**Grado en Ingeniería Informática**

# Estadística

# SEGUNDO PARCIAL

6 de junio de 2018

| | |
|---|---|
| Apellidos, nombre: | |
| Grupo: | Firma: |

## Instrucciones

1. Rellenar la información de cabecera del examen.
2. Responder a cada pregunta en la hoja correspondiente.
3. Justificar todas las respuestas.
4. No se permiten anotaciones personales en el formulario.
5. No se permite tener teléfonos móviles encima de la mesa. Sobre la mesa sólo se permite el DNI, calculadora, útiles de escritura, las tablas y el formulario.
6. No desgrapar las hojas.
7. Todas las preguntas puntúan lo mismo (sobre 10).
8. Se debe firmar en las hojas que hay en la mesa del profesor al entregar el examen. Esta firma es el justificante de la entrega del mismo.
9. Tiempo disponible: **2 horas**

**1.** La empresa **AIRFLY** S.A. dispone de un avión tipo A-340 para realizar sus rutas transoceánicas. Cuando se adquirió esta aeronave hace 5 años el fabricante señalaba que el consumo medio de combustible era de 6900 kilogramos de fuel por hora (kg/h) con una desviación típica de 180 kg/h. Tras analizar un total de 50 rutas efectuadas por este avión, **AIRFLY** ha estimado un consumo medio de 7200 kg/h con una desviación típica de 200 kg/h. Se pide:

**a)** La empresa considera que este avión no consume lo mismo, en promedio, que cuando se adquirió. Justifica, planteando y resolviendo un contraste de hipótesis con un riesgo de primera especie del 5%, si la empresa está en lo cierto. Nota: no utilices un intervalo de confianza para resolver este apartado.

*(4 puntos)*

**b)** Sin realizar cálculos adicionales, justifica si crees que cambiaría la respuesta del apartado anterior si el riesgo de primera especie fuese del 1%.

*(2 puntos)*

**c)** Calcula un intervalo de confianza al 95% para la desviación estándar poblacional actual. Teniendo en cuenta el intervalo obtenido, ¿consideras que la variabilidad que presentaba el consumo de combustible en las rutas transoceánicas ha cambiado? *(4 puntos)*

**2.** Un fabricante de baterías (pilas) para dispositivos móviles produce dos tipos diferentes de unidades (A y B). Para analizar el efecto que la temperatura tiene en la duración de las mismas, el fabricante realiza un experimento donde se mide la duración de ambos tipos de pila bajo 3 condiciones diferentes de temperatura (5 ºC, 20 ºC y 35 ºC). Para ello en cada una de las 6 combinaciones (tipo de pila y Tª) estudia la duración de 3 pilas. A continuación se muestra el resultado del análisis estadístico realizado. Teniendo en cuenta la información obtenida contesta a las siguientes preguntas:

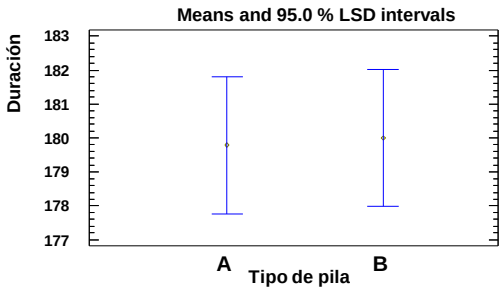
**a)** Completa la Tabla Resumen del ANOVA con aquellos valores que te puedan hacer falta para contestar el resto de apartados de esta pregunta. Justifica los cálculos realizados. *(3 puntos)*

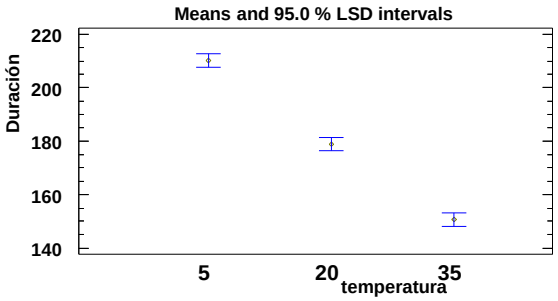
**Análisis de Varianza para Duración - Suma de Cuadrados Tipo III**

| *Fuente* | *Suma de Cuadrados* | *Gl* | *Cuadrado Medio* | *F-ratio* | *P-valor* |
|---|---|---|---|---|---|
| EFECTOS PRINCIPALES | | | | | |
| A: Tipo Pila | 0,222222 | | | | 0,9067 |
| B: Temperatura | 10630,8 | | | | 0,0000 |
| INTERACCIONES | | | | | |
| AB | 1270,78 | | | | 0,0000 |
| RESIDUOS | 186,0 | | | | |
| TOTAL (CORREGIDO) | 12087,8 | | | | |

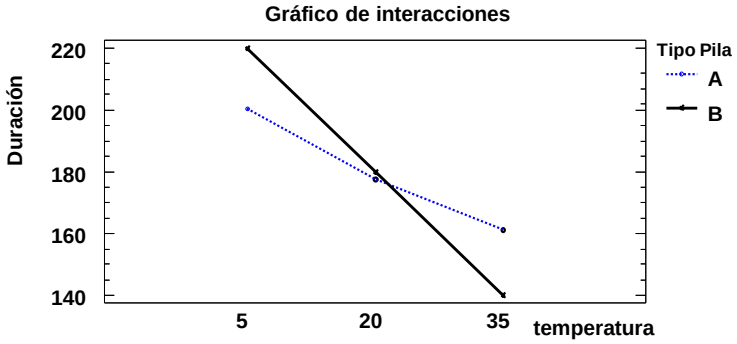
**b)** Estudia la significación estadística de los efectos simples de ambos factores y de su interacción (asume un $\alpha$ =5% para la significación). *(2 puntos)*

**c)** Interpreta los gráficos de intervalos LSD para los factores tipo de batería y temperatura. Describe la naturaleza del efecto del tipo de batería y de la temperatura. *(2 puntos)*




**d)** Interpreta el gráfico de la interacción en coherencia con las conclusiones derivadas de los apartados anteriores. Justifica el motivo por el que la interacción ha resultado estadísticamente significativa o no en la tabla del ANOVA. ¿Qué recomendarías, teniendo en cuenta este gráfico, a un futuro comprador para maximizar la duración de las baterías? *(3 puntos)*



**3.** La empresa **AIRFLY** S.A. ha diseñado un experimento con 10 pruebas para medir el efecto de un cierto tipo de aditivo en el tiempo de secado de pintura que aplica a sus aeronaves, obteniéndose los siguientes resultados:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Concentración_Aditivo (%) | 4.0 | 4.2 | 4.4 | 4.6 | 4.8 | 5.0 | 5.2 | 5.4 | 5.6 | 5.8 |
| Tiempo_Secado (h) | 8.7 | 8.8 | 8.3 | 8.7 | 8.1 | 8.0 | 8.1 | 7.7 | 7.5 | 7.2 |

A partir de los datos anteriores la empresa ha obtenido los siguientes resultados tras aplicar ciertas herramientas estadísticas:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Tiempo_Secado
Independent variable: Concentracion_Aditivo
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        12,1933       0,523576        23,2885         0,0000
Slope           -0,833333      0,106126       -7,85234         0,0000
------------------------------------------------------------------------------


                           Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                      2,29167      1      2,29167      61,66       0,0000
Residual                  0,297333      8    0,0371667
------------------------------------------------------------------------------
Total (Corr.)               2,589      9

