Histogramas

Resumen

El **Histograma** ilustra la distribución de lo valores de una variable numérica agrupando los datos en intervalos y graficando barras en las cuales la altura es proporcional al numero de observaciones en cada grupo. Para una muestra relativamente grande, la grafica da una buena idea de la forma de la distribución de la cual los datos fueron muestreados. El usuario puede también elegir gráficos de frecuencias como de polígonos en lugar de grafico de barras.

Ejemplo StatFolio: *histogram.sgp*

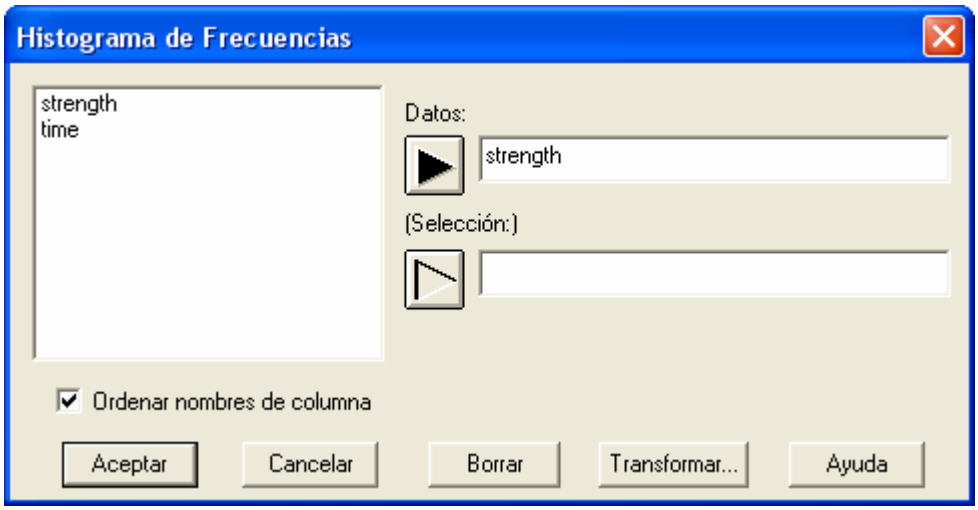
Datos del Ejemplo:

El archivo *bottles.sf3* contiene mediciones de la fuerza a la ruptura de 100 botellas de cristal, similar a un conjunto de datos contenido en Montgomery (2005). La tabla de abajo muestra una lista parcial de los datos de este archivo:

Strength (Fuerza)
255
232
282
260
255
233
240
255
254
259
235
262

Entrada de Datos

Los datos que son analizados consisten de de una sola columna numérica que contiene $n = 2$ o mas observaciones.



- **Datos:** Columna numérica que contiene los datos que serán analizados.
- **Selección:** Selección de un subconjunto de los datos.

Resumen del Análisis

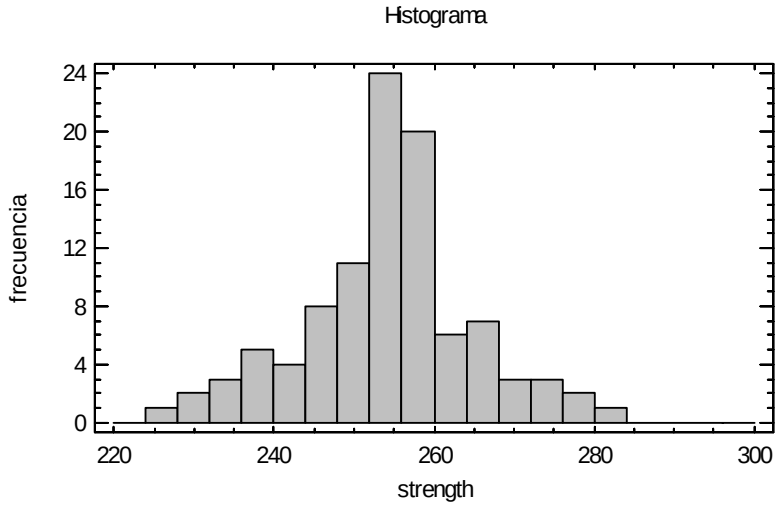
El *Resumen del Análisis* muestra el número de observaciones en la columna de datos.

Histograma - strength
Datos/Variable: strength
100 valores con rango desde 225.0 a 282.0

También son desplegados los valores más grande y más pequeño.

Histograma

Este panel muestra el histograma de frecuencias.



La grafica es construida de la siguiente manera:

- El rango de los datos es dividido entre *m* intervalos adyacentes que no se traslapan y de igual tamaño. El numero de barras por defecto es determinado usando la regla especificada en la sección *EDA* de la caja de dialogo de ***Preferencias*** del menú ***Edición***. Puedes incrementar el número de barras usando ***Opciones del Panel***.

