ESTADÍSTICA CON EXCEL

1. INTRODUCCIÓN

La estadística es la rama de las matemáticas que se dedica al análisis e interpretación de series de datos, generando unos resultados que se utilizan básicamente en dos contextos: la toma de decisiones y la proyección de situaciones futuras.

La estadística descriptiva sirve para recoger, analizar e interpretar los datos, generando una tabla sobre la que se realizan cálculos para obtener diversas medidas. De esta forma, se obtiene por ejemplo la altura media de los alumnos de una clase.

Una hoja de cálculo es una de las herramientas más adecuadas para introducir tablas de valores, obteniendo resultados y gráficas que faciliten su representación.

2. VARIABLES, MUESTRAS Y TABLAS DE DATOS

Al ser tratados con Excel, los valores de las **variables cualitativas** aparecerán normalmente como textos, mientras que **las cuantitativas** serán números, enteros o con decimales en el caso discreto, o continuo.

Tablas estadísticas

Una vez determinada la población, las características que quieren analizarse y seleccionada la muestra, llega el momento de recoger los datos y de organizarlos en tablas.

Las tablas de frecuencias resumen numéricamente, la información sobre el carácter estadístico que queremos estudiar.

Antes de construir una tabla de frecuencias, vamos a definir los elementos que suelen aparecer en ella:

- **La frecuencia absoluta f_i**, de un valor x_i es el número de veces que se repite dicho valor.

- **La frecuencia relativa h_i** del valor x_i es el cociente entre la frecuencia absoluta del x_i y el número total de valores, n.

- **La frecuencia absoluta acumulada F_i** del valor x_i, es la suma de todas las frecuencias absolutas de todos los valores anteriores a x_i, más la frecuencia absoluta de x_i.

$$F_i = f_1 + f_2 + \ldots + f_i$$

- **La frecuencia relativa acumulada H_i** del valor x_i es la suma de todas las iHixfrecuencias relativas de todos los valores anteriores a x_i, más la frecuencia relativa de x_i.

$$H_i = h_1 + h_2 + \ldots + h_i$$

- **El porcentaje p_i** de un valor x_i se obtiene multiplicando por 100 la frecuencia relativa del valor x_i.

Tabla de frecuencias de una variable cualitativa o cuantitativa discreta con Excel:

- Introducimos **en la primera columna (A)** las distintas modalidades si el carácter es cualitativo (Figura 2), o bien, los valores de la variable estadística discreta. (Figura 3).

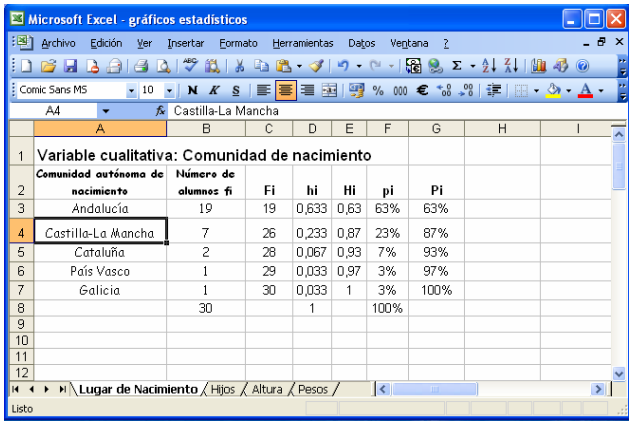


Figura 2

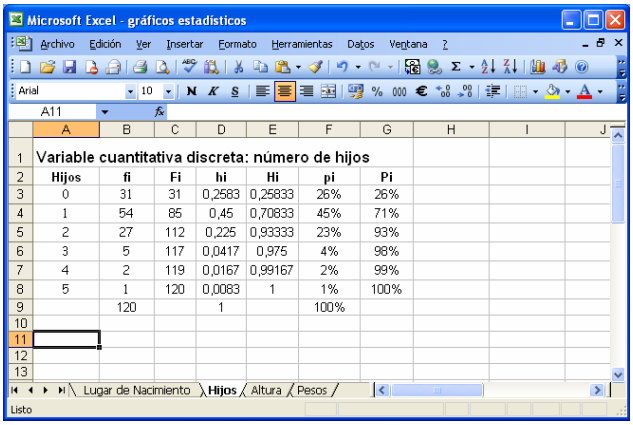


Figura 3

- **En la segunda columna (B)** introducimos los valores de la frecuencia absoluta f_i.
- **En la tercera columna (C)** vamos a colocar la frecuencia absoluta acumulada (F_i), pero en lugar de hacer nosotros los cálculos, será el programa el que se encargue de hacerlos.¿Cómo?

