UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Aplicación del Análisis de la Varianza para estudiar el tiempo de acceso en las aulas informáticas

Apellidos, nombre	Capilla Romá, Carmen[1] (ccapilla@eio.upv.es)
Departamento	[1]Estadística e Investigación Operativa Aplicadas y Calidad
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

En este artículo vamos a presentar la aplicación del Análisis de la Varianza (ANOVA) a datos del tiempo de acceso a ficheros pdf en las aulas informáticas donde se realizan las prácticas de Estadística. Forma parte de los objetos de aprendizaje que se pueden reutilizar en contextos educativos distintos.

Se trata de análisis que los alumnos pueden realizar en grupo o individualmente. Se utiliza el programa Statgraphics para realizar los cálculos y gráficos. Los datos se obtienen midiendo con el cronómetro del teléfono móvil el tiempo que se tarda desde la conexión de los equipos hasta que se puede visualizar el texto pdf del guión de la práctica. Se dispone de muestras en dos aulas distintas y en diferentes turnos y grupos, lo que permite realizar comparaciones de medias con el ANOVA.

2 Introducción

El ANOVA es la técnica de inferencia estadística más aplicada. Permite estudiar el efecto que uno o más factores tienen sobre la media de una variable aleatoria continua. Consiste en descomponer la variabilidad total observada en las componentes asociadas a los distintos efectos, y compararlas con la denominada variabilidad residual. La variabilidad residual se calcula como la diferencia entre cada dato y la media de la población en que está. Cuantifica mediante el cuadrado medio residual la varianza de cada población, que es resultado de todos los efectos no controlados.

Si un factor no tiene efecto, al calcular la posible varianza asociada a él mediante su cuadrado medio, éste no difiere significativamente del cuadrado medio residual. Si por el contrario hay efecto del factor sobre la media de la variable respuesta, su cuadrado medio es significativamente mayor que el residual. Para determinar esta cuestión se aplica un contraste de hipótesis con la distribución F que permite comparar varianzas.

En la aplicación que aquí se presenta, la variable respuesta es el tiempo de acceso en minutos. En una primera etapa de análisis se utilizan métodos de inferencia básica en una población normal, para estudiar los 19 valores de una muestra de dicho tiempo. En la siguiente etapa se aplica el ANOVA con una sólo factor a dos niveles: dos aulas informáticas distintas a distintos horarios. Se comparan las mediciones de las dos aulas y se aplican métodos gráficos para analizar los residuos. Finalmente, se utiliza el ANOVA para estudiar el efecto de dos factores: dos turnos de prácticas y dos grupos distintos. En este caso se incluye en los cálculos la interacción doble también.

3 Objetivos

Una vez que el alumno haya leído con detenimiento este documento y reproducido los análisis que plantea, será capaz de:

- Aplicar los conceptos relativos a las distribuciones en el muestreo, e inferencia básica en una población Normal con las opciones correspondientes de Statgraphics.

- Realizar los cálculos y gráficos necesarios con el software Statgraphics para el ANOVA con un factor y con dos factores.

4 Desarrollo

4.1 Inferencia en una población normal

En dos sesiones de aula informática se midió el tiempo (minutos) de acceso al fichero pdf disponible en red con el guión de la práctica. Los valores son:

3,43 4,07 3,77 3,35 3,37 4,88 5,05 4,5 3,65 4,02 3,55 3,02 3,38 3,87 3,5 3,85 4,8 3,87 3,6

Las fases de análisis de estos datos son:

- Introducir en una columna del editor Statgraphics los 19 valores (Figura 1).
- Estudiar si los datos siguen la distribución normal con la representación en papel probabilístico normal (Figura 2).
- Analizar si se puede admitir una media de 3,5 minutos en la población (Figuras 3, 4 y 5). Usar frases bien estructuradas y conectadas entre sí.
- Calcular los intervalos de confianza para la media y la desviación típica (Figuras 6 y 7).