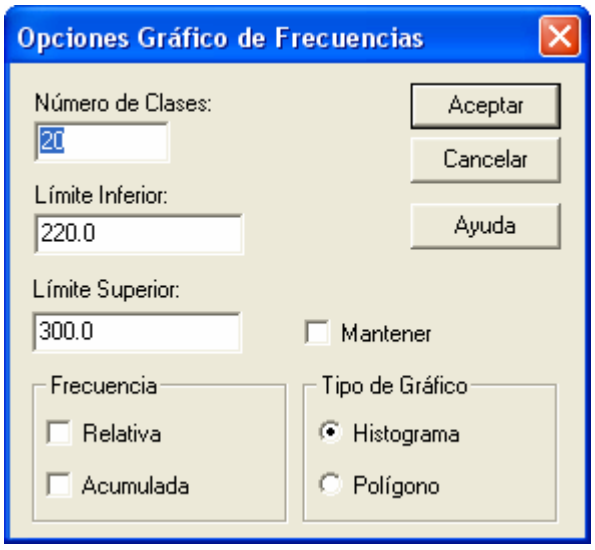
- El número de observaciones en cada intervalo es contado, una observación se considera estar en el intervalo si es mayor que el límite inferior de este intervalo y menor o igual que el límite superior.

- Por defecto, las barras son graficadas con alturas proporcionales al número de observaciones en cada intervalo. Otras opciones están disponibles en *Opciones del Panel*, como se describen más adelante.

La grafica anterior muestra una distribución relativamente simétrica con pico en la parte de en medio. Esta forma es típica de datos de una distribución normal.

Nota:
Esta grafica también es usada en el procedimiento *Análisis de una Variable*, junto con una tabla de frecuencias que muestra los conteos dentro de cada intervalo.

Opciones del Panel



- **Numero de clases**: El numero de intervalos entre los cuales los datos serán divididos. Intervalos adyacentes y de igual tamaño.

- **Limite Inferior**: Limite inferior del primer intervalo.

- **Limite Superior**: Limite superior del último intervalo.

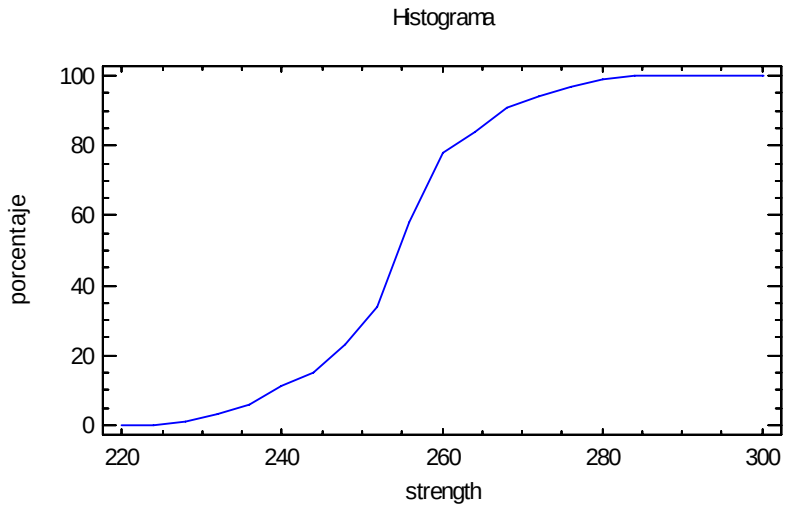
- **Mantener**: Mantiene el número seleccionado de intervalos y límites aunque la fuente de datos cambien. Por defecto, el número de clases y los límites son recalculados cuando los datos cambian. Esto es necesario ya que algunos datos pueden caer fuera de los límites originales.

- **Conteos**: Si se selecciona *Relativo*, la altura de las barras representan las observaciones en un solo intervalo. Si se selecciona *Acumulativo*, la altura representan las observaciones en el intervalo indicado y de todos los intervalos a la izquierda.

- **Tipo de Grafico**: Si se selecciona *Histograma*, las clases de frecuencias son desplegadas como un grafico de barras. Si se selecciona *Polígono*, las frecuencias son desplegadas usando un grafico de líneas conectadas.

Ejemplo – Polígono de Frecuencias Acumuladas

Fijando el *Tipo de Grafico* a *Polígono* y seleccionando la caja *Acumulativo* nos muestra un grafico de distribución acumulada de los datos:



La grafica anterior muestra el porcentaje de observaciones por debajo del límite superior de cada intervalo entre los cuales los datos fueron agrupados. Se puede ver que alrededor del 50% de los datos caen por debajo de 255 p.s.i.

Nota:
El numero de intervalos entre los cuales los datos son agrupados por defecto es fijado por la regla especificada en la sección de *EDA* de la caja de dialogo de *Preferencias* en el menú *Edición*. Cada regla determina el numero de intervalos *m* como una función del tamaño de muestra. Las reglas son:

Regla de Sturges: $m = ceiling(1 + 3.322\ log(n))$ (1)

10 log10(n): $m = ceiling(10\ log(n))$ (2)

Regla de Scott: $m = ceiling[\ (max-min) / (3.5\ s / n^{1/3})\ ]$ (3)

Regla de Freedman-Diaconis: $m = ceiling[\ (max-min) /(2.0\ IQR/ n^{1/3})\ ]$ (4)

Numero Fijo: m = *numero preespecificado* (5)

donde *min* es igual al valor mas pequeño en la muestra, *max* es igual al valor mas grande de los datos, *s* es igual a la desviación estándar muestral, *IQR* es igual al rango intercuantil muestral, y la función *ceiling* redondea al entero mas pequeño que es mayor o igual a su argumento. Puedes experimentar con diferentes reglas para determinar cual regla da un buen número de intervalos para tu tipo de datos.