

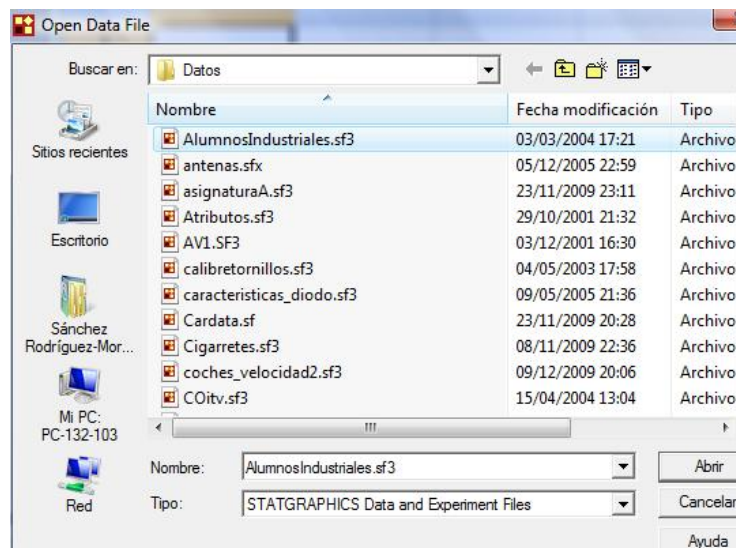
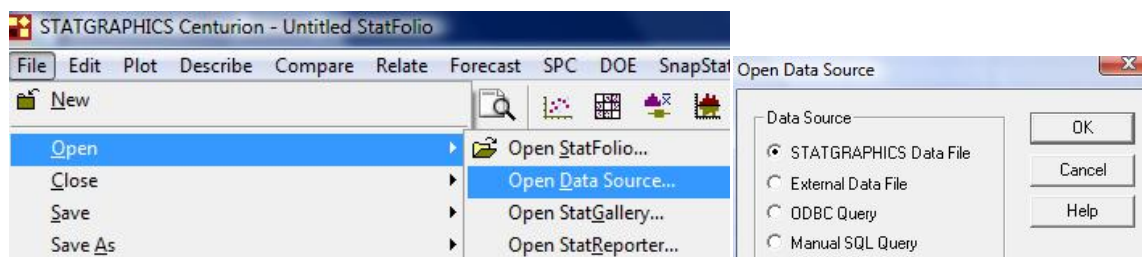
# Estadística Descriptiva de una variable con STATGRAPHICS CENTURION


Ficheros empleados: AlumnosIndustriales.sf3,

## 1. Introducción

El objetivo de este documento es la utilización de las técnicas de estadística descriptiva más habituales para resumir la información de un conjunto de datos de una variable. Se usará el programa Statgraphics Centurion. Los datos que se utilizan en este guión se encuentran en el fichero AlumnosIndustriales.sf3. Los datos corresponden a 91 estudiantes de Ingeniería Industrial, a los que se les ha preguntado sobre variables tales como estatura, peso, número de hermanos, etc. Emplearemos así una base de datos muy sencilla que nos ayude a entender las posibilidades del programa Statgraphics.

En primer lugar leemos el fichero de datos.



O bien pulsamos la tecla  y seleccionamos el fichero.

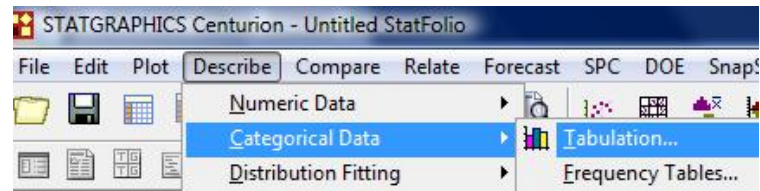
## 2. Descripción de variables cualitativas

La variable residencia corresponde al lugar de residencia de los alumnos. Esta variable es cualitativa. Su codificación es

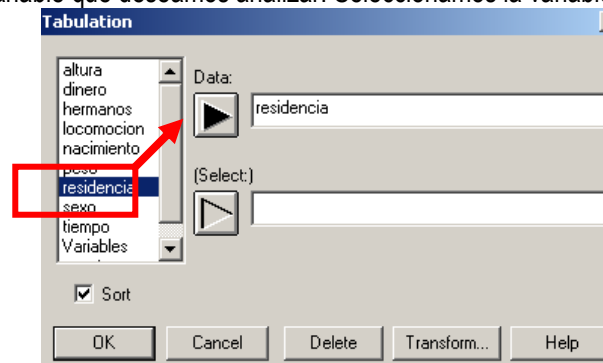
1-Madrid Sur

- 2-Madrid Centro
- 3-Madrid-otros
- 4-Fuera de Madrid

Para describir esta variable haremos primeramente un análisis gráfico y luego una tabla de frecuencias. Todas estas opciones están en Describe/Categorical Data/Tabulation...

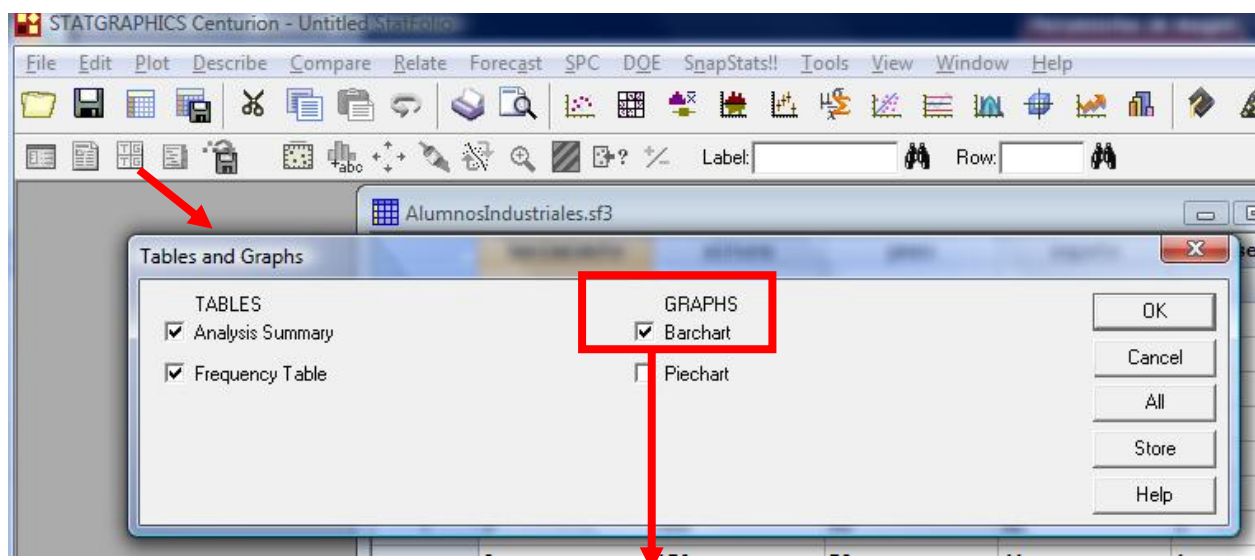


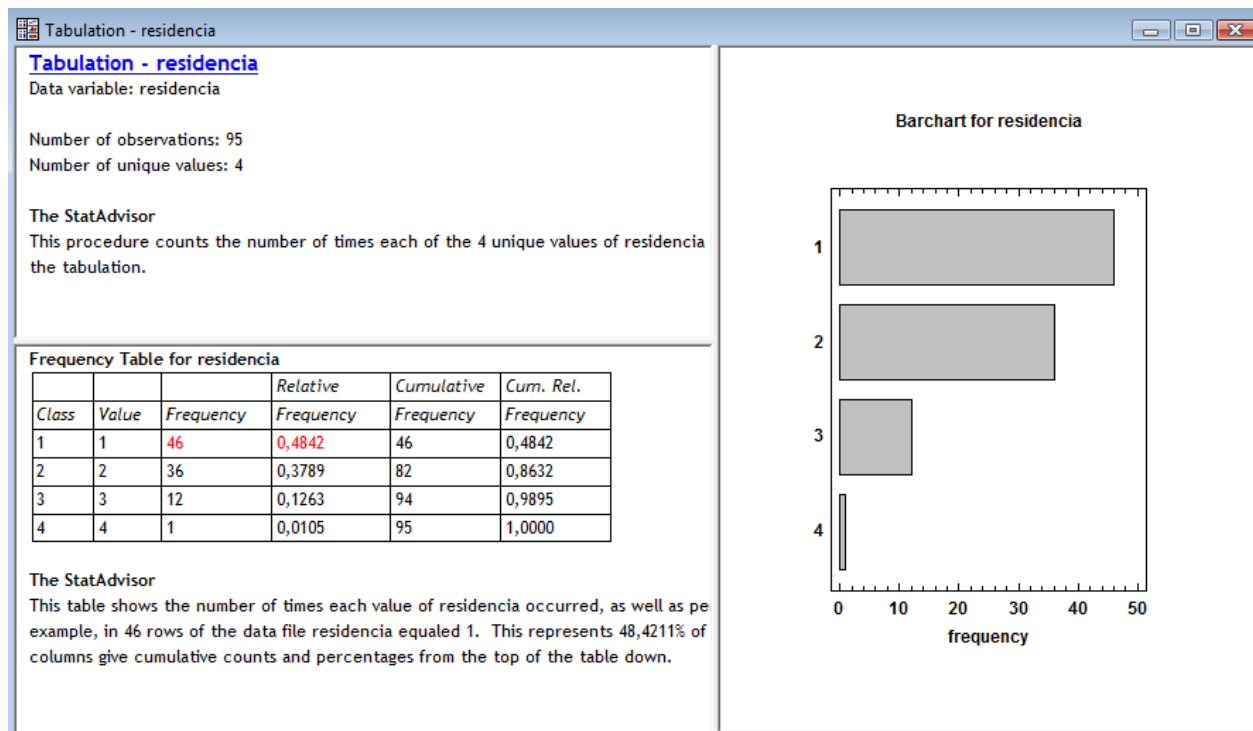
donde nos preguntan por la variable que deseamos analizar. Seleccionamos la variable 'residencia'.