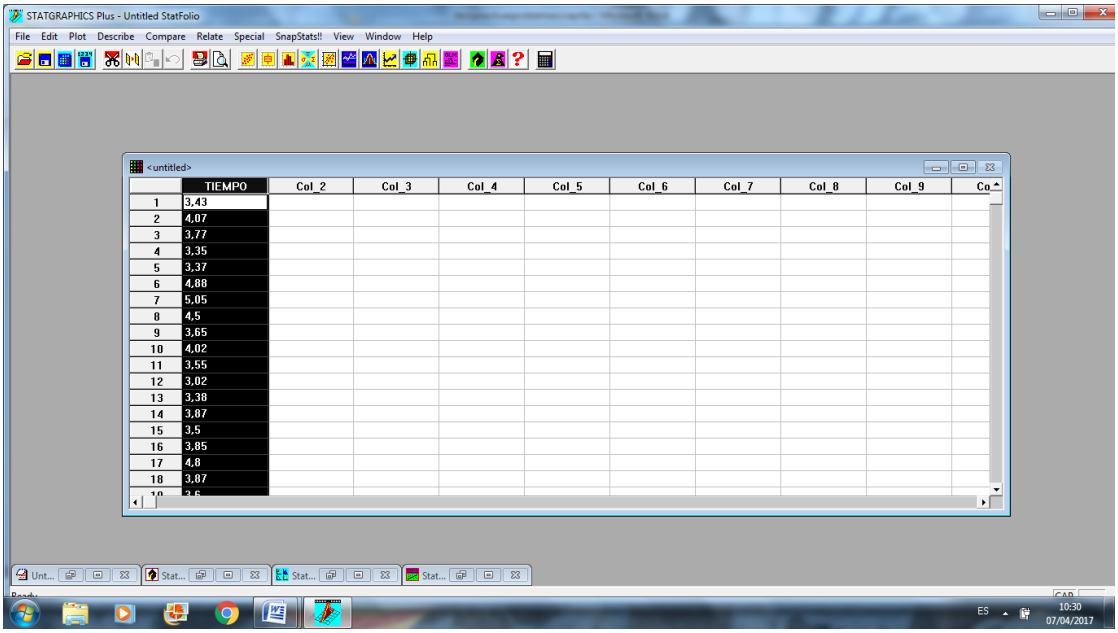

Figura 1. Datos de tiempo (minutos)

La Figura 1 muestra la columna con los 19 valores del tiempo. Esta es la variable respuesta, de tipo cuantitativo continua. La muestra son los 19 datos, medidos en la población de equipos disponibles en las aulas informáticas y que están con conexión a red. La Figura 2 representa estos valores en papel probabilístico normal.

Se ha obtenido con la opción **Describe...Numeric Data…One-Variable Analysis**. Dentro de ella, y con **Graphical Options**, se representa el **Normal Probability Plot.** Se aprecia que los datos están próximos a una recta, por lo que se puede admitir distribución normal.

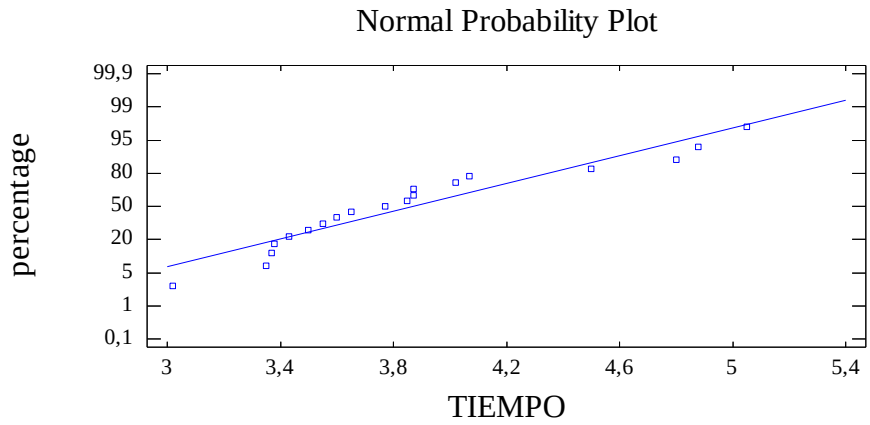

Figura 2. Representación en papel probabilístico normal

Para estudiar si se puede admitir la hipótesis de tiempo medio en la población igual a 3,5 minutos, se plantea el contraste:

$$H_0 : m=3,5$$

$$H_1 : m\neq 3,5$$

Para realizar este contraste de hipótesis con Statgraphics, en la opción de menú **Describe…Numeric Data…One-Variable Analysis**, se selecciona con **Tabular Options**  **Hypothesis Test** (contraste de hipótesis, Figura 3)