  - En la celda **C3** escribimos **=B3** y en la celda **C4** escribimos **=C3+B4**. A continuación copiamos la fórmula, situando el puntero del ratón en la esquina inferior derecha de esta celda y cuando el puntero del ratón se convierta en + y arrastramos hasta la casilla última casilla.

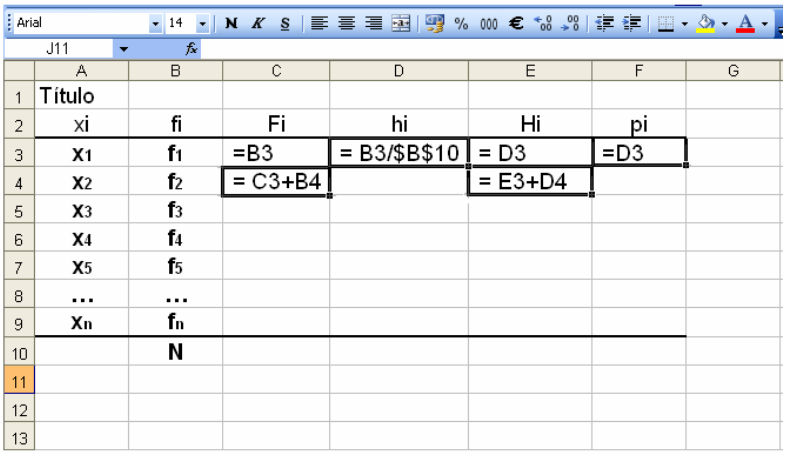
  - Para completar la columna de la frecuencia relativa (h_i), basta con escribir en la celda **D3=B3/$B$8**. (Con el símbolo $, lo que hacemos el fijar el valor de la celda que no varía).

  - En la columna de la frecuencia relativa acumulada (H_i), en **E3, escribimos =D3**; en **E4, =E3 + D4** y copiamos la fórmula.

  - Para el porcentaje, en **F3, se escribe =D3** y pulsamos el botón 

. El paso siguiente es copiar la expresión de la celda anterior.

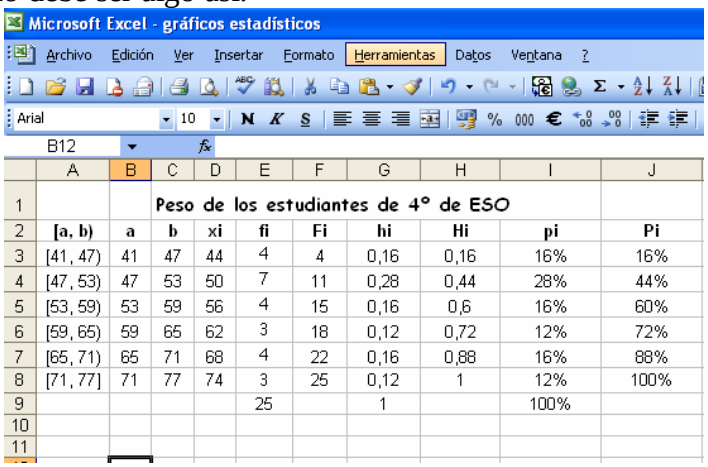
En resumen, la tabla de frecuencias se construye así:



	A	B	C	D	E	F	G
1	Título						
2	xi	fi	Fi	hi	Hi	pi	
3	x1	f1	=B3	= B3/$B$10	= D3	=D3	
4	x2	f2	= C3+B4		= E3+D4		
5	x3	f3					
6	x4	f4					
7	x5	f5					
8	...	...					
9	xn	fn					
10		N					
11							
12							
13							

