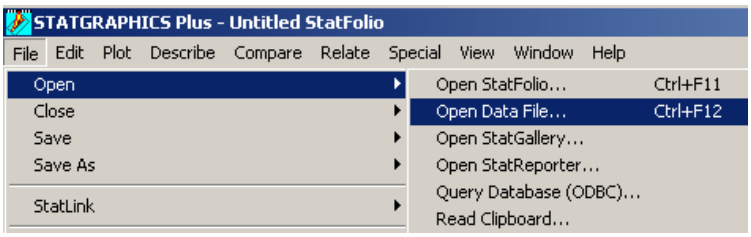
Estadística Descriptiva de una variable con STATGRAPHICS

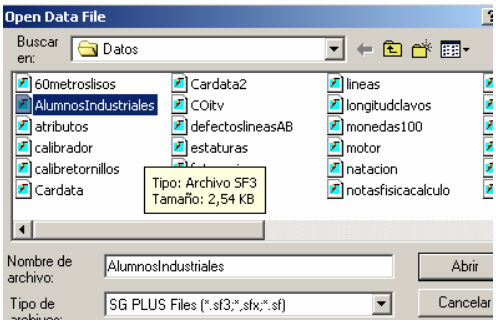
Ficheros empleados: AlumnosIndustriales.sf3,

1. Introducción

El objetivo de este documento es la utilización de las técnicas de estadística descriptiva más importantes para resumir la información de un conjunto de datos. Se usará el programa Stagraphics (v 4.1). 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.





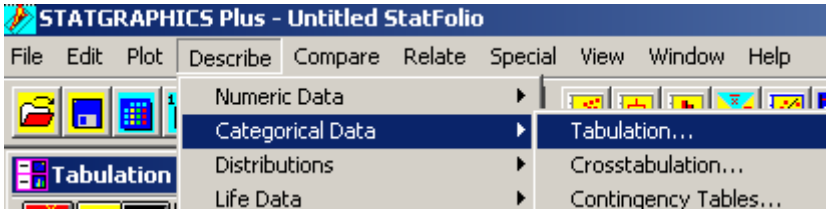
O bien pulsamos la tecla  y seleccionamos el fichero.

2. Descripción de variables cualitativas

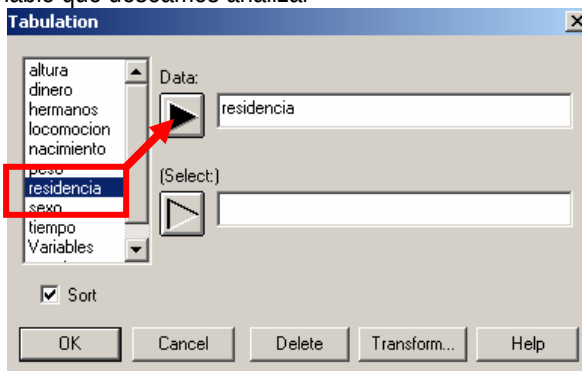
La variable residencia corresponde al lugar de residencia de los alumnos. Su codificación es

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...