## 2.1 Gráfico de barras

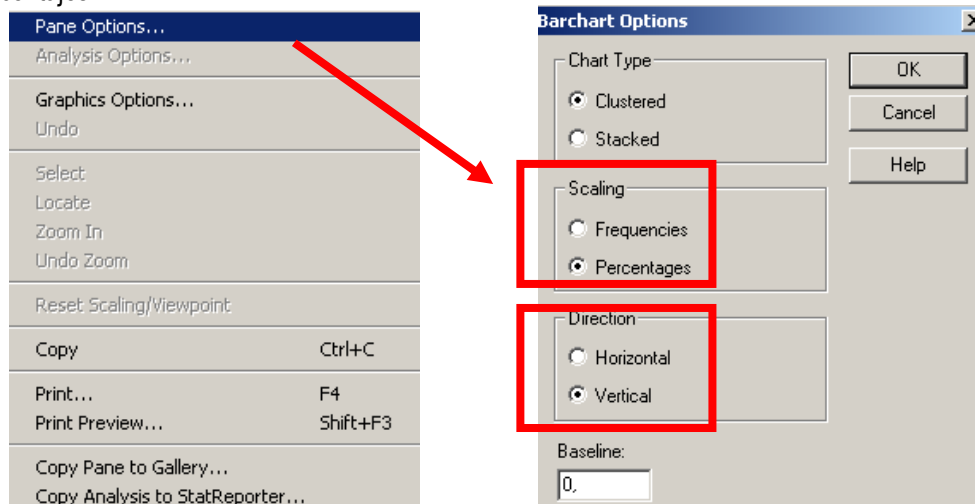
En las opciones gráficas seleccionamos barchart





donde puede verse que la población más grande de alumnos son los procedentes de Madrid Sur, con casi 50 alumnos.

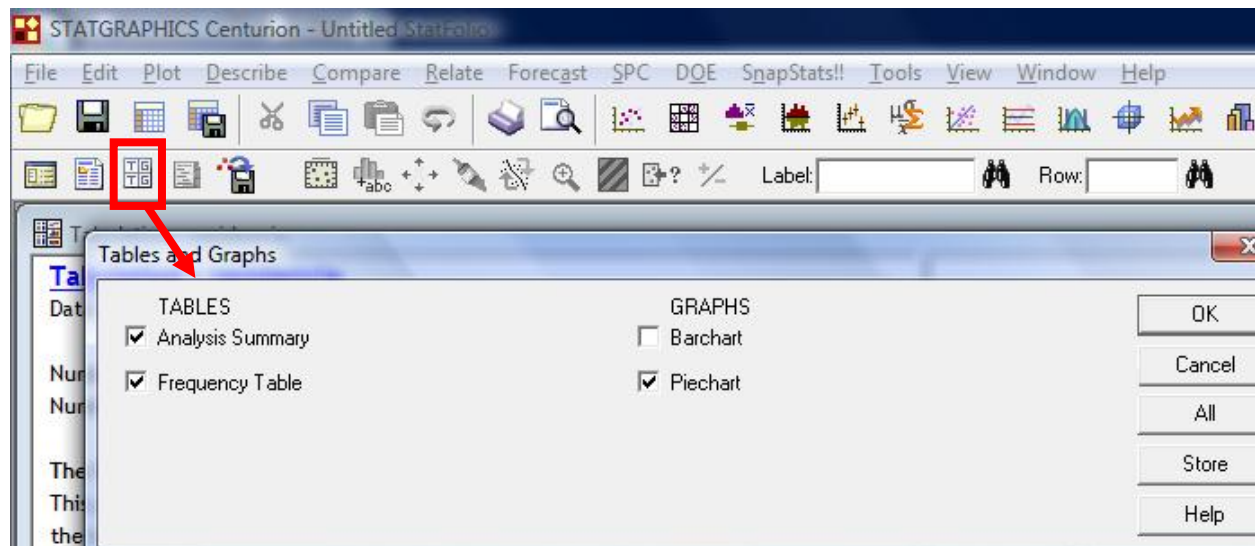
Si queremos cambiar el aspecto del gráfico nos colocamos sobre él, pulsamos el botón derecho del ratón y seleccionamos Pane Options. Seleccionamos, por ejemplo, que las barras sean verticales y que las frecuencias sean en porcentajes.



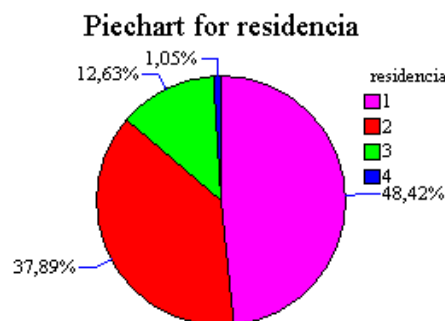
Al estar escala de las barras en % podemos ver que casi el 50% de los alumnos proceden del sur de Madrid.

## 2.2 Gráfico de tarta o porciones

Dentro de las opciones gráficas (GRAPHS), seleccionamos Piechart



y obtenemos



## 2.3 Tabla de frecuencias

En las opciones numéricas (TABLES ) encontramos la tabla de frecuencias, que nos da la información numérica que antes hemos representado en gráficos.. La tabla de frecuencias resultante es

**Frequency Table for residencia**


			Relative	Cumulative	Cum. Rel.
Class	Value	Frequency	Frequency	Frequency	Frequency
1	1	46	0,4842	46	0,4842
2	2	36	0,3789	82	0,8632
3	3	12	0,1263	94	0,9895
4	4	1	0,0105	95	1,0000

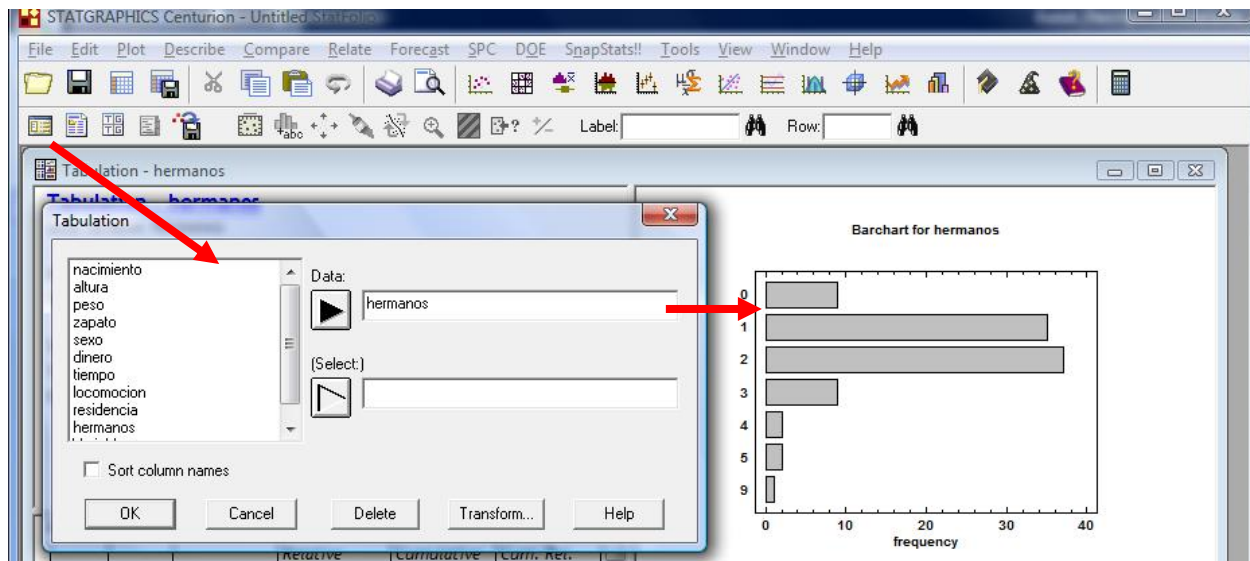
donde puede verse que entre el sur (48.4%) y el centro (37.9%) de Madrid, se abarca al 86.3% de los estudiantes.

## 3. Descripción de variables cuantitativas

### 3.1 Análisis gráfico de variables discretas con pocos valores

En el caso de variables cuantitativas discretas con pocos valores, el análisis gráfico es igual que para variables cualitativas. Podemos suministrar un gráfico de barras. También la tabla de frecuencias sería igual que para el caso de variables cualitativas.

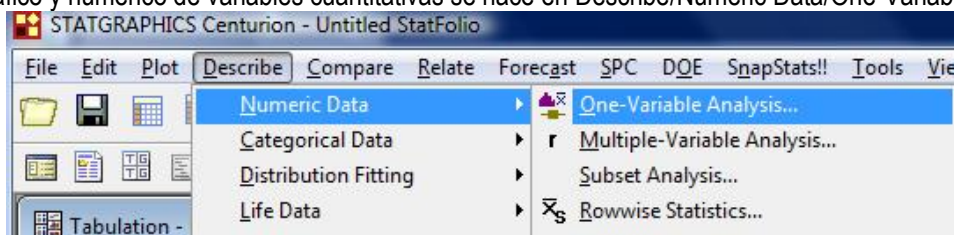
Por ejemplo, la variable hermanos proporciona el número de hermanos que tiene cada alumno. Para seleccionar una variable nueva dentro de un mismo análisis, basta con pulsar el icono que aparece en primer lugar de la segunda fila de iconos .