Tabla de frecuencias para una variable cuantitativa continua con Excel:

-En la primera **columna (A)** escribimos los intervalos **[a, b)**, en la **columna B** el valor **"a"** y en la C el valor **"b"**. En la **columna D**, vamos a calcular la marca de clase, escribimos la fórmula **=(B3+C3)/2** y la copiamos.
-La primera columna, la utilizamos para la representación gráfica y las dos siguientes B y C, para calcular la marca de clase.
-En la siguiente **columna E**, introducimos **la frecuencia absoluta (f_i)**, en la siguiente introducimos la fórmula para el cálculo de la frecuencia absoluta acumulada de forma análoga a los ejemplos anteriores y así sucesivamente hasta terminar de construir la tabla. El resultado debe ser algo así:



	A	B	C	D	E	F	G	H	I	J
1			Peso de los estudiantes de 4º de ESO							
2	**[a, b)**	**a**	**b**	**xi**	**fi**	**Fi**	**hi**	**Hi**	**pi**	**Pi**
3	[41, 47)	41	47	44	4	4	0,16	0,16	16%	16%
4	[47, 53)	47	53	50	7	11	0,28	0,44	28%	44%
5	[53, 59)	53	59	56	4	15	0,16	0,6	16%	60%
6	[59, 65)	59	65	62	3	18	0,12	0,72	12%	72%
7	[65, 71)	65	71	68	4	22	0,16	0,88	16%	88%
8	[71, 77]	71	77	74	3	25	0,12	1	12%	100%
9					25		1		100%	
10										
11										

3. GRÁFICOS ESTADÍSTICOS

Según el tipo de variable, la representación gráfica más utilizada en cada caso es…

- **Variable cualitativa:** diagrama de sectores (En Excel… circular)
- **Variable cuantitativa discreta:** diagrama de barras (columnas).
- **Variable cuantitativa continua:** histograma (columnas)

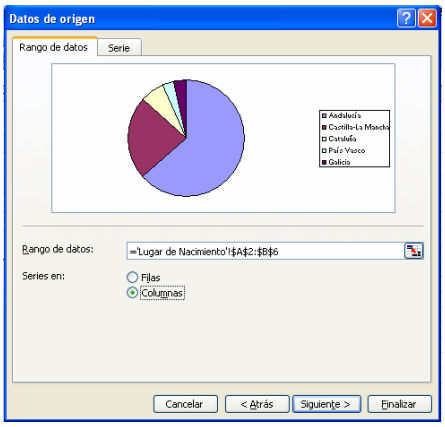
Veamos ahora como podemos hacer un gráfico estadístico, utilizando la herramienta

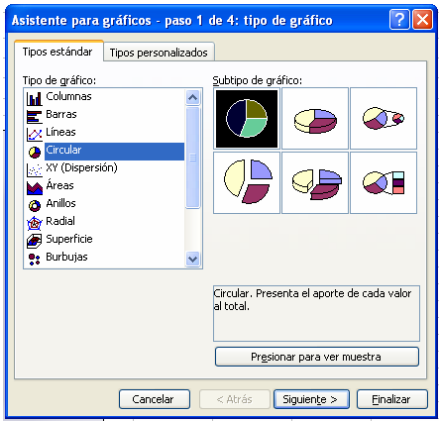
Asistente para gráficos  que nos guiará a lo largo de toda la creación del gráfico.

- **Diagrama de sectores**, de la siguiente forma:

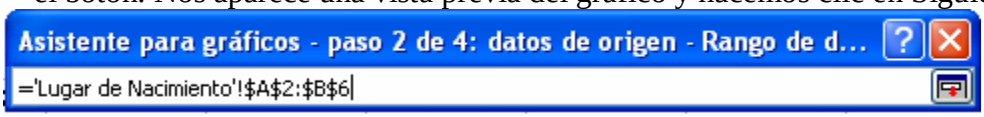
1. Hacemos clic en el botón Asistente para gráficos de la barra de herramientas. 

