Tutorial para el análisis exploratorio de datos univariados con el programa informático Statistica v.8

José Alberto Montoya-Márquez

Introducción

El primer paso en una investigación científica es observar, resumir y graficar los datos provenientes ya sea de un muestreo o de un experimento, con el fin de discernir el comportamiento general de las variables de estudio y observar el tipo de distribución de ellas. Esto se logra a través del llamado Análisis Exploratorio de Datos (del cuál Tukey fue el fundador), se trata de obtener los estadísticos descriptivos y gráficos que nos permitan observar el comportamiento de nuestros datos incluyendo la identifiación de puntos aberrantes.

Se pueden separar las técnicas de análisis exploratorio en dos grupos: Herramientas gráficas y herramientas numéricas. Entre las primeras se pueden mencionar, entre otras, a los histogramas, ojivas de frecuencia, diagramas de dispersión, diagramas de tallo y hoja, diagramas de caja y bigote.

Entre las herramientas numéricas se pueden considerar a las tablas de frecuencia, estimación de las medidas de tendencia central (media, moda, mediana, etc.), medidas de dispersión (varianza, desviación estándar, coeficiente de variación, etc), medidas de posición (cuartiles, octiles, percentiles, etc.). Entre estas herramientas, un análisis importante es la identificación de datos aberrantes o extremos (outlier en inglés) y que por definición son datos que no siguen la distribución del conjunto de valores; a este respecto una de las técnicas más sencillas es la propuesta por Tukey con base a los cuartiles, el intervalo intercuartílico, y los límites donde no hay puntos aberrantes (superior e inferior).

Estas técnicas han tenido un gran avance y mayor uso, principalmente por el desarrollo de las computadoras; en este sentido, en la actualidad se disponen de un conjunto de paquetes estadísticos que nos facilitan la tarea de calcular estas herramientas, un programa potente y amigable es el Statistica (StatSoft 2008).

Además de auxiliarnos en el análisis exploratorio de datos, el programa Statistica presenta una gran variaedad de rutinas que incluyen un sinnúmero de gráficos y pruebas estadísticas inferenciales, bi y multivariadas, haciendo de este programa una herramienta muy poderosa en la investigación científica y en la evaluación de recursos naturales.

El presente tutorial tiene como objetivo presentar, en una serie de pasos, la realización y exposición del análisis exploratorio de datos, así como la edición y presentación de sus resultados en el programa Microsoft Word. Esto incluye también la exportación de una base de datos de Microsoft Excel al programa Statistica.

El tutorial se divide en las siguientes partes: 1) presentación del problema, 2) importar un archivo de Microsoft Excel al programa Statistica, 3) cálculo de las medidas

Universidad del mar Campus Puerto Ángel, Oaxaca, Instituto de Recursos, Carretera a Zipolite, Ciudad Universitaria s/n, Distrito de San Pedro Pochutla C.P. 70902. México. Correo electrónico amontoya@angel.umar.mx

de tendencia central (MTC) de dispersión (MD), de posición (MP), intervalos de confianza (95%) para la media, 4) identificación de puntos aberrantes, 5) creación de un histograma, 6) edición de gráficos y tablas, 7) exportar tablas y gráficos de Statistica o Excel y/o Word.

Desarrollo y procedimiento

1. Presentación del problema

En un estudio llevado a cabo en la bahía de Puerto Ángel durante un año (datos ficticios), se registraron los datos que se presentan en la Tabla I.

Tabla I.- Datos de temperatura mensual en la bahía de Puerto Ángel (datos ficticios)

Temperatura
26.82
26.3
26.82
25.6
26.45
27.1
25.4
25.4
26.9
25.7
25.8
25.4

Con los datos de la Tabla I se desea calcular: medidas de tendencia central (MTC: media, mediana y moda), medidas de dispersión (MD: desviación estándar, varianza, recorrido, recorrido intercuartílico y coeficiente de variación), medidas de posición (MP: primer y tercer cuatiles, el sesgo y la curtosis), los intervalos de confianza de la media muestral (95%), identificar si hay puntos aberrantes y por último realizar el histograma de frecuencias con cinco intervalos y comenzando con el valor menor de los datos.

