Estadística II
Examen final enero 19/1/17
Curso 2016/17 – Soluciones

Duración del examen: 2 h y 15 min

---

1. (3 puntos) El Instituto para la Diversificación y Ahorro de la Energía (IDAE) ha publicado un estudio sobre la energía consumida anualmente por distintos electrodomésticos. En dicho estudio se afirma que un aspirador doméstico consume en promedio 46 kilovatios-hora por año. En la realización de otro estudio de 12 hogares, los aspiradores consumen un promedio de 42 kilovatios-hora anuales, con una cuasidesviación típica de 11,9 kilovatios-hora. Se pide que contestes las siguientes preguntas, indicando los supuestos que debas asumir:

   a) (1,5 puntos) ¿Implican estos datos que, para un nivel de significación de 0,05, los aspiradores domésticos consumen en promedio menos de 46 kilovatios-hora por año? Indica tus hipótesis nula y alternativa, y calcula el p-valor del contraste. Incluye una breve interpretación de tu conclusión.

   b) (1 punto) Calcula un intervalo de confianza al 95 % para la desviación típica de la población de consumos de energía por aspiradores.

   c) (0,5 puntos) Queremos aplicar la siguiente regla de decisión para un contraste basado en la muestra de hogares anterior: la afirmación de que la varianza de la población de consumos de aspiradores es menor o igual a 120 se rechaza siempre que la cuasivarianza de la muestra de los 12 valores supera 144. Calcula la probabilidad de un error de Tipo I al aplicar esta regla.

   Solución: Debemos suponer que la población de "consumos de energía anuales en kilovatios-hora de aspiradores domésticos" sigue una distribución normal, y que tenemos una muestra aleatoria simple.

   a) De nuestros datos,
   $$n = 12, \quad \bar{x} = 42, \quad s = 11{,}9$$
   El contraste se define como
   $$\begin{aligned} H_0 &: \ \mu \geq 46 \equiv \mu_0 \\ H_1 &: \ \mu < 46 \end{aligned}$$
   El estadístico del contraste y su distribución son
   $$T = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} \sim t_{n-1}.$$
   Su valor para nuestra muestra es
   $$t = \frac{42 - 46}{11{,}9/\sqrt{12}} = -1{,}16,$$
   y de las tablas $0{,}1 <$ p-valor $< 0{,}15$.
   Podemos concluir que, para un nivel de significación $\alpha = 0{,}05$, no disponemos de evidencia que nos permita afirmar que los aspiradores consumen en promedio menos de 46 kilovatios-hora anualmente.

   b) El intervalo de confianza para la varianza se define como
   $$\mathrm{IC}_{0,95}(\sigma^2) = \left[\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \,;\, \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}} \right] = \left[\frac{(11)11{,}9^2}{21{,}9} \,;\, \frac{(11)11{,}9^2}{3{,}82} \right] = [71{,}13 \,;\, 407{,}78] \,.$$
   Y para la desviación típica tenemos
   $$\mathrm{CI}_{0,95}(\sigma) = [8{,}43 \,;\, 20{,}19] \,.$$

c) El contraste que nos describen viene dado por

$$\begin{aligned} H_0 &: \sigma^2 \leq 120 \equiv \sigma_0^2 \\ H_1 &: \sigma^2 > 120 \end{aligned}$$

Si la regla de decisión implica el rechazo de H_0 cuando $s^2 > 188$, se cumple que

$$\begin{aligned} P(\text{Type I error}) &= P(\text{reject } H_0 \,|\, H_0 \text{ is true}) = P(s^2 > 144 \,|\, \sigma^2 = 120) \\ &= P\left(\frac{(n-1)s^2}{\sigma^2} > \frac{(n-1)144}{\sigma^2} \,\middle|\, \sigma^2 = 120\right) = P\left(\chi^2_{11} > \frac{11 \times 144}{120}\right) \\ &= P\left(\chi^2_{11} > 13{,}2\right) \approx 0{,}28 \end{aligned}$$

2. (3 puntos) Un investigador cree que las personas que consumen café tienden a hacerlo con más frecuencia durante periodos de tensión. En un grupo de 10 estudiantes seleccionados aleatoriamente entre consumidores habituales de café, se ha comparado el número de tazas de café que se toman en un periodo habitual (sin exámenes) de dos semanas, con el número de tazas que se toman en las dos semanas de un periodo de exámenes. La siguiente tabla muestra los datos recogidos:

Estudiante	1	2	3	4	5	6	7	8	9	10
Tazas habituales	9	16	23	12	30	8	14	21	11	14
Tazas exámenes	13	15	23	17	32	16	18	23	15	14

a) (1,5 puntos) Plantea un contraste unilateral para determinar si el número medio de tazas de café consumido habitualmente en dos semanas aumenta significativamente en los periodos de dos semanas de exámenes. Indica tus hipótesis nula y alternativa, los supuestos para el contraste, el estadístico a utilizar, su distribución y la regla de decisión que utilizarías para obtener una conclusión para el contraste. Utiliza un nivel de significación del 1 %.

b) (0,5 puntos) Calcula el p-valor del contraste y explica tu conclusión.

c) (1 punto) Para una variante en la que se quiere contrastar si las medias de los consumos en los dos periodos son iguales, se ha obtenido la siguiente tabla de Excel (de la que se han eliminado algunos valores). Indica el valor de la casilla marcada como "XXXXX", y explica cuál es tu conclusión en base a ese valor.

