2, indican la bondad del ajuste. El modelo será más fiable cuanto más próximo a 1 o a –1 sea el coeficiente de correlación r, o cuanto más próximo a 1 sea el coeficiente de determinación r^2.

- **REGRESIÓN NO LINEAL**

A veces el modelo lineal no es adecuado al fenómeno que se analiza. Hay otros métodos basados en el mismo principio de ajuste que la regresión lineal (minimizar la suma de los cuadrados de los residuos), en muchos casos reducibles al modelo lineal (transformando las variables). Algunos ejemplos son los siguientes:

Regresión exponencial:	$y = a \cdot e^{b x}$
Regresión potencial:	$y = a \cdot x^b$
Regresión logarítmica:	$y = a + b \cdot \ln(x)$

- **REGRESIÓN POLINÓMICA**

La regresión polinómica consiste en encontrar una función polinómica (una parábola, una cúbica, …) de grado n que ajuste la nube de puntos: $y = a + b x + c x^2 + ... + m x^n$

- **NUBE DE PUNTOS**

Consideremos por ejemplo la siguiente variable bidimensional (X, Y):

X	3	2	5	6	7
Y	6	6	7	9	9

Para comprobar si hay relación o no entre las variables, representamos gráficamente la nube de puntos asociada con STATGRAPHICS de la siguiente forma:

- Inicia el programa y crea un archivo de datos con las dos variables X e Y de la tabla.
- Selecciona **Gráficos / Gráficos de Dispersión / Gráfico X–Y**.
- En la siguiente ventana introduce la variable Y en la caja **Y** y la variable X en la caja **X**.
- Haz clic en el botón **Aceptar**. Aparece el gráfico solicitado. Haz doble clic sobre el panel gráfico para maximizarlo.

A pesar de tratarse de pocas observaciones, el gráfico muestra que puede haber una relación lineal fuerte entre las dos variables.

- **CÁLCULO DE LAS VARIANZAS, CORRELACIONES Y COVARIANZA**

Con STATGRAPHICS es posible calcular rápidamente las varianzas de cada variable por separado (varianzas marginales) y la covarianza. Para ello:

- Selecciona **Descripción / Datos Numéricos / Análisis Multidimensional**.
- Introduce en el cuadro de diálogo que aparece las dos variables y haz clic sobre el botón **Aceptar**.
- Haz clic sobre el botón **Opciones Tabulares**.
- Haz clic en la casilla **Covarianzas** para obtener las varianzas y la covarianza.
- Haz clic sobre la casilla **Correlaciones** para obtener las correlaciones.
- Haz clic, además, sobre la casilla **Resumen Estadístico**, para obtener las medidas estadísticas de cada variable (distribución marginal).
- Haz clic en el botón **Aceptar**.
- Haz doble clic sobre el panel de covarianzas.

Obtendrás los datos en forma de matriz. En ella aparece la varianza (en realidad, la cuasivarianza) de la primera variable X (en este caso, 4,3), la varianza de la segunda Y (en este caso, 2,3) y la covarianza entre las variables X e Y (en este caso, 2,95).

- Haz doble clic sobre el panel de las correlaciones.

Obtendrás, también en forma de matriz, la correlación entre ambas variables (en este caso, 0,938, que es bastante alta).

- Haz doble clic sobre el panel **Resumen Estadístico**.

Obtendrás los parámetros estadísticos de las dos variables X e Y, tal como se indica en la siguiente tabla:

	X	Y
Frecuencia	5	5
Media	4,6	7,4
Varianza	4,3	2,3
Desviación típica	2,07364	1,51658
Mínimo	2,0	6,0
Máximo	7,0	9,0
Asimetría tipificada	–0,214994	0,287879
Curtosis tipificada	–0,896085	–1,40641
Suma	23,0	37,0

En caso de que necesites otro parámetro estadístico que no esté en esta lista, puedes hacer clic con el botón derecho del ratón sobre el panel y seleccionar Opciones de Ventana en el menú contextual. En el siguiente cuadro de diálogo puedes elegir el parámetro que necesites.