2. Cómo importar un archivo de Microsoft Excel al programa Statistica

El programa Statistica trabaja en hojas de cálculo (Spreadsheet) terminación .sta, hay varias maneras de crear una hoja de cálculo en el programa, una de ellas es copiar y pegar las variables (columnas) y los renglones (casos), sin embargo en el presente tutorial explicaré el procedimiento para importar de Excel al Statistica debido a que, generalmente se tienen bases de datos muy grandes y éstas son realizadas en Excel por su versatilidad y facilidad de manejo. Es importante que se incluya, en el primer renglón, el nombre de las variables, ya que Statistica tiene la opción de importar archivos considerando el nombre de las variables e inlcuso el de los renglones.

Una vez listo el archivo de temperatura en Excel, se debe guardar en la versión 97-2003.

Abrir el programa Statistica seleccionar File, en la barra de comandos, seleccionar Open, aparecerá una ventana de búsqueda de archivos; en el campo inferior desplegar las opciones de Tipo y seleccionar All files, buscar el archivo de Excel (terminación.xls), oprimir el botón Abrir (Fig. 1).

En la siguiente ventana seleccionar Import select sheet to a Spreadsheet y escoger Hoja 1 (donde están los datos en Excel), dar clic en OK. En la siguiente ventana se indican el número de renglones y de columnas, que son leídos automáticamente por el programa. Elegir Get variable names from first row (para el nombre de la variable) y clic en OK (Fig. 1).

A continuación se despliega la base de datos con el mismo nombre del archivo original con terminación.sta, el programa lo guarda en la misma carpeta del archivo fuente.

3. Procedimiento para obtener las MTC, MD, MP e intervalos de confianza de la media muestral (95%)

Activar en el menú principal el botón Statistics ir a Basic Statistics/Tables y dar clic en Descriptive statistics. A continuación se muestra la ventana del comando ejecutado, seleccionar la variable Temperatura. Ir

a la carpeta Advanced y elegir los estadísticos requeridos (MTC, MD, MP), oprimir el botón summary: statistics (Fig.2).

El programa guarda los resultados de una sesión en un libro de trabajo (Workbook) (Fig. 3), la tabla puede también copiarse y pegarse en Word o Excel para su edición.

4. Identificación de puntos aberrantes

Éste es un paso importante en el análisis descriptivo de los datos, pues estos valores sesgan la estimación de muchos estadísticos, lo cual puede afectar las inferencias que de ellos se obtengan y por ende las conclusiones que se deriven. Una de las formas de identificar estos valores es a través de la construcción del gráfico de caja y bigotes (Box and Whiskers Plot), considerando los cuartiles, el recorrido intercuartílico y los límites donde no hay puntos aberrantes. En el programa Statistica hacer lo siguiente:

Ir al menú principal y seleccionar Graphs, luego 2D graphs y Box Plots; se selecciona la variable haciendo clic en el botón de: Variables dentro de la ventana del módulo 2D Box Plots, la ventana de Grouping vairable debe ir vacía (Fig. 4). En la carpeta: Advanced ir a Box ecoger: Percentiles y en Coefficient: 25, en Whiskers: Non-outliers range, en Outliers: Out & Extremes y por último en Coefficient: 1.5 (todo esto lo selecciona el programa por default) (Fig. 5) dar clic en OK para obtener el gráfico (Fig.6).