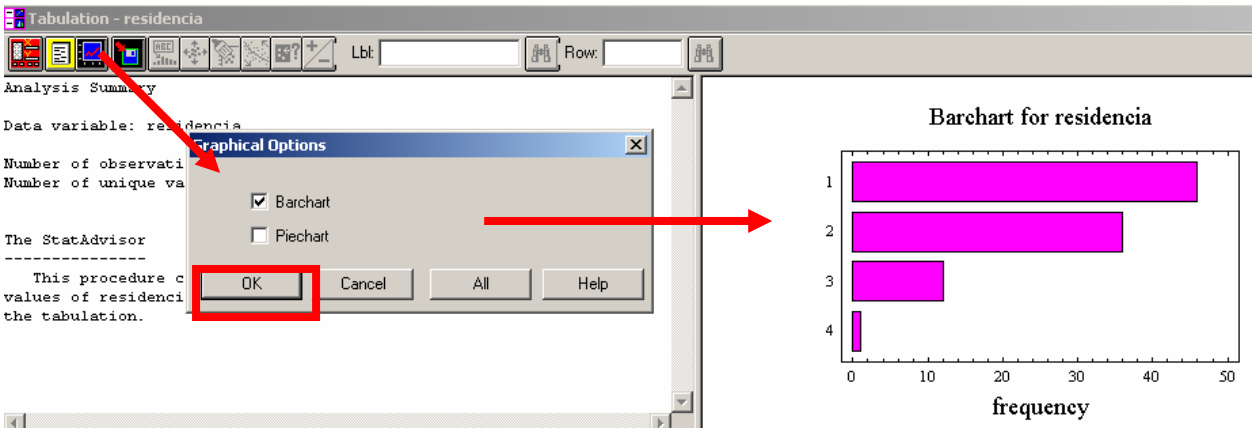


donde nos preguntan por la variable que deseamos analizar



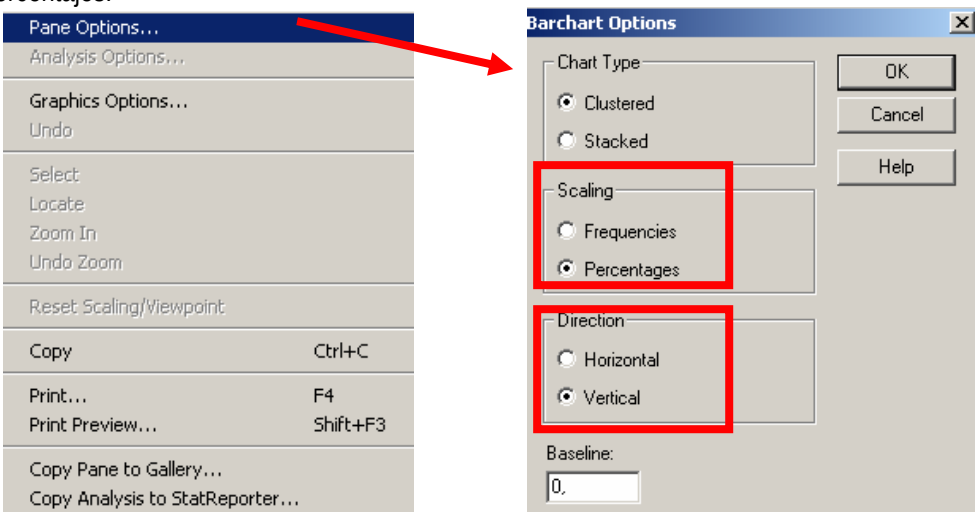
2.1 Gráfico de barras

En las opciones gráficas seleccionamos barchart

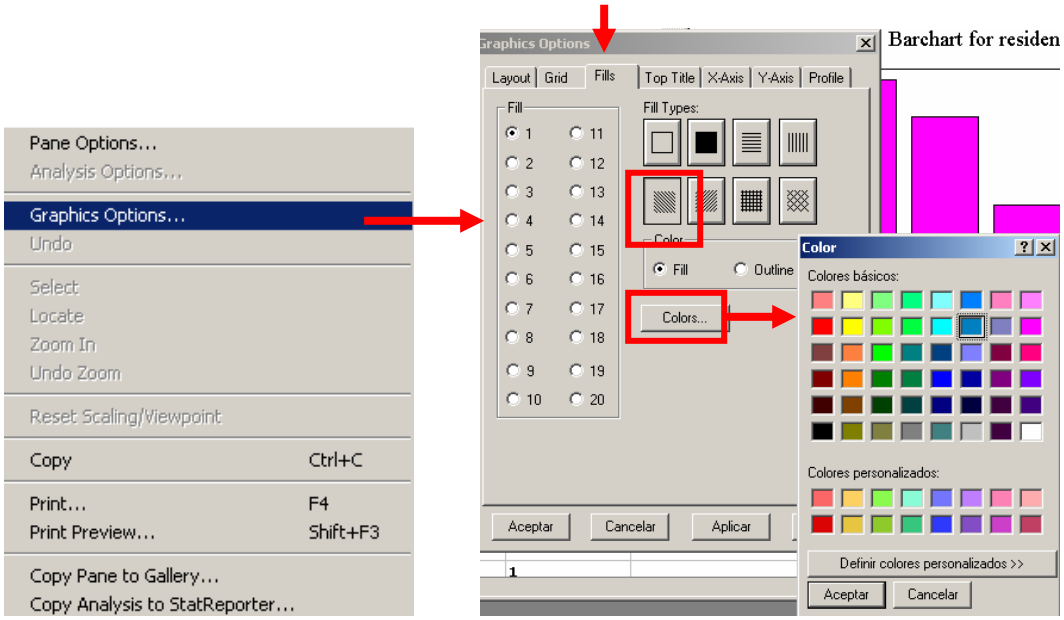


donde puede verse que la población más grande de alumnos son los procedentes de Madrid Sur, con casi 50 alumnos.