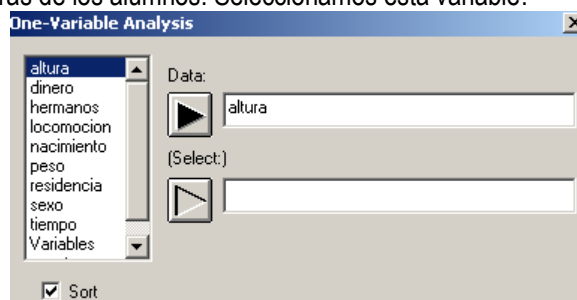
donde puede verse que las familias más frecuentes (entre las que tienen hijos cursando Ingeniería Industrial) son las de 2 y 3 hijos (1 y 2 hermanos).

### 3.2 Análisis gráfico de variables cuantitativas

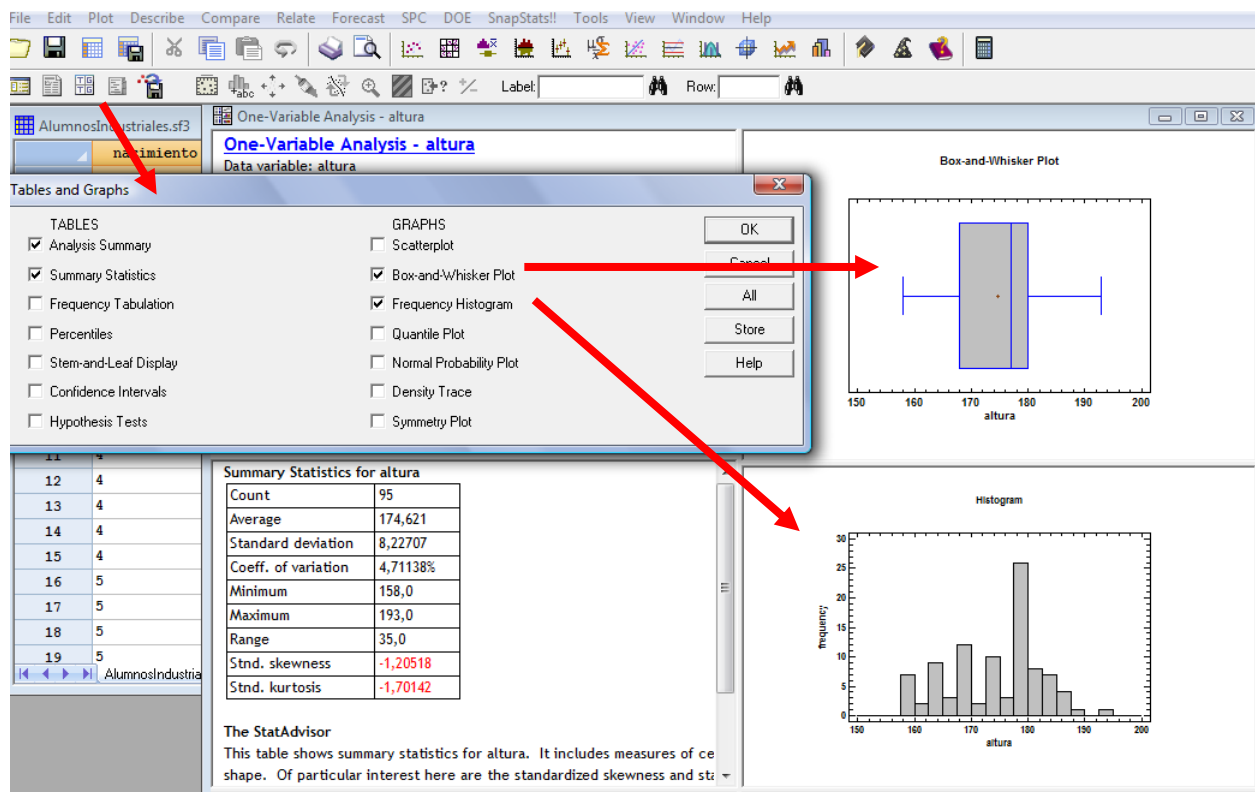
El análisis gráfico y numérico de variables cuantitativas se hace en Describe/Numeric Data/One-Variable Analysis.



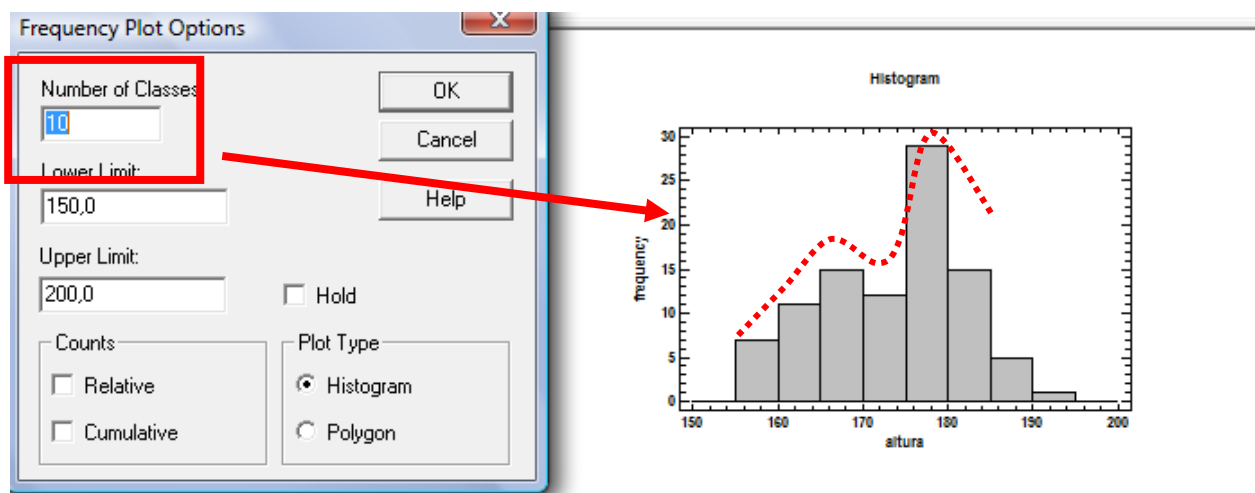
La variable altura tiene las alturas de los alumnos. Seleccionamos esta variable.



Vamos a hacer su histograma y su diagrama box-plot. Seleccionamos las opciones gráficas que queremos.



La media aritmética es la cruz roja que aparece dentro de la caja. El Box-plot nos muestra que la distribución de las alturas es algo asimétrica. La caja central muestra una asimetría negativa, si bien las colas de la distribución no son muy largas. Este efecto se ve también en el histograma. Vamos a cambiar el número de clases del histograma. Como tenemos 95 observaciones, tomaremos  $\sqrt{95} \sim 10$  clases. Nos posicionamos en el histograma y con el botón derecho del ratón seleccionamos Pane Options.

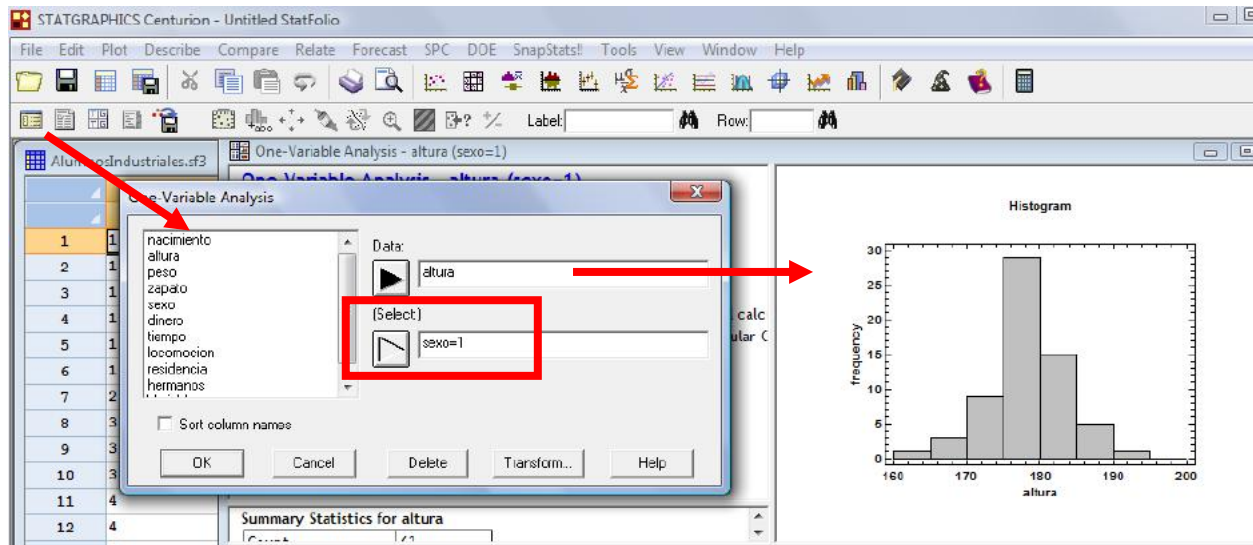


Este mayor número de clases nos muestra una bimodalidad que es imposible de visualizar en un boxplot. Hay una moda en torno a 165cm. Y otra en torno a 178 cm. Esas dos modas sugieren que la población no es homogénea. Es muy posible que se deba a las alturas de chicos y chicas (ver guión de análisis de varias variables).