2. Paso 1 de 4: tipo de gráfico. Nos aparece un cuadro de diálogo con dos fichas, en la ficha Tipos estándar (que es la que aparece por defecto), hacemos clic en **Circular**, elegimos el Subtipo que queramos (elegimos el Circular o el Circular en 3D) y hacemos clic en siguiente.



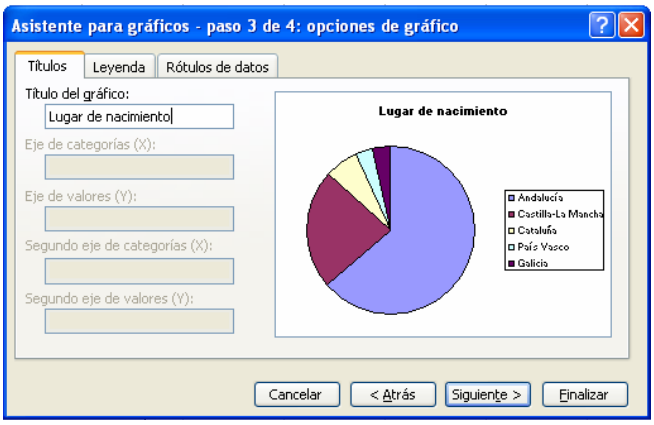


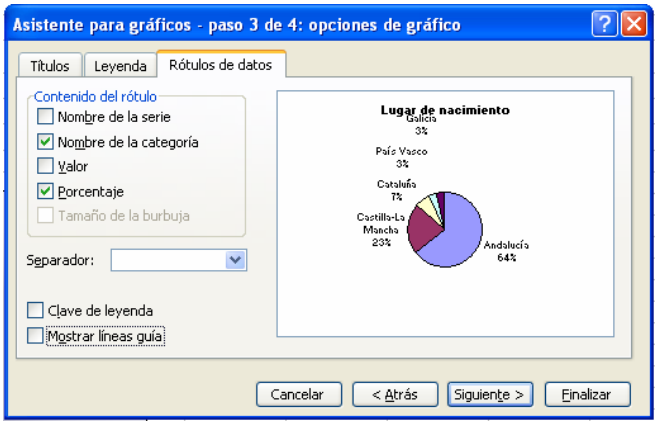
3. Paso 2 de 4: datos de origen. En este paso debemos indicar los datos que vamos a representar, para ello, hacemos clic en el botón  que aparece al final del cuadro Rango de datos, y seleccionamos el rango A2:B6 (pinchamos y arrastramos desde la celda A2 hasta la B6), una vez hecho esto, hacemos clic en el botón. Nos aparece una vista previa del gráfico y hacemos clic en Siguiente.



Nos aparece una vista previa del gráfico y hacemos clic en Siguiente.

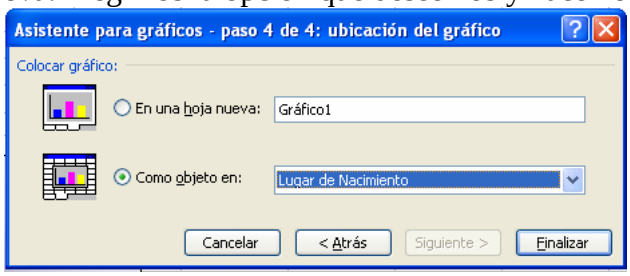
4. Paso 3 de 4: opciones de gráfico. En este paso podremos ponerle un Título al gráfico, quitar con cambiar el lugar de la Leyenda y modificar el Rótulo del gráfico, haciendo clic en la ficha correspondiente y eligiendo lo que queramos.





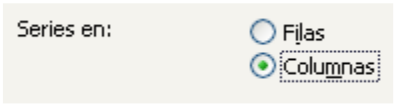
Hacemos clic en Siguiente.

