Estadística II
Examen final junio 27/6/17
Curso 2016/17 – Soluciones

Duración del examen: 2 h y 15 min

1. (3 puntos) Los responsables de un aeropuerto afirman que el retraso medido en minutos en el tiempo de salida de los vuelos sigue un distribución normal, $N(\mu, \sigma)$, con $\mu = 30$ y $\sigma = 5$. Una organización de consumidores, aunque se muestra de acuerdo con el retraso medio en dicho tiempo de salida, afirma que la variabilidad es mayor a la publicitada por el aeropuerto.

   Se ha recogido una muestra de 15 observaciones de retrasos en salidas de vuelos, cuyo resumen se indica a continuación:

		cuartiles					
media	desv est	0 %	25 %	50 %	75 %	100 %	n
38,46	6,56	23,5	35,8	39,2	42,4	51,5	15

   Se pide que:

   a) (1 punto) Realices un contraste de hipótesis ($\alpha = 5\,\%$), para determinar si existe suficiente evidencia empírica que soporte la afirmación de dicha organización de consumidores (en relación con la variabilidad de los retrasos). Indica las hipótesis correspondientes, así como la conclusión del contraste.

   b) (0,5 puntos) Acotes el p-valor asociado al contraste anterior: ¿qué podemos concluir para el contraste si el nivel de significación es el 1 %?

   c) (1 punto) Calcules la potencia del contraste en el supuesto de que la desviación típica del retraso en el tiempo de salida fuese de 15 minutos.

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

   1) En un contraste de la media con $H_0 : \mu \geq 5$ y $\alpha = 0{,}05$, si el valor real del parámetro μ es $\mu_1 = 3$, se cumple que $1 - \beta > 0{,}05$.
   2) En un contraste de la desviación típica con $H_0 : \sigma = 0{,}5$, $H_1 : \sigma \neq 0{,}5$, $n = 25$ y $\alpha = 0{,}1$, la región crítica viene dada por $\{t : t < \chi^2_{24;0,95}\} \cup \{t : t > \chi^2_{24;0,05}\}$, donde t denota el estadístico del contraste $t = 24s^2/0{,}5^2$.

   Solución:

   a) Planteamos un contraste sobre la desviación típica (o análogamente sobre la varianza), unilateral por la derecha. En concreto, las hipótesis nula y alternativa del contraste serían:

$$\begin{aligned} H_0 &: \sigma^2 = 5^2 = \sigma_0^2 \\ H_1 &: \sigma^2 > 5^2 \end{aligned}$$

   donde el estadístico de contraste, bajo hipótesis de normalidad, es el siguiente:

$$T = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

   Por la tabla de estadísticos que nos proporciona el enunciado, sabemos que:

$$n = 15, \quad (n - 1 = 14), \quad \sigma^2 = 6{,}56^2.$$

   Además, por el contraste planteado $\sigma_0^2 = 5^2$.
   Si miramos las tablas de la distribución Chi cuadrado, obtenemos como punto crítico:

$$\chi^2_{14;0,05} = 23{,}7.$$

   Por tanto, nuestra región de rechazo vendrá dada por aquellos valores del estadístico T tales que:

$$\mathrm{RR}_{0,05} = \{t : t > \chi^2_{14;0,05}\} = \{t : t > 23{,}7\}$$

Nuestra muestra proporciona un valor particular de $t = 14 \times 6{,}56^2/5^2 = 24{,}07$, por lo que rechazamos la hipótesis nula y podemos concluir que la afirmación realizada por la asociación de consumidores es correcta, y el retraso de vuelos sigue una distribución $N(\mu = 30; \sigma > 5)$.

b) El valor de nuestro estadístico de contraste, que hemos calculado en el apartado anterior, es:

$$t = \frac{14 \times 6{,}56^2}{5^2} = 24{,}07,$$

y este estadístico se distribuye según una χ^2_{14}.
Si observamos la tabla de la Chi cuadrado, vemos que nuestro valor (24,07) lo podemos acotar entre los valores de probabilidad (0,025; 0,05), y a un nivel de significación $\alpha = 0{,}01$ no podemos rechazar H_0, ya que el p-valor obtenido es superior al nivel de significación.

c) Recordemos que la potencia de un contraste está definida como:

$$\text{Potencia} = P(\text{Rechazar } H_0 | H_0 \text{ falsa}) = P(\text{Rechazar } H_0 | \sigma_1^2 = 15^2) = 1 - \beta.$$