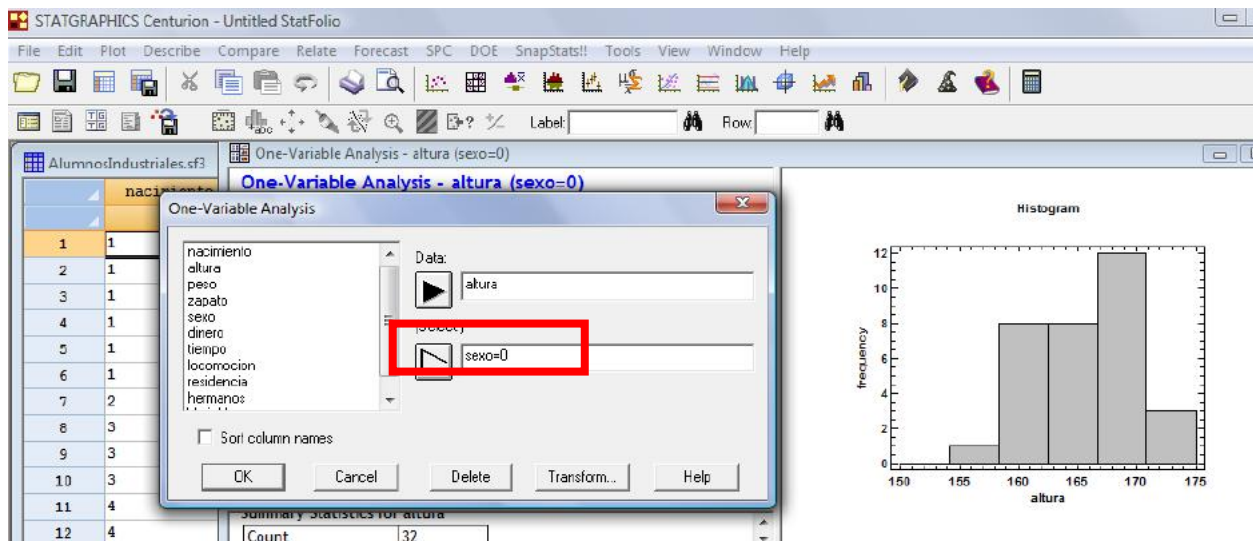
La variable sexo tiene el sexo de los alumnos (1=chico, 0=chica). Vamos a emplear esa variable para seleccionar la altura de los chicos o de las chicas y ver así si cada grupo se concentra alrededor de modas diferentes.



Si queremos seleccionar sólo a los chicos hacemos lo siguiente (el histograma se ha hecho con 8 clases):



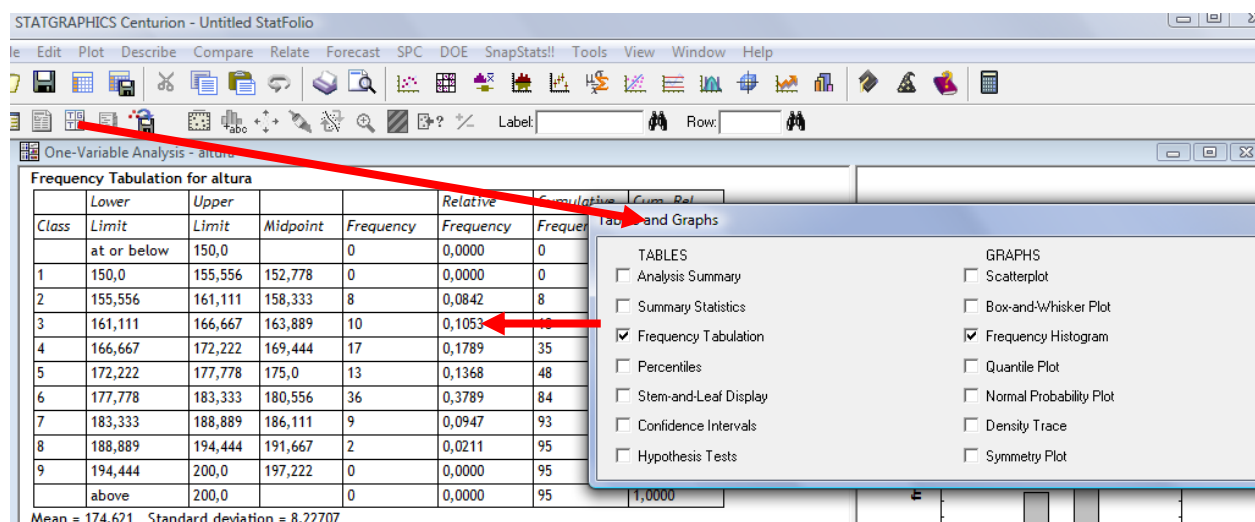
y vemos que con sólo los chicos, la distribución es muy simétrica, unimodal, con moda en los 180 cm, y alta concentración alrededor de ella. Si lo repetimos para las chicas tenemos (histograma con 6 clases)



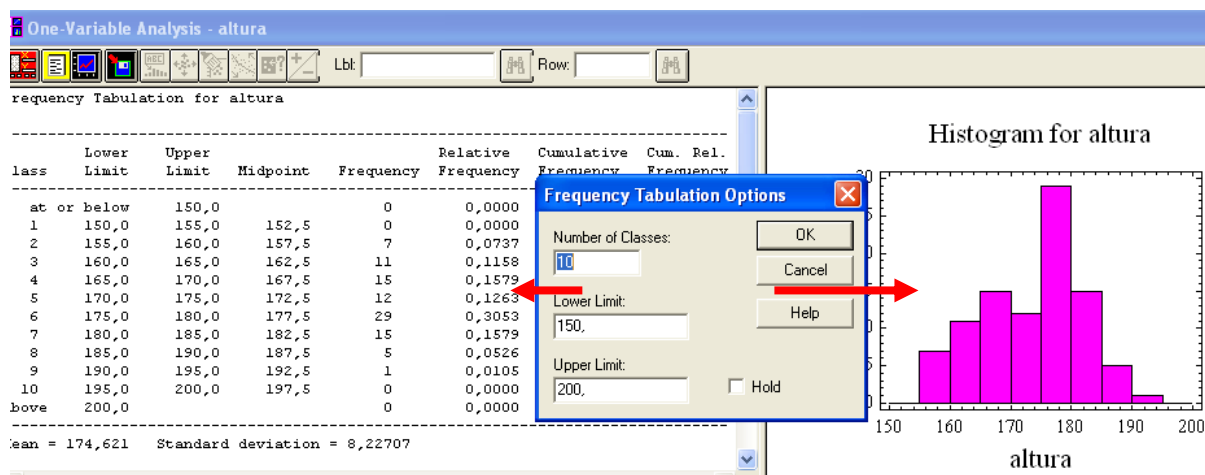
La distribución de las chicas es más uniforme. No es con forma de campana unimodal como la de los chicos. Tal vez sea porque hay menos datos (sólo 32) o porque las chicas de esa titulación sean realmente más heterogéneas entre si.

### 3.3 Tabla de Frecuencias

La distribución de frecuencias mediante una tabla nos proporciona la misma información que un histograma, pero nos permite ver los valores numéricos de las frecuencias de cada intervalo. Para hacer la tabla de frecuencias vamos a las opciones numéricas (TABLES) (Se han seleccionado 9 clases)



La tabla de frecuencias tienen las mismas opciones que el histograma. Podemos cambiar el número de clases o limitar el rango de valores. Para acceder a estas opciones nos colocamos sobre la ventana de resultados y pulsamos el botón derecho del ratón. Seleccionamos entonces Pane Options. En la ventana que obtenemos seleccionamos 10 clases. Los cambios que propongamos para la tabla de frecuencias también afectan al histograma de frecuencias

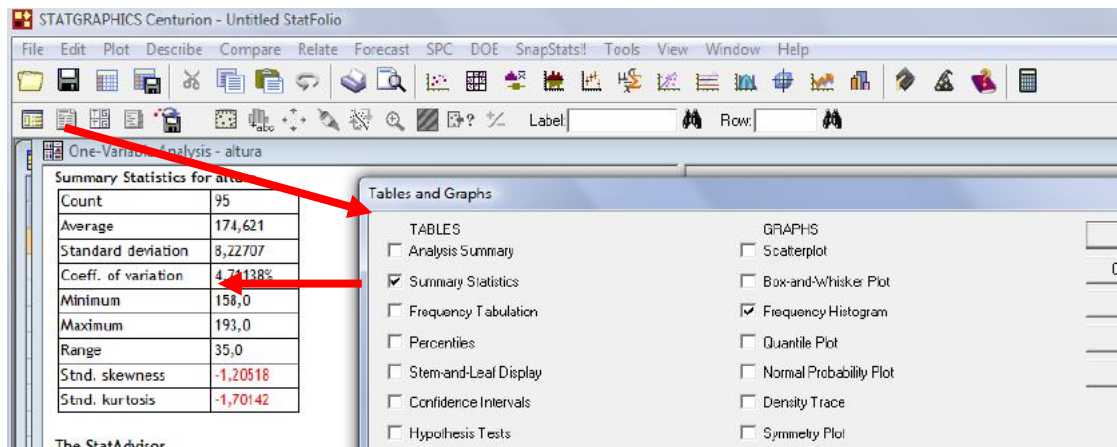


Esta tabla muestra que los dos intervalos modales son alrededor de los valores (midpoint), 167.5 y 177.5 y que el intervalo más frecuente, el centrado en 177.5 contiene a más del 30% de los alumnos.