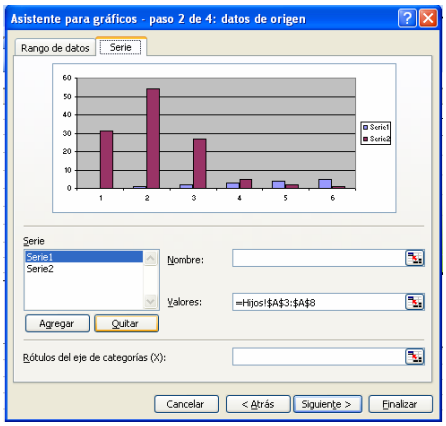
5. Paso 4 de 4: Ubicación del gráfico. En este último paso elegiremos si queremos insertar el gráfico como objeto, en esta misma hoja, o si queremos insertarlo en una hoja nueva. Elegimos la opción que deseemos y hacemos clic en Finalizar.

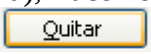


- **Diagrama de barras de una variable estadística discreta**, se procede de forma similar, veamos el ejemplo del número de hijos de los 120 trabajadores de una fábrica:

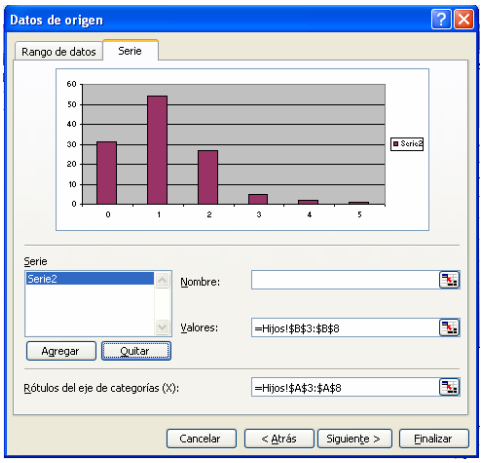
1. Hacemos clic en el botón Asistente para gráficos  de la barra de herramientas.
2. Paso 1 de 4: tipo de gráfico. Nos aparece un cuadro de diálogo con dos fichas, en la ficha Tipos estándar (que es la que aparece por defecto), hacemos clic en Columnas, elegimos el Subtipo que queramos (elegimos el primero, Columnas apiladas) y hacemos clic en siguiente.
3. Paso 2 de 4: datos de origen. Indicamos los datos que vamos a representar; hacemos clic en el botón  que aparece al final del cuadro Rango de datos, y seleccionamos el rango A2:B8 (pinchamos y arrastramos desde la celda A2 hasta la B8), una vez hecho esto, hacemos clic en el botón  .
   A continuación hacemos clic en el botón Columnas





Activamos la ficha Serie (haciendo clic sobre ella), hacemos clic sobre la Serie 1 y la quitamos haciendo clic sobre el botón 

A continuación hacemos clic sobre el botón  del cuadro correspondiente a Rótulos del eje de categorías (X): para seleccionar los valores de la variable que aparecerán en el gráfico, después seleccionamos el rango A3:A8
Debe quedar algo así:

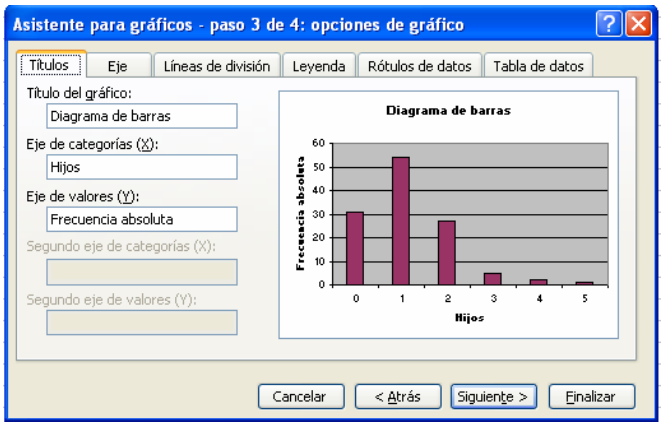


Hacemos clic en Siguiente.