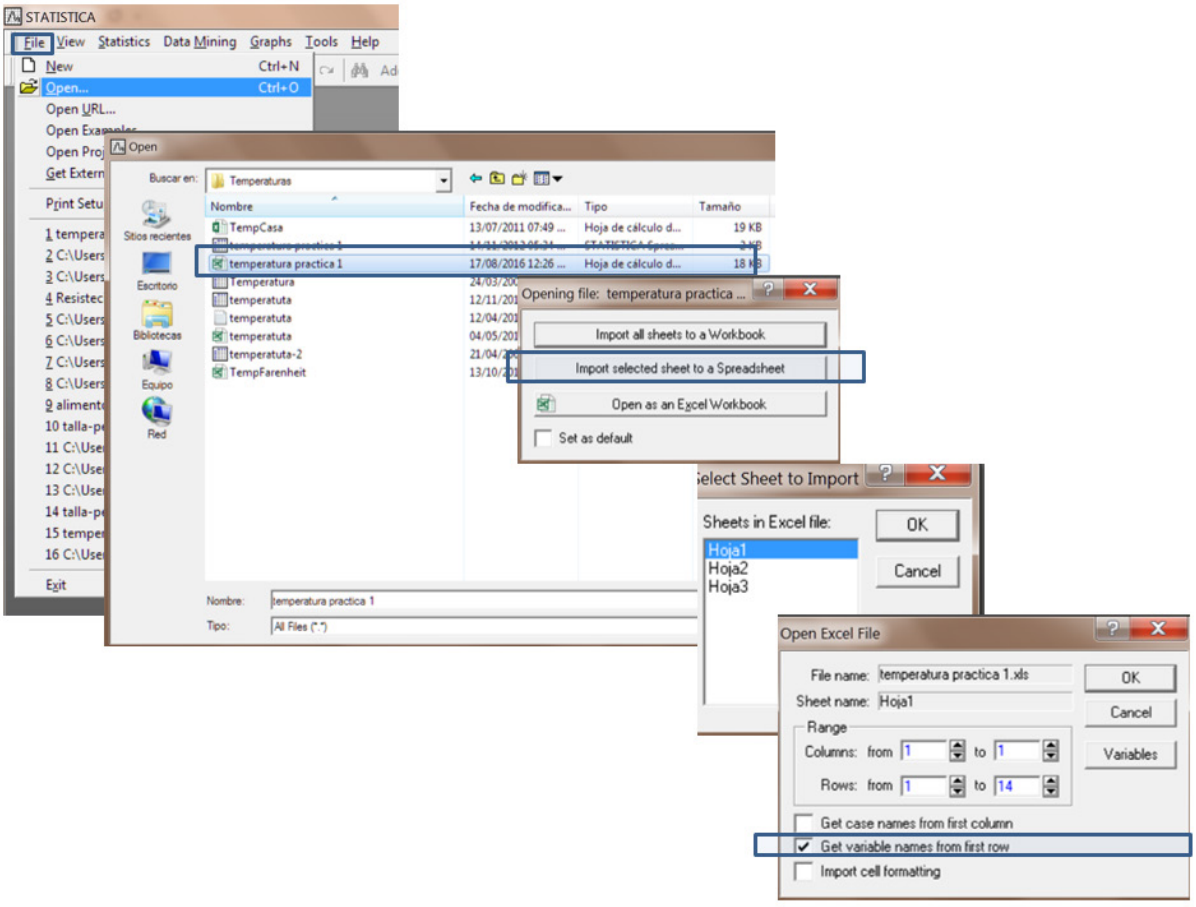


Figura 1.- Pasos para importar un archivo (base de datos) al programa Statistica(StatSoft 2008)

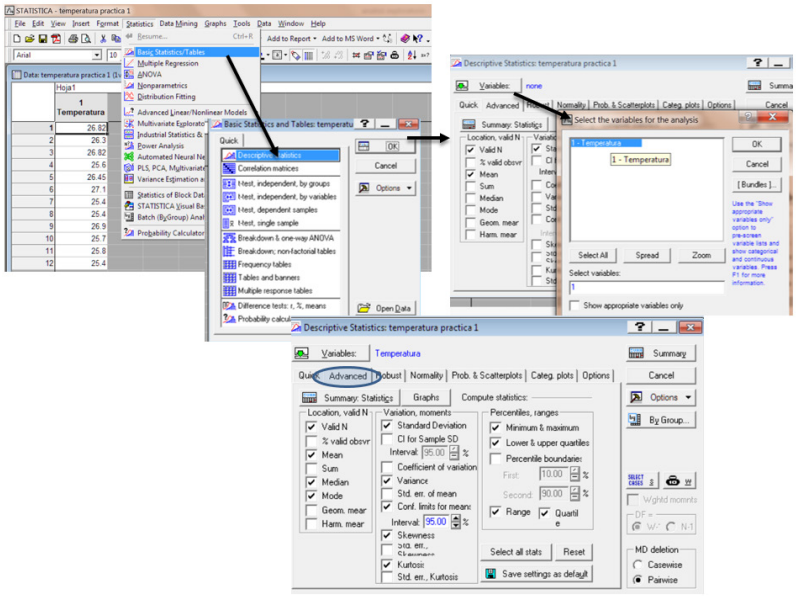
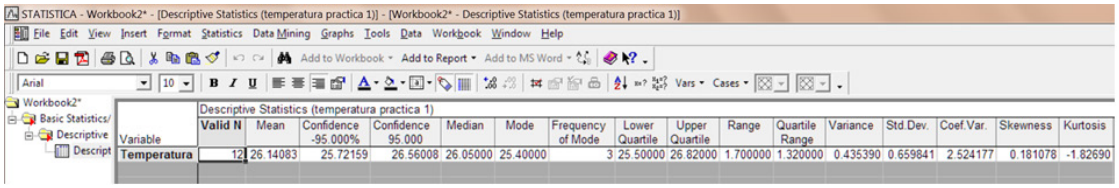