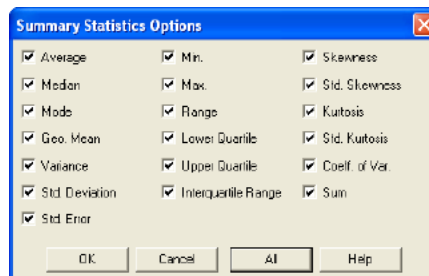
### 3.4 Medidas características de variables cuantitativas

Para calcular las medidas características de la variable altura vamos a las opciones numéricas (TABLES).





Podemos seleccionar las medidas que deseemos. Nos posicionamos en la ventana de resultados y seleccionamos Pane Options. Aparece una ventana con todos los estadísticos univariantes que calcula el Statgraphics. Si los seleccionamos todos obtenemos los siguientes resultados:



One-Variable Analysis - altura

Summary Statistics for altura

Count	95
Average	174,621
Median	177,0
Mode	180,0
Geometric mean	174,427
5% Trimmed mean	174,696
5% Winsorized mean	174,526
Variance	67,6847
Standard deviation	8,22707
Coeff. of variation	4,71138%
Standard error	0,844079
5% Winsorized sigma	8,34133
MAD	5,0
Sbi	8,86827
Minimum	158,0
Maximum	193,0
Range	35,0
Lower quartile	168,0
Upper quartile	180,0
Interquartile range	12,0
1/6 sextile	165,0
5/6 sextile	182,0
Intersextile range	17,0
Skewness	-0,302876
Std. skewness	-1,20518
Kurtosis	-0,855173
Std. kurtosis	-1,70142
Sum	16589,0
Sum of squares	2,90315E6

Es necesario hacer algunas puntualizaciones sobre estas medidas características:

- Al ser la altura una medida continua, LA MODA NO TIENE SENTIDO. La moda es el valor más frecuente, y en una variable continua podría suceder que no se repitiese ningún valor. En esos casos, el programa nos devolvería el primer valor que leyese en el fichero de datos. En este tipo de variables sólo tiene sentido hablar de intervalo modal de un histograma.
- La varianza que se calcula es realmente la cuasivarianza, de expresión

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

en lugar de

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

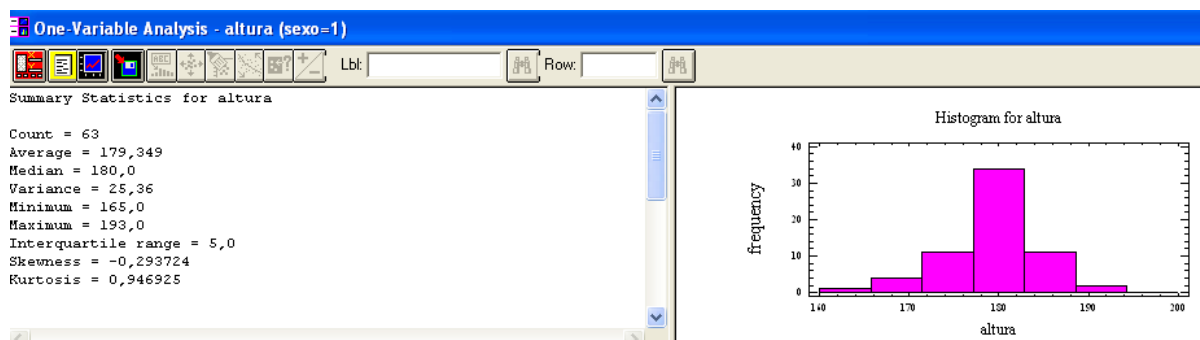
La conveniencia de dividir por  $n-1$  en lugar de  $n$  no es inmediata, y su justificación teórica se verá en temas más avanzados.

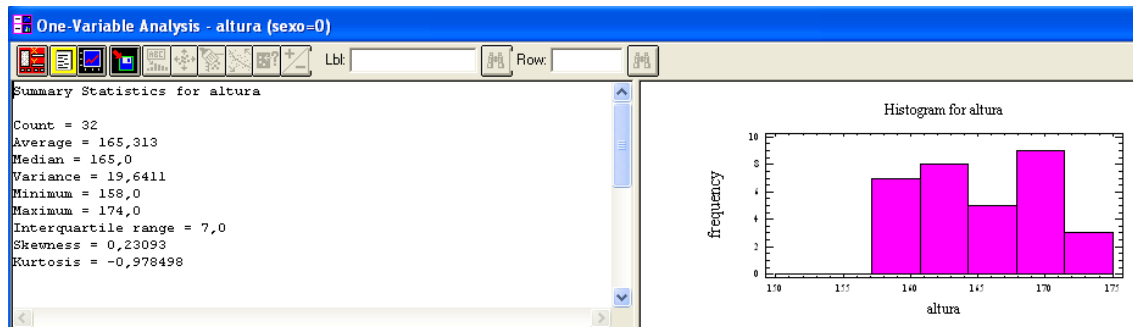
- La desviación típica usa también esta misma formulación, dividiéndose por  $n-1$ . Es por tanto la cuasidesviación típica.
- Las siguientes medidas  
Standard Error  
Std. Skewness  
Std. Kurtosis  
no son propiamente de estadística descriptiva, sino de inferencia. Por tanto no se cubren en este documento.
- El coeficiente de curtosis que calcula el Statgraphics es realmente el 'Exceso de curtosis', definido como

$$K = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n \times s^4} - 3.$$

Por tanto, para una variable que tenga forma de campana, la curtosis es 0.

A continuación se muestra la comparación entre chicos y chicas mediante algunas medidas características:



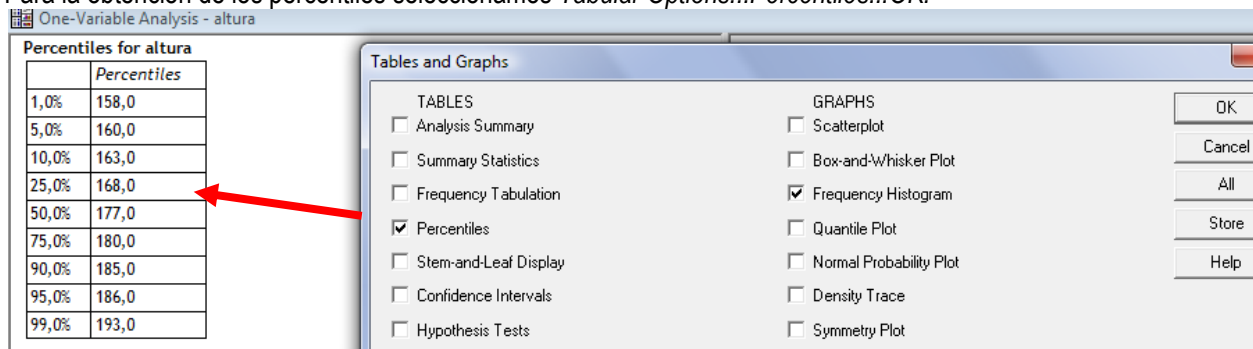


Puede verse ahora que los chicos de esta muestra son, por término medio, más altos que las chicas. La estatura media de los chicos es 179 mientras que para las chicas es sólo 165. Los chicos se concentran mucho alrededor de esa media. Puede verse que la media es casi igual que la moda, lo que se ve también fácilmente en la simetría del histograma y el bajo valor del coeficiente de asimetría. Esa concentración cerca de la media se ve también en el rango intercuartílico. El 50% de los chicos colocados en las posiciones centrales sólo se diferencian en un máximo de 5 centímetros, mientras que para las chicas ese rango intercuartílico es de 7 cm.

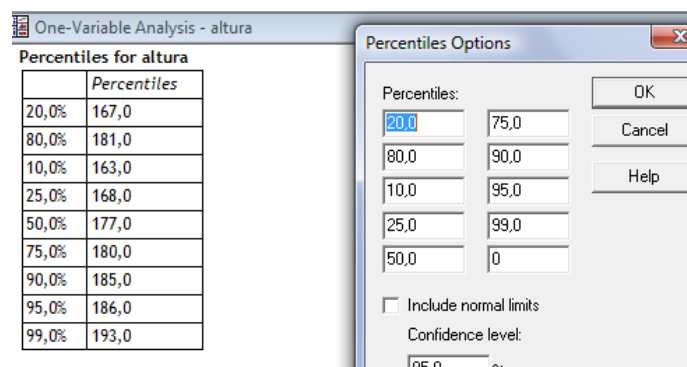
Otra medida que resume esa mayor concentración es la curtosis. En los chicos es positiva y alta, mientras que en las chicas es negativa. En este caso, la curtosis es la medida que mejor resume la diferencia entre esos dos histogramas.

### 3.5 Percentiles

Para la obtención de los percentiles seleccionamos *Tabular Options...Percentiles...OK*.



Para seleccionar algún percentil concreto, nos ponemos sobre la ventana de resultados, pulsamos el botón derecho del ratón y seleccionamos *Pane Options*. Vamos a calcular el percentil 20 y 80.



El 20% de los alumnos mide menos de 167 cm, y el 80% mide menos de 181 cm.