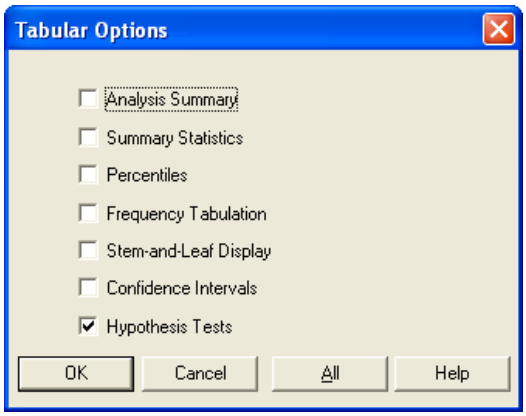

Figura 3. Selección de contraste de hipótesis

Para introducir los datos del contraste, se selecciona **Pane Options** (Figura 4), con el botón derecho del ratón sobre la ventana abierta. En el cuadro de diálogo que aparecerá, se especifica la siguiente información:

- **Mean:** Valor de la media de la población a contrastar en la hipótesis nula (3,5).
- **Alpha:** riesgo de 1ª especie. Se toma inicial el valor 0,05.

- **Alt. Hypothesis:** Si se rechaza la hipótesis nula m=3,5, se acepta la hipótesis alternativa que puede ser: **Not Equal** (m ≠ 3,5), **Less Than** (m < 3,5) o **Greater Than** (m > 3,5)

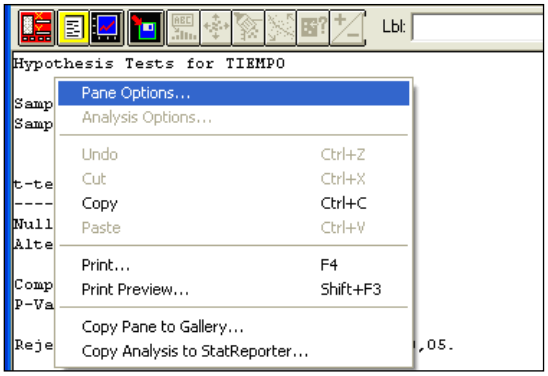

*Figura 4. Venta para seleccionar **Pane Options***

Tras pulsar **OK**, Statgraphics calcula la información relativa al **Test-t** (Figura 5). Se muestran los valores de m_0 (`Null hypothesis: mean = 3,5`) y α (`alpha = 0,05`) introducidos previamente. Los cálculos resultantes son el valor de la **t calculada** (**Computed t statistic**) y el **p-valor** (**P-Value**).

```
t-test
------
Null hypothesis: mean = 3,5
Alternative: not equal

Computed t statistic = 2,84502
P-Value = 0,0107462
Reject the null hypothesis for alpha = 0,05
```

Figura 5. Resultados del contraste de hipótesis

Se observa que se rechaza la media 3,5 minutos. Sin embargo si el riesgo de primera especie alfa fuese del 1%, se aceptaría ya que **p-valor** > 0,01.

Otra manera de obtener conclusiones sobre la población, es con los intervalos de confianza. Para obtenerlos con Statgraphics, en la ventana de resultados del análisis de una variable (**Describe…Numeric Data…One-Variable Analysis**), con **Tabular Options**  , se selecciona **Confidence Intervals** (Intervalos de Confianza, Figura 6). Con clic en el botón derecho sobre el panel de intervalos de confianza y **Pane Options**, se puede cambiar el nivel de confianza 1-a (**Confidence Level**) de los intervalos. La Figura 7 muestra el resultado con un nivel de confianza del 90%.

