Estadística II
Examen final junio 19/6/19
Curso 2018/19 – Soluciones

Duración del examen: 2 h y 15 min

1. (2,5 puntos) Una empresa verifica regularmente que el peso las unidades elaboradas de uno de sus productos no sea inferior a 240 gramos. En caso de detectar una desviación significativa la empresa deberá reajustar sus procesos de producción, con el coste correspondiente.

   Suponemos que este peso sigue una distribución normal, con una desviación típica de 20 gramos.

   a) (1 punto) Se ha tomado una muestra de $n = 36$ unidades del producto, obteniendo un peso promedio de $\bar{x} = 234$ gramos. Realiza un contraste de hipótesis para determinar si la empresa debe reajustar sus procesos. Calcula el p-valor del contraste e indica tu conclusión.

   b) (0,5 puntos) Indica si las afirmaciones siguientes son verdaderas o falsas, razonando tu respuesta:

   i. Si aumenta el valor de la desviación típica de la población manteniendo los demás valores constantes, el p-valor del contraste anterior aumentará.

   ii. Si para el contraste anterior aumenta el tamaño muestral pero se mantienen todos los valores restantes, incluyendo el nivel de significación α, el tamaño de la región de rechazo definida en términos del valor del estadístico del contraste Z, disminuirá.

   Se quiere aplicar la siguiente regla de decisión: seleccionar una muestra de tamaño n de este producto y rechazar la hipótesis nula si el promedio de los pesos en la muestra es menor de 230 gramos.

   c) (1 punto) Supongamos que el valor correcto del peso promedio es de 220 gramos. Si se emplean muestras de tamaño $n = 9$, ¿cuál es la probabilidad de un error de Tipo II para esta regla de decisión?

   Solución:

   a) Tenemos que $n = 36$ y $\bar{x} = 234$. Si denotamos por μ el peso promedio del producto, el contraste que se quiere llevar a cabo es:

   $$H_0 : \mu \geq 240 = \mu_0$$
   $$H_1 : \mu < 240.$$

   El estadístico del contraste es

   $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

   y el p-valor de este contraste vendrá dado por

   $$\text{p-valor} = P\left(Z < \frac{234 - 240}{20/\sqrt{36}}\right) = P(Z < -1{,}8) = 0{,}0359.$$

   Por tanto, rechazaríamos la hipótesis nula para cualquier nivel de significación superior al 3,6 %, y en particular la rechazaríamos para un nivel de significación del 5 % (pero no lo haríamos para un 1 %, por ejemplo).

   b) Las respuestas son:

   i. Verdadero. Si z denota el valor del estadístico del contraste, en el caso indicado se cumple que $0 > z_1 > z_0$, donde z_0 es el valor correspondiente a la desviación típica dada en el enunciado, y z_1 es el valor cuando aumenta dicha desviación típica (teniendo en cuenta que $z_i < 0$). Como consecuencia, p-valor$(z_0) = P(Z < z_0) < P(Z < z_1) =$ p-valor(z_1).

   ii. Falso. En nuestro caso, el valor crítico que define la región de rechazo es $z_{1-\alpha}$, que no depende de n, por lo que la región de rechazo no cambiaría.

c) Para $n = 9$ y $\mu = 220 < \mu_0$, la probabilidad de un error de Tipo II vendrá dada por

$$\begin{aligned} P(\text{no rechazar } H_0 \mid \mu = 220) &= P(\bar{X} > 230 \mid \mu = 220) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{230 - 220}{20/\sqrt{9}}\right) \\ &= P(Z > 1{,}5) = 1 - \Phi(1{,}5) = 0{,}0668, \end{aligned}$$

donde $Z \sim N(0, 1)$ y Φ denota la función de distribución de Z.

2. (2,5 puntos) En los ensayos de un nuevo producto antes de lanzarlo al mercado se ha detectado que una de las características del mismo excede los estándares de diseño fijados.

   Los responsables del proceso de fabricación han identificado dos posibles modificaciones que podrían solucionar este problema. Para escoger la mejor alternativa se han fabricado 14 unidades del producto empleando la primera modificación y 9 empleando la segunda. Se ha medido el valor de la característica de interés y se ha obtenido la siguiente información:

Mod 1	*Mod 2*
26	25
22	19
23	23
30	20
20	25
25	18
26	25
25	25
24	20
27	
27	
24	
30	
30	

Prueba F para varianzas de dos muestras

	Mod 1	*Mod 2*
Media	25,64285714	22,22222222
Varianza	9,17032967	8,694444444
Observaciones	14	9
Grados de libertad	13	8
F	1,054734403	
P(F<=f) una cola	*****	
Valor crítico para F (una cola)	3,259019235	