# Descripción simultánea de varias variables con STATGRAPHICS CENTURION

Ficheros empleados: AlumnosIndustriales.sf3, Rotura.sf3

## 4. Introducción

En muchas ocasiones nos interesará comparar varias variables, o comparar los valores de una variable en dos o más grupos de individuos. En esos casos es más interesante producir gráficos y resúmenes estadísticos conjuntamente, que faciliten esa comparación, que realizar el análisis univariante por separado de cada variable. Por ejemplo, nos interesará hacer diagramas box-plot de cada variable pero en un mismo gráfico.

## 5. Box-plot Múltiple

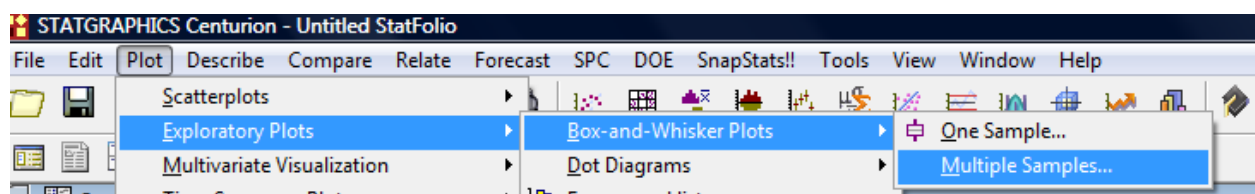
Nuestro objetivo es una representación gráfica que tenga los diagramas box-plot de varias variables. Este gráfico permitirá una mejor comparación de esas variables. El Statgraphics proporciona varios lugares para hacer box-plot múltiples.

### 2.1 Una variable que se subdivide en subgrupos

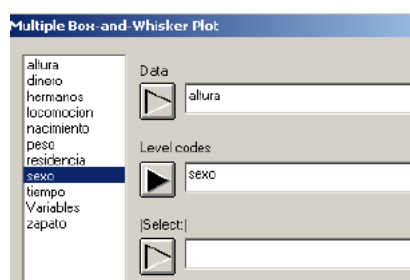
Nos interesa analizar cómo es la distribución de valores de una variable cuando el conjunto de datos lo subdividimos en subgrupos de acuerdo a algún criterio. Por ejemplo las alturas de un grupo de estudiantes en función de su sexo. Necesitamos tener dos columnas:

- Columna con los datos de la variable
- Columna con códigos que nos permita hacer los subgrupos. Por ejemplo, con la variable sexo, basta con que tenga valores 1 y 0. Estos valores son sólo para distinguir a los miembros de cada grupo, por lo que su valor es irrelevante. Podrían ser -1 y 1, 33 y 34, o incluso caracteres.

Hay varios lugares para hacer este tipo de Box-plots. El primer lugar es en Plot/Exploratory Plots/Multiple Box-and-Whisker Plot

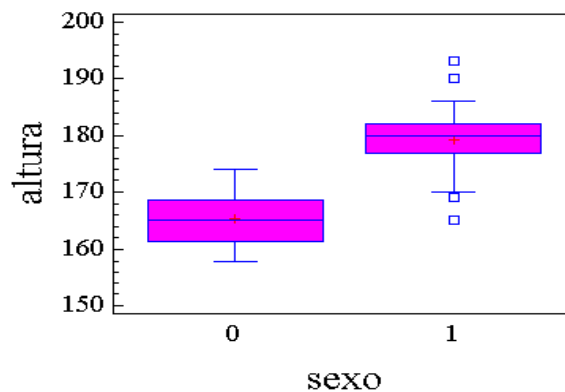


Veamos un ejemplo con el fichero AlumnosIndustriales.sf3. que tiene algunos datos de un conjunto de estudiantes de Ingeniería Industrial. Vamos a comparar la estatura de los chicos y de las chicas. Se nos pregunta entonces por el nombre de la variable que tiene los datos -altura- y la variable que nos ayudará a formar los dos grupos chicos/chicas -sexo- (Level Codes)



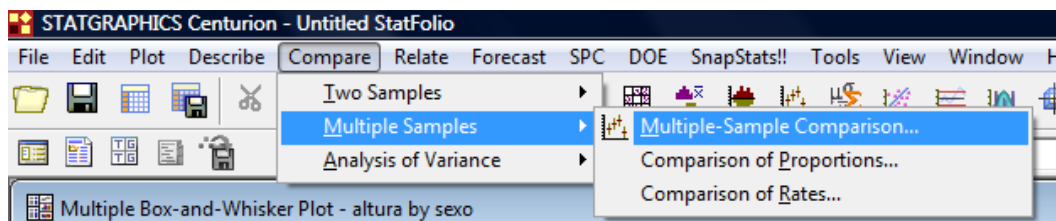
El resultado es el siguiente

## Box-and-Whisker Plot



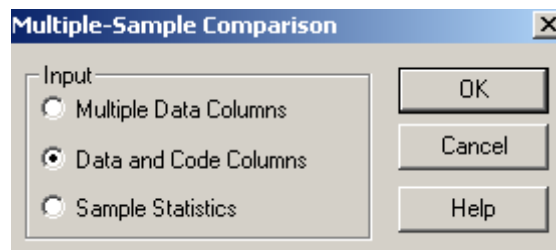
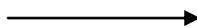
donde chicos=1, chicas=0. Puede verse que los chicos son, en general, más altos que las chicas. Viendo el solapamiento de los bigotes de ambos box-plots podemos interpretar que, aproximadamente, sólo el 25% de las chicas más altas tienen estaturas comparables al 25% de los chicos más bajos.

Los Box-plots múltiples también se pueden hacer en Compare/Multiple Samples/Multiple-Sample Comparison.

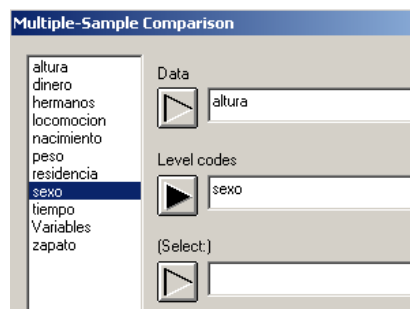


Esta opción es mucha más flexible que la anterior, pues permite no solo comparar una variable en varios subgrupos como comparar varias variables diferentes (varias columnas). Se nos ofrecen dos posibilidades, donde la que nos interesa en este momento es la de Data and Code Columns:

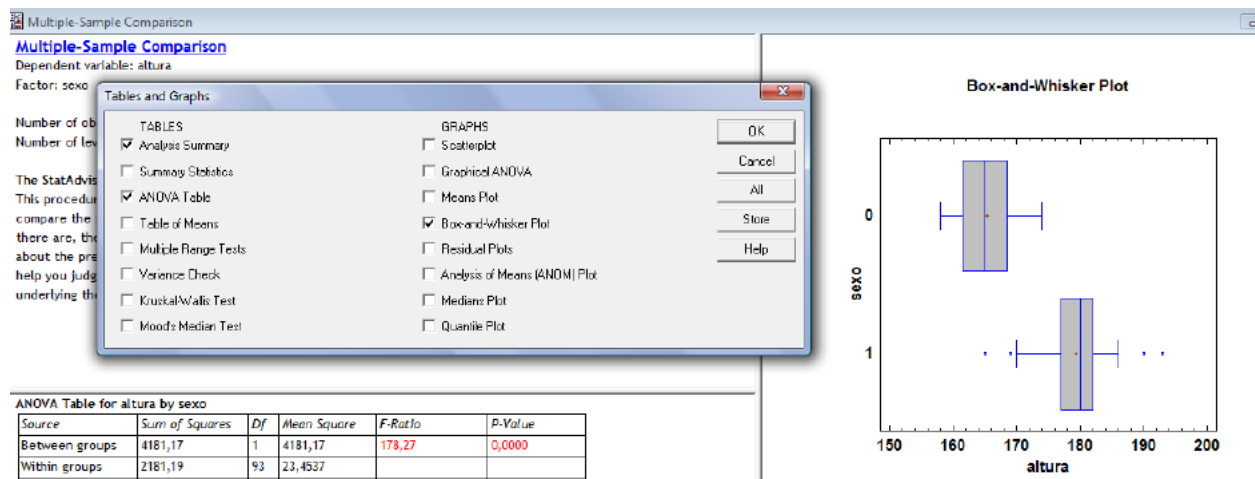
Similar a la anterior



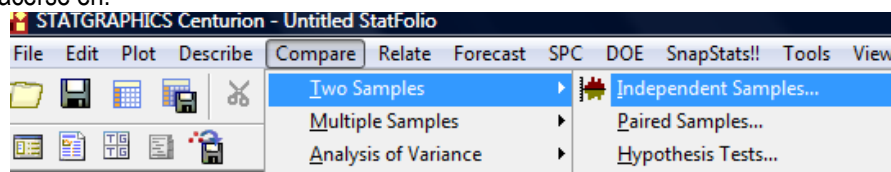
La opción *Data and Code Columns* permite usar datos en el mismo formato que antes: una variable con los datos y otra con la información para hacer subgrupos. Si seleccionamos esta opción llegamos a una ventana similar a la que vimos antes donde nos preguntan los nombres de las variables



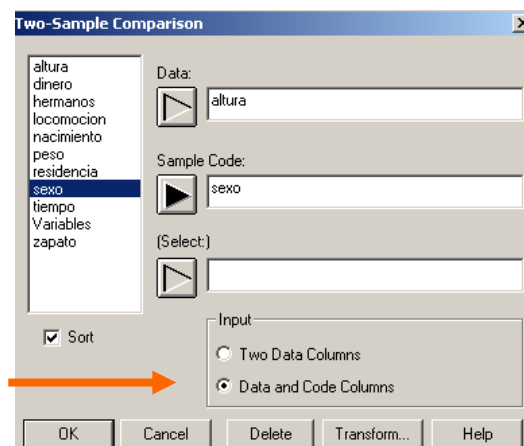
Para visualizar el box-plot tenemos que seleccionarlo en las opciones gráficas, como se muestra a continuación