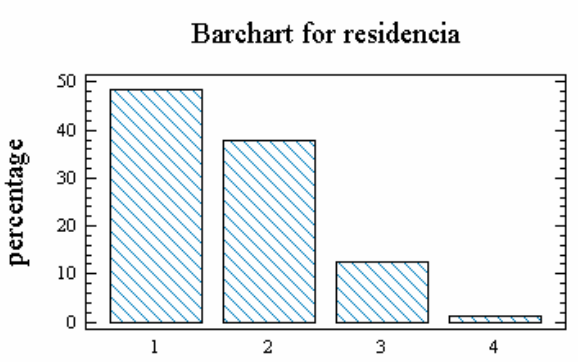
Si queremos cambiar el aspecto del gráfico nos colocamos sobre él, pulsamos el botón derecho del ratón y seleccionamos Pane Options. Seleccionamos entonces que las barras sean verticales y que las frecuencias sean en porcentajes.



También vamos a cambiar la estética de los rectángulos. En lugar de Pane Options seleccionamos Graphics Options. Cambiaremos el relleno de las barras (fills)



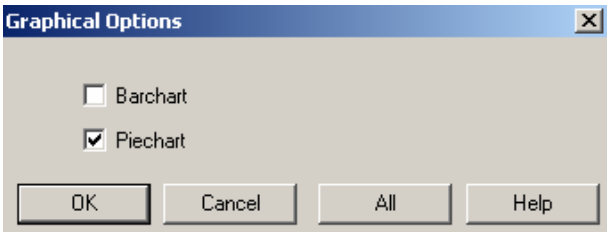
y obtenemos



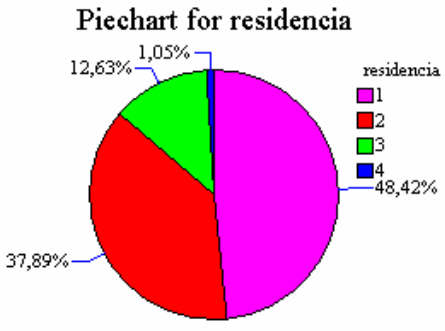
Al estar la altura de las barras en % podemos ver que casi el 50% de los alumnos proceden del sur de Madrid

2.2 Gráfico de tarta o porciones

Dentro de las opciones gráficas,  seleccionamos Piechart

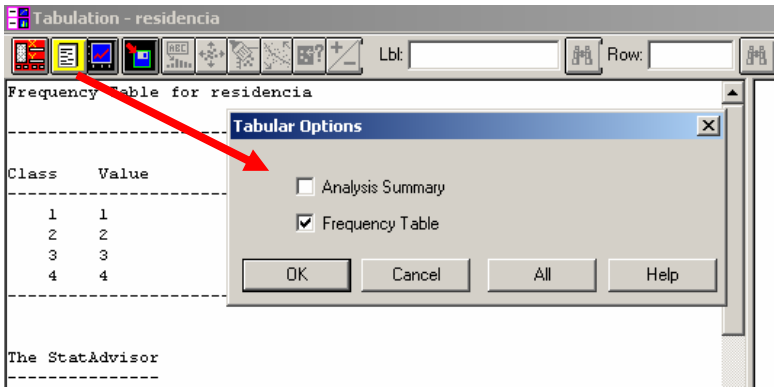


y obtenemos



2.3 Tabla de frecuencias

En las opciones numéricas (Tabular Options ) encontramos la tabla de frecuencias, que nos da la información numérica que antes hemos representado en gráficos.



y la tabla de frecuencias resultante es

Frequency Table for residencia

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. 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

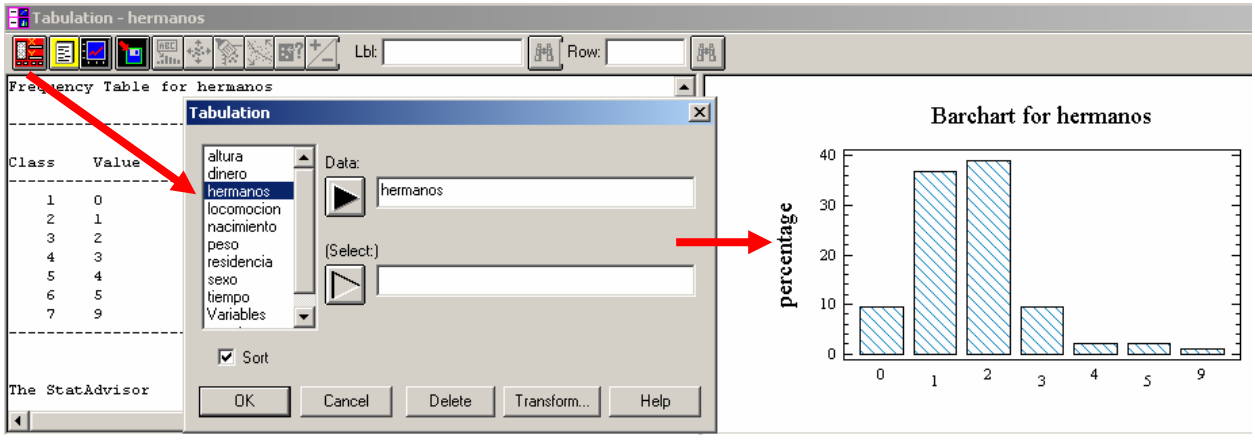
donde puede verse que entre el sur (48.4%) y el centro (37.9%) de Madrid, se abarca al 86.3% de los estudiantes

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.

