Grado en Ingeniería Informática

Estadística

SEGUNDO PARCIAL

29 de mayo de 2019

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. Una industria elabora cierto modelo de condensadores empleados en circuitos electrónicos de equipos informáticos. El valor de la capacitancia es un parámetro de calidad que se controla periódicamente. El departamento de control de calidad asume que este parámetro sigue una distribución Normal de media 20 nF y desviación típica 0,2 nF. Se asume que el proceso funciona correctamente cuando la capacitancia de un condensador elegido al azar está comprendida entre 19,5 y 20,5 nF. En caso contrario, se considera que el proceso está fuera de control estadístico y necesita ser revisado.

a) Si hiciéramos un contraste de hipótesis aplicado a este caso, ¿qué interpretación tendría el riesgo de primera especie? Justifica tu respuesta.

(2 puntos)

b) Indica el rango de valores en el cual fluctúa habitualmente:

b.1) El riesgo de primera especie (justifica tu respuesta) *(1 punto)*

b.2) El p-valor (justifica tu respuesta) *(1 punto)*

c) La empresa decide realizar un cambio en el proceso de fabricación, que se sospecha podría afectar a los parámetros del modelo estadístico de distribución de la capacitancia. Para estudiarlo, se toma una muestra aleatoria de 15 condensadores, obteniéndose un valor medio de capacitancia de 20,15 nF y una varianza muestral de 0,09 nF^2.

c.1) A partir de esta información y considerando α=0,05, ¿existe suficiente evidencia para afirmar que el cambio efectuado en el proceso ha alterado la capacitancia media a nivel poblacional existente inicialmente? Para responder a esta pregunta, formula previamente el test de hipótesis que se plantea en este caso, resuelve dicho test e interpreta los resultados en este contexto, justificando tu respuesta. *(3 puntos)*

c.2) ¿Es admisible una desviación típica poblacional de 0,2? Contesta a esta pregunta utilizando la herramienta del intervalo de confianza, considerando un nivel de confianza del 95%. *(3 puntos)*

2. Una biblioteca pública dispone de seis ordenadores para las consultas de los usuarios, tres situados en la sala A y otros tres en la B. Los equipos se diferencian, entre otras cosas, en la velocidad del procesador, que puede ser 2, 3 o bien 4 GHz. Se toma una muestra aleatoria del tiempo de consulta de 24 usuarios: 8 de ellos emplean un equipo de 2 GHz, otros 8 con un equipo de 3 GHz y otros 8 con uno de 4 GHz. La mitad de usuarios realizan la consulta en la sala A, y la otra mitad en la sala B. Los valores experimentales obtenidos de tiempo (medido en minutos), se indican a continuación, así como la tabla resumen del ANOVA.

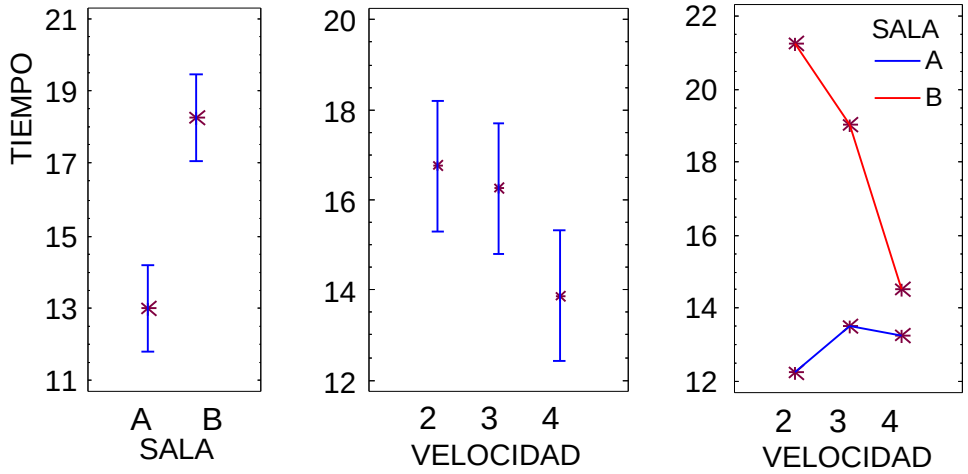
Sala A	
velocidad	tiempo
2 GHz	9; 12; 13; 15
3 GHz	11; 13; 16; 14
4 GHz	14; 16; 12; 11