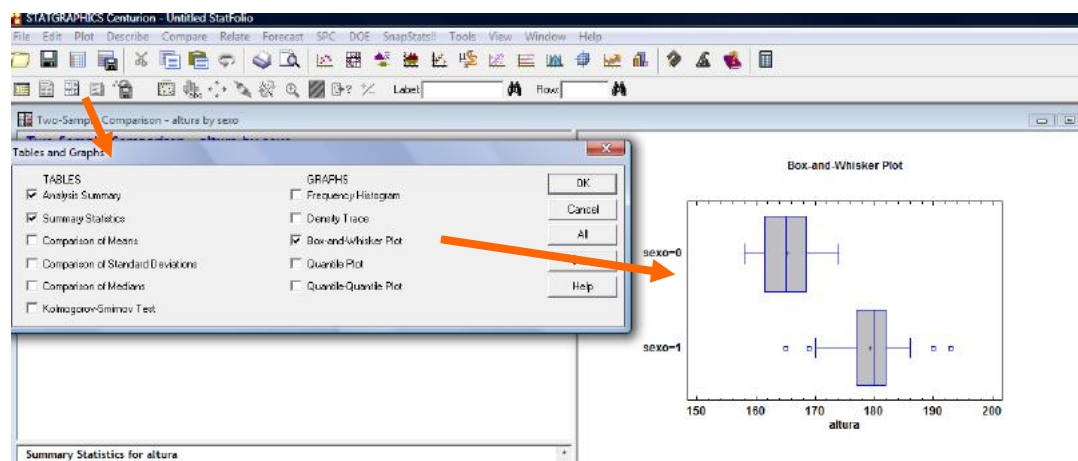
Cuando, como en el ejemplo de las estaturas de chicos y chicas, sólo tenemos dos subgrupos, los box-plot también pueden hacerse en:



A continuación hay que especificar que nuestros datos, la variable altura, están en una columna (Data Column) y que los dos subgrupos se forman con la variable sexo (Code Column)



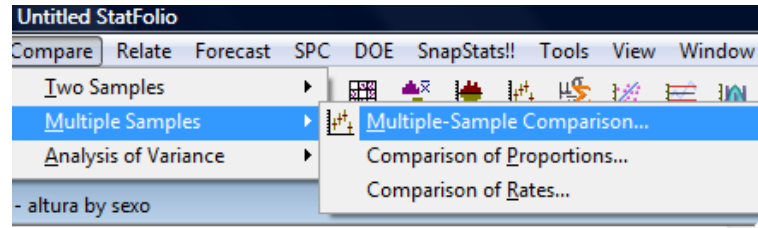
Entonces seleccionamos la opción gráfica de Box-Plot



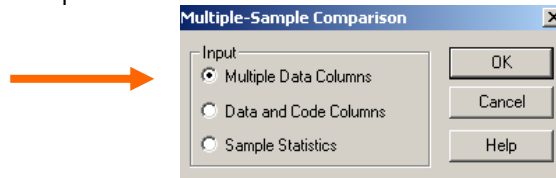


## 2.2 Variables en columnas diferentes

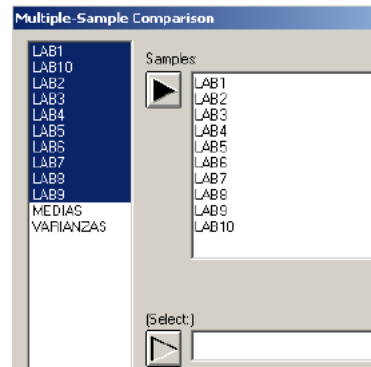
Si tenemos varias columnas, de una misma magnitud o de magnitudes diferentes, podemos también hacer box-plot múltiples de esas variables en un mismo gráfico. Como ejemplo usaremos los datos del fichero rotura.sf3. Este fichero contiene la tensión de rotura (presión ejercida en el momento en que se produce la rotura) de un conjunto de piezas idénticas con el objetivo de probar la resistencia del material empleado. El fichero contiene 10 variables, LAB1 a LAB10, donde cada una muestra las tensiones de roturas obtenidas en 10 laboratorios diferentes. En cada laboratorio se rompieron 100 piezas diferentes, y se anotaron las correspondientes tensiones de rotura. Vamos a ver el box-plot de las 10 variables. En primer lugar vamos a



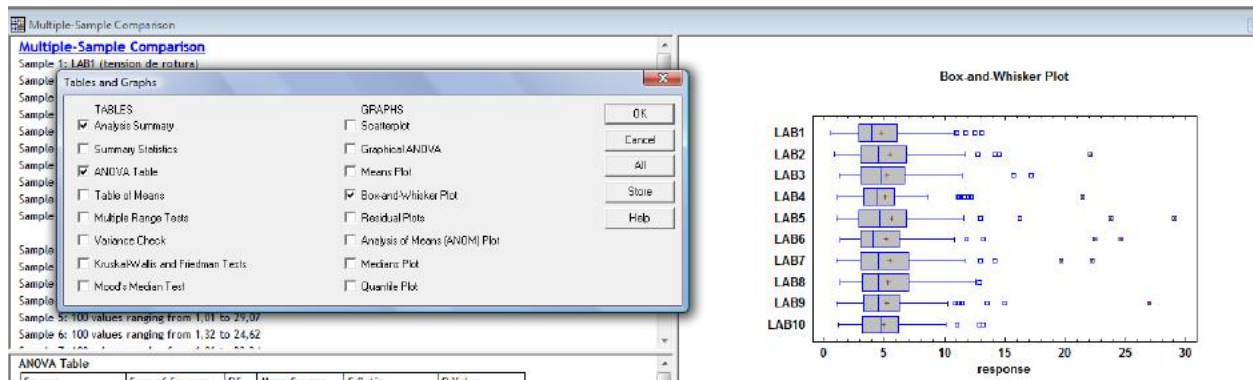
donde ahora seleccionamos que tenemos varias variables:



e introducimos los nombres de las variables

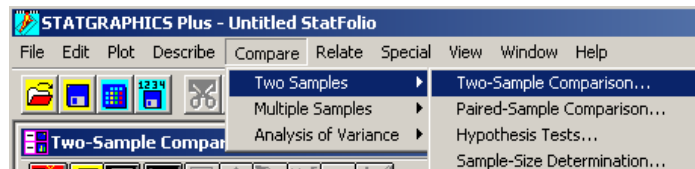


A continuación seleccionamos la opción gráfica de box-plot.

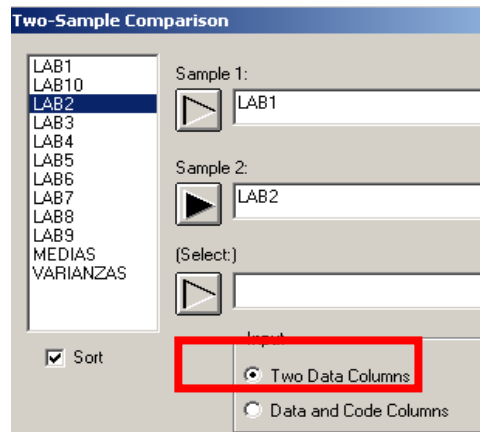


Puede verse que los resultados de los 10 laboratorios son muy similares. A partir de los bigotes del tercer cuartil y de los atípicos se aprecia que, en general, las distribuciones son asimétricas positivas. Por tanto, esos valores no son propiamente atípicos, sino la cola de la distribución por la derecha.

Si sólo tuviésemos dos variables, las podemos comparar con Box-plots en:



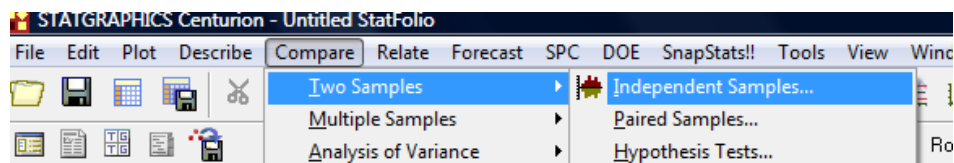
A continuación habría que especificar los datos, que en nuestro caso serían los datos de dos de los laboratorios que quisiésemos comparar.