Figura 2.- Selección de estadísticos descriptivos



STATISTICA - Workbook2* - [Descriptive Statistics (temperatura practica 1)] - [Workbook2* - Descriptive Statistics (temperatura practica 1)]

Descriptive Statistics (temperatura practica 1)

Variable	Valid N	Mean	Confidence -95.000%	Confidence 95.000	Median	Mode	Frequency of Mode	Lower Quartile	Upper Quartile	Range	Quartile Range	Variance	Std.Dev.	Coef.Var.	Skewness	Kurtosis
Temperatura	12	26.14083	25.72159	26.56008	26.05000	25.40000	3	25.50000	26.82000	1.700000	1.320000	0.435390	0.659841	2.524177	0.181078	-1.82690

Figura 3.- Tabla de resultados en el libro de trabajo (Workbook)

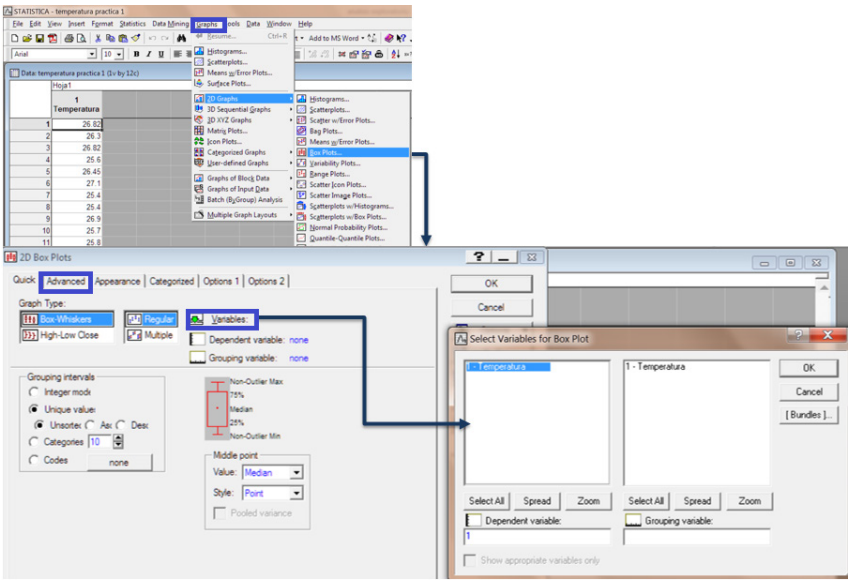


Figura 4.- Procedimiento para realizar un gráfico de caja y bigotes (Box &Whiskers Plot)