Correlation Coefficient = -0,940827
R-squared = 88,5155 percent
Standard Error of Est. = 0,192787
```

Teniendo en cuenta los resultados obtenidos, se pide:

**a)** La empresa **AIRFLY** S.A. considera que no existe ningún tipo de relación entre el tiempo de secado y la aplicación del tipo de aditivo, de modo que los técnicos han decidido emplear la máxima cantidad de aditivo para maximizar el tiempo de secado. ¿Qué opinas al respecto?. Justifica tu respuesta y plantea los test de hipótesis necesarios. *(2,5 puntos)*

**b)** Calcula el residuo que se cometería al utilizar el modelo estimado para un nivel de concentración de aditivo del 4%. ¿Qué representa este valor? *(3 puntos)*

**c)** Si se cumpliesen las hipótesis necesarias para aplicar el modelo de regresión lineal, ¿cómo podríamos modelizar la variable "tiempo de secado" para un nivel de concentración de aditivo de 4,8%? Estima también los parámetros de dicho modelo de distribución. *(3 puntos)*

**d)** ¿Cuáles son las hipótesis necesarias para aplicar el modelo de regresión lineal? *(1,5 puntos)*

## SOLUCIÓN

**1a)** El contraste de hipótesis que se pretende estudiar es si la media poblacional del consumo de fuel vale 6900 (valor que supuestamente tenía cuando se adquirió el avión) o bien si se ha modificado; $H_0$: m=6900; $H_1$: m≠6900.

$$\frac{\bar{x}-m}{s/\sqrt{n}} \approx t_{n-1};\quad \frac{7200-6900}{200/\sqrt{50}} = 10{,}61$$

El estadístico de contraste sigue una distribución t de Student con 49 grados de libertad (n=50); el valor obtenido es muy poco probable para esta distribución, ya que el 95% de sus valores varían entre -2,01 y 2,01. Por tanto la empresa está en lo cierto: hay suficiente evidencia para afirmar que el avión consume significativamente más que cuando se adquirió.

**1b)** Considerando un riesgo de primera especie α=1%, resulta que el 99% de valores de la distribución $t_{49}$ fluctúan entre -2,68 y 2,68. El estadístico de contraste está fuera de este intervalo, por lo cual no cambia la respuesta del apartado anterior.

**1c)** $\sigma^2 \in \left[\frac{(n-1)\cdot s^2}{g_2}; \frac{(n-1)\cdot s^2}{g_1}\right]$ siendo $g_1$ y $g_2$ el intervalo de valores de una distribución $\chi^2$ con 49 grados de libertad que comprende el 95% de valores. La tabla de esta distribución no indica estos valores directamente pero sí para 50 grados de libertad, de lo cual resultaría un intervalo aproximado:

$$\sigma^2 \in \left[\frac{49\cdot 200^2}{71{,}42}; \frac{49\cdot 200^2}{32{,}357}\right];\quad \sigma^2 \in [27443;\ 60574];\ \text{raíz cuadrada: } \sigma \in [165{,}7;\ 246]$$

Los valores críticos para 49 grados de libertad pueden obtenerse interpolando entre 45 y 50 grados de libertad, de lo cual se obtiene el intervalo en cuestión:

$$\sigma^2 \in \left[\frac{49\cdot 200^2}{70{,}22};\ \frac{49\cdot 200^2}{31{,}55}\right];\quad \sigma^2 \in [27912;\ 62124];\ \sigma \in [167{,}1;\ 249{,}2]$$

Dado que este intervalo incluye el valor 180 (desviación típica que el consumo de combustible tenía hace 5 años), se concluye que no hay suficiente evidencia para afirmar que la variabilidad que presentaba el consumo haya cambiado.

**2a)** La tabla del ANOVA completada se muestra a continuación:

| *Fuente* | *Suma de Cuadrados* | *Gl* | *Cuadrado Medio* | *F-ratio* | *P-valor* |
|---|---|---|---|---|---|
| EFECTOS PRINCIPALES | | | | | |
| A: Tipo Pila | 0,222222 | **1** | **0,2222** | **0,0143** | 0,9067 |
| B: Temperatura | 10630,8 | **2** | **5315,4** | **342,93** | 0,0000 |
| INTERACCIONES | | | | | |
| AB | 1270,78 | **2** | **635,39** | **40,99** | 0,0000 |
| RESIDUOS | 186,0 | **12** | **15,5** | | |
| TOTAL (CORREGIDO) | 12087,8 | **17** | | | |