- Haz clic con el botón derecho del ratón sobre el panel **Resumen Estadístico**.
- En el menú de contexto selecciona **Opciones de Ventana**.
- En la siguiente ventana haz clic sobre las casillas **Mediana**, **Moda**, **Primer Cuartil**, **Tercer Cuartil**, **Coeficiente de Variación** y haz clic en el botón **Aceptar**. Observa el resultado.

- **RECTA DE REGRESIÓN**

Una vez comprobada la existencia de correlación lineal entre las dos variables, podemos proceder a efectuar el análisis de regresión. Para obtener la ecuación de la recta de regresión y el coeficiente de correlación, sigue los siguientes pasos:

- Selecciona **Dependencia / Regresión Simple**.
- Introduce la variable dependiente Y en la caja **Y**. Introduce la variable independiente X en la caja **X**.
- Haz clic sobre el botón **Aceptar**.

En el panel de la ventana de análisis se pueden observar los datos más relevantes:

* La ecuación de la recta de regresión buscada es: $Y = 4{,}24419 + 0{,}686047 \cdot X$, ya que a=Ordenada y b=Pendiente.
* El coeficiente de correlación es r = 0,938045 (dato que ya conocíamos de apartados anteriores), e indica una fuerte relación lineal entre las variables.
* El valor del coeficiente de determinación es $r^2 = 87{,}9929$, lo que indica que una gran parte de una variable está explicada por la otra (prácticamente el 88%).

- Haz clic en el botón **Opciones Gráficas**.
- Selecciona la opción **Gráfico del Modelo Ajustado** y haz clic en el botón **Aceptar**.
- Haz doble clic sobre el panel gráfico para maximizarlo.

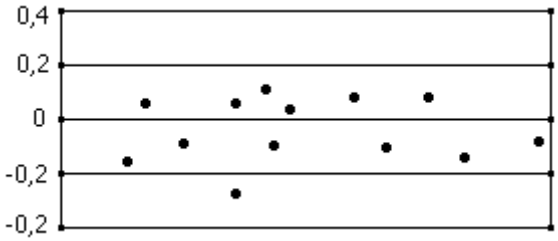
Observa que en el gráfico aparece la recta de regresión que ajusta a la nube de puntos. Además, se incluyen las denominadas “bandas de confianza” y las “bandas de predicción”. La forma de estas bandas sugiere que la predicción y la confianza en dicha predicción es mayor cuanto más próximo al centro de los datos, que es el punto $(\overline{x}, \overline{y})$. Cuando nos alejamos de este punto (a izquierda y derecha) la predicción es peor y la confianza se diluye.

- Haz clic con el botón derecho del ratón sobre el panel gráfico y en el menú de contexto elige **Opciones de Ventana**.
- En el campo **Incluir** de la siguiente ventana, desactiva las opciones **Límites de Predicción** y **Límites de Confianza**.
- Haz clic en el botón **Aceptar**.

De esta forma eliminamos del gráfico las bandas de confianza y de predicción.

- **RESIDUOS**

El gráfico de residuos (diferencia entre los datos reales y los proporcionados por el modelo) permite ver si el modelo es adecuado y comprobar que éstos tienen un comportamiento aleatorio y no presentan ninguna tendencia anómala. Por ejemplo, si los residuos se muestran como en la siguiente figura, el modelo es adecuado a los datos.



En cambio, si los residuos se distribuyen como en la siguiente figura, el modelo puede no ser el adecuado.

Hay distintos residuos susceptibles de representarse gráficamente. Para representar los correspondientes al eje OX sigue los siguientes pasos:

- Haz clic sobre el botón **Opciones Gráficas**.
- Haz clic sobre la casilla **Resíduos frente a X**.
- Haz clic sobre el botón **Aceptar**.

En el panel que aparece se observa un comportamiento no demasiado aleatorio de los residuos, lo que sugiere que quizás hay un modelo (exponencial, multiplicativo, …) que se adecue mejor a los datos observados (teniendo en cuenta que hay muy pocas observaciones para que cualquier modelo pueda considerarse el adecuado).