Sala B	
velocidad	tiempo
2 GHz	23; 25; 20; 17
3 GHz	23; 15; 21; 17
4 GHz	14; 17; 15; 12

Analysis of Variance for TIEMPO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:VELOCIDAD	37,75	2	18,875	2,49	0,1107
B:SALA	165,375	1	165,375	21,85	0,0002
INTERACTIONS					
AB	60,25	2	30,125	3,98	0,0370
RESIDUAL	136,25	18	7,56944		
TOTAL (CORRECTED)	399,625	23			

A continuación se muestran los gráficos de medias con intervalos LSD para los dos factores (con un nivel de confianza del 95%) y el gráfico de la interacción.

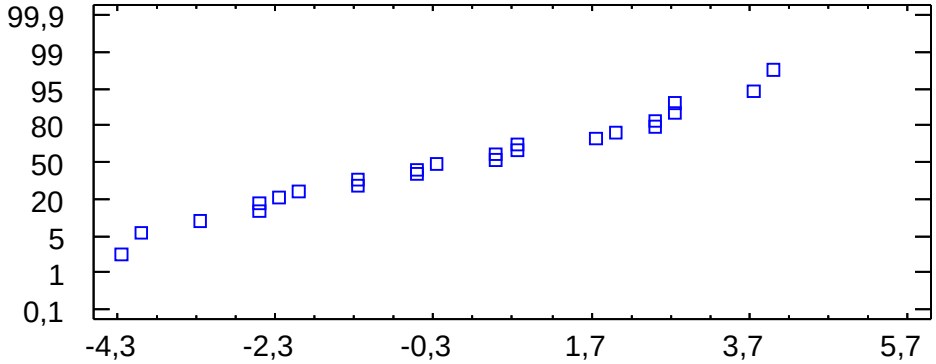


a) Con el equipamiento informático disponible, la biblioteca considera que, a nivel poblacional, el tiempo de consulta no depende de la velocidad del procesador. ¿Estás de acuerdo? Justifica convenientemente la respuesta con la(s) herramienta(s) estadísticas apropiadas y considerando $\alpha = 5\%$. *(2 puntos)*

b) Teniendo en cuenta lo que se deduce de los dos gráficos de medias con intervalos LSD, ¿qué información adicional aporta en este caso el gráfico de la interacción para describir cómo afecta el factor "sala" y la velocidad del procesador sobre el tiempo de consulta, a nivel poblacional? Considera α = 1%. *(2,5 puntos)*

c) Si un nuevo usuario desea elegir la sala y el ordenador donde previsiblemente los tiempos de consulta sean más bajos, ¿cuál sería la recomendación, considerando α=1%? ¿Cuál sería el tiempo medio estimado en dichas condiciones? *(2,5 puntos)*

d) La siguiente figura se ha obtenido con los residuos del modelo:

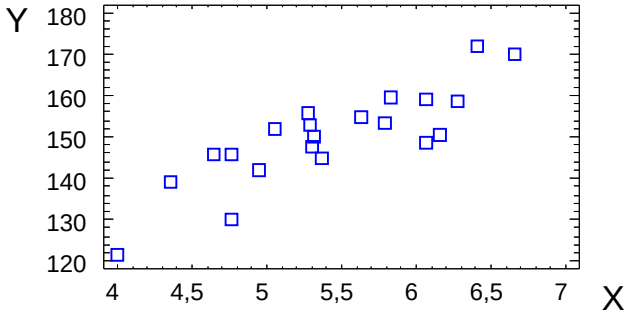


d.1) ¿Qué hipótesis del ANOVA se comprueba habitualmente a partir de esta figura? ¿Consideras que dicha hipótesis se cumple razonablemente en este caso? Justifica tu respuesta. *(1 punto)*

d.2) ¿Qué recomendarías hacer en caso de que dicha hipótesis no se cumpliese? *(1 punto)*

d.3) Además de esta hipótesis, ¿qué otros supuestos deben cumplir los datos para que sean fiables las conclusiones que se deducen del ANOVA? *(1 punto)*

3. Una empresa fabricante de cereales para el desayuno desea conocer la relación que permita predecir las ventas en función de los gastos en publicidad infantil en televisión (ambas variables, medidas en miles de euros). Se realiza un estudio en el que se reúnen los datos mensuales correspondientes a los últimos 21 meses. A partir de estos datos se ha obtenido el siguiente gráfico:



a) ¿Cuál sería la variable independiente X y la variable dependiente Y en este contexto? ¿Qué se deduce a nivel muestral a partir del gráfico? Justifica razonadamente tu respuesta. *(2 puntos)*

b) Con el programa *Statgraphics* se ha obtenido la siguiente tabla de resultados de un modelo de regresión lineal ajustado con los datos:

```
Regression Analysis - Linear model: Y = a + b*X
-------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
-------------------------------------------------------------------------------
Intercept        75,0224       10,7008        7,01093         0,0000
Slope            13,8411        1,9562         7,0755         0,0000
-------------------------------------------------------------------------------

                          Analysis of Variance
-------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio       P-Value
-------------------------------------------------------------------------------
Model                     1919,29      1      1919,29      50,06        0,0000
Residual                  728,415     19      38,3376
-------------------------------------------------------------------------------
Total (Corr.)              2647,7     20
```

¿Qué porcentaje de la variabilidad en las ventas mensuales está explicado por otras variables que no son el gasto mensual en publicidad, por ejemplo: precio, nivel de ventas y precio de productos competidores, etc.? *(2,5 puntos)*

c) Escribe la ecuación matemática del modelo de regresión planteado, estimando el valor de sus parámetros y su significación estadística (α = 5%). ¿Cuál es la interpretación práctica de los valores obtenidos para los parámetros estimados, dentro del contexto del problema? *(2.5 puntos)*

d) Suponiendo que se cumplen todas las hipótesis de un modelo de regresión lineal simple, ¿cuál es la probabilidad de obtener unas ventas mensuales superiores a 140.000 euros si se realiza un gasto mensual de 4.000 euros en publicidad infantil en televisión? ¿Cuál es el modelo estadístico de distribución condicional que es necesario considerar para calcular esta probabilidad?

(3 puntos)

SOLUCIÓN

1a) El riesgo de primera especie (α) se define como la probabilidad de rechazar la hipótesis nula cuando es cierta. En este caso, la hipótesis nula es que el proceso funciona correctamente, siendo la capacitancia (C) una variable con distribución Normal de media 20 y $\sigma = 0.2 \rightarrow H_0$: $C \approx N(20; 0.2)$. Esta hipótesis se rechaza cuando C<19.5 o bien si C>20.5. Por tanto, en este caso:

$\alpha = P[N(20; 0.2) < 19.5] + P[N(20; 0.2) > 20.5] = 2 \cdot P[N(20; 0.2) > 20.5] =$

$= 2 \cdot P[N(0; 1) > 0.5/0.2] = 2 \cdot 0.0062 =$ **0.0124**

El valor resultante es pequeño, lo cual tiene sentido (generalmente $\alpha \leq 5\%$).

1b1) El riesgo de primera especie (α) es un valor que tiene que decidirse a la hora de interpretar los resultados estadísticos. Dado que es la probabilidad de cometer un error, esa probabilidad se establece en un valor pequeño. Los valores más habituales son: 5% y 1%, y en alguna ocasión (en caso de tener pocos datos) podría admitirse hasta 10%, pero nunca superior. Por el contrario, si la cantidad de datos es muy grande, pueden considerarse valores menores, como 0.1%. En resumen, el rango más frecuente es: $\mathbf{0.001 \leq \alpha \leq 5\%}$.

1b2) El p-valor, llamado también "nivel de significación observado", es una probabilidad y por tanto fluctúa entre 0 y 1. Se define como el nivel de significación (α) más pequeño posible que puede escogerse, para el cual todavía se rechazaría la hipótesis nula con las observaciones actuales. Si p-valor < α se rechaza H_0 y en caso contrario se acepta. En la práctica nos podemos encontrar con cualquier p-valor entre 0 y 1, no es posible establecer a priori un rango de valores frecuentes, pues es muy habitual encontrar diferencias que no resultan estadísticamente significativas. No obstante, cuando se trabaja con grandes cantidades de datos, las diferencias tienden a ser estadísticamente significativas, de modo que en ese caso es más probable encontrar que $p < 0.05$.

1c1) Test de hipótesis: H_0: m = 20; H_1: m ≠ 20.

Cálculo del estadístico de contraste: $t_{calc} = \frac{\bar{x}-m_0}{s/\sqrt{n}} = \frac{20.15-20}{\sqrt{0.09}/\sqrt{15}} = 1.936$

Este parámetro sigue una distribución t de Student con 14 grados de libertad. El 95% de valores de esta distribución están comprendidos entre -2.145 y 2.145. Dado que t_{calc} queda dentro de este intervalo, se acepta H_0. Es decir, no hay suficiente evidencia, a partir de la muestra obtenida, para afirmar que el cambio efectuado en el proceso haya alterado la capacitancia media a nivel poblacional.

1c2) Intervalo de confianza para la desviación típica poblacional (α=5%):

$$\sigma \in \left[\sqrt{\frac{(n-1)\cdot s^2}{g_2}}; \sqrt{\frac{(n-1)\cdot s^2}{g_1}}\right] = \left[\sqrt{\frac{14\cdot 0.09}{26.119}}; \sqrt{\frac{14\cdot 0.09}{5.629}}\right] = [0.219; 0.473]$$

Siendo g_1 (5.629) y g_2 (26.119) el intervalo de valores de una distribución chi-cuadrado con 14 grados de libertad que comprende el 95% de los datos.

H_0: $\sigma = 0.2$; H_1: $\sigma \neq 0.2$; Como el valor de 0.2 está fuera del intervalo de confianza obtenido se rechaza la hipótesis nula: **no se puede admitir $\sigma = 0.2$**.

2a) El p-valor de la interacción es 0.037, inferior a α =0.05, de modo que efecto de la interacción resulta estadísticamente significativo. Esto implica que el efecto de la velocidad del procesador en el tiempo de consulta es distinto para cada sala. El gráfico de la interacción indica que en la sala A apenas se han observado diferencias muestrales en el tiempo para las tres velocidades ensayadas, mientras que en la sala B se observa un efecto lineal decreciente. Por tanto, la biblioteca no tiene razón: no es admisible considerar que, a nivel poblacional, el tiempo de consulta no dependa de la velocidad del procesador. Sí que depende, pues en la sala B a mayor velocidad cabe esperar un menor tiempo en promedio. Sin embargo, esto no sucede en la sala A.

2b) Del primer gráfico se deduce que el tiempo de consulta en la sala B es significativamente mayor que en la sala A. La tabla del ANOVA indica que el efecto simple del factor velocidad no resulta estadísticamente significativo (p=0.11 > 0.01), lo cual es coherente con el segundo gráfico pues los tres intervalos LSD se solapan. Considerando α=1%, el efecto de la interacción no es estadísticamente significativo ya que p=0.037 > 0.01. Por tanto, aunque a nivel muestral se observan diferencias en los tiempos entre las dos salas en función de la velocidad, no hay evidencia suficiente a partir del ANOVA para afirmar que estas diferencias se correspondan también a nivel poblacional, de modo que el efecto de la velocidad en el tiempo debe considerarse igual en ambas salas (es decir, como si se tratase de dos rectas paralelas en el gráfico de la interacción a nivel poblacional). En definitiva, el ANOVA revela que sólo el efecto simple del factor sala es estadísticamente significativo; esta conclusión es coherente con los dos gráficos de medias con intervalos LSD, de modo que el gráfico de la interacción no aporta ninguna información adicional a nivel poblacional.

No obstante, si los datos fueran analizados con regresión lineal múltiple, es posible que el efecto de la interacción resultase estadísticamente significativo para α=1%, de modo que la interpretación de resultados sería distinta.

2c) Esta pregunta considera también α=1% por lo que partimos del razonamiento anterior. Cabe esperar tiempos más bajos a nivel poblacional en la sala A. Ya que la interacción no se considera significativa ni tampoco el efecto simple de la velocidad del procesador, no hay evidencia suficiente para afirmar que a mayor velocidad los tiempos serán significativamente menores. De modo que la recomendación es emplear cualquier ordenador de la sala A. En dichas condiciones, el tiempo medio estimado será la media de todos los tiempos medidos en la sala A, que es de 13 minutos según el gráfico de medias.

2d1) En la figura se representa un papel probabilístico normal de los residuos del modelo. Este gráfico se emplea habitualmente para verificar la hipótesis de normalidad, ya que el ANOVA asume que todas las poblaciones involucradas en el estudio siguen un modelo normal de distribución estadística. En este caso los puntos se ajustan "razonablemente bien" a una línea recta, de modo que es admisible asumir una distribución normal de los residuos: no hay evidencia suficiente para rechazar la hipótesis de normalidad. Por otra parte, no se detecta ningún dato anómalo que requiera ser descartado.