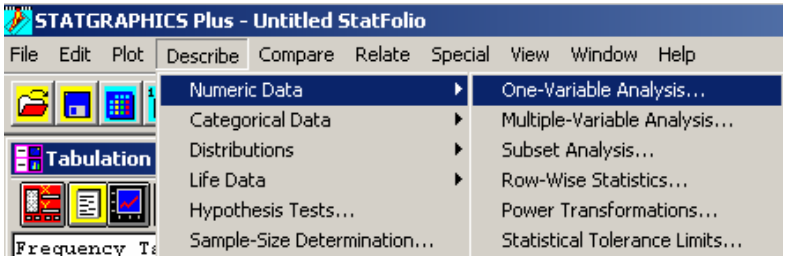
Por ejemplo, la variable hermanos proporciona el número de hermanos que tiene cada alumno. Para selecionar una variable nueva dentro de un mismo análisis, basta con pulsar la tecla denominada Input Dialog  .



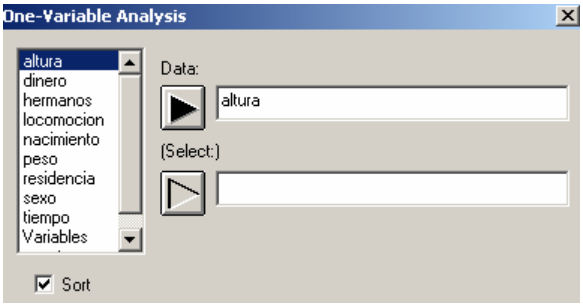
donde puede verse que las familias más frecuentes (entre las que tienen hijos cursando Ingeniería Industrial) son las de 2 y 3 hijos (1 y 2 hermanos)

3.2 Análisis gráfico de variables cuantitativas

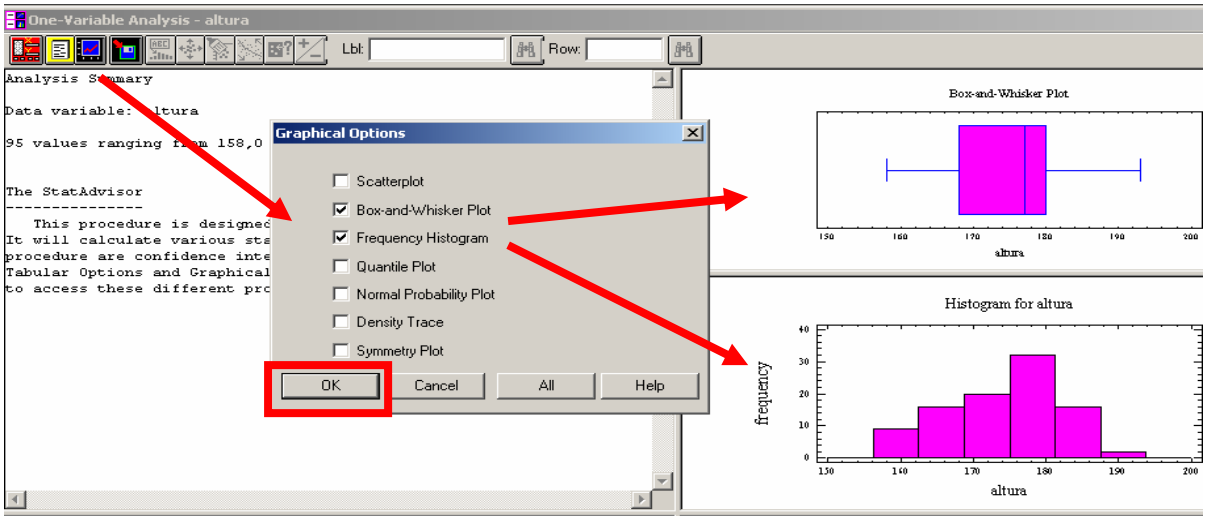
El análisis gráfico y numérico de variables cuantitativas se hace en Describe/Numeric Data/One-Variable Analisys