Nº total de datos = 6 combinaciones x 3 pilas ensayadas en cada combinación = 18. Grados de libertad totales = N-1 = 17; G.l. de cada factor = nº variantes-1; G.l. interacción = 2·1= 2; Cuadrado medio= SC / g.l.;
F-ratio = CM / $CM_{resid.}$

**2b)** El p-valor asociado al factor "tipo de pila" es muy superior a 0,05 de modo que se acepta la hipótesis nula: el efecto simple de este factor no resulta estadísticamente significativo. Es decir, se acepta la hipótesis nula de que la media a nivel poblacional de la duración es la misma para ambos modelos de batería: $m_A = m_B$.

El p-valor asociado al factor "temperatura" es muy inferior a 0,05 de modo que se rechaza la hipótesis nula: el efecto simple de este factor resulta estadísticamente significativo. Es decir, se rechaza la hipótesis nula de que la media poblacional de la duración sea la misma para las tres temperaturas.

El p-valor asociado a la interacción es muy inferior a 0,05 de modo que se rechaza la hipótesis nula: el efecto de la interacción entre ambos factores resulta estadísticamente significativo.

**2c)** Factor "tipo de pila": el gráfico de los intervalos LSD es coherente con la conclusión anterior: los intervalos se solapan, de modo que se acepta la hipótesis nula: $m_A = m_B$. Este factor no tiene efecto en el valor medio de la duración a nivel poblacional. Es decir, cambiar entre el modelo A o B de batería no causa una modificación, en promedio, de la duración.

Factor "temperatura": el gráfico muestra que los intervalos no se solapan, lo cual revela un efecto estadísticamente significativo. Este efecto es aproximadamente lineal ya que puede ajustarse una línea recta que atraviesa los intervalos LSD. Se trata de un efecto lineal negativo, pues a mayor temperatura la duración es menor. Este efecto puede modelizarse con regresión lineal, lo cual permitiría estudiar si existe adicionalmente un efecto cuadrático.

**2d)** El gráfico de la interacción sugiere un efecto lineal decreciente de la temperatura en la duración de las pilas, pero este efecto es distinto para cada tipo (es decir, la pendiente no es la misma). El hecho de que las líneas correspondientes a cada tipo de pila no sean paralelas indica una interacción, lo cual es coherente con la tabla del ANOVA, ya que el efecto de la interacción resulta estadísticamente significativo ya que el p-valor es muy bajo. Es decir, hay suficiente evidencia para afirmar que la interacción observada en la gráfica a nivel muestral se corresponde también con una interacción a nivel poblacional.

Recomendación: Si es posible, se aconseja reducir la temperatura de trabajo para maximizar la duración de las pilas. Si la Tª es algo fijo que no se puede elegir, la recomendación es la siguiente. En caso de que la Tª sea de unos 5ºC, interesan las pilas de tipo B. En caso de trabajar a unos 35ºC se recomiendan las pilas de tipo A. Pero tratándose de baterías para dispositivos móviles, resulta poco probable que la temperatura de trabajo sea habitualmente tan elevada. Si la Tª es de unos 20ºC, indistintamente puede emplearse un tipo u otro (se recomienda la más económica).

**3a)** Afirmación 1: La empresa considera que no existe ningún tipo de relación entre el tiempo de secado y la concentración de aditivo. Esta afirmación es falsa, ya que la correlación entre ambas variables es estadísticamente significativa. El test de hipótesis que se plantea ($H_0$: b=0; $H_1$: b≠0) está asociado a la pendiente de la recta de regresión Y = a+b·x. Dado que el p-valor asociado a la pendiente es muy bajo (prácticamente cero) se rechaza la hipótesis nula, y puede afirmarse que la pendiente es significativamente distinta de cero a nivel poblacional, lo cual permite afirmar que la correlación también es estadísticamente significativa.