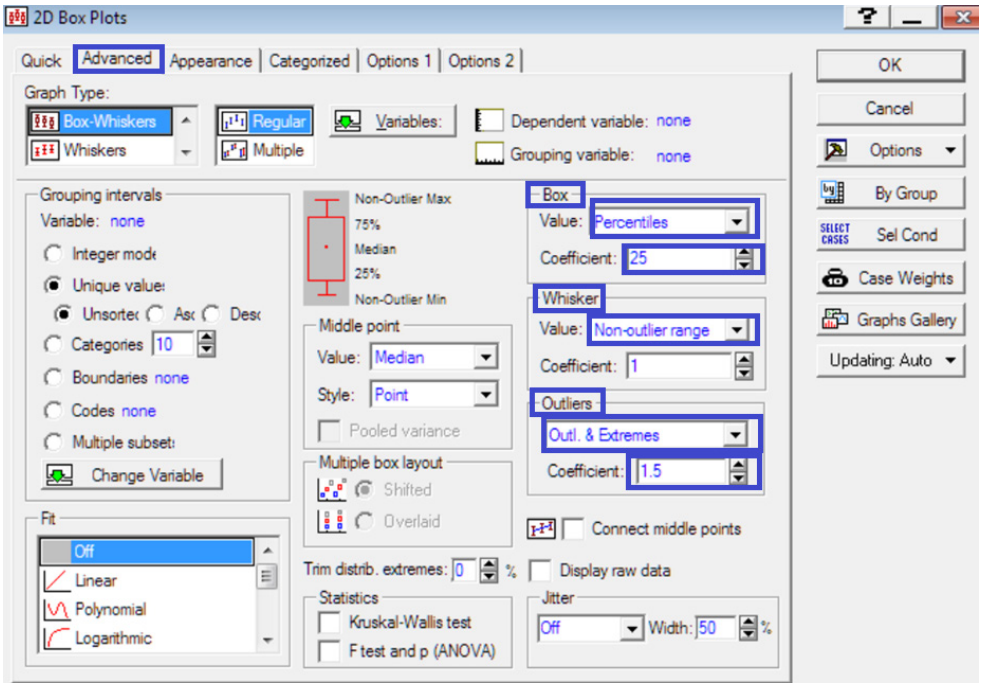


Figura 5.- Selección de las características del diagrama de caja y bigotes

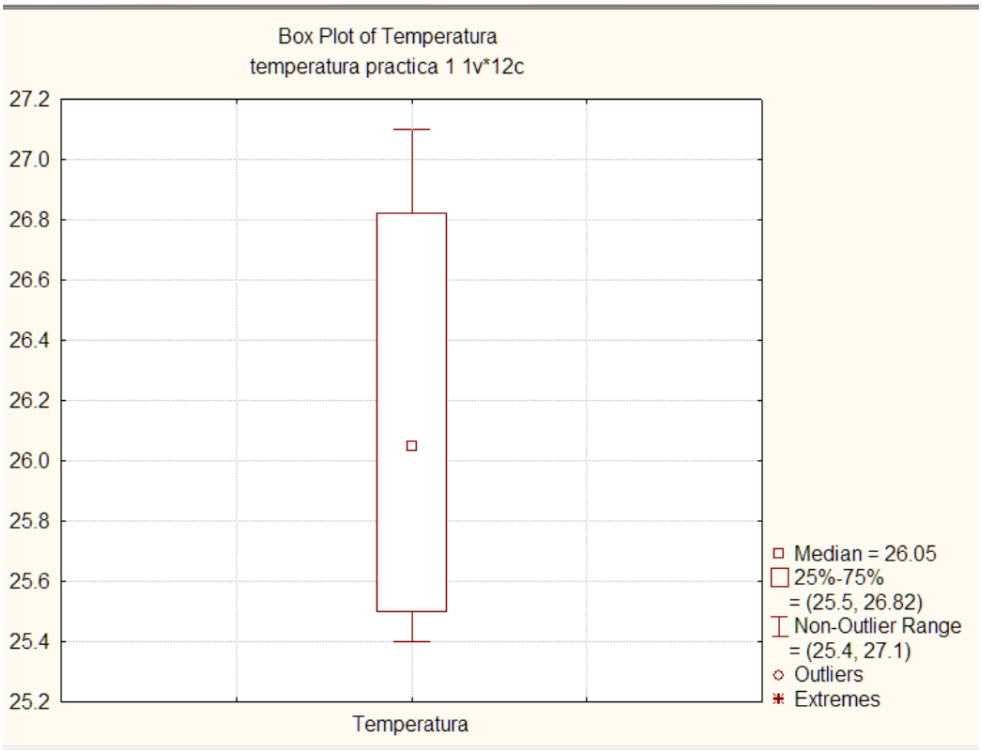


Figura 6.- Identificación de valores aberrantes con el gráfico de caja y bigotes, en este ejemplo no hay, en caso de que sí los hubiera se marcan con círculos vacíos y con asteriscos

5. Crear un histograma

Para realizar un histograma con cinco intervalos y el primero inciciando en el valor mínimo; ir de nuevo al menú de graficos (Graphs) seleccionar Histograms, entrar al módulo de 2D Histograms, ir a la carpeta Advanced seleccionar la variable en el botón del mismo nombre. En la opción de Fit type (ajuste de la distribución) escoger Off para que no tener ninguna curva ajustada y sólo observar la distribución de la variable de interés; en Categories seleccionar 5 (Fig. 7), y clic en OK para obtener el histograma (Fig. 8). Este histograma se puede copiar a Word.

6. Edición de tablas y gráficos

Ya sea en la elaboración de un informe o artículo, las tablas y gráficos se deben editar antes de su publicación. En el caso de las tablas recomiendo que se copien (inciso 7 de ese tutorial) y pegen en Excel y ahí editarlas antes de incluirlas en el documento de Word; en el caso de los gráficos se deben editar en el programa Statistica. Los pasos para editar un gráfico son los siguientes:

Considérese el gráfico de caja y bigotes del inciso tres. La edición de títulos y nombre de los ejes es con doble clic sobre el título o nombres que se deseen cambiar o eliminar (dando suprimir) (Fig. 9). Para cambiar el color del gráfico y otras opciones dar doble clic en cualquier área del gráfico, aparece la ventana All Options dentro de la cual se pueden hacer los cambios requeridos (Fig. 10).