Es conveniente guardar los residuos para posteriores análisis. Esto se traduce en la creación automática de una nueva variable (columna) en la ventana de hoja de cálculo.

- Haz clic sobre el botón **Guardar resultados** de la barra de herramientas de análisis. Aparecen todos los valores que pueden almacenarse en variables (cuyos nombres aparecen a la derecha y se pueden modificar).
- Haz clic sobre la casilla **Resíduos**.
- Haz clic sobre el botón **Aceptar**.

Si maximizas o restauras la ventana de hoja de cálculo podrás comprobar la existencia de una nueva variable con los residuos.

Podemos crear dos nuevas variables a partir de los datos, que permitan obtener los residuos. De esta forma comprobaremos cómo se han obtenido los residuos.

- Maximiza o restaura la ventana de hoja de cálculo.
- Haz clic sobre la etiqueta de la columna **Col_4** para seleccionarla.
- Pulsa el botón derecho del ratón sobre dicha columna y selecciona **Modificar Columna** en el menú de contexto.
- Escribe el nombre de la nueva columna, por ejemplo, EstimacionY, y haz clic en el botón **Aceptar**.
- Pulsa el botón derecho del ratón sobre la columna de nuevo y selecciona **Generar Datos** en el menú de contexto.
- Introduce la fórmula para la nueva variable: 4,24419 + 0,686047 * X, y haz clic sobre el botón **Aceptar**. La nueva variable generada permitirá comprobar el grado de aproximación entre el valor Y real y el valor Y estimado por el modelo.
- Haz clic sobre la etiqueta de la columna **Col_5** para seleccionarla.
- Pulsa el botón derecho del ratón sobre dicha columna y selecciona **Modificar Columna** en el menú de contexto.
- Escribe el nombre de la nueva columna, por ejemplo, Residuos, y haz clic en el botón **Aceptar**.
- Pulsa el botón derecho del ratón sobre la columna de nuevo y selecciona **Generar Datos** en el menú de contexto.
- Introduce la fórmula para la nueva variable: Y−EstimacionY, y haz clic sobre el botón **Aceptar**. Comprueba que los resultados coinciden con los de la columna **Resíduos**.

- **ESTIMACIONES**

El modelo lineal construido con STATGRAPHICS permite predecir automáticamente el valor de Y para un valor dado de X. Esto se puede hacer incluyendo el nuevo valor de X en la columna correspondiente de la hoja de datos y generando de nuevo la columna EstimacionY. Pero el programa permite obtener la estimación automáticamente:

- Haz clic sobre el botón **Opciones Tabulares**.
- Haz clic sobre la casilla **Predicciones**.
- Haz clic sobre el botón **Aceptar**.
- Haz doble clic sobre el nuevo panel para maximizarlo.
- Pulsa el botón derecho del ratón sobre el panel y selecciona **Opciones de Ventana** del menú contextual.
- En la siguiente ventana introduce el valor (o valores) de X para el que se quiere predecir el valor de Y. En nuestro caso, introduce los valores 3; 4,5; 5,5.
- Haz clic sobre el botón **Aceptar**. En el nuevo panel se visualizan los valores de Y estimados por el modelo.

- **COMPARACIÓN CON OTROS MODELOS**

Podemos comparar el modelo obtenido con otros y comprobar si existen o no líneas de regresión que ajusten mejor la nube de puntos de los datos que se están estudiando.

- Haz clic sobre el botón **Opciones**.
- Haz clic sobre la casilla **Comparación de Modelos Alternativos**.
- Haz clic sobre el botón **Aceptar**.

En el panel que aparece se puede comprobar que, desde el punto de vista del coeficiente de determinación, el modelo que mejor se ajusta a los datos es el Inverso–Y, en el que el valor de la variable dependiente se obtiene con la ecuación $Y = \frac{1}{a + bX}$, seguido del Exponencial, en el que la ecuación de regresión es $Y = e^{a + b X}$.

