PRÁCTICA 8. ANÁLISIS DE LA VARIANZA

Objetivo

El objeto de la presente sesión de práctica informática es complementar y afianzar los conceptos relativos a la técnica de Análisis de la Varianza (ANOVA) vistos en clase (apartado 3 de la UD5). Asimismo se pretende que el alumno se familiarice con las opciones que el programa Statgraphics ofrece al respecto.

NOTA: Se recomienda que, durante el trabajo no presencial del alumno, los resultados obtenidos a partir del Statgraphics en esta práctica se calculen "a mano" y se cotejen con los mismos.

Ejemplo

Una planta clasificadora de residuos sólidos dispone de tres líneas diferentes de clasificación. Dichas líneas fueron fabricadas por compañías diferentes y en momentos distintos, por ello la tecnología de la cámara que se utiliza en cada una de las líneas, para la obtención de las imágenes de los objetos que pasan, es diferente. La línea 3 está dotada de una cámara espectral, la línea 2 utiliza una cámara de infrarrojos y la línea 1 una cámara normal.

Las imágenes obtenidas por cada una de las líneas son los datos de entrada a un programa que aplica técnicas de reconocimiento de patrones para realizar la clasificación del objeto. Dicho programa analiza la imagen de entrada, aplica un modelo y da como salida una clasificación del tipo de objeto que está pasando por la línea en ese instante. Tras esta clasificación un brazo actuador se encarga de ubicar el objeto en el lugar idóneo.

El resultado de la clasificación depende de la calidad de la imagen de entrada que viene determinada por el tipo de cámara y por tanto de la **línea**. Asimismo depende de un parámetro de gran importancia (λ) que interviene en el modelo.

El objetivo del estudio que se plantea en esta práctica es evaluar los aciertos en la clasificación final en función del parámetro λ (para el que se han probado 3 valores) y de la línea por la que ha transitado el objeto. Para ello, para cada una de las 9 combinaciones posibles se ha probado dos veces el clasificador durante una hora, obteniéndose la proporción de errores en la clasificación. Los resultados se recogen en la siguiente tabla en la que un resultado mayor indica mayor error (los resultados han sido debidamente transformados, puesto que se trata de proporciones).

	λ = 0,2	λ = 0,5	λ = 0,8
LINEA = 1	33,21 31,95	25,84 23,57	25,10 22,79
LINEA = 2	22,79 21,97	21,97 20,27	21,13 19,37
LINEA = 3	20,26 21,97	19,37 18,43	18,43 17,46

A la vista de los datos, se pide:

1) ¿Cuál es la variable respuesta?, ¿cuáles son los factores?, indica de qué tipo son.

2) Indica en qué consiste un tratamiento en este experimento y de qué Plan Factorial se trata.

3) Introduce los datos en Statgraphics. (Fichero "Clasificadora.sf3" disponible en

PoliformaT: Recursos / 05 | Prácticas / 032 | Ficheros de datos)

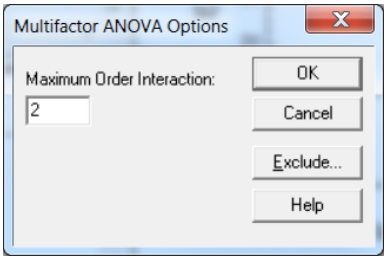
<untitled>

	Linea	Lamda	Errores	Col_4
1	1	0,2	33,21	
2	1	0,2	31,95	
3	2	0,2	22,79	
4	2	0,2	21,97	
5	3	0,2	20,26	
6	3	0,2	21,97	
7	1	0,5	25,84	
8	1	0,5	23,57	
9	2	0,5	21,97	
10	2	0,5	20,27	
11	3	0,5	19,37	
12	3	0,5	18,43	
13	1	0,8	25,1	
14	1	0,8	22,79	
15	2	0,8	21,13	
16	2	0,8	19,37	
17	3	0,8	18,43	
18	3	0,8	17,46	
19				
20				

4) Realiza un ANOVA para determinar si la LINEA, el parámetro λ o su interacción tienen algún efecto sobre la proporción de aciertos (α=0,05).

[*Compare → Analysis of Variance → Multifactor ANOVA*]

Para indicar que se incluya la interacción [*Botón derecho → Analysys Options*]



5) Analiza el efecto del factor LINEA mediante los intervalos LSD.

6) Estudia la naturaleza del factor λ a nivel descriptivo mediante los intervalos LSD o gráfico de medias. ¿Existen indicios de una posible relación lineal o cuadrática entre los aciertos y el factor λ?

7) Interpreta gráficamente la interacción que haya resultado significativa.

8) ¿Qué tipo de línea y qué valor de λ serían los óptimos con el fin de obtener, en promedio, el mínimo número de errores en el clasificador?.