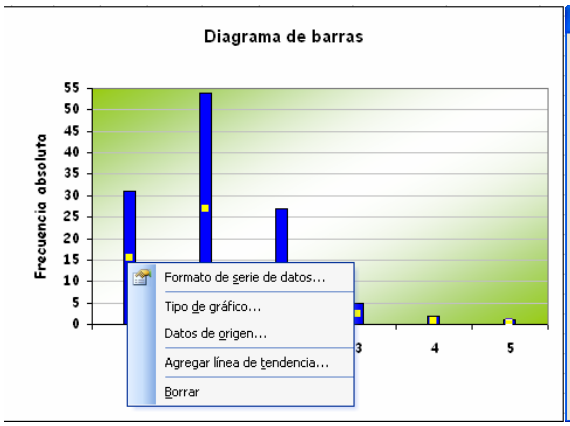
4. Paso 3 de 4: opciones de gráfico; escribimos el título del gráfico; En el Eje de categorías (X) la variable estadística que representamos, en este caso Hijos, y en el Eje de valores (Y): la frecuencia absoluta, relativa, o lo que estemos representando.

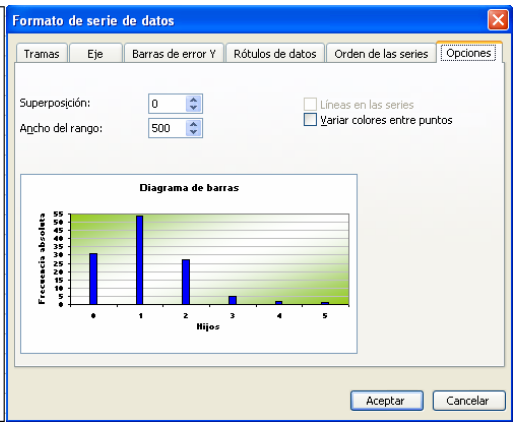
En la ficha Leyenda, quitamos la leyenda.

5. Paso 4 de 4: ubicación de gráfico insertamos el gráfico como objeto es la misma hoja, o en otra hoja nueva. Hacemos clic en Finalizar



Una vez que tenemos hecho el gráfico, podemos poner más finas las barras, hacemos clic en la primera barra y a continuación hacemos clic con el botón derecho del ratón sobre dicha barra, elegimos el menú Formato de serie de datos…, hacemos clic sobre la ficha Opciones y en Ancho de rango: seleccionamos 500. Hacemos clic en Aceptar.

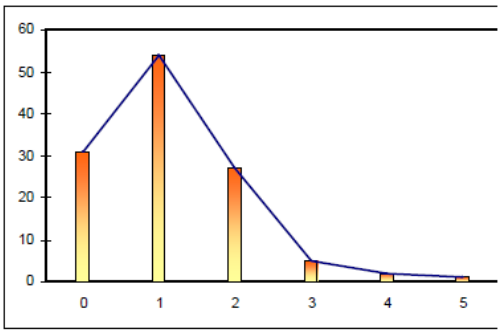




Después podemos cambiar los colores de las barras, el color del fondo, la escala de los ejes, etc.

Para construir el histograma, hacemos lo mismo que para el caso de la variable estadística discreta. La única diferencia es que ahora tenemos que hacer las barras más gruesas, así que Ancho de rango: seleccionamos 0.

El polígono de frecuencias se obtiene uniendo la parte superior de las barras del diagrama (los puntos medidos de los rectángulos del histograma). Investiga cómo se podría hacer, es decir, que tipo de gráfico de los que incorpora Excel, nos permite obtener este gráfico.



Una pista: inténtalo en Tipos personalizados

4. MEDIDAS DE CENTRALIZACIÓN Y DE DISPERSIÓN

Las tablas de frecuencias nos permiten resumir la información y adquirir una idea general sobre el significado de los datos. Su finalidad, sin embargo, no es la de aportar indicadores sobre los datos, para este fin existen distintos tipo de medidas: las de centralización y las de dispersión.

Las medidas de centralización, tratan de dar un valor central en torno al cual se distribuyen los datos; son tres:

-**Moda (M_o):** es el valor de la variable con mayor frecuencia absoluta. Se puede calcular para cualquier tipo de variable. Para calcularla basta con observar la columna de frecuencias absolutas (f_i). Una distribución estadística puede tener más de una moda.

-**Media** $(\overline{x})$**:** es la medida de centralización más conocida, pero no se puede calcular para variables cualitativas y es muy sensible para valores extremos, por lo que no siempre es la mejor medida de centralización. Se calcula: $\overline{x} = \frac{\sum x_i \cdot f_i}{N}$

En la tabla de frecuencias añadiremos una nueva columna con los productos $f_i \cdot x_i$

-**Mediana (Me):** es el valor de la variable que deja por encima y por debajo, el mismo número de datos, es decir, es el valor central de la variable. No existe una fórmula para calcular la medida, sino una serie de normas. Tampoco se puede calcular para variables cualitativas.

Cálculo de la mediana: Se busca en la columna de la frecuencia absoluta acumulada el primer valor que supere la mitad de los datos (n/2), la mediana será el valor que se corresponda con esta frecuencia absoluta acumulada, o también se busca en la columna de frecuencias relativas acumuladas el primer valor que supera a 0,5.

Ejemplo:

	A	B	C	D	E	F
1	Hijos de los trabajadores de una fábrica					
2	xi	fi	Fi	hi	Hi	xi·fi
3	0	31	31	0,26	0,26	0
4	1	54	85	0,45	0,71	54
5	2	27	112	0,23	0,93	54
6	3	5	117	0,04	0,98	15
7	4	2	119	0,02	0,99	8
8	5	1	120	0,01	1	5
9		120		1		136
10						
11						
12	Moda	1				
13	Mediana	1				
14	Media	= F9/$B$9				

El cálculo de las medidas de centralización y de dispersión para variables continuas o agrupadas en intervalos se hace de la misma forma que para variables discretas, tomando, la marca de clase para la media.

Las medidas de dispersión nos dan una idea de en que medida los datos están más o menos juntos (concentrados) o más o menos dispersos, y cual es la fiabilidad de las medidas de centralización. Son:

- Para calcular **la varianza** en Excel, añadimos la columna correspondiente en la tabla de frecuencias y después sumamos y dividimos entre n. La varianza también se puede calcular con esta otra expresión

$$s^2 = \frac{\sum x_i^2 \cdot f_i}{N} - \overline{x}^2$$

- **Desviación típica (S):** es la raíz cuadrada positiva de la varianza. $s = +\sqrt{s^2}$
  Para calcular la desviación típica, calculamos la raíz cuadrada de la varianza con la función =RAIZ()

- **Coeficiente de variación (CV):** es el cociente entre la desviación típica y la media de una variable estadística.

$$CV = \frac{s}{\overline{x}}$$

  El coeficiente de variación se utiliza para comparar la dispersión de dos o más distribuciones; a menor coeficiente de variación menor dispersión de los datos (o mayor concentración).

En la figura siguiente verás las columnas que añadimos para calcular estas medidas de centralización y de dispersión. Veamos:

	A	B	C	D	E	F
1	**Título**					
2						
3	**xi**	**fi**	**Fi**	**xi·fi**	**|xi - x|·fi**	**(xi - x)^2·fi**
4	**x1**	**f1**	=B4	=A4*B4	=ABS(A4-$B$14)*B4	=(A4-$B$14)^2*B4
5	**x2**	**f2**	=C4+B5			
6	**x3**	**f3**				
7	**x4**	**f4**				
8	**…**	**…**	…	…	…	…
9	**xn**	**fn**				
10		**N**		**=SUMA(D4:D9)**	**=SUMA(E4:E9)**	**=SUMA(F4:F9)**
11						
12						
13	**Medidas de centralización**					
14	Media =	=D10/B10				
15	Mediana =	xi	(la encuentro mirando en la columna Fi)			
16	Moda =	xi	(la encuentro mirando en la columna fi)			
17						
18	**Medidas de dispersión**					
19	DM=	=E10/B9				
20	**s2** =	=F10/B9				
21	**s** =	=RAIZ(B20)				
22	**CV** =	= B21/B14				

Para calcular la mediana observamos la columna de las frecuencias absolutas acumuladas, F_i, y para el valor de la variable se supere la mitad de los datos (n/2) tenemos esta medida de centralización. La moda la obtenemos al observar la columna de frecuencias absolutas f_i. Para el cálculo de la media añadimos la columna $x_i \cdot f_i$.

Una vez calculada la media, añadimos las otras dos columnas para la desviación media y la varianza.