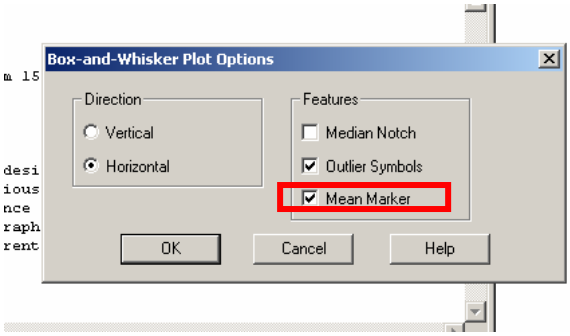
La variable altura tiene las alturas de los alumnos. Seleccionamos esta variable

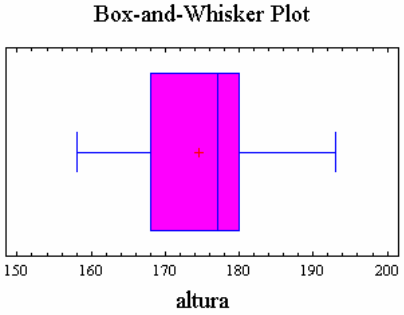


Vamos a hacer su histograma y su diagrama box-plot. Seleccionamos las opciones gráficas que queremos.

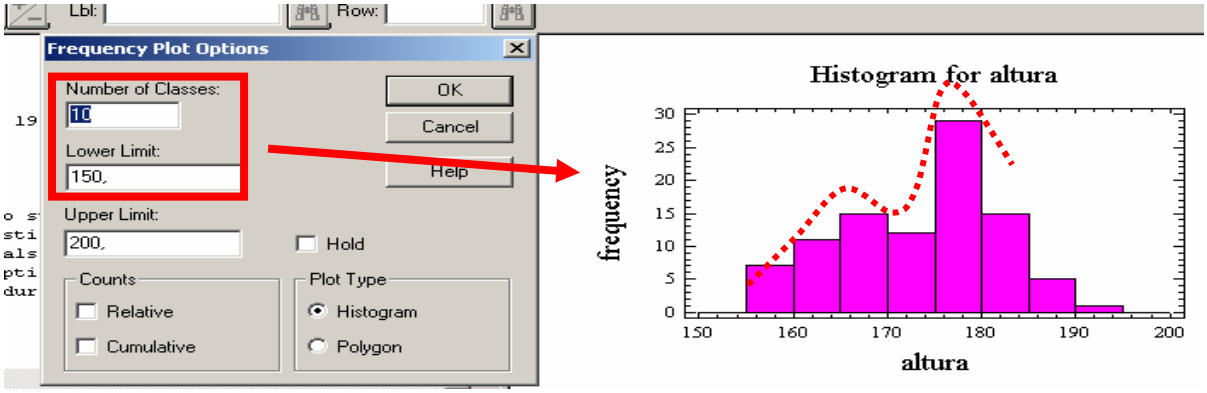


Vamos a añadir en el Box-plot la marca de la posición de la media muestral. Nos colocamos en ese gráfico, pulsamos el botón derecho del ratón, y seleccionamos Pane Options





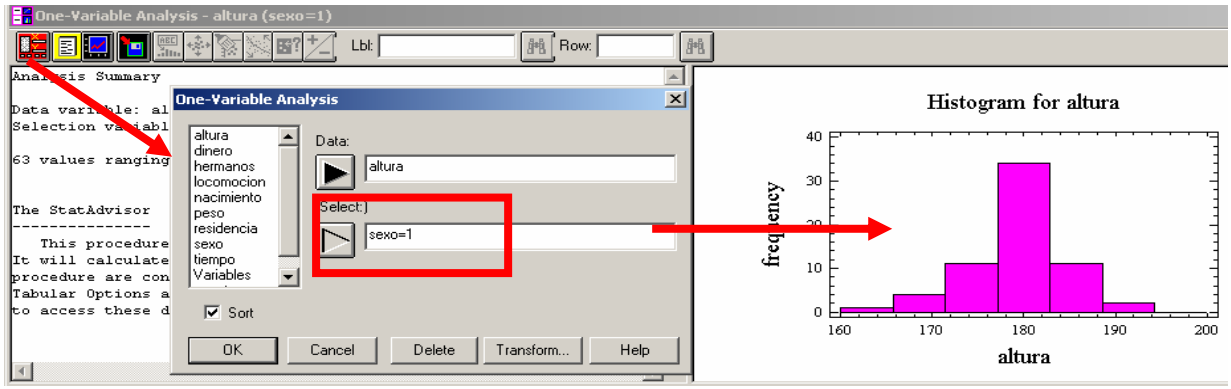
El Box-plot nos muestra que la distribución de las alturas es algo asimétrica. La caja central muestra una asimetría negativa, si bien las colas de la distribución no son muy largas. Este efecto se ve tambi´çen en el histograma. Vamos a cambiar el número de clases del histograma. Como tenemos 95 observaciones,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



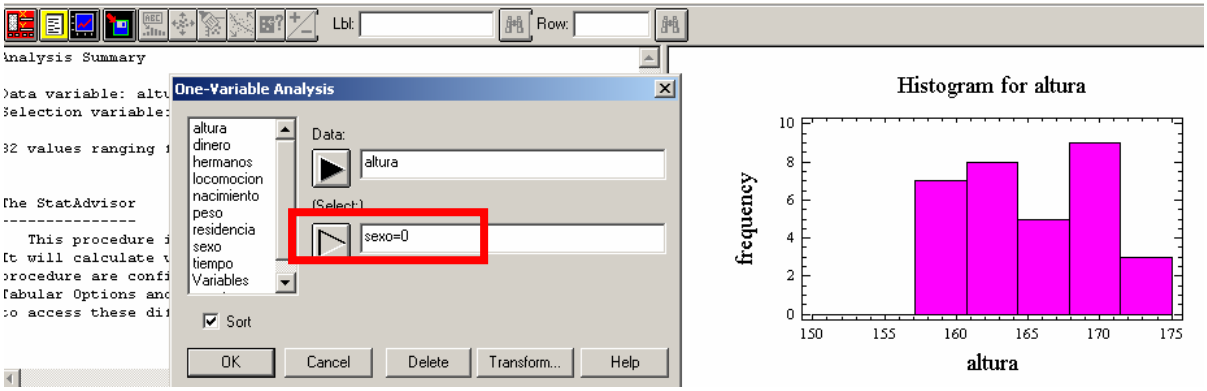
Este mayor número de clases nos muestra una bimodalidad que es imposible de visualizar en un boxplot. Hay una moda en torno a 165cm. Y otra en torno a 178 cm. Esas dos modas sugieren que la población no es homogénea. Es muy posible que se deba a las alturas de chicos y chicas (ver guión de análisis de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

Si queremos seleccionar sólo a los chicos hacemos lo siguiente:



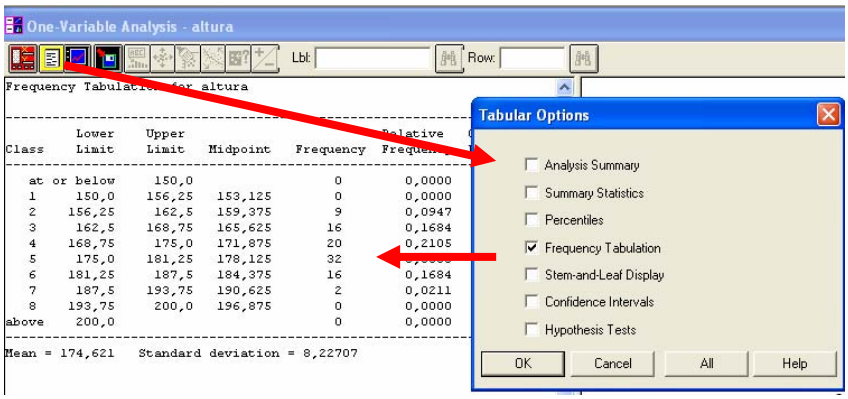
y vemos que con sólo los chicos, la distribución es muy simétrica, unimodal, con moda en los 180 cm. Si lo repetimos para las chicas tenemos



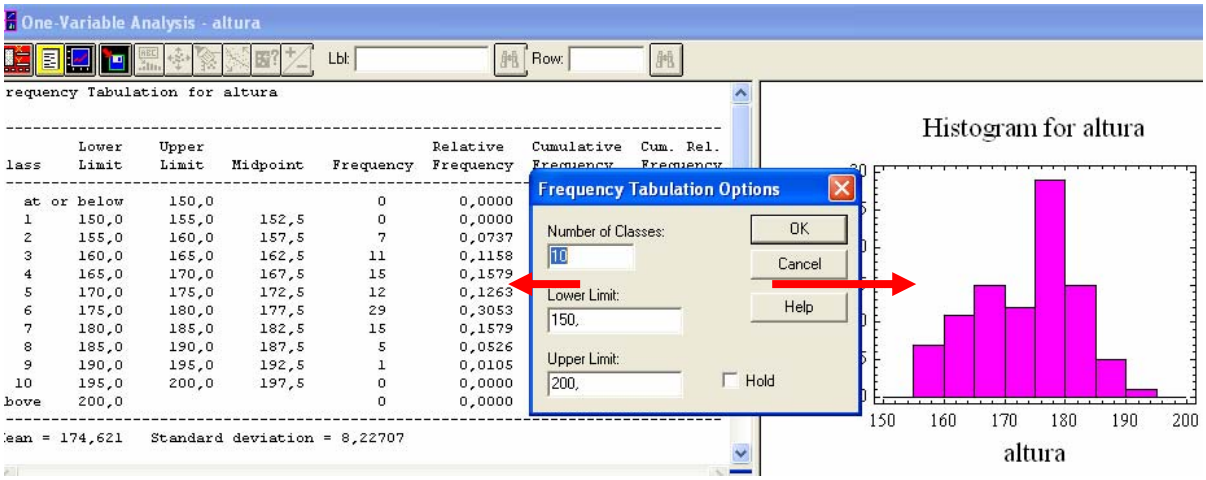
La distribución de las chicas es más uniforme. No es con forma de campana como la de los chicos. Tal vez sea porque hay menos datos (sólo 32) o porque las chicas de esa titulación sean realmente más heterogéneas.

3.3 Tabla de Frecuencias

La distribución de frecuencias mediante una tabla nos proporiocna 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 (Tabular Options)



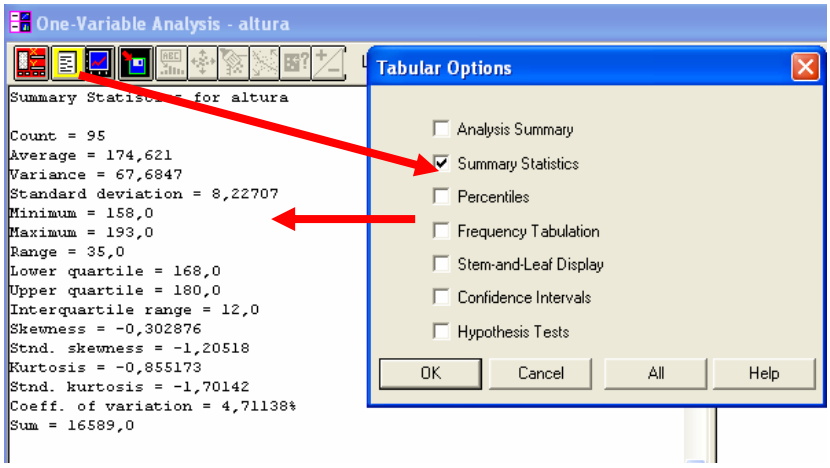
La tabla de frecuencias tienen las mismas opciones que el histograma. Podemos cambiar el número de clases o limitar el rango de valores. Para acceder a estas opciones nos colocamos sobre la ventana de resultados y pulsamos el botón derecho del ratón. Seleccionamos entonces Pane Options. En la ventana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



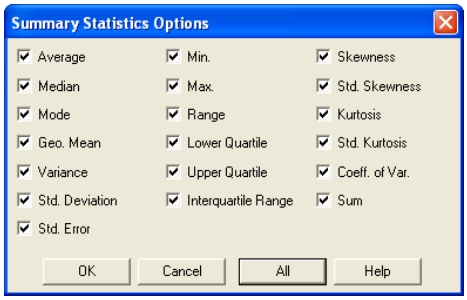
Esta tabla muestra que los dos intervalos modales son alrededor de los valores (midpoint), 167.5 y 177.5 y que el intervalo más frecuente, el centrado en 177.5 contiene a más del 30% de los alumnos.

3.4 Medidas características de variables cuantitativas

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



Podemos selecionar 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 selecionamos todos obtenemos los siguientes resultados:



```
Summary Statistics for altura

Count = 95
Average = 174,621
Median = 177,0
Mode = 180,0
Geometric mean = 174,427
Variance = 67,6847
Standard deviation = 8,22707
Standard error = 0,844079
Minimum = 158,0
Maximum = 193,0
Range = 35,0
Lower quartile = 168,0
Upper quartile = 180,0
Interquartile range = 12,0
Skewness = -0,302876
Stnd. skewness = -1,20518
Kurtosis = -0,855173
Stnd. kurtosis = -1,70142
Coeff. of variation = 4,71138%
Sum = 16589,0
```

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 está calculada con la expresión