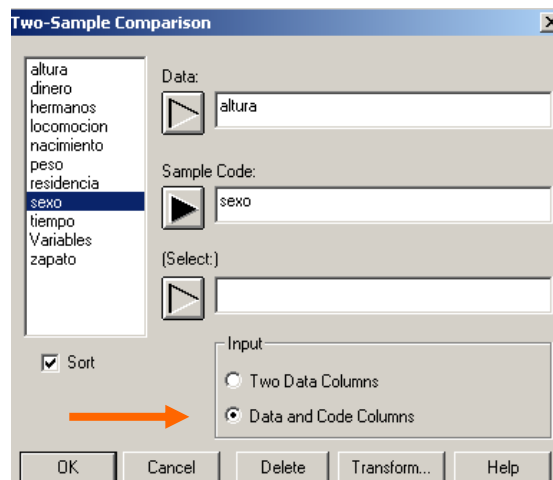
y haríamos el box-plot como se hizo anteriormente

## 6. Dos histogramas superpuestos

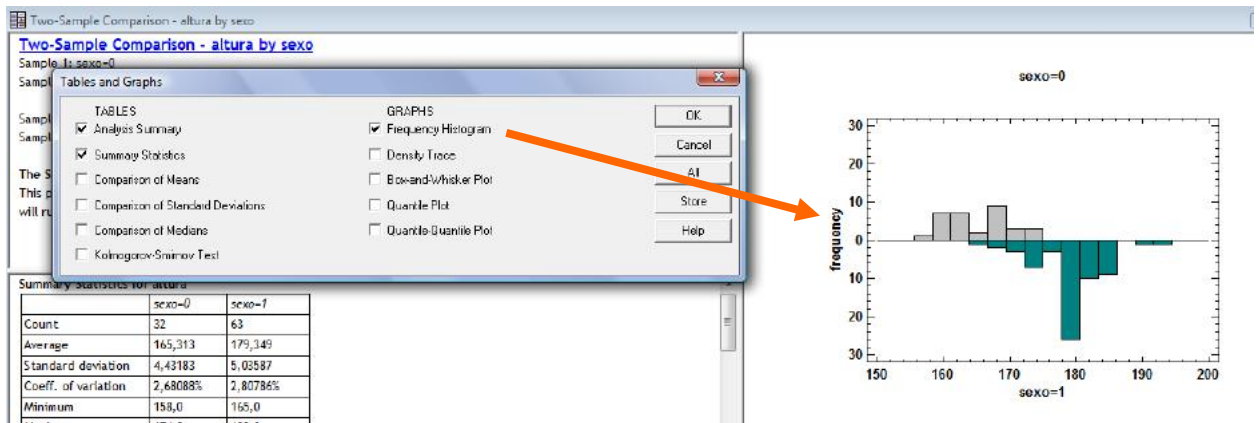
Si queremos comparar dos histogramas, es útil colocarlos en la misma figura. Esta opción gráfica se hace en Compare/Two Samples/Two-sample Comparison, Tanto si tenemos una variable dividida en dos grupos, como dos variables en dos columnas diferentes



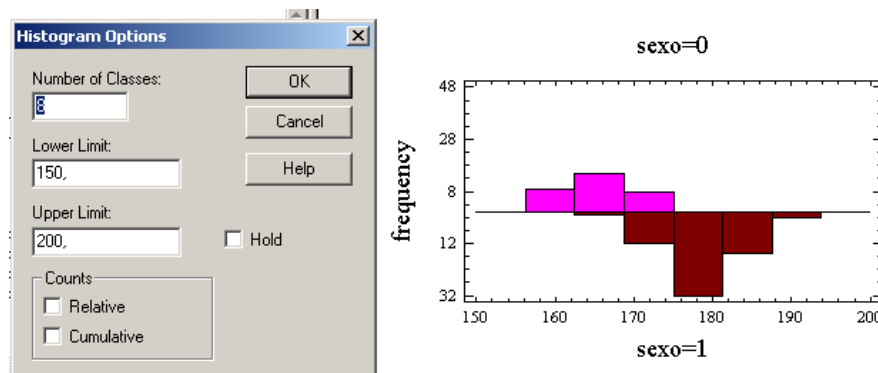
Vamos a aplicarlo al caso de las estaturas de chicos y chicas. El Statgraphics nos pregunta el formato de nuestros datos: dos columnas de datos, o una columna con datos y otra con códigos. En nuestro caso la opción a seleccionar es la segunda:



y seleccionamos entonces los histogramas. El resultado es el siguiente

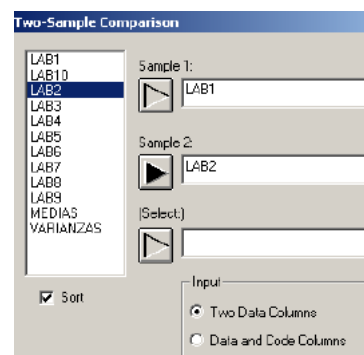


Vemos que para las chicas salen muy pocas clases. Podemos aumentar el número de clases colocándonos en la ventana del histograma y pulsando el botón derecho del ratón. En Pane Options podemos cambiar el número de clases.

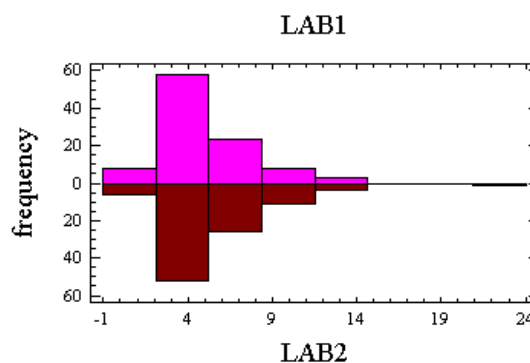


Con este número de clases se aprecia mejor que dentro de cada grupo, las alturas son bastante simétricas, y que los chicos tienen estaturas más altas.

Si queremos hacer dos histogramas con las tensiones de rotura de dos laboratorios (fichero rotura.sf3) tenemos que introducir los siguientes datos



y el resultado es



donde puede apreciarse que las distribuciones en los laboratorios 1 y 2 son muy similares: unimodales, con el mismo intervalo modal y con fuerte asimetría positiva.

## 7. Medidas características

En general, el análisis de los datos se comienza mediante gráficos. Posteriormente buscaremos medidas características que nos cuantifiquen los aspectos que más nos interesen. **El Statgraphics proporciona medidas características de varias variables en casi los mismos lugares en los que proporciona Box-plots múltiples.** Por ejemplo, si queremos conocer las medias y las asimetrías de las tensiones de rotura de los 10 laboratorios (archivo rotura.sf3) podemos ir a Compare/Multiple Samples/Multiple Sample Comparison. Allí seleccionamos Tabular Options y Summary Statistics

	Count	Average	Standard deviation	Coeff. of variation	Minimum	Maximum	Range
LAB1	100	4,7659	2,70082	56,6698%	0,6	13,09	12,49
LAB2	100	5,4516	3,3187	60,8757%	0,88	22,07	21,19
LAB3	100	5,2422	2,94087	56,0998%	1,27	17,16	15,89
LAB4	100	5,0333	2,87773	57,1739%	1,08	21,46	20,38
LAB5	100	5,5892	4,23106	75,7007%	1,01	29,07	28,06
LAB6	100	5,1124	3,51978	68,8478%	1,32	24,62	23,3
LAB7	100	5,451	3,4907	64,0377%	1,06	22,24	21,18
LAB8	100	5,2868	2,77769	52,5401%	1,36	12,88	11,52
LAB9	100	5,2225	3,42222	65,5283%	1,03	26,92	25,89
LAB10	100	4,9422	2,27174	45,9663%	1,17	13,21	12,04
Total	1000	5,20971	3,19308	61,291%	0,6	29,07	28,47

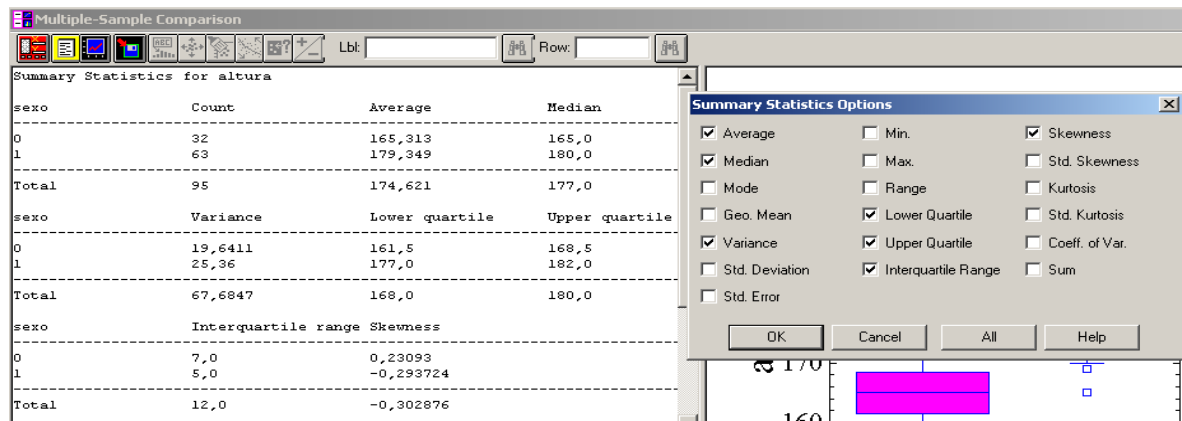
Si nos posicionamos en la ventana de resultados y pulsamos el botón derecho de ratón, podemos seleccionar Pane Options, y allí escoger las medidas características que nos interesen:

	Count	Average	Skewness
LAB1	100	4,7659	1,12998
LAB2	100	5,4516	1,88386
LAB3	100	5,2422	1,35638
LAB4	100	5,0333	2,51899
LAB5	100	5,5892	2,81372
LAB6	100	5,1124	3,1749
LAB7	100	5,451	2,17543
LAB8	100	5,2868	0,997454
LAB9	100	5,2225	3,14587
LAB10	100	4,9422	1,20103
Total	1000	5,20971	2,42129

Puede verse que, efectivamente, las distribuciones son muy asimétricas

En el caso de las alturas de chicas y chicos tenemos el siguiente resultado:

Información de las medidas características



Los resultados muestran que en media, los chicos son  $179.349 - 165.313 = 14.036$  cm. más altos que las chicas. Llama la atención que el 50% de los chicos mide (declara medir) más de 180 cm. La varianza de los chicos es mayor que la de las chicas, pero su rango intercuartílico es menor. Ambas distribuciones son bastante simétricas, pues los coeficientes de asimetría son pequeños.