Por tanto,

$$\begin{aligned} 1-\beta &= P(t > \chi^2_{14;0,05} | \sigma_1^2 = 15^2) = P\left(\frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{14;0,05} | \sigma_1^2 = 15^2\right) \\ &= P\left(\frac{(n-1)s^2}{\sigma_1^2}\frac{\sigma_1^2}{\sigma_0^2} > \chi^2_{14;0,05} | \sigma_1^2 = 15^2\right) = P\left(\chi^2_{14} > \chi^2_{14;0,05}\frac{\sigma_0^2}{\sigma_1^2} | \sigma_1^2 = 15^2\right) \\ &= P\left(\chi^2_{14} > 23{,}7\frac{5^2}{15^2}\right) = P\left(\chi^2_{14} > 2{,}63\right) \approx 1. \end{aligned}$$

d) Las respuestas son:

1) Verdadera. Si el valor real del parámetro μ es $\mu_1 = 3 < 5 = \mu_0$, estamos en la región de cumplimiento de H_1 para dicho contraste. La potencia del contraste $1 - \beta = P(\text{rechazar } H_0 | H_0 \text{ es falso})$ es igual a $0{,}05 = \alpha$ para $\mu = \mu_0$, y es una función creciente conforme nos alejamos de ese valor en H_1 (para valores menores de μ). Por tanto, $1 - \beta > 0{,}05$ para $\mu_1 = 3$.
2) Verdadera. Tanto el estadístico como la región crítica son correctos.

2. (3 puntos) Se conjetura que la media de las puntuaciones de mujeres directivas en una evaluación de habilidades de negociación es superior a la media de las puntuaciones de hombres que ocupan cargos directivos similares. Para realizar este contraste, se han seleccionado dos muestras aleatorias de 12 mujeres directivas y 9 hombres, respectivamente. Las puntuaciones obtenidas en la evaluación de habilidades de negociación han sido:

Mujeres (X)	Hombres (Y)
50	45
54	50
56	43
57	55
49	56
56	41
50	42
55	44
59	40
53	
60	
58	

Las puntuaciones de la variable "habilidad negociadora" se expresan en una escala entre 0 y 100, donde valores próximos a 0 indican poca habilidad y valores altos mucha habilidad. Asumiendo la normalidad de las poblaciones, responde a las siguientes preguntas (considera un nivel de significación del 5 % para todos los apartados):

a) (1 punto) Si para el contraste de la razón de varianzas con los datos anteriores has obtenido los resultados en Excel que se recogen en la tabla siguiente, ¿podemos asumir la igualdad de varianzas poblacionales? Razona tu respuesta.

Prueba F para varianzas de dos muestras		
	Variable 1	*Variable 2*
Media	54,75	46,22222222
Varianza	13,29545455	35,94444444
Observaciones	12	9
Grados de libertad	11	8
F	0,369888998	
P(F<=f) una cola	0,064522268	
Valor crítico para F (una cola)	0,33921414	

b) (1,5 puntos) Contrasta la hipótesis de que las mujeres puntúan más alto que los hombres en la evaluación de habilidad negociadora, indicando:

- Tus hipótesis para el contraste
- El estadístico de contraste
- La región crítica o de rechazo de la hipótesis nula.
- El p-valor del contraste
- Tu decisión
- Su interpretación

c) (0,5 puntos) A la vista de los resultados obtenidos, si construyéramos un intervalo de confianza al 90 % para la diferencia de puntuaciones medias, ¿contendría al cero? Razona la respuesta.

Solución:

a) La tabla de Excel muestra que el p-valor del contraste (unilateral) es 0,0645, y por tanto el p-valor del contraste bilateral es $2 \times 0{,}0645 = 0{,}129$, superior al nivel de significación indicado (5 %). Por tanto, no rechazamos la hipótesis de igualdad de varianzas, esto es, aceptamos que las varianzas de las dos poblaciones pueden ser iguales.

b) A continuación se indican las respuestas a cada uno de los puntos solicitados:

- Tus hipótesis para el contraste serían

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

  o bien

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

- El estadístico de contraste, debido al resultado obtenido del apartado 2*a* y a nuestro supuesto de normalidad, será el correspondiente al caso con varianzas iguales,

$$T = \frac{\bar{Y} - \bar{X} - (\mu_Y - \mu_X)}{s_P\sqrt{\frac{1}{n_Y} + \frac{1}{n_X}}} \sim t_{n_Y+n_X-2}.$$

  El valor del estadístico para las muestras recogidas será

$$\begin{aligned} s_P^2 &= \frac{(n_Y-1)s_Y^2 + (n_X-1)s_X^2}{n_Y+n_X-2} = \frac{11 \times 13{,}30 + 8 \times 35{,}94}{19} = 22{,}83 \\ t &= \frac{54{,}75 - 46{,}22}{4{,}778\sqrt{\frac{1}{12} + \frac{1}{9}}} = 4{,}05. \end{aligned}$$

  Los valores muestrales son los indicados en la salida de Excel incluida en el apartado anterior.

- La región crítica o de rechazo de la hipótