Grado en Ingeniería Informática

Estadística

EXAMEN FINAL

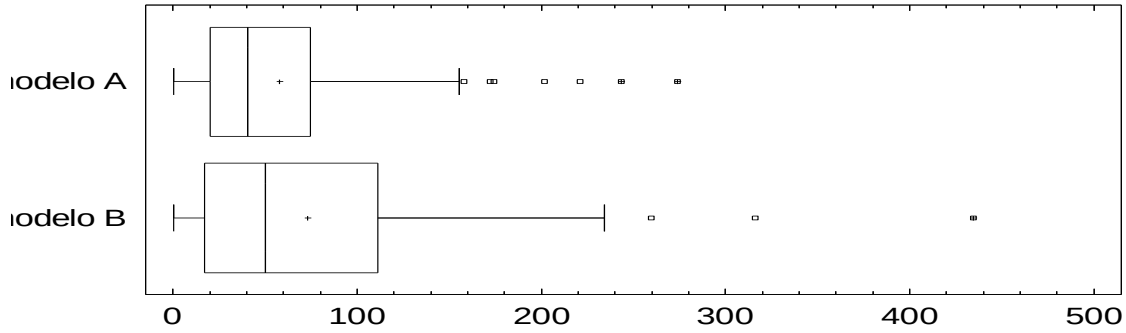
12 de junio de 2019

Apellidos y nombre:		
Grupo:	Firma:	
Marcar las casillas de los parciales presentados	P1 ☐	P2 ☐

Instrucciones

1. **Rellenar** la cabecera del examen: **nombre, grupo y firma**.
2. Responder a cada pregunta en la hoja correspondiente.
3. **Justificar todas las respuestas**.
4. No se permiten anotaciones personales en el formulario. Sobre la mesa sólo se permite el DNI, calculadora, útiles de escritura, las tablas y el formulario.
5. **No desgrapar** las hojas.
6. El examen consta de 6 preguntas, 3 correspondientes al primer parcial (50%) y 3 del segundo (50%). El profesor corregirá los parciales que el alumno haya señalado en la cabecera del examen. **En cada parcial, todas las preguntas puntúan lo mismo** (sobre 10).
7. 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.
8. Tiempo disponible: **3 horas**

1. (1er Parcial) Una empresa fabrica dos tipos de piezas distintas (modelo A y modelo B). La empresa está interesada en analizar el tiempo que transcurre hasta que se produce un fallo en la fabricación de estas piezas. Para ello, ha recopilado los tiempos hasta el fallo (en horas) registrados en los últimos años. Estos datos se representan en la siguiente figura:



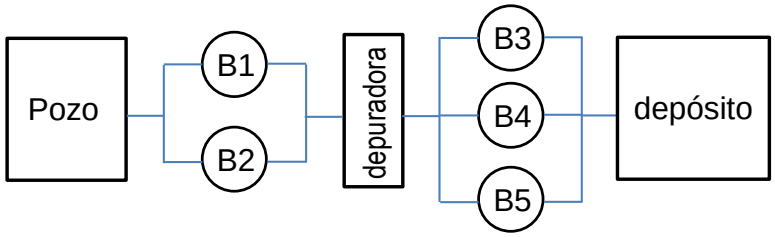
a) ¿Cuál es la población (o poblaciones) en estudio? ¿Cuáles son los individuos? ¿Cuál es la variable aleatoria? *(3 puntos)*

b) La empresa decide invertir en la mejora de la manufactura de uno de los dos modelos. ¿En cuál de los dos (A o B) recomendarías realizar esta inversión? Justifica razonadamente la respuesta. *(2 puntos)*

c) Indica si la siguiente afirmación es verdadera o falsa, justificando convenientemente la respuesta: "Para el modelo B, el 50% de los fallos corresponden a un tiempo menor o igual a 72 horas, aproximadamente".
(2 puntos)

d) ¿Qué distribución estadística podría ajustarse razonablemente a los datos del modelo B? Calcula los parámetros de dicha distribución. A partir de ésta, calcula la probabilidad de que el tiempo hasta el fallo en el modelo B sea superior a 500 horas. *(3 puntos)*

2. (1er Parcial) En una planta industrial, dos bombas B1 y B2 en paralelo conducen agua desde un pozo a una depuradora, y posteriormente otras tres bombas (B3, B4 y B5, también en paralelo) la trasladan a un depósito, tal como se indica en la figura. Los tiempos de vida (tiempo de funcionamiento hasta el fallo) de la depuradora y de las bombas son variables aleatorias independientes con distribución normal y desviación típica de 6 mil horas, siendo 25 mil horas la vida media de la depuradora y 20 mil horas la de cada bomba.



a) Si T_D es la variable aleatoria "tiempo de funcionamiento hasta el fallo de la depuradora", calcular el percentil 7 de esta distribución. *(2 puntos)*

b) Calcular la probabilidad de que llegue agua desde la salida de la depuradora hasta el depósito después de 15 mil horas de funcionamiento. *(2 puntos)*

c) Calcular la probabilidad de que llegue agua desde el pozo al depósito después de 15 mil horas de funcionamiento. Para resolver este apartado, define en primer lugar los sucesos involucrados en el cálculo de la probabilidad. *(3 puntos)*

d) Si se desea que la probabilidad del apartado anterior sea exactamente del 90%, ¿qué vida media debería tener la depuradora, considerando la misma desviación típica? *(3 puntos)*

3. (1er Parcial) Los códigos CRC utilizados en el envío de paquetes a través de la red son capaces de corregir como máximo 7 errores por paquete. En caso de superarse este valor, el paquete es rechazado. Se sabe que el número medio de errores por paquete en un determinado envío es igual a 3.

a) ¿Cuál es la probabilidad de que un paquete sea rechazado por no poder ser corregido? *(2 puntos)*

b) Sabiendo que un paquete tiene menos de 10 errores, ¿cuál es la probabilidad de ser rechazado? *(3 puntos)*

c) Si se toman al azar 5 paquetes consecutivos, ¿cuál es la probabilidad de que exactamente dos de ellos sean rechazados? *(2 puntos)*

d) Si se toman al azar 5 paquetes consecutivos, ¿cuál es la probabilidad de que el número total de errores sea menor que 18? Resuelve esta pregunta aproximando con la distribución normal. *(3 puntos)*

4. (2º Parcial) Se sospecha que un sensor de temperatura (S_1) está mal calibrado. Para estudiarlo, se toma un sensor de referencia (S_2), debidamente calibrado por una empresa de metrología. Se colocan ambos sensores dentro de una cámara climática de ambiente variable, y se toma una medida cada hora. Los valores obtenidos se indican en la siguiente tabla:

S_1	18.7	19.4	20.2	21.6	23.8	24.1	26.3	25.7	24.9	24.1	22.9	21.5
S_2	18.4	19.3	20.2	21.8	23.7	24.1	26.5	25.5	24.8	24.1	22.6	21.3
S_1-S_2	0.3	0.1	0	-0.2	0.1	0	-0.2	0.2	0.1	0	0.3	0.2

Media..... 22.7667 (S_1); 22.6917 (S_2); 0.075 (S_1-S_2)
Desviación típica... 2.4810 (S_1) ; 2.5350 (S_2); 0.1658 (S_1-S_2)

a) A partir de los valores de S_1-S_2, ¿es admisible la hipótesis nula de que el valor medio de esta variable es nulo a nivel poblacional? Resuelve este test de hipótesis con la técnica del intervalo de confianza, considerando α=5%.
(3,5 puntos)

b) A partir de los resultados obtenidos en el apartado anterior, ¿hay suficiente evidencia para afirmar que el sensor S_1 requiere ser calibrado? *(1,5 puntos)*

c) Asumiendo normalidad e independencia entre los valores obtenidos experimentalmente, obtener un intervalo de confianza para la varianza poblacional del sensor S_2, con un nivel de confianza del 95%. *(3 puntos)*

d) Si se realiza una nueva medición en el interior de la cámara climática con el sensor S_2, ¿entre qué valores cabe encontrar dicho valor, con una probabilidad del 95%? *Nota*: se asume independencia en los datos y que estos se ajustan a un modelo normal cuya media y desviación típica son los valores indicados en el enunciado para S_2. *(2 puntos)*

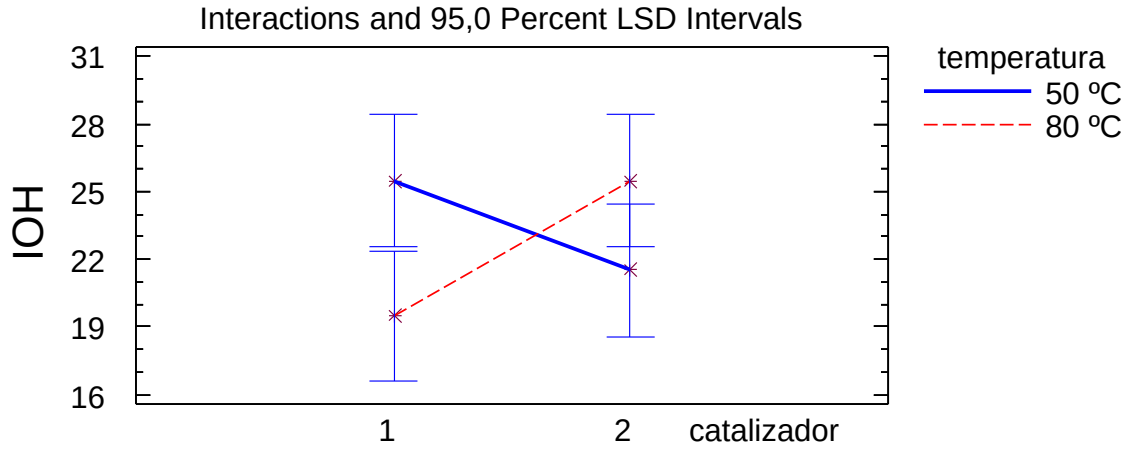
5. (2º Parcial) Una empresa petroquímica elabora un determinado tipo de polímero (poliol) por medio de un proceso de fabricación por lotes. El índice de hidroxilo (IOH) es un parámetro de calidad que se mide en el producto final, el cual interesa que sea lo más elevado posible. Se realiza un diseño de experimentos para estudiar el posible efecto de dos factores en el parámetro IOH: la temperatura en el interior del reactor (dos niveles ensayados: 50ºC y 80ºC), y la concentración de catalizador (1 g/m^3 o 2 g/m^3). Los datos obtenidos experimentalmente se muestran en la siguiente tabla:

catalizador	2	2	2	2	1	1	1	1
temperatura	50	50	80	80	50	50	80	80
IOH	20	23	24	27	24	27	18	21

Analysis of Variance for IOH - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:catalizador	2,0	1	2,0	0,44	0,5415
B:temperatura	2,0	1	2,0	0,44	0,5415
INTERACTIONS					
AB	50,0	1	50,0	11,11	0,0290
RESIDUAL	18,0	4	4,5		
TOTAL (CORRECTED)	72,0	7			

A partir de los datos obtenidos, se ha llevado a cabo un Análisis de la Varianza (ANOVA), incluyendo en el modelo la interacción doble. El gráfico de la interacción se muestra a continuación:



a) Considerando un nivel de significación del 1%, ¿qué temperatura y qué concentración de catalizador recomendarías para conseguir a nivel poblacional un producto final con el mayor valor posible de IOH? *(3 puntos)*

b) En caso de considerar α=5%, ¿qué temperatura y qué concentración de catalizador recomendarías? *(3 puntos)*

c) Describe detalladamente el procedimiento que habitualmente se utiliza en este tipo de análisis para detectar la presencia de datos anómalos. *(2 puntos)*

d) ¿Qué información útil podría obtenerse en este caso a partir del gráfico de medias con intervalos LSD para el factor "catalizador" con el objetivo de estudiar el efecto simple de dicho factor sobre la variable IOH a nivel poblacional? *(2 puntos)*

6. (2º Parcial) La resistencia de cierto tipo de polímero empleado en la fabricación de equipos informáticos está correlacionada con la temperatura media durante su fabricación. Se dispone de 100 datos de resistencia y temperatura, los cuales se han analizado por medio de regresión lineal simple obteniéndose los siguientes resultados con el programa *Statgraphics*:

```
-------------------------------------------------------------------------------
Multiple Regression Analysis -  Dependent variable: RESISTENCIA
-------------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
-------------------------------------------------------------------------------
CONSTANT                 13,2566       9,99004        1,32698         0,1876
TEMPERATURA              1,69615      0,164561        10,3071         0,0000
-------------------------------------------------------------------------------

R-squared = 52,0164 percent
R-squared (adjusted for d.f.) = 51,5268 percent
Standard Error of Est. = 47,3658
```

Con el objetivo de estudiar si un modelo de regresión lineal múltiple ajusta mejor los datos, se ha obtenido también esta tabla:

```
-------------------------------------------------------------------------------
Multiple Regression Analysis - Dependent variable: RESISTENCIA
-------------------------------------------------------------------------------
                                        Standard          T
Parameter               Estimate          Error       Statistic        P-Value
-------------------------------------------------------------------------------
CONSTANT                -8,61657          16,214      -0,531429         0,5963
TEMPERATURA              2,85408        0,699248        4,08165         0,0001
TEMPERATURA^2         -0,0108588     0,00637671       -1,70288         0,0918
-------------------------------------------------------------------------------

R-squared = 53,4092 percent
R-squared (adjusted for d.f.) = 52,4486 percent
Standard Error of Est. = 46,9132
```

a) A la vista de todos los resultados, discutir si interesa emplear el modelo de regresión lineal simple (primera tabla) o múltiple (segunda tabla), considerando un nivel de significación del 5%. Justifica convenientemente tu respuesta.

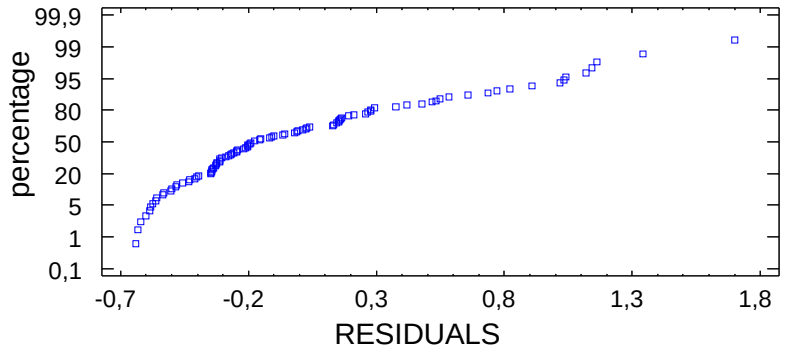
(2,5 puntos)

b) Calcular el coeficiente de correlación entre resistencia y temperatura a partir del primer modelo. ¿Dicho coeficiente será positivo o negativo? ¿Por qué?

(2 puntos)

c) En el primer modelo, calcular la probabilidad de obtener una resistencia superior a 70 en caso de que la temperatura haya sido de 70ºC. *(3 puntos)*

d) Los residuos del primer modelo se han representado en un papel probabilístico normal, el cual se muestra a continuación. A la vista de la figura, ¿puede afirmarse que existe una relación no lineal (cuadrática) entre las variables "*residuals*" y "*percentage*"? ¿Qué información útil se deduce del gráfico? ¿Qué recomendarías en este caso? Justifica tus respuestas. *(2,5 puntos)*



SOLUCIÓN

1a) En este caso hay dos poblaciones. Una población es el conjunto de todos los fallos que se producen en la fabricación de piezas del modelo A: los que ya se han producido en el pasado y también los que se producirán en el futuro. La otra población es el conjunto de fallos en la fabricación de las piezas de B.
La población es el conjunto de individuos, de modo que cada individuo corresponde a un fallo del proceso de fabricación de estas piezas.
La variable aleatoria es el tiempo (en horas) que transcurre hasta que se produce un fallo en la fabricación de estas piezas, es decir, tiempo desde la última vez que ocurrió un fallo (tiempo entre fallos).

1b) En un diagrama box-whisker, la media se representa como un punto en el interior de la caja, y la mediana es la línea representada dentro de la caja. A partir de la figura, se deduce que tanto la media como la mediana del tiempo para el modelo A es inferior que en B. Esto indica que el sistema de producción de A falla más frecuentemente que B, pues los tiempos entre fallos (que es lo que mide la variable aleatoria) son menores en promedio. Por tanto, se recomienda invertir en la fabricación del modelo A para conseguir incrementar estos tiempos.

1c) La afirmación es falsa, ya que para el modelo B, el 50% de los fallos corresponden a un tiempo menor o igual a 50 horas, aproximadamente. Este valor es la mediana.

1d) Dado que la variable es continua con mínimo en cero y la distribución es muy asimétrica positiva, los tiempos entre fallos podrían ajustarse a un modelo exponencial, el cual se emplea a menudo para modelizar tiempos entre fallos. Este modelo sólo tiene un parámetro, que es la inversa de la media. En este caso, la media vale 72 h aproximadamente, de modo que: $X \approx \exp(\alpha = 1/72)$.
$P(X>500) = e^{-\alpha \cdot 500} = e^{-500/72} =$ **0.0010**

2a) Considerando las unidades como miles de horas: $T_D \approx N(m=25, \sigma=6)$. Por definición, el percentil 7 es el valor que debajo por debajo el 7% de los datos. Si denominamos k_7 a este valor:

$P(T_D < k_7) = 0.07$; $P[N(25; 6) < k_7] = 0.07$; $P[N(0; 1) < (k_7-25)/6] = 0.07$

A partir de la tabla N(0;1) se obtiene que $P[N(0; 1) < -1.476] = 0.07$

$(k_7-25)/6 = -1.476 \rightarrow$ **k_7 = 25-6·1.476 = 16.145 mil horas**

2b) Definimos la variable aleatoria T_{Bi}: tiempo de funcionamiento hasta el fallo de la bomba "i". $T_{Bi} \approx N(m=20, \sigma=6) \rightarrow P(T_{Bi} >15) = P[N(20; 6) >15] =$
$= P[N(0; 1) > (15-20)/6] = P[N(0;1) > -0.833] = 1-0.202 = 0.798$

Suceso B_i: la bomba "i" funciona más de 15 mil horas.
P(agua de depuradora a depósito) = $P(B_3 \cup B_4 \cup B_5) = 1 - P(\overline{B_3 \cup B_4 \cup B_5}) =$
(aplicando una de las leyes de De Morgan y asumiendo sucesos independientes)
$= 1 - P(\overline{B_3} \cap \overline{B_4} \cap \overline{B_5}) = 1 - P(\overline{B_3}) \cdot P(\overline{B_4}) \cdot P(\overline{B_5}) = 1 - (1 - 0.798)^3 =$ **0.9917**

2c) Suceso B_i: la bomba "i" funciona más de 15 mil horas.
Suceso B_{1-2}: la bomba B1 o la B2 funcionan más de 15 mil horas.
Suceso B_{3-4-5}: alguna de las tres bombas en paralelo funcionan más de 15 mil h.
Suceso D: la depuradora funciona más de 15 mil horas.
Suceso p→d: el agua llega del pozo al depósito después de 15 mil horas.

$P(B_{3\text{-}4\text{-}5}) = 0.9917$ (cálculo del apartado anterior)
$P(B_{1\text{-}2})$ = (cálculo análogo al apartado anterior sustituyendo el exponente 3 por 2) =
$= P(B_1 \cup B_2) = 1 - (1 - 0.798)^2 = 0.9591.$
P(D) = P[N(25; 6) >15] = P[N(0; 1) > (15-25)/6] = P[N(0;1) >-1.667] = 0.952

Asumiendo que los sucesos son independientes:
$P(p \rightarrow d) = P(B_{1-2} \cap D \cap B_{3-4-5}) = P(B_{1\text{-}2}) \cdot P(D) \cdot P(B_{3\text{-}4\text{-}5}) = 0.9591 \cdot 0.952 \cdot 0.9917 =$
$= \mathbf{0.9057}$

2d) $P(B_{1\text{-}2}) \cdot P(D) \cdot P(B_{3\text{-}4\text{-}5}) = 0.9591 \cdot P(D) \cdot 0.9917 = 0.90 \rightarrow P(D) = 0.9462$
P(D) = P[N(m; 6) >15] = 0.9462; P[N(0; 1) > (15-m)/6] = 0.9462;
A partir de la tabla N(0; 1) se obtiene:
(15-m)/6 = -1.609 → m = 15+1.609·6 → m = **24.65 mil horas**
Dado que la probabilidad del apartado anterior es casi del 90%, la vida media de la depuradora es casi la que tenía, es decir, de 25 mil horas.

3a) Sea X: nº de errores en un paquete. En esta variable aleatoria discreta el valor mínimo es cero y el máximo no está acotado, de modo que puede modelizarse como un modelo Poisson, cuyo parámetro coincide con la media: X ≈ Ps (λ=3). P(rechazar paquete) = P(X>7) = 1-P(X≤7) = 1- 0.988 = **0.012**
Este valor se obtiene del ábaco, levantando una vertical para λ=3 (eje horizontal), que corta la curva ν=7.

3b) Nos piden una probabilidad condicional que se obtiene con la fórmula:

$$P[(X > 7)/(X < 10)] = \frac{P[(X>7) \cap (X<10)]}{P(X<10)} = \frac{P(X=8)+P(X=9)}{P(X<10)} = \frac{0.008101+0.0027}{0.9989} = \mathbf{0.0108}$$

P(X<10) = P(X≤9) = 0.9989 (valor obtenido del ábaco, con λ=3).
Aplicando la función de probabilidad para la distribución Poisson se obtiene:
$P(X=8) = e^{-3} \cdot 3^8/8! = 0.0081$; $P(X=9) = e^{-3} \cdot 3^9/9! = 0.0027$

3c) Sea Y: nº de paquetes rechazados de un conjunto de 5 paquetes. En esta variable discreta el valor mínimo es cero y el máximo es 5, de modo que sigue una distribución Binomial de parámetros n=5 y p = probabilidad de que un paquete sea rechazado = 0.012 (del apartado *a*) → Y ≈ Bi (n=5, p=0.012)
Nos piden la probabilidad de que esta variable tome el valor 2, que se obtiene aplicando la función de probabilidad:

$$P(Y=2) = P[Bi(5;0.012) =2] = \binom{5}{2} \cdot 0.012^2 \cdot (1 - 0.012)^{(5-2)} = \mathbf{0.00139}$$

El número combinatorio es: 5! / (2! · 3!) = 120/2·6 = 10

3d) Sea Z: nº total de errores contenidos en 5 paquetes: $Z = X_1 + X_2 + ... + X_5$.
La suma de variables Poisson independientes sigue una distribución del mismo tipo, cuyo parámetro es la suma de λ_i: Z ≈ Ps (λ=5·3 = 15). Como este parámetro es mayor de 9, la función de probabilidad se asemeja a una normal, cuya media y varianza valen 15. Aplicando la corrección de continuidad:

$$P(Z<18) = P[Ps\ (15) < 18] \approx P\left[N\left(15, \sqrt{15}\right) < 17.5\right] = P\left[N(0;1) < \frac{17.5-15}{\sqrt{15}}\right] =$$
$$= P[N(0; 1) < 0.6455] = 1 - 0.259 = \mathbf{0.741}$$

4a) El test de hipótesis es H_0: $m_{S1\text{-}S2} = 0$; H_1: $m_{S1\text{-}S2} \neq 0$.
Media muestral de S_1-S_2 = 0.075; desv. típica = 0.1658; n=12. ; α=0.05
El valor crítico de una t-Student con 11 grados de libertad que deja por encima un área de 0.025 es 2.201.

$m \in \left[\bar{x} - t_{n-1}^{\frac{\alpha}{2}} \cdot s/\sqrt{n}\,;\ \bar{x} + t_{n-1}^{\frac{\alpha}{2}} \cdot s/\sqrt{n}\right]$; $m \in \left[0.075 \pm 2.201 \cdot 0.1658/\sqrt{12}\right]$;

$m \in [\mathbf{-0.0304}; \mathbf{0.1804}]$ Dado que el valor cero está contenido en este intervalo, es admisible la hipótesis nula de que la media poblacional de S_1-S_2 es cero.

4b) Al aceptarse la hipótesis de que $m_{S1\text{-}S2}$ = 0, de ahí se deduce: $m_{S1} = m_{S2}$; es decir, no hay suficiente evidencia para afirmar que la media a nivel poblacional de las medidas del sensor S_1 sea distinta de la media poblacional de las medidas de S_2. En definitiva, no hay suficiente evidencia para afirmar que el sensor S_1 requiera ser calibrado, ya que la calibración es necesaria cuando los valores de un sensor, en promedio, difieren significativamente del sensor calibrado.

4c) Intervalo de confianza para la varianza poblacional de S_2, con 1-α=95%:
$\sigma^2 \in [(n-1) \cdot s^2/g_2\,;(n-1) \cdot s^2/g_1]$;$\sigma^2 \in [11 \cdot 2.535^2/21.92\,;11 \cdot 2.535^2/3.816]$
$\boldsymbol{\sigma^2} \in [\mathbf{3.225}\,;\mathbf{18.525}]$

Siendo g_1=3.816 y g_2=21.920 el intervalo de valores de una distribución chi-cuadrado con 11 grados de libertad que comprende el 95% de los valores. La desviación típica muestral es s = 2.535 (valor indicado en el enunciado).

4d) En una distribución normal, el intervalo comprendido entre $m \pm 1.96 \cdot \sigma$ comprende el 95% de los valores. Considerando m=22.3917 y σ=2.535, resulta:
$22.6917 \pm 1.96 \cdot 2.535$ → **[17.723; 27.660]**
Este intervalo comprenderá previsiblemente el 95% de todas las medidas realizadas en la cámara climática con el sensor S_2.

5a) Considerando α=1%, la interacción no resulta estadísticamente significativa ya que el p-valor asociado a ésta (0.029) es mayor de 0.01. Los efectos simples de los dos factores tampoco son estadísticamente significativos, por la misma razón. Por tanto, no hay evidencia suficiente para afirmar que sea mejor emplear una temperatura de 50 o 80ºC, ni tampoco una concentración de 1 o 2 g/m^3 de catalizador, a efectos de conseguir un mayor valor de IOH a nivel poblacional. Se recomendarían las condiciones operativas más económicas o favorables: una concentración de 1 g/m^3 (pues lógicamente será más barato que emplear el doble) y una temperatura de 50ºC en el interior del reactor, ya que ésta será más económica que tener que calentar hasta 80ºC.

5b) Considerando α=5%, la interacción resulta estadísticamente significativa (p-valor = 0.029 < 0.05). Esto implica que el efecto de cambiar de temperatura es significativamente distinto en función de la concentración de catalizador. Para conc=1, los intervalos LSD (obtenidos con 1-α=95%) no se solapan entre ambas temperaturas, de modo que para maximizar la variable respuesta IOH a nivel poblacional conviene elegir una temperatura de 50ºC. El intervalo LSD de conc=1 y Temp=50 se solapa con los dos intervalos para conc=2, de modo que no hay evidencia suficiente para afirmar que conc=1 y Temp=50 sea mejor que conc=2 a nivel poblacional. No obstante, dado que conc=1 es la condición más

económica por razones obvias, ésta sería la condición recomendada (conc=1 y Temp=50).

5c) El procedimiento para detectar datos anómalos en este tipo de análisis (ANOVA con dos factores) es el siguiente:
1) Si la interacción doble resulta estadísticamente significativa, hay que incluirla en el modelo. Si no fuera significativa, habría que descartarla, así como también los factores cuyo efecto simple no fuera significativo.
2) Posteriormente, hay que obtener los residuos del ANOVA, y representar estos sobre un papel probabilístico normal.
3) Si los puntos se ajustan razonablemente bien a una recta pero se observa algún valor extremo que se separa claramente de la recta, corresponderá a un dato anómalo.
4) Si los puntos no se ajustan a una recta y revelan una distribución asimétrica de los residuos, hay que transformar los datos originales y repetir este protocolo.

5d) Ninguna: dado que el efecto simple del factor "catalizador" no resulta estadísticamente significativo (p-valor = 0.54 > α), se sabe que en el gráfico de medias con intervalos LSD estos intervalos se van a solapar, de modo que este gráfico no aporta ninguna información adicional en este caso.

6a) El modelo de regresión múltiple es: $Y = \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot Temp^2$. Esta ecuación será más apropiada que el modelo lineal simple si puede garantizase que el coeficiente β_2 es distinto de cero a nivel poblacional. Se plantea la hipótesis nula $H_0: \beta_2 = 0$ frente a la alternativa $H_1: \beta_2 \neq 0$. El p-valor asociado a este test de hipótesis vale 0.0918 que es mayor a 0.05 (nivel de significación), de modo que se acepta la hipótesis nula. En la ecuación cuadrática, al considerar nulo el coeficiente β_2, equivale al modelo de regresión lineal simple. Dado que no hay evidencia suficiente para afirmar un efecto cuadrático a nivel poblacional, interesa emplear el primer modelo.

6b) En un modelo de regresión lineal simple, el coeficiente de correlación es la raíz cuadrada del coeficiente de determinación: $r = \sqrt{R^2} = \sqrt{0.52016} = \mathbf{0.7212}$
El coeficiente es positivo en este caso porque la pendiente de la recta es positiva (en caso de que la pendiente fuese negativa, habría que poner signo negativo). Esto indica que a mayor temperatura media de fabricación, en promedio se obtiene un polímero de mayor resistencia.

6c) Asumiendo la hipótesis de normalidad y homocedasticidad, la distribución condicional de la resistencia (Y) cuando la temperatura es de 70ºC, será un modelo normal cuya media viene dada por la ecuación de la recta y cuya desviación típica será 47.3658 (valor indicado en la tabla como *standard error of est.*): Ecuación de la recta: $\mathbf{Y = 13.2566 + 1.69615 \cdot Temp}$. No hay que eliminar la constante aunque no sea estadísticamente significativa, pues habría que volver a ajustar el modelo y el nuevo valor estimado de la pendiente se desconoce.

$E(Y/X=70) = 13.2566 + 1.69615 \cdot 70 = 131.99$

$P[N(131.99, 47.366) > 70] = P[N(0; 1) > (70-131.99)/47.366] =$
$= P[N(0;1) > -1.309] = 1-0.0951 = \mathbf{0.9047}$

Este cálculo es aproximado ya que la figura mostrada en el apartado siguiente indica una distribución asimétrica positiva de los residuos; es decir, no se cumple la hipótesis de normalidad.

6d) 1. No tiene sentido afirmar una relación cuadrática entre "residuals" y "percentage" porque en el papel probabilístico normal se representan los datos de una sola variable aleatoria unidimensional, este gráfico no pretende visualizar la relación entre dos variables (como sería el caso de un gráfico de dispersión). En el eje vertical se representa la frecuencia acumulada (en porcentaje) de los residuos en una escala especial.

2. Información útil: En el gráfico los puntos no aparecen alineados sino siguiendo una curvatura. Esto indica una distribución asimétrica positiva de los residuos. No hay información suficiente para afirmar que existan claramente datos anómalos que requieran ser descartados.

3. En este caso se recomienda transformar los datos de resistencia, repetir el modelo de regresión, volver a calcular los residuos, repetir de nuevo el papel probabilístico normal y comprobar si los residuos tienden a ser normales y si aparece algún dato anómalo. Las transformaciones más habituales son la raíz cuadrada, logaritmo o raíz cuarta.