

## Exámen de Análisis de Datos (E58). Febrero 2002

### Problema 1 [2.5 puntos]. Inspección de circuitos.

En un experimento se investigan tres procedimientos para la inspección de circuitos. Los tres procedimientos se basaron en tres condiciones llamadas standard (1), overlay (2) y alternating (3). Por otra parte, tres tipos de defectos ocurrieron en los circuitos fill (1), gap (2), short (3). El experimento se llevó a cabo de la siguiente forma. Cada condición fue asignada a 6 inspectores y cada inspector evaluó tres paneles de circuitos, cada uno con un defecto distinto. La variable dependiente fue el tiempo en segundos requerido para detectar el defecto. Los datos vienen en una Tabla adjunta. Se pide justificando cualquier respuesta:

1. Determinar la estructura formal del modelo estadístico que utilizaríais para evaluar si los tiempos de detección dependen de las condiciones y/o de los tipos de defectos.
2. Comprobar la significación de los efectos principales y de la interacción.
3. Independientemente de la condición, qué tipo de defecto es más fácil de detectar y difiere significativamente de los otros?
4. Qué método de inspección (condición) recomendaríais para detectar defectos tipo fill? Y para detectar defectos tipo gap?

### Problema 2 [2.5 puntos]. Porcentaje de ingestión de oxígeno.

En un curso de educación física se tomaron diferentes medidas sobre 31 hombres. Las variables tomadas fueron (ver Tabla adjunta): Edad (years), Peso (Weight, kg), Porcentaje de ingestión de oxígeno (Oxygen intake rate in ml), Tiempo en correr 1.5 millas (RunTime in minutes), Pulsaciones en reposo (RestPulse), Pulsaciones durante la carrera (RunPulse), Máximo número de pulsaciones durante la carrera (MaxPulse). La variable de interés es la ingestión de oxígeno. Proponer un modelo lineal para modelizar la dependencia de la ingestión de oxígeno en función de las otras variables medidas. Se pide:

1. Contrastar la hipótesis a un 95% de que las 6 variables conjuntamente no expliquen significativamente la ingestión de oxígeno. En caso de que sí expliquen la variabilidad en la toma de oxígeno, especificar qué variables forman parte del modelo y cuáles no. Definir el modelo lineal en cuestión y comprobar sus hipótesis básicas.
2. Cuál es el porcentaje de oxígeno predicho para un hombre de 40 años que pesa 90kg, le cuesta 11 minutos correr 1.5 millas y tiene 65 pulsaciones en reposo, 170 en carrera y un máximo de 185 pulsaciones?
3. Imaginemos que un hombre gana 4 kg en los próximos 5 años, pero que esto no tiene efecto en sus diferentes tipos de pulsaciones. ¿Cuánto más rápido tendrá que correr 1.5 millas para que sea considerado que está en igual forma física que hoy?

4. Construir intervalos de confianza al 90% sobre los parámetros asociados a las variables RestPulse y MaxPulse. Se puede considerar que el parámetro asociado a Weight es nulo significativamente-probarlo mediante un contraste de hipótesis.

### Problema 3 [2 puntos].

Los datos de la Tabla adjunta representan medidas sobre la calidad del agua basada en concentración de nitrógeno en miligramos por litro para 20 cuencas de ríos. Conjuntamente se midieron otras variables como el área de la cuenca (en Km cuadrados), porcentaje de tierra para uso agrícola, forestal, residencial y para propósitos comerciales. Los investigadores están interesados en llevar a cabo los siguientes análisis. Ayudadles en la medida de vuestros conocimientos.

1. Analizar la existencia de grupos homogéneos entre las 20 cuencas, describiendo y justificando el método estadístico seguido.
2. Con los grupos homogéneos obtenidos en el apartado anterior, tenemos una estructura de covariables clara para proceder a un análisis discriminante? Justificar esta respuesta con descripción de fiabilidad de la(s) funciones discriminantes.
3. Confirmar los grupos obtenidos mediante un análisis de la varianza para evaluar la existencia de diferencias significativas para la variable concentración de nitrógeno en función de los grupos.

### Problema 4 [1.5 puntos]. Cuestiones teóricas sobre regresión

a) Determinar las ecuaciones normales y los estimadores de los parámetros en un modelo de regresión polinomial de segundo grado.

b) Sea la ecuación lineal de regresión  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$  para  $i = 1, \dots, n$  con  $\varepsilon \sim N(0, \sigma^2)$  y satisfaciéndose las siguientes condiciones:  $\sum x_{i1} = \sum x_{i2} = 0$ ,  $\sum x_{i1}^2 = \sum x_{i2}^2 = n$  y  $\sum x_{i1} x_{i2} = k > 0$ . Demostrar que:

1. Bajo un diseño ortogonal ( $k = 0$ ) se tiene que  $X^t X = \begin{pmatrix} n & 0 & 0 \\ 0 & n & 0 \\ 0 & 0 & n \end{pmatrix}$

2. Para cualquier diseño, i.e.  $k$  no necesariamente nulo, se tiene que  $(X^t X)^{-1} = \begin{pmatrix} 1/n & 0 & 0 \\ 0 & n/(n^2 - k^2) & -k/(n^2 - k^2) \\ 0 & -k/(n^2 - k^2) & n/(n^2 - k^2) \end{pmatrix}$

3. Las varianzas de los estimadores  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  son  $\sigma^2/n, n\sigma^2/(n^2 - k^2), n\sigma^2/(n^2 - k^2)$  respectivamente.

4. Demostrar que la expresión para los estimadores viene dada por  $\hat{\beta} = \begin{pmatrix} \bar{y}_n \\ \frac{1}{n^2 - k^2} [n \sum x_{i1} y_i - k \sum x_{i2} y_i] \\ \frac{1}{n^2 - k^2} [n \sum x_{i2} y_i - k \sum x_{i1} y_i] \end{pmatrix}$

Problema 5 [1.5 puntos]. Cuestiones teóricas sobre anova

Un experimento factorial  $2 \times 3$  fue llevado a cabo mediante una estructura completamente aleatorizada con 5 repeticiones para estudiar el efecto de un factor A con dos niveles y otro factor B con 3 niveles sobre una variable dependiente Y. El modelo estadístico para este experimento se plantea como  $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ . Se pide:

1. Esquematizar la tabla ANOVA enfatizando las fuentes de variabilidad y los grados de libertad.
2. Si el cuadrado medio del error es  $MS_{Residual}=5.0$ , dar el valor concreto del punto de comparación en el test de Scheffé al 95% para comparar medias entre los niveles del Factor A, B e interacción.
3. Supongamos que las sumas de cuadrados del efecto interacción son  $SS(Int.)=23$ . Comprobar que la interacción no es significativa. Decidimos eliminarla del modelo. Como cambia la tabla ANOVA sin el efecto interacción?