7. Exportar tablas y gráficos de Statistica a Excel y/o Word

Las tablas se pueden copiar y pegar de manera sencilla: seleccionar la tabla y escoger copy with headers, pegar la tabla en Word o Excel. En el caso de un gráfico dar clic con el botón derecho del ratón sobre cualquier área del gráfico y seleccionar en copy graph (Fig. 11) abrir el documento de word y seleccionar el comando pegar, de esta manera se tendrá el gráfico en word.

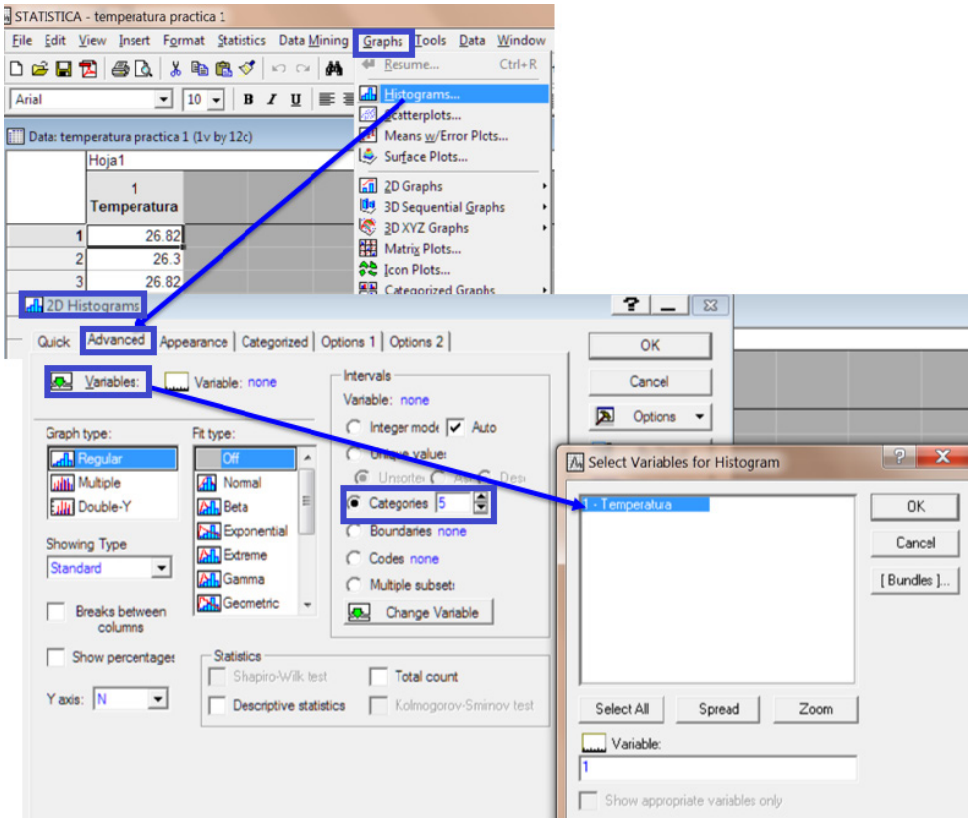


Figura 7.- Pasos a seguir para realizar un Histograma

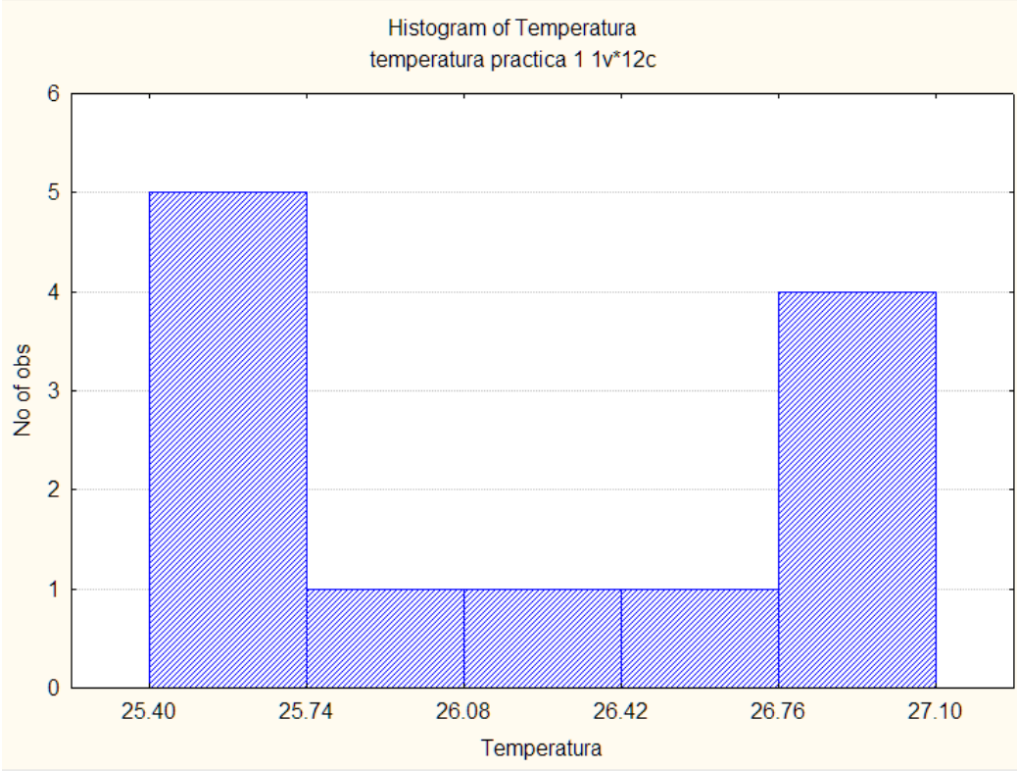


Figura 8.- Histograma con cinco intervalos comenzando a partir del valor mínimo (25.4)