Afirmación 2: los técnicos han decidido emplear la máxima cantidad de aditivo para maximizar el tiempo de secado. En primer lugar, llama la atención que se pretenda maximizar el tiempo, ya que en un proceso industrial habitualmente se pretende minimizar los tiempos de proceso. En cualquier caso, la empresa es incoherente: si considera que no existe correlación, no tendría sentido pensar que al aumentar la cantidad de aditivo también aumente el tiempo de secado. Por último, esta relación es errónea ya que la pendiente es negativa, de modo que a mayor cantidad de aditivo, el tiempo será menor.

**3b)** La ecuación obtenida es: Y = 12,1933 - 0,833 ·X. Los p-valores de los parámetros del modelo, al ser muy bajos indican que ambos parámetros del modelo resultan estadísticamente significativos. Para una concentración del 4%, la predicción del modelo es: $Y_{predicho}$ = 12,1933 - 0,833 · 4 = 8,8613.
El valor observado de tiempo para un 4% de aditivo es de 8,7 (primer dato en la tabla). Por definición, **residuo** = $Y_{observado}$ - $Y_{predicho}$ = 8,7 - 8,8613 = **-0,161**.

El residuo representa el error del modelo, es decir, la diferencia entre el valor observado y el estimado. Este error representa el efecto sobre la variable respuesta (tiempo de secado) de todos los factores restantes no incluidos en el modelo de regresión. Si se conociesen absolutamente todas las variables que intervienen en la predicción del tiempo de secado, seríamos capaces de estimar este valor con total exactitud y el residuo sería nulo. Esto sucede con muchos modelos deterministas habituales en ciencias físicas. Pero en este caso la ecuación de regresión sólo se construye con una variable explicativa, de modo que el resto de factores no incluidos en el modelo causan una perturbación aleatoria en cada observación, que se asume como distribución normal, que es lo que mide el residuo. En este caso, al ser negativo, significa que el valor de Y observado es menor que el estimado por el modelo.

**3c)** Si se cumplen las hipótesis del modelo (ver apartado 3d), el tiempo de secado para una concentración de aditivo X=4,8 puede modelizarse de acuerdo a una distribución normal, cuyos parámetros son los siguientes:
- Media: si X=4,8, el modelo predice el siguiente valor, que será el tiempo medio para este valor de X: Y = 12,1933 - 0,833 · 4,8 = **8,1949**.
- Varianza: coincide con el cuadrado medio residual = 0,03717.
- Desviación típica: raíz cuadrada del valor anterior, que aparece en la tabla como *standard error of est*. = **0,1928**.

**3d)** Las hipótesis para aplicar un modelo de regresión lineal son tres, si bien las dos primeras son las más importantes:

- Hipótesis de normalidad: la variable en cuestión sigue una distribución normal bivariante, lo cual implica que: (1) las distribuciones marginales de X e Y son normales, (2) la distribución condicional de Y para un valor concreto de X sigue un modelo normal, y (3) los residuos del modelo son normales y no existen datos anómalos.

- Hipótesis de homocedasticidad: la varianza de la distribución condicional de Y para un valor de X es constante, no depende de X.

- Hipótesis de independencia: los individuos de la población se han elegido aleatoriamente, de modo que todos ellos tienen la misma probabilidad de pertenecer a la muestra. Esta hipótesis a veces no se cumple cuando las observaciones se generan a lo largo del tiempo, de modo que el valor obtenido en cierto instante de tiempo puede depender parcialmente de los valores obtenidos con anterioridad. Si no se cumple esta hipótesis hay que emplear modelos de regresión avanzados que tienen en cuenta las series temporales.