2d2) Puede ser que la hipótesis de normalidad no se cumpla por varios motivos:
(**a**) Si los residuos se ajustan razonablemente bien a una recta en el papel probabilístico normal pero algunos pocos datos se alejan claramente de la recta, se consideran datos anómalos que es necesario eliminar, o bien corregir la anomalía si es posible.
(**b**) Si los residuos sugieren una distribución asimétrica se recomienda transformar los datos. En caso de asimetría positiva, la transformación logaritmo o raíz cuadrada son las más habituales. En caso de asimetría negativa, hay que conseguir primero una asimetría positiva (por ejemplo, cambiando el signo y adicionando un valor constante apropiado).
(**c**) No siempre puede conseguirse que se cumpla la hipótesis de normalidad, en caso de que los residuos revelen mezcla de distribuciones, presencia de datos truncados, etc.

2d3) Además de la hipótesis de normalidad, el ANOVA asume la hipótesis de homocedasticidad (las distintas poblaciones involucradas en el estudio tienen la misma varianza), y la hipótesis de independencia (las distintas observaciones experimentales son independientes entre sí, es decir, se han obtenido aleatoriamente de modo que cada observación tiene la misma probabilidad de aparecer en la muestra).

3a) Las ventas mensuales dependen de los gastos en publicidad, de modo que la variable dependiente Y será "ventas mensuales de la empresa" y la variable independiente X será "gastos en publicidad infantil en televisión".
A nivel muestral se deduce una relación directa entre ambas variables (correlación positiva): a mayor valor de gasto en publicidad cabe esperar un mayor valor de las ventas en promedio, lo cual tiene sentido. El grado de correlación puede describirse como "moderado". Se observa un efecto lineal que muy probablemente es estadísticamente significativo, si bien sería necesario realizar un análisis de regresión simple para verificar esta hipótesis.

3b) El porcentaje de variabilidad de la variable Y explicada por la variable X se calcula como SS_{modelo} / SS_{total}. Por tanto, el porcentaje de variabilidad de Y **no** explicado por X será: $SS_{residual}$ / SS_{total} = 728.415/2647.7 = 0,2751 = **27.51%**.

3c) La ecuación matemática del modelo de regresión planteado es:

$$\text{Ventas mensuales} = 75.022 + 13.8411 \cdot \text{gastos}_{public}$$

El valor estimado de los dos parámetros (pendiente y ordenada en el origen) viene dado en la tabla de resultados de Statgraphics. Para ambos parámetros, su p-valor es prácticamente cero (menor que α) y por tanto ambos resultan estadísticamente significativos, es decir, pueden considerarse distintos de cero a nivel poblacional.

- Interpretación práctica de la ordenada: es el valor estimado de las ventas mensuales que cabe esperar en promedio si la empresa no gasta en publicidad infantil en televisión. No obstante, el valor X=0 queda bastante alejado del rango de valores de X ensayados (de 4 a 6.5), de modo que la predicción del modelo para X=0 no es fiable.

- Interpretación práctica de la pendiente: por cada mil euros que la empresa decida incrementar cada mes en gastos de publicidad infantil en televisión, cabe esperar un incremento promedio de ventas mensuales de 13841.1 euros.

3d) Asumiendo que se cumplen todas las hipótesis del modelo, la distribución de las ventas condicionado a un gasto de 4 mil euros (es decir, X=4 ya que las unidades son en miles de euros), será una distribución normal cuya media se obtiene a partir del modelo de regresión y cuya varianza es la varianza residual:

$E(Y/X=4) = 75.022 + 13.8411 \cdot 4 = 130.3868$

$\sigma^2(Y/X=4) = s^2_{residual}$ = Cuadrado Medio residual = 38.3376 (ver tabla). Aplicando la raíz cuadrada: $s_{resid} = 6.1917$.

Por tanto: $\mathbf{Y/X=4 \approx N\ [m=130.39,\ \sigma = 6.19]}$.

En estas condiciones, la probabilidad de que las ventas superen 140 mil euros:

$P[(Y>140)/(X=4)] = P[N(130.39; 6.19) > 140] =$

$= P[N(0;1) > (140-130.39)/6.19] = P[N(0; 1)> 1.5526] =$ **0,0603**