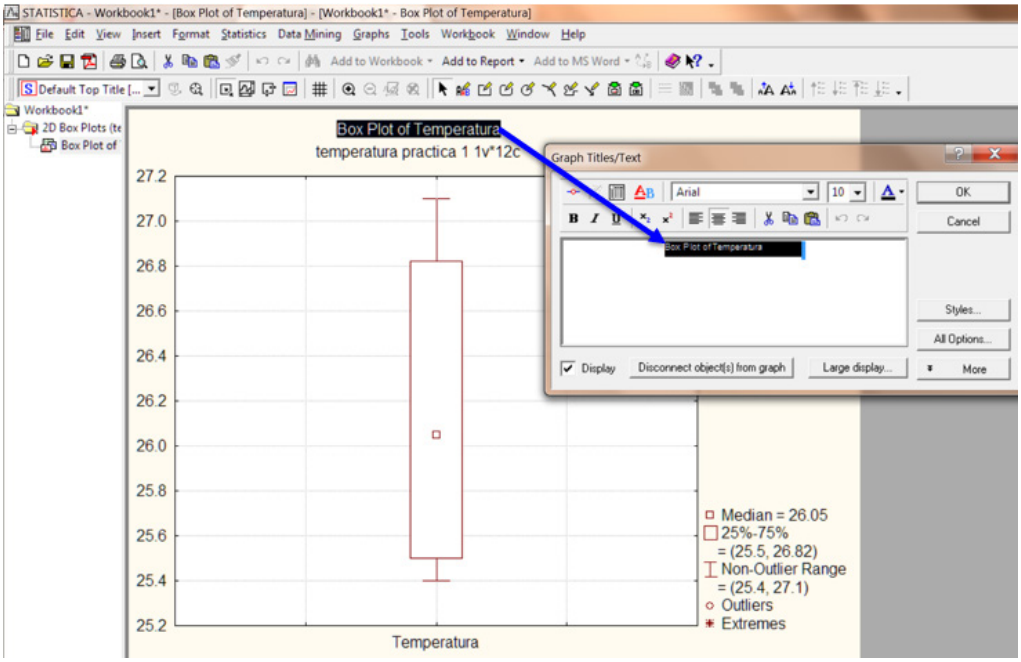


Figura 9.- Edición de títulos del gráfico de caja y bigotes

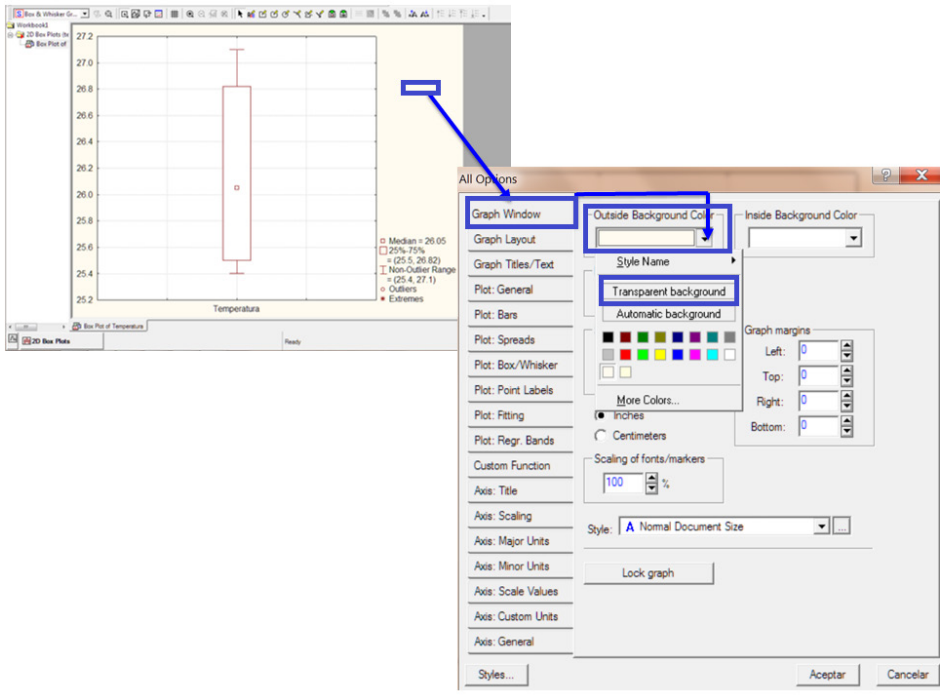


Figura 10.- Edición de color y otras opciones del gráfico de caja y bigotes