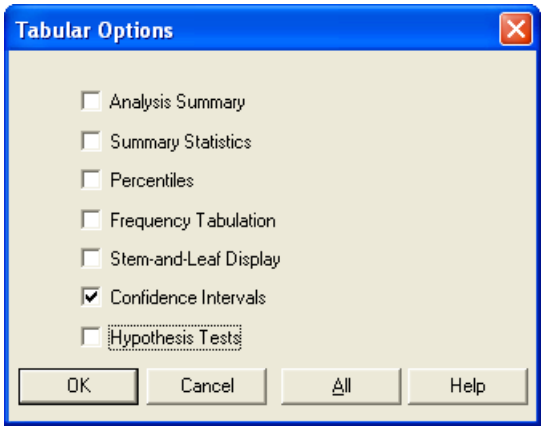

Figura 6. Selección de intervalos de confianza

```
Confidence Intervals for tiempo
------------------------------
90,0% confidence interval for mean: 3,87 +/- 0,225518
[3,64448;4,09552]
90,0% confidence interval for standard deviation:
[0,447622;0,784848]
```

Figura 7. Intervalos de confianza para la media y la desviación típica

Con un riesgo de 1ª especie a=10%, ¿difiere la media significativamente de 4 minutos? La respuesta es que no pues el valor 4 está dentro del intervalo de confianza de la media. Tomando el mismo nivel de confianza de la pregunta anterior, ¿es admisible la hipótesis nula de que la desviación típica de la población es σ=1 minuto? No es admisible pues el valor 1 no está en el intervalo de confianza de la desviación típica

4.2 ANOVA con un factor

En dos sesiones de prácticas de días diferentes, se midió en los laboratorios A y B, el tiempo de acceso (minutos) al fichero pdf del guión disponible red:

Lab A, hora 9:30 a 11

4,13 3,07 4,1 4,02 8,32 5,02 5,1 4,33

Lab B, hora 11:30 a 13

6,1 7,25 8,23 6,18 8,47 10,08 9,82 9,8 12 22,28

Las etapas de este análisis son:

- Introducir los datos en el editor Statgraphics, con una columna para los tiempos y otra para los laboratorios-hora.

- Estudiar con el ANOVA (Figura 8) si la media del tiempo difiere significativamente entre los dos laboratorios (riesgo de 1ª especie a=5%). La opción es **Compare….Analysis of Variance….Multifactor ANOVA**.
- El tiempo medio de acceso, ¿en qué laboratorio es significativamente mayor? Utilizar los intervalos LSD (Figura 9) para contestar esta pregunta.
- Analizar los residuos con un gráfico (Figura 10). Con **Graphical Options**  seleccionar: **Residuals versus Factor Levels**.

La tabla de ANOVA resulta:

Analysis of Variance for tiempo_acceso - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:laboratorio	122,955	1	122,955	9,15	0,0080
RESIDUAL	214,989	16	13,4368		
TOTAL (CORRECTED)	337,945	17			

All F-ratios are based on the residual mean square error.

Figura 8. Tabla del ANOVA

Se observa que el P-Value=0,008 < alfa=0,05. Por tanto el tiempo medio de acceso difiere significativamente entre las dos situaciones en que se ha medido (laboratorio-hora). Para estudiar con más detalle esta cuestión, se representan los intervalos LSD (Figura 9). En dicho gráfico se aprecia que el tiempo medio es significativamente mayor en el laboratorio B.

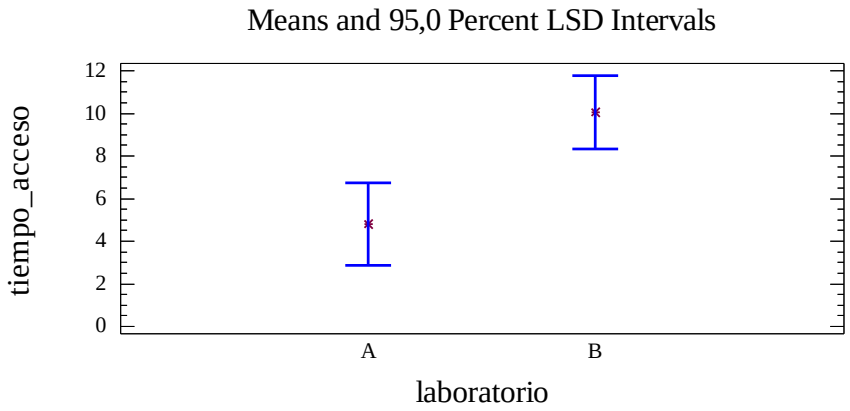

Figura 9. Intervalos de comparación de medias

Finalmente se analizan los residuos. Se calculan como la diferencia entre cada observación y la media de la muestra en que está. Hay en total 18 residuos.

Representados frente al factor se obtiene el gráfico de la Figura 10. Se aprecia que hay una observación anómala en el laboratorio B, que corresponde al valor 22,28.