Respuestas a las preguntas propuestas

Pregunta 1

- La **variable respuesta** (o variable dependiente) es la **proporción de errores** (debidamente transformada)

- Los **factores** son:

  - el tipo de **línea**: **cualitativo**
  - el parámetro λ: **cuantitativo**

Pregunta 2

- Un **tratamiento** consiste en una combinación de **línea** y **lamda** (λ)

- **Plan Factorial**:

  - Con dos **factores**:
    - el tipo de **línea** con 3 **variantes** (1, 2 y 3)
    - el parámetro λ con 3 **niveles** (0,2, 0,5 y 0,8)
  - 9 **tratamientos** distintos
  - 2 **pruebas** para cada tratamiento (**equilibrado**)
  - Cada **prueba** consiste en probar el clasificador durante una hora para cada uno de los 9 tratamientos y anotar la proporción de errores (variable repuesta)

Pregunta 4

La **Tabla Resumen del ANOVA** que se obtiene se muestra a continuación:

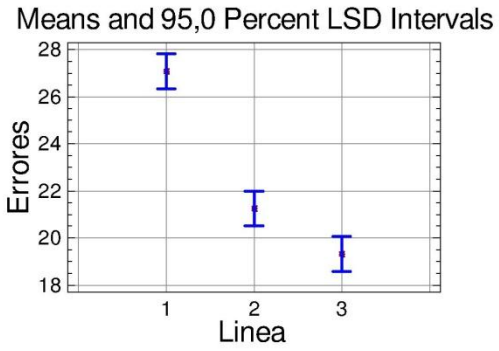
```
Analysis of Variance for Errores - Type III Sums of Squares
--------------------------------------------------------------------------------
Source        Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:Lamda          73,2642 SCF1    2    36,6321          28,08     0,0001
 B:Linea         195,682  SCF2    2    97,8408          74,99     0,0000

INTERACTIONS      33,3391 SCF1xF2 4     8,33478          6,39     0,0102

RESIDUAL          11,7426 SCR     9     1,30473
--------------------------------------------------------------------------------
TOTALCORRECTED) 314,028   SCT     17
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Source	Origen de la variación
Sum of Squares	Suma de cuadrados
Df	Grados de Libertad
Mean Square	Cuadrado Medio
F-Ratio	F-ratio
P-Value	P-valor

Como *p-value* es menor que 0,05 (a), podemos concluir que, para este nivel de significación, los efectos de los dos factores (Línea y Lamda), así como el de su interacción, han resultado significativos (efecto sobre el promedio de errores del clasificador).

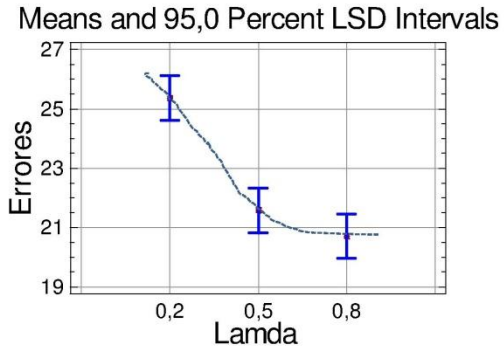
Sin el *p-value*, se tendría que comparar el F-ratio con los respectivos valores críticos buscados en la Tabla F.

Pregunta 5



A la vista del anterior gráfico con los **Intervalos LSD** se puede concluir que, con un nivel de confianza del 95%, el promedio de errores del clasificador es significativamente distinto entre las tres líneas (los respectivos intervalos LSD NO se solapan). El promedio de errores más alto se obtiene al utilizar la línea 1, el más bajo, usando la línea 3.

Pregunta 6

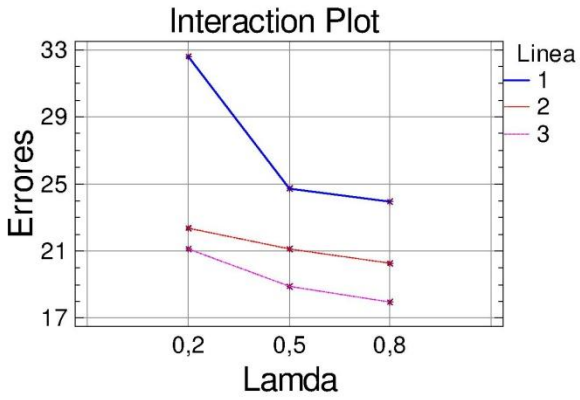


En este caso, como se trata de un factor cuantitativo, no tiene sentido concluir simplemente que el promedio de errores es peor para el valor 0,2 del parámetro λ, por lo que hay que analizar el "tipo" de relación entre λ y la variable respuesta.

A nivel puramente descriptivo podríamos decir que a mayores valores del parámetro λ, menor es el promedio de errores que se obtiene durante la clasificación de residuos, pero la disminución de los errores no parece ser lineal .

Por tanto, podríamos decir que la relación entre el valor de lamda y el promedio de errores es inversamente proporcional y con indicios de que dicha relación sea de naturaleza cuadrática.

Pregunta 7



Estudiando el gráfico anterior, podemos decir que el comportamiento (relación) entre el parámetro lamda y el promedio de errores del clasificador es similar para las líneas 2 y 3 (inversamente proporcional y de carácter más o menos lineal). Sin embargo, este comportamiento es diferente para la línea 1, a mayores valores de lamba también se producen menos errores en promedio, pero no de modo lineal.

Pregunta 8

El **mínimo número de errores (promedio)** en el clasificador se obtiene utilizando la **línea 3**, dotada con cámara espectral, y los valores 0,5 o 0,8 del parámetro lamda.

Para elegir entre estos dos valores deberíamos atender a otros criterios (económicos, técnicos, etc) diferentes de los estadísticos.