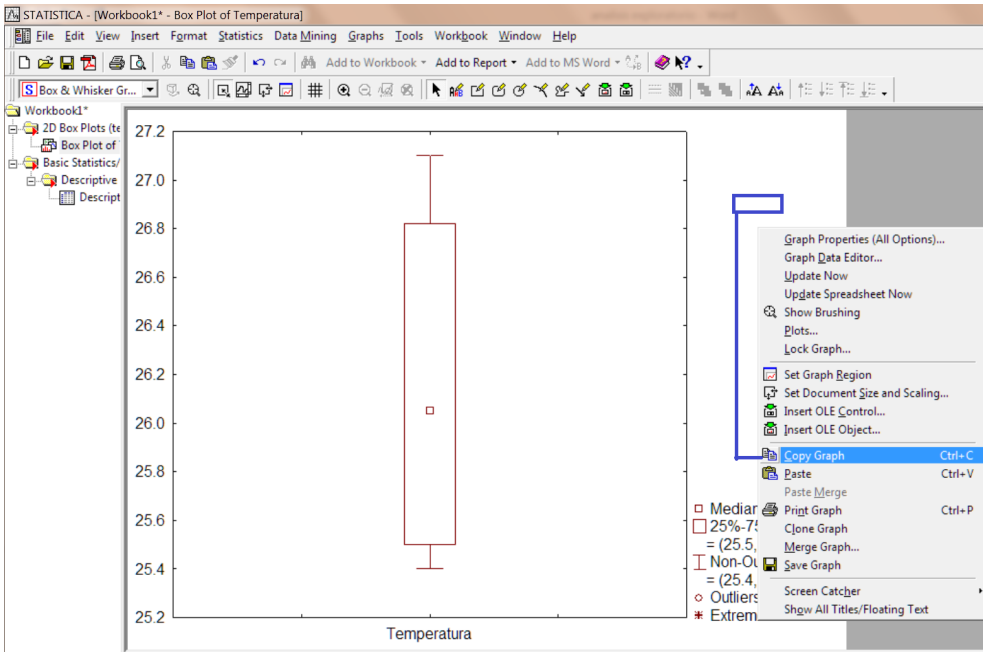


Figura 11.- Copiando un gráfico

Referencias

StatSoft, Inc. 2008. STATISTICA (data analysis software system), version 8.0. www.statsoft.com.

Recibido: 19 de agosto del 2016

Aceptado: 25 de agosto del 2016