- Pulsa el botón derecho del ratón sobre cualquier panel y selecciona la opción **Opciones de Análisis** en el menú de contexto.
- En la siguiente ventana, elige el modelo **Inversa de Y** y haz clic en **Aceptar**.
- Haz doble clic sobre el nuevo panel para maximizarlo.

Observa que la nueva ecuación de regresión es: $Y = \frac{1}{0{,}19869 - 0{,}0128276 * X}$, ya que a=Ordenada y b=Pendiente. El coeficiente de correlación es r=–0.955648 y el coeficiente de determinación es $r^2 = 91{,}3263$, es decir, el modelo explica un 91% de la variabilidad de los datos.

- Haz clic en el botón **Opciones Gráficas**.
- En la siguiente ventana, selecciona **Gráfico del Modelo Ajustado** y haz clic en **Aceptar**.
- Haz doble clic sobre el nuevo panel gráfico para maximizarlo.
- Haz clic con el botón derecho del ratón sobre el panel y selecciona **Opciones de Ventana** del menú contextual.
- En el campo **Incluir** desactiva las opciones **Limites de Predicción** y **Limites de Confianza**.
- Haz clic en **Aceptar**. Observa que la nueva curva de regresión es más próxima a los datos que en el modelo lineal.

- **REGRESIÓN EXPONENCIAL**

Veamos un ejemplo proporcionado por el estadístico Snedecor:

En un experimento se observó el peso (en gramos) de un embrión de pollo desde el sexto día de su nacimiento hasta el decimosexto. Los resultados se muestran en la siguiente tabla:

Días	6	7	8	9	10	11	12	13	14	15	16
Peso	0,029	0,052	0,079	0,125	0,181	0,261	0,425	0,738	1,130	1,882	2,812

Intenta encontrar el mejor modelo de regresión que se ajuste a los datos.

- Crea un archivo de datos con las dos variables, días y peso. Guárdalo con el nombre de EMBRIONES.
- Efectúa con los datos un análisis de regresión lineal.
- Comprueba que, con el modelo lineal, la recta que se ajusta a los datos tiene por ecuación:

  $$\text{Peso} = -1{,}88453 + 0{,}235073 * \text{días}$$

  y que el coeficiente de correlación es 0,862656, es decir, el modelo explica el 74% de la variabilidad de los datos ($r^2 = 0.744175$)
- Obtén el gráfico de los residuos (**Resíduos frente a X**) y deduce del mismo que el modelo lineal no es el adecuado, debido a la falta de aleatoriedad de los residuos.
- Compara el modelo lineal con otros modelos de regresión (**Comparación de Modelos Alternativos**) y deduce que el mejor modelo para los datos es el Exponencial.
- Efectúa con los datos un análisis de regresión exponencial.
- Comprueba que la curva de regresión exponencial tiene por ecuación

  $$\text{Peso} = e^{-6{,}19211 + 0{,}451033 * \text{Dias}}$$

  y que tanto el coeficiente de correlación (r=0,999165) como el coeficiente de determinación (r^2=99,83%) son próximos a 1 y a 100 respectivamente, lo que indica que la correlación es muy fuerte y que prácticamente el 100% de la variabilidad del peso se explica por el modelo.
- Obtén el gráfico de dispersión con la curva de ajuste (**Gráfico del Modelo Ajustado**) y comprueba que, efectivamente, dicha curva está muy próxima a los puntos.
- Obtén la gráfica de los residuos (**Resíduos frente a X**) y observa como éstos muestran un mejor comportamiento frente al modelo lineal.
- Almacena el statfolio con el nombre de REG_EXP**.**SGP.

El hecho de que el modelo ajustado, el exponencial, sea reducible al lineal, se puede comprobar matemáticamente si se toman logaritmos neperianos en la igualdad $\text{peso} = e^{a + b * \text{dias}}$. Resultará así que $\ln(\text{peso}) = a + b * \text{dias}$ y ésta es la ecuación de una recta.