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

en lugar de

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

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

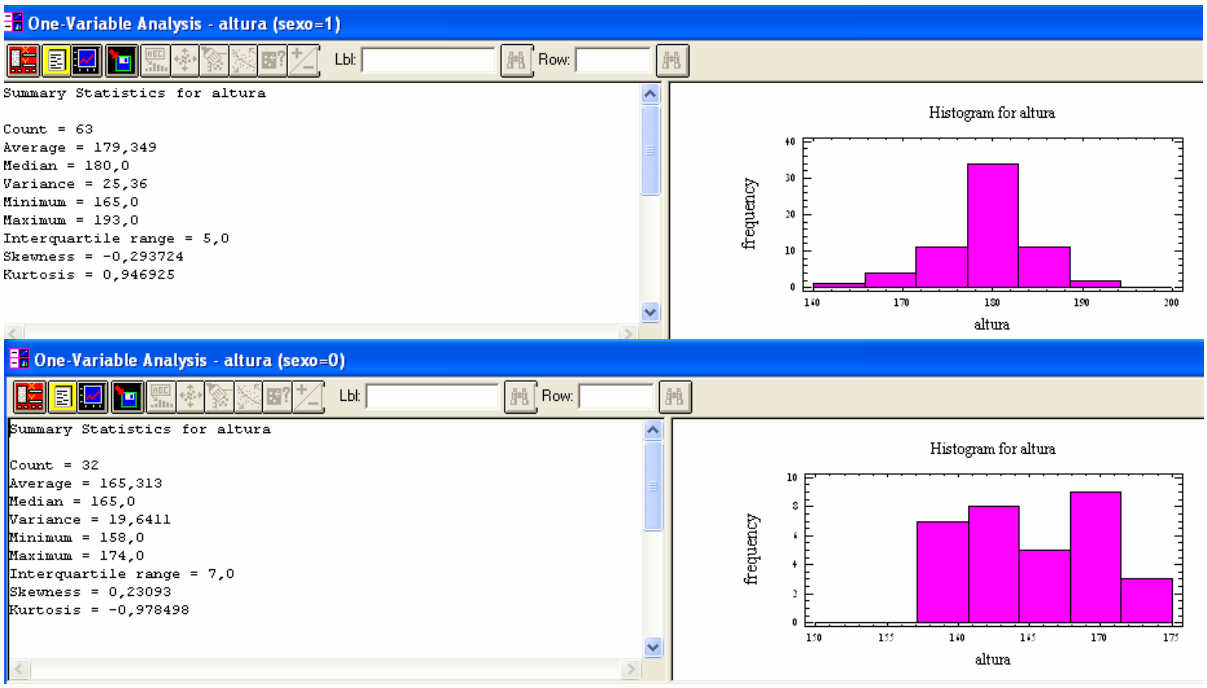
- La desviación típica usa también esta misma formulación, dividiéndose por n-1.
- Las siguientes medidas
  - Standard Error
  - Stnd. Skewness
  - Stnd. Kurtosis

  no son propiamente de estadística descriptiva, sino de inferencia. Por tanto no se cubren en este documento.
- El coeficiente de curtosis es

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

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

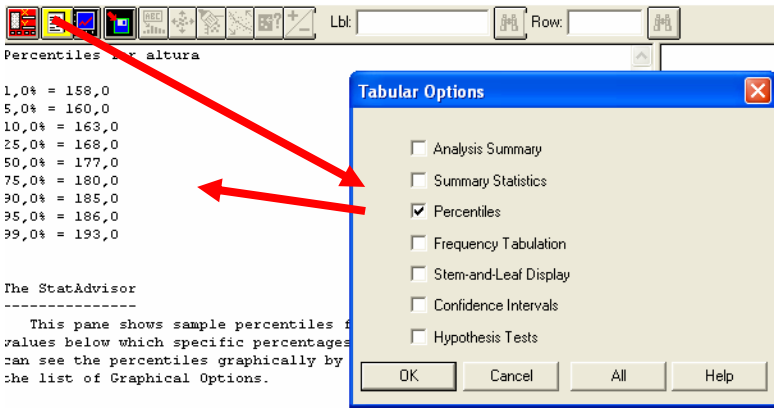
A continuación se muestra la comparación entre chicos y chicas mediante algunas medidas características



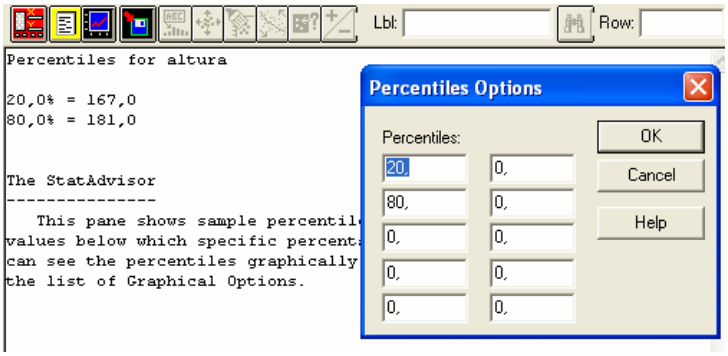
Puede verse ahora que los chicos son más altos que las chicas. La estaura media de los chicos es 179 mientras que para las chicas es sólo 165. Los chicos se concentran mucho alrededor de esa media. Puede verse que la media es casi igual que la moda, lo que se ve también fácilmente en la simetría del histograma y el bajo valor del coeficiente se 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 selecionamos Pane OIptions. 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.