   Se desea que los valores de esta característica sean reducidos. Suponiendo que el valor medido en cada producto sigue una distribución normal, se pide que contestes a las preguntas siguientes:

   a) (1 punto) ¿Podríamos asumir la igualdad de varianzas de las dos modificaciones con una confianza del 90 %?

   b) (1 punto) Si el objetivo deseado es reducir el valor de la característica, realiza un contraste de hipótesis con un nivel de significación del 5 % para comprobar si en promedio la segunda modificación proporciona valores menores que la primera modificación. Indica las hipótesis del contraste, el estadístico de contraste, la región crítica o de rechazo de la hipótesis nula, tu decisión y tu interpretación.

   c) (0,5 puntos) De las siguientes afirmaciones, contesta cuál es la correcta:

      I. La probabilidad del error Tipo I representa:
         A. La probabilidad de rechazar la hipótesis nula siendo verdadera.
         B. La probabilidad de aceptar la hipótesis nula siendo verdadera.
         C. La probabilidad de aceptar la hipótesis nula siendo falsa.
      II. La probabilidad del error Tipo II representa:
         A. La probabilidad de rechazar la hipótesis nula siendo verdadera.
         B. La probabilidad de aceptar la hipótesis nula siendo verdadera.
         C. La probabilidad de aceptar la hipótesis nula siendo falsa.

Solución:

a) Hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Estadístico de contraste:

$$F = \frac{s_1^2}{s_2^2} = \frac{9{,}17}{8{,}69} = 1{,}055$$

Región crítica o de rechazo de H_0:

$$\begin{aligned} \mathrm{RR}_{0,1} &= \{F : F < F_{n_x-1,n_y-1;1-\alpha/2}\} \cup \{F : F > F_{n_x-1,n_y-1;\alpha/2}\} \\ \mathrm{RR}_{0,1} &= \{F : F < F_{13,8;0,95}\} \cup \{F : F > F_{13,8;0,05}\} \\ &= \{F : F < F_{13,8;0,95}\} \cup \{F : F > 3{,}259\}. \end{aligned}$$

Decisión: Como $1 < 1{,}055 < 3{,}259$, no rechazamos la hipótesis nula. Obsérvese que la cola izquierda no es relevante para este análisis porque el valor del estadístico es mayor que 1.

Interpretación: para un nivel de confianza del 90 %, podemos asumir la igualdad de las varianzas poblacionales.

b) Planteamiento de las hipótesis: Si el segundo proceso es más eficaz será porque, en media, el valor de la característica de interés es más pequeño que con el primer proceso.

Hipótesis: concluimos que el segundo proceso es mejor cuando rechazamos la hipótesis nula.

$$\begin{aligned} H_0 &: \mu_1 \leq \mu_2 \\ H_1 &: \mu_1 > \mu_2 \end{aligned}$$

Estadístico de contraste: Se trata de un contraste para medias, con muestras independientes y varianzas poblacionales desconocidas pero supuestamente iguales (por el apartado anterior). Por tanto,

$$t = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim t_{n_x+n_y-2}$$

De los datos de las muestras tenemos que

$$s_P^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2} = \frac{13 \times 9{,}170 + 8 \times 8{,}694}{21} = 8{,}989$$

Este estadístico toma el valor (bajo H_0):

$$t = \frac{25{,}64 - 22{,}22 - 0}{\sqrt{8{,}989}\sqrt{\frac{1}{14} + \frac{1}{9}}} = 2{,}67$$

Región crítica o de rechazo de H_0:

$$\mathrm{RR}_{0,1} = \{t : t > t_{n_x+n_y-2;\alpha}\} = \{t : t > t_{21;0,05}\} = \{t : t > 1{,}721\}$$

Decisión: Como $2{,}67 > 1{,}721$, tenemos evidencia suficiente para poder rechazar la hipótesis nula.

Interpretación: Con una confianza de 95 %, podemos afirmar que el segundo proceso es más eficiente que el primer proceso.

c) Las respuestas correctas son:

I. La probabilidad del error Tipo I representa:

A. La probabilidad de rechazar la hipótesis nula siendo verdadera.

II. La probabilidad del error Tipo II representa:

C. La probabilidad de aceptar la hipótesis nula siendo falsa.

3. (5 puntos) Se quiere explicar la demanda mensual de un bien en una localidad (variable Y) en función del gasto en publicidad en determinados medios (variable X_1). El modelo estimado a partir de una muestra de valores observados en un periodo de 16 meses ($n = 16$) es el siguiente:

$$\hat{Y} = -0{,}73 + 3{,}59X_1,$$

siendo 0,579 y 1,347 las medias respectivas de X_1 e Y, así como 0,0155 y 0,221 sus respectivas cuasivarianzas. La varianza residual del modelo es 0,0230.

a) (1 punto) Completa la siguiente tabla correspondiente a salidas de Excel obtenidas para la regresión de Y (variable dependiente) sobre X_1 (variable independiente).

	Coeficientes	*Error típico*
Intercepción		
X1		

b) (0,5 puntos) Contrasta que el modelo de regresión lineal de Y sobre X_1 es estadísticamente significativo, para un nivel de significación del 1 %.

c) (1 punto) Completa la tabla ANOVA para este modelo de regresión lineal simple, rellenando las casillas de la tabla siguiente:

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*
Regresión				
Residuos				
Total				

d) (0,5 puntos) Indica si las siguientes afirmaciones son verdaderas o falsas, justificando tu respuesta:

I. En el modelo de regresión lineal simple, si aumenta la variabilidad de la variable explicativa pero la variabilidad de los residuos no cambia, la varianza del estimador de la pendiente aumenta.

II. Un cambio en las unidades de medida de la variable X_1 no afecta a la estimación de la ordenada en el origen.

Se ha llevado a cabo una regresión múltiple en la que se añadieron dos nuevas variables explicativas: X_2, que representa el número de puntos de venta donde se puede adquirir el producto, y X_3, que mide la evolución del precio del producto. Los resultados obtenidos empleando Excel para este modelo de regresión lineal múltiple son los siguientes:

ANÁLISIS DE VARIANZA

	Grados de Libertad	*Suma de Cuadrados*	*Cuadrados Medios*	*F*
Regresión	3	3,2891	1,09636	494,40
Residuos	12	0,0266	0,00222	
Total	15	3,3157		

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	-0,09691	0,42423	-0,22843	0,82315
X1	0,77780	0,36655	2,12197	0,05534
X2	1,71298	0,15054	11,37856	0,00000
X3	-0,28718	0,15255	-1,88256	0,08422

e) (0,5 puntos) Te indican que el p-valor asociado al test F de este modelo es $7{,}81\,10^{-13}$ ¿Qué interpretación tiene este valor?

f) (0,5 puntos) Calcula un intervalo de confianza al 95 % para el parámetro asociado a la variable X_3. Interpreta el resultado.

g) (1 punto) Calcula el valor del coeficiente de determinación del modelo, y valora si la incorporación de las variables X_2 y X_3 ha mejorado el grado de explicatividad del mismo. *Para esta comparación utiliza los valores del coeficiente de determinación R^2, aunque en la práctica sería más adecuado comparar los valores del coeficiente de determinación ajustado $\bar{R}^2$.*

Solución:

a) Los valores pedidos se pueden obtener a partir de la información que nos han proporcionado para el modelo de regresión indicado. En particular:

$$\hat{\beta}_0 = -0{,}732, \quad \hat{\beta}_1 = 3{,}590,$$

También nos dan los valores de

$$s_R^2 = 0{,}0230, \quad \bar{x} = 0{,}579, \quad s_x^2 = 0{,}0155.$$

Con estos valores podemos calcular los errores típicos de los estimadores, dados por

$$\begin{aligned} s(\hat{\beta}_0) &= \sqrt{s_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)} = \sqrt{0{,}023 \left(\frac{1}{16} + \frac{0{,}579^2}{15 \times 0{,}0155} \right)} = 0{,}186 \\ s(\hat{\beta}_1) &= \sqrt{\frac{s_R^2}{(n-1)s_x^2}} = \sqrt{\frac{0{,}023}{15 \times 0{,}0155}} = 0{,}315 \end{aligned}$$

y con todos ellos completamos la tabla indicada:

	Coeficientes	*Error típico*
Intercepción	-0,732	0,186
X1	3,590	0,315

b) El contraste de significación del modelo que nos piden es

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

El estadístico para dicho contraste es

$$T = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t_{n-2}$$

Bajo la hipótesis nula y para la muestra indicada, su valor se puede obtener como el valor del coeficiente $\hat{\beta}_1$ dividido por su error típico,

$$t = \frac{3{,}590}{0{,}315} = 11{,}40$$

La región de rechazo de este contraste viene dada por

$$\mathrm{RR}_{0,01} = \{t < -t_{14;0,005}\} \cup \{t > t_{14;0,005}\} = \{t < -2{,}977\} \cup \{t > 2{,}977\}.$$

El valor del estadístico está en esta región de rechazo, $11{,}40 > 2{,}977$, por lo que concluimos que para el nivel de significación indicado el modelo de regresión es significativo.

c) Para completar la tabla ANOVA comenzamos por calcular los valores de las sumas de cuadrados,

$$\begin{aligned} \mathrm{SCT} &= \sum_i (y_i - \bar{y})^2 = (n-1)s_y^2 = 15 \times 0{,}221 = 3{,}316 \\ \mathrm{SCR} &= \sum_i e_i^2 = (n-2)s_R^2 = 14 \times 0{,}0230 = 0{,}322 \\ \mathrm{SCM} &= \sum_i (\hat{y}_i - \bar{y})^2 = \mathrm{SCT} - \mathrm{SCR} = 2{,}993 \end{aligned}$$

A partir de estos valores, y teniendo en cuenta que los grados de libertad a considerar son $n - 1 = 15$ los totales, $n - 2 = 14$ para los residuos y 1 para el modelo, obtenemos la tabla

ANÁLISIS DE VARIANZA				
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*
Regresión	1	2,993	2,993	130,010
Residuos	14	0,322	0,023	
Total	15	3,316		

d) Las respuestas a las dos preguntas planteadas son:

I. Falso. La varianza del estimador de la pendiente (el cuadrado de su error típico) viene dada por

$$s^2(\hat{\beta}_1) = \frac{s_R^2}{(n-1)s_x^2}$$

Si aumenta la variabilidad de X, y por tanto s_x^2 aumenta, la varianza del estimador disminuye.

II. Verdadero. La ordenada en el origen se estima como

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \bar{y} - \frac{\text{cov}(x,y)}{s_x^2}\bar{x}.$$

Un cambio en las unidades de medida de X afecta a los valores de $\text{cov}(x,y)$, de s_x^2 y de $\bar{x}$, pero todos estos cambios se cancelan entre sí, ya que $\text{cov}(x,y)$ y $\bar{x}$ dependen linealmente de X, y s_x^2 lo hace cuadráticamente.

El valor de $\hat{\beta}_0$ está directamente asociado a valores de Y (cuando $X = 0$), y solo debiera depender de las unidades de medida de Y.

e) El p-valor asociado al test F es el p-valor del contraste de significación global del modelo de regresión lineal múltiple. Un valor tan pequeño indica que este modelo de regresión lineal es significativo para (casi) cualquier nivel de significación.

f) El intervalo a calcular se puede obtener del valor estimado del parámetro y su error típico (que están indicados en la salida de Excel) como

$$\hat{\beta}_3 \pm t_{n-k-1;\alpha/2}\, s(\hat{\beta}_3),$$

Teniendo en cuenta que el estimador $\hat{\beta}_2$ sigue una distribución t de Student con $n-k-1 = 16-3-1 = 12$ grados de libertad y que $t_{12;0,025} = 2{,}179$, el intervalo que se obtiene es

$$-0{,}287 \pm 2{,}179 \times 0{,}153 = [-0{,}620; 0{,}045]$$

Como el valor 0 está dentro del intervalo, podemos concluir que la variable X_3 no es estadísticamente significativa para un nivel de significación del 5 %.

g) El coeficiente de determinación (múltiple) del modelo de regresión múltiple se puede obtener como

$$R_m^2 = 1 - \frac{\text{SCR}_m}{\text{SCT}_m} = 1 - \frac{0{,}0266}{3{,}3157} = 0{,}992$$

Como consecuencia, la proporción de la variabilidad en la respuesta explicada por el modelo de regresión es el 99,2 %.

Si comparamos este valor con el correspondiente al modelo de regresión simple (empleando una fórmula similar con los valores obtenidos de la tabla ANOVA para el modelo de regresión lineal simple), tenemos que

$$R_s^2 = 1 - \frac{\text{SCR}_s}{\text{SCT}_s} = 1 - \frac{0{,}322}{3{,}316} = 0{,}903$$

Por tanto, hay una mejora importante en la explicatividad del modelo, porque el porcentaje de variabilidad no explicada se ha reducido del 10 % a menos del 1 %.

Aunque no nos lo piden, al tratarse de modelos con diferentes números de parámetros sería conveniente emplear el coeficiente de determinación ajustado, en el que corregimos por la diferencia en grados de libertad de los modelos. Se define como

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$

En nuestro caso tenemos que

$$\bar{R}_m^2 = 1 - (1 - 0{,}992)\frac{15}{12} = 0{,}990, \quad \bar{R}_s^2 = 1 - (1 - 0{,}903)\frac{15}{14} = 0{,}896$$

y la reducción en la variabilidad no explicada cuando incorporamos las variables explicativas adicionales sigue siendo muy elevada si la medimos en función de este coeficiente ajustado.