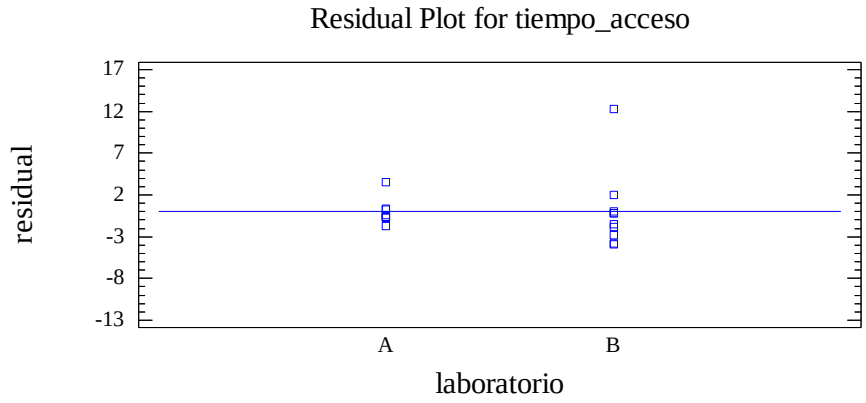

Figura 10. Diagrama de dispersión de los residuos frente al factor

4.3 ANOVA con dos factores

En este caso se analizan los tiempos de acceso al fichero pdf del guión de la práctica en red (**tiempoacc,** minutos), observados en dos turnos de prácticas (**turno**), de dos grupos A y B (**grupo**), en un aula informática.

Grupo A Turno 1: 3,25 3,83 4,4 3,18 3,88 4,52 3,37 3,13

Turno 2: 4 3,87 4,03 4 4,5 4,33 5,38 4,43

Grupo B Turno 1: 3,65 4,02 3,55 3,02 3,38 3,87 3,5 3,85

Turno 2: 3,43 4,07 3,77 3,35 3,37 4,88 5,05 4,5

Las etapas de este caso son:

- Introducir los datos en el editor Statgraphics, con una columna para los tiempos, otra para el grupo y una tercera para el turno.
- Estudiar con el ANOVA (alfa=5%), si hay diferencias significativas en el tiempo medio entre los dos grupos y entre turnos. Analizar también la interacción entre estos dos factores. Para realizar el ANOVA se utiliza la opción Statgraphics **Compare….Analysis of Variance….Multifactor ANOVA** (Figura 11). Se incluye la interacción con **Analysis Options** (botón derecho ratón, Figura 12)
- Representar los gráficos LSD para los efectos que han resultado significativos en el ANOVA (Figura 14).