Prueba t para medias de dos muestras		
	Tazas habituales	*Tazas exámenes*
Media	15,8	18,6
Varianza	47,95555556	34,04444444
Observaciones	10	10
Coeficiente de correlación de Pearson	0,921762236	
Diferencia hipotética de las medias	0	
Grados de libertad	9	
Estadístico t	-3,230769231	
P(T<=t) una cola	0,005154969	
Valor crítico de t (una cola)	2,821437925	
P(T<=t) dos colas	XXXXX	
Valor crítico de t (dos colas)	-	

Solución.

a) Sean X: número regular de tazas de café consumidas en dos semanas habituales, e Y: número de tazas consumidas en las dos semanas de exámenes.

El contraste pedido se puede plantear como sigue:

- Hipótesis nula y alternativa:

$$\begin{aligned} H_0 &: \mu_Y = \mu_X \\ H_1 &: \mu_Y > \mu_X \end{aligned}$$

- Supuestos para realizar el contraste: Dos m.a.s. La diferencia $D = Y - X$ se distribuye como una normal.
- Estadístico a utilizar:
$$\frac{\bar{D} - \mu_D}{s_D/\sqrt{n}}$$
- Distribución del estadístico: t_{n-1}
- Regla de decisión: Rechazar H_0 si y solo si $\frac{\bar{D}-\mu_D}{s_D/\sqrt{n}} > t_{n-1;\alpha}$.

b) Calculamos las diferencias $d_i = y_i - x_i$:

Estudiante	1	2	3	4	5	6	7	8	9	10
X	9	16	23	12	30	8	14	21	11	14
Y	13	15	23	17	32	16	18	23	15	14
D	4	-1	0	5	2	8	4	2	4	0

Obtenemos $\bar{D} = 28/10 = 2{,}8$, $s_D^2 = 7{,}511$ y $s_D = 2{,}741$.
El valor del estadístico del contraste es:
$$t = \frac{2{,}8}{2{,}741/\sqrt{10}} = 3{,}231$$

El p-valor viene dado por p-valor $= P(t_9 > 3{,}231)$ y ese valor en las tablas está en el intervalo $(0{,}005; 0{,}01)$.
Como conclusión, este p-valor es menor que el nivel de significación del 1 % y por tanto podemos rechazar la hipótesis nula y concluir que en las dos semanas de exámenes el número de tazas de café consumidas aumenta respecto a su consumo habitual.

c) El valor que nos piden es el p-valor de un contraste bilateral, y conocemos por la salida de Excel el p-valor del contraste unilateral. El valor pedido será el doble del p-valor unilateral, es decir:
$$\text{XXXXX} = P(|T| > t) = P(|t_9| > t) = 2P(t_9 > t) = 2 \times 0{,}005154969 = 0{,}010309938$$

Dado este p-valor, rechazamos la igualdad de los consumos medios para cualquier nivel de significación superior a 0,0103 (al 1,03 %). En particular, y para $\alpha = 0{,}01$, no rechazamos la hipótesis nula, y concluimos que no tenemos suficiente evidencia para pensar que el consumo de café cambia en el periodo de exámenes.

3. (4 puntos) El Centro de Consulta Estadística de una universidad analizó, a petición de un departamento de la universidad, datos de gasto en rebajas frente a ingresos totales de familias. Las variables de interés fueron el gasto en rebajas de cada familia (Y en euros) y sus ingresos anuales totales (X en miles de euros). Uno de los estudios de interés consistía en obtener una ecuación de regresión lineal (Y en función de X) y determinar si existe una relación lineal significativa entre el gasto en rebajas y los ingresos totales. Con este fin, se han recogido 19 datos de unidades familiares para cada una de las variables, obteniéndose los siguientes resultados:
$$\sum_{i=1}^{19} x_i = 226{,}3, \quad \sum_{i=1}^{19} y_i = 58{,}97, \quad \sum_{i=1}^{19} x_i^2 = 2740{,}15, \quad \sum_{i=1}^{19} y_i^2 = 202{,}14715$$
$$\sum_{i=1}^{19} x_i y_i = 700{,}2435, \quad \sum_{i=1}^{19} e_i^2 = 19{,}0225.$$

a) (1 punto) Obtén la tabla ANOVA para la variable Y.

b) (0,5 puntos) Lleva a cabo un contraste al 5 % de significación para analizar la influencia de los ingresos totales en los gastos en rebajas.

c) (0,5 puntos) Calcula el coeficiente de determinación e interprétalo.

d) (1 punto) Estima el gasto en rebajas puntual para el caso en que los ingresos totales sean de 11,5 miles de euros. Proporciona, con un nivel de confianza del 95 %, un intervalo de confianza para dicha predicción.

e) (0,5 puntos) Justificando tu respuesta de forma precisa, responde a la siguiente pregunta: ¿Cambiarían los estimadores de mínimos cuadrados del modelo anterior si los ingresos totales se expresaran en euros y el gasto en rebajas se midiera en miles de euros?

f) (0,5 puntos) Indica si las siguientes afirmaciones sobre el modelo de regresión lineal múltiple $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ son verdaderas o falsas; razona tu respuesta.

1) La significación global del modelo de regresión se contrasta a partir de un estadístico que sigue una distribución F de Fisher con 2 y $n-3$ grados de libertad.
2) La significación de la variable x_2 en el modelo se contrasta con un estadístico que sigue una distribución t de Student con $n-2$ grados de libertad.

Solución

a) De los datos tenemos que SCR = 19,0225. También se cumple que:

$$\text{SCT} = \sum_{i=1}^{19} (y_i - \bar{y})^2 = (19-1)\, s_y^2 = \sum_{i=1}^{19} y_i^2 - 19 \times \bar{y}^2 = 19{,}1229.$$

De esta manera, la tabla ANOVA queda como:

Fuente	Suma de cuadrados	G.L.	Cuadrado medio	Razón-F
Modelo	0,10035	1	0,10035	0,0898
Residuos	19,0225	17	1,11897	
Total	19,1229	18		

El valor crítico del contraste de significación de este modelo es $F_{1,17;0,05} = 4{,}451$.

b) Dado que razón-F $< F_{1,17;0,05}$, se puede concluir que no podemos rechazar la hipótesis nula a un nivel de significación del 5 %, lo que quiere decir que no hay influencia entre el gasto en rebajas y los ingresos totales.

c) El coeficiente de determinación viene dado por:

$$R^2 = \frac{\text{SCM}}{\text{SCT}} = \frac{0{,}10035}{19{,}1229} = 0{,}00525$$

Por lo tanto podemos decir que con los ingresos totales sólo explicamos un 0,52 % de la variabilidad del gasto en rebajas.

d) Los parámetros de la recta de regresión son:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\text{cov}\,(x,y)}{s_x^2} = \frac{\sum_{i=1}^{19} x_i y_i - 19\bar{x}\bar{y}}{\sum_{i=1}^{19} x_i^2 - 19\bar{x}^2} = -0{,}0473 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 3{,}6674. \end{aligned}$$

La estimación puntual pedida para $x_0 = 11{,}5$ será:

$$\hat{y}_0 = 3{,}6674 - 0{,}0473 \times 11{,}5 = 3{,}1235.$$

Para obtener el intervalo de confianza, primero recordamos que la varianza residual también aparece en la tabla ANOVA, y su valor es

$$s_R^2 = \frac{\sum_{i=1}^{19} e_i^2}{19-2} = 1{,}119.$$

Ahora, empleamos la fórmula correspondiente al intervalo para una predicción:

$$\begin{aligned} \text{IC}_{0,05}(y_0) &= \hat{y}_0 \pm t_{17;0,025} \sqrt{s_R^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)} \\ &= 3{,}1235 \pm 2{,}1098 \sqrt{1{,}119 \left(1 + \frac{1}{19} + \frac{(11{,}5 - 11{,}91)^2}{(19-1) \times 2{,}489}\right)} \\ &= (0{,}8296; 5{,}4174)\,. \end{aligned}$$

e) Las nuevas variables serían $x^* = ax$ y $y^* = by$. Las nuevas estimaciones de los parámetros serían:

$$\hat{\beta}_1^* = \frac{b}{a}\hat{\beta}_1 \text{ y } \hat{\beta}_0^* = b\hat{\beta}_0.$$

Si x^* se mide en euros, entonces $a = 1000$ ya que x está en miles de euros. Además, si y^* se mide en miles de euros, entonces $b = 1/1000$ ya que y está en euros. Por lo tanto, los valores de los estimadores quedan como:

$$\hat{\beta}_1^* = 10^{-6}\hat{\beta}_1 \text{ y } \hat{\beta}_0^* = 10^{-3}\hat{\beta}_0.$$

4. Las respuestas a las preguntas anteriores son:

a) VERDADERO. La significación global del modelo de regresión múltiple se contrasta a partir del valor de la razón-F obtenida en la tabla ANOVA, por ejemplo, como

$$\text{Razón-F} = \frac{\text{SCM}/k}{\text{SCR}/(n-k-1)}.$$

Este estadístico sigue una distribución F de Fisher con k y $n-k-1$ grados de libertad. En nuestro caso, como $k = 2$, los grados de libertad son 2 y $n-3$.

b) FALSO. La significación de la variable x_2 en el modelo se verifica con el contraste $H_0 : \beta_2 = 0$. El estadístico de este contraste es $\hat{\beta}_2/s(\hat{\beta}_2)$, que sigue una distribución t de Student con $n-k-1$ grados de libertad. En nuestro caso, como $k = 2$, el número de grados de libertad es $n-3$.