- Vuelve a efectuar un análisis de regresión lineal con las variables dias y log(peso), para lo que tendrás que transformar previamente la variable peso en log(peso).
- Comprueba que el coeficiente de correlación del nuevo modelo lineal coincide con el correspondiente al modelo exponencial.
- Comprueba que la ecuación de la recta de regresión obtenida se traduce al eliminar los logaritmos en la ecuación del modelo exponencial.

De forma análoga podemos proceder para otros modelos reducibles al lineal, utilizando para ello la función de transformación adecuada.

- **STATGALLERY**

Tanto los paneles de texto como los gráficos se pueden presentar juntos en una única ventana: la denominada **StatGallery**. De esta forma se puede imprimir de forma conjunta los análisis estadísticos y gráficos.

- Abre el statfolio REG_EXP**.**SGP, si no lo está ya.
- Pulsa el botón derecho del ratón sobre uno de los paneles de este statfolio.
- Selecciona la opción **Copiar Ventana a Galería**.
- Abre la ventana **StatGallery** y observa que está dividida en cuatro paneles. En la parte superior tienes unos botones que permiten ir de una página a la siguiente, a la primera o a la última página. Haz clic con el botón derecho del ratón sobre el primer panel y selecciona el comando **Copiar** del menú de contexto. Observa el resultado.
- Minimiza la ventana **StatGallery** y repite el procedimiento con otros paneles. Incluye también algún panel gráfico. Observa que los paneles se van situando en la ventana **StatGallery** de manera ordenada, de acuerdo con la opción activada en cada panel.
- Cuando hallas trasladado a la ventana **StatGallery** dos paneles de texto y dos gráficos, pulsa el botón derecho del ratón sobre la ventana y elige el comando **Organizar Ventanas**.
- Selecciona la opción **Arriba y Abajo** y haz clic en **Aceptar**.
- Repite el procedimiento para probar cada una de las opciones disponibles en el comando **Organizar Ventanas**. Deja la que más te guste.
- Haz clic con el botón derecho del ratón sobre el panel de la ventana **StatGallery** que contiene un gráfico y selecciona el comando **Cortar** del menú contextual.
- Sitúa el cursor sobre un panel vacío de la ventana StatGallery y haz clic con el botón derecho del ratón. En el menú contextual selecciona **Pegar**. De esta forma hemos cambiado la posición del gráfico dentro de la ventana StatGallery. Observa el resultado.
- Cuando hayas situado el gráfico en la posición que te interese, haz clic con el botón derecho del ratón sobre el panel de la ventana **StatGallery** que lo contiene y selecciona **Imprimir** del menú contextual.
- Haz clic en **Aceptar** y se empezará a imprimir la ventana **StatGallery**.
- Cierra el statfolio sin almacenar los cambios. Cierra el archivo de datos. Cierra el programa STATGRAPHICS.

ACTIVIDADES

- **CALIFICACIONES**

Las calificaciones obtenidas por un grupo de estudiantes en Biología y Química son las de la siguiente tabla:

B	5	6	6	7	5	7	8	3	5	4	8	5	5	8	8	8	5
Q	5	5	8	7	7	9	10	4	7	4	10	5	7	9	10	5	7

a) Halla la covarianza y el coeficiente de correlación.

b) Representa la nube puntos correspondiente.

c) Halla la línea de regresión que mejor ajusta los datos. ¿Cuál es su ecuación?.

d) Predice la nota de Química para un estudiante que ha obtenido un 8 en Biología.

- **PRECIO DE LA VIVIENDA**

En la siguiente tabla se compara el precio medio de la vivienda (en euros por m^2) en Valencia y Barcelona durante los años 1988 y 1999 (fuente: revista DADES)

1er trimestre	1010	1918
2º trimestre	1049	1973
3er trimestre	1078	2039
4º trimestre	1096	2069

1er trimestre	1119	2150
2º trimestre	1147	2262
3er trimestre	1159	2382
4º trimestre	1223	2533

Dibuja el diagrama de dispersión de los datos y halla el coeficiente de correlación. Si se conoce el precio del metro cuadrado en Valencia, ¿se puede predecir el precio en Barcelona?.