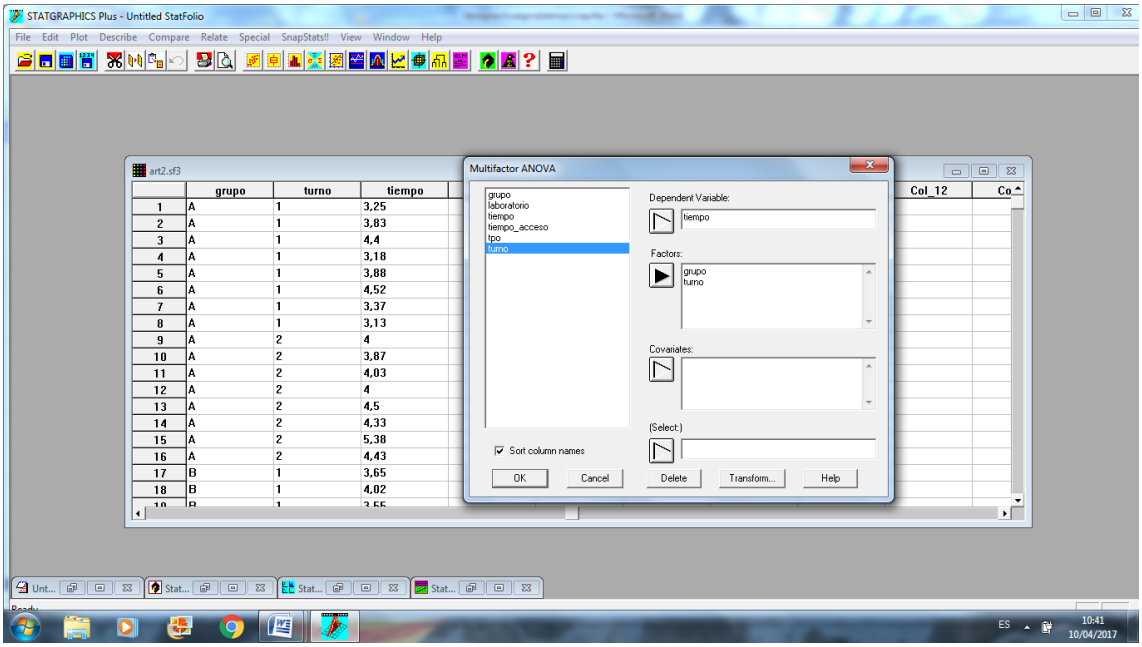

Figura 11. Opción del ANOVA para dos factores

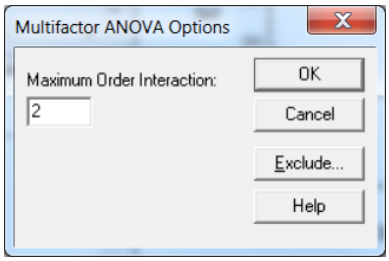

Figura 12. Opción para incluir la interacción doble

La Figura 13 recoge la salida Statgraphics con los cálculos del ANOVA: Factor grupo, p-value=0,3495 > alfa=0,05 ⇒ No hay diferencia en el tiempo medio entre grupos (nivel de confianza 95%). Factor turno, p-value=0,0078 < 0,05 ⇒ Hay diferencia significativa en el tiempo medio entre el primer y segundo turno de prácticas (nivel de confianza 95%). Interacción doble, p-value=0,6427 > 0,05 ⇒ No hay interacción significativa (nivel de confianza 95%)

Analysis of Variance for tiempo - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:grupo	0,25205	1	0,25205	0,91	0,3495
B:turno	2,2898	1	2,2898	8,22	0,0078
INTERACTIONS					
AB	0,06125	1	0,06125	0,22	0,6427
RESIDUAL	7,7961	28	0,278432		
TOTAL (CORRECTED)	10,3992	31			

Figura 13. Cálculos del ANOVA con Statgraphics

El gráfico LSD para el efecto de turno (Figura 14), indica que el tiempo medio de acceso es significativamente menor en el segundo turno de prácticas:

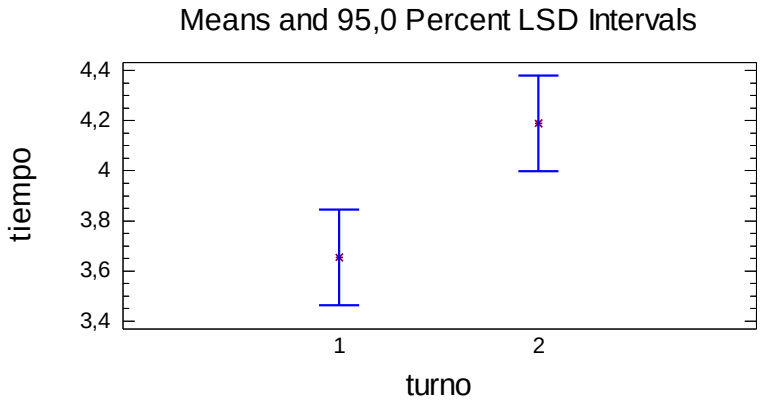

Figura 14. Intervalos de comparación de medias para el factor turno

5 Cierre

A lo largo de este objeto de aprendizaje hemos visto cómo aplicar diversos métodos de inferencia para analizar el tiempo de acceso a un fichero en red en las aulas informáticas.

Para comprobar qué realmente has aprendido los procedimientos se recomienda reproducir los análisis que se han presentado.

6 Bibliografía

Peña, D: "Estadística modelos y métodos. Vol 1 y 2", Ed. Alianza Universidad, 1995.

Romero Villafranca, R.; Zúnica Ramajo, L: "Métodos estadísticos para ingenieros", Ed. Universidad Politécnica de Valencia, 2013.

Stagraphics Plus 5: "User Manual", Ed. Manugistics Inc., Maryland, USA, 2000.