- **PRODUCCIÓN**

A partir de los siguientes datos referentes a horas trabajadas en un taller (X) y unidades producidas (Y), obtén las ecuaciones de las rectas de regresión de Y sobre X y de X sobre Y.

	80	79	83	84	78	60	82	85	79	84	80	62
	300	302	315	330	300	250	300	340	315	330	310	240

a) ¿Cuál será el número de horas trabajadas si se han producido 320 unidades?.

b) ¿Cuántas unidades se esperan producir si se trabaja 87 horas?.

c) ¿Son fiables los resultados obtenidos?.

d) Efectúa un análisis de los residuos e investiga cuál es el mejor modelo que se ajusta a los datos anteriores.

- **COLOCACIONES**

En la siguiente tabla se indica el número de colocaciones en la ciudad de Valencia durante el cuarto trimestre de 1999, según titulación y sexo (fuente: revista DADES):

Analfabetos y sin estudios	765	339
Certificado de escolaridad	17786	11132
EGB / Secundaria	16767	15384
FP1 / FP2 / Ciclos formativos	3605	5490
BUP / COU / Bachillerato	6275	8921
Titulación Grado Medio	1342	3350
Titulación Grado Superior	1605	3373
Total	48145	47989

Dibuja la nube de puntos y calcula el coeficiente de correlación entre el número de colocaciones de hombres y mujeres. Efectúa un análisis de regresión.

- **VEHÍCULOS**

En la siguiente tabla se muestra el número de turismos y de motocicletas matriculados en la ciudad de Valencia durante los años 1996, 1997, 1998 y 1999 (fuente: revista DADES):

		Turismos matriculados	Motocicletas matriculadas
1997	1er. trimestre	4994	197
	2º trimestre	5955	268
	3er. trimestre	5638	261
	4º trimestre	5776	227
	Total	**22363**	**953**
1998	1er. trimestre	5622	226
	2º trimestre	6937	355
	3er. trimestre	6747	328
	4º trimestre	7247	310
	Total	**26553**	**1219**
1999	1er. trimestre	7228	310
	2º trimestre	7854	406
	3er. trimestre	7894	434

Dibuja el diagrama de dispersión y halla el coeficiente de correlación lineal. Efectúa un análisis de regresión. Estudia los residuos e investiga cuál puede ser el mejor modelo de ajuste.

- **GASOLINA Y GASÓLEO**

La evolución del precio de la gasolina y el gasóleo en el período 1983 a 1995 viene dada en la siguiente tabla:

FECHA	GASOLINA	GASÓLEO
26-07-1983	13,5	7,40
02-03-1984	20	10,50
24-08-1986	28	14
03-07-1989	46	21
05-12-1990	61	34
09-01-1995	93	58
10-07-1995	93	62
11-12-1995	87	62

a) Calcula el coeficiente de correlación.

b) Interpreta el resultado.

c) ¿Qué dependencia existe entre las variables?.

d) ¿Qué precio debería tener la gasolina si se desea bajar el precio del gasóleo a 50 céntimos de euro?.

ANÁLISIS DE DATOS Y ANÁLISIS DE REGRESIÓN CON FUNCIONES PARA WINDOWS

1. Introducción

Funciones para Windows permite representar los datos expresados en una tabla y encontrar la función que más se ajuste a ellos, para realizar estudios de interpolación y regresión. En las siguientes actividades veremos como utilizar el programa para analizar datos estadísticos.

2. Interpolación y regresión

Ejemplo.– Los siguientes datos muestran la relación entre el número de habitantes (X, en millares) y el número de fincas con más de 12 alturas (Y) en siete ciudades. Dibuja la nube de puntos, halla la recta de regresión y estima la cantidad de edificios con más de 12 alturas que puede haber en una ciudad de 300000 habitantes.

X	**150**	**500**	**800**	**250**	**500**	**750**	**950**
Y	**4**	**31**	**42**	**9**	**20**	**55**	**73**

- Inicia el **Explorador de Windows**, abre la carpeta **Func27** y haz doble clic sobre el archivo **Fw27.exe**.

- Haz clic dos veces en **OK**. En el cuadro de diálogo **Entrada de dades**, haz clic en el botón **Funció numèrica**. Haz clic en el botón **Regressió** y haz clic en **OK**.

- En el cuadro de diálogo **Introduir valors**, escribe los datos de la tabla anterior, ordenados de menor a mayor. Para pasar de cada pareja a la siguiente, haz clic en el botón **Següent**. Así:

	X	F(X)
1º	150	4
2º	250	9
3º	500	20
4º	500	31
5º	750	55
6º	800	42
7º	950	73

- Observa la barra inferior de la ventana y fíjate que ofrece diferentes modelos de regresión: **Lineal**, **Quadràtic**, $a \cdot b^X$ (**exponencial**), $a \cdot X^b$ (**potencial**), $a \cdot X/(b + X)$. Selecciona el modelo **Lineal**, haciendo clic en dicho botón.

- Haz clic en el botón **Coeficients**. En la siguiente ventana se muestran los valores extremos de las dos variables, así como los parámetros estadísticos y la covarianza. Además, se muestran los coeficientes de correlación correspondientes a cada uno de los modelos. De esta forma, podemos comparar su bondad de ajuste. En este caso, vemos que el modelo exponencial presenta una mayor correlación.

- Haz clic en el botón **OK**. En la siguiente ventana, el programa pregunta si deben modificarse o no las escalas de los ejes.

- Haz clic en el botón **Yes**. En el cuadro de diálogo **Entrada de dades**, introduce los siguientes valores: **Origen eix X: –2**, **Final eix X: 1000**, **Origen eix Y: 0**, **Final eix Y: 70**. Haz clic en el botón **OK** y comprueba que se dibuja la nube de puntos, junto con la recta de regresión.

- Selecciona el comando **1 fu / Equació de regressió**. Aparece la ventana **Copiar a la carpeta?,** donde se muestra el coeficiente de correlación y la ecuación de la recta de regresión. Haz clic en el botón **Yes**, para añadir este resultado a la carpeta de trabajo.

- Selecciona el comando **Opcions / Mostrar valor unitats eixos**. Observa el resultado.

- Selecciona el comando **Opcions / Pissarra**. Observa como cambia la pantalla. Selecciona el comando **Opcions / Trama** y fíjate cómo se muestra la cuadrícula. Desactiva las opciones **Pissarra** y **Trama**.

- Selecciona el comando **1 fu / Imatge**. En el cuadro de texto introduce el valor **x = 300** y haz clic en el botón **D'accord**. Observa que se señala el punto en la gráfica y el valor de Y correspondiente es 12.45587. Por tanto, en una ciudad de 300000 habitantes, esperamos encontrar 12 edificios con más de doce alturas.

ACTIVIDADES

- **FAMILIAS**

La siguiente tabla muestra la relación entre el número de matrimonios (X) y el número de nacimientos (Y) entre los años 1984 y 1993, ambos inclusive:

X	268	271	261	262	258	246	221	202	193	196
Y	688	669	678	656	637	602	571	533	516	485

a) Halla el coeficiente de correlación y la ecuación de la recta de regresión.
b) Dibuja la nube de puntos, junto con la recta de regresión.

- **APARTAMENTOS SOCIALES**

Un ayuntamiento está convencido de que la venta de pisos sociales está íntimamente relacionado con el precio de venta; pero no entiende el motivo de que en los últimos meses se haya estancado la venta de pisos en una urbanización situada en una zona muy agradable. El concejal de urbanismo ordena un estudio para poder analizar si es que el precio es muy alto y por eso se venden menos pisos. Estos son los resultados del estudio:

Precio venta (miles de euros)	26	44'5	36	40	31	54	58	62'5	71'5	49
Pisos vendidos	25	10	20	16	22	8	7	7	2	8

¿A qué conclusión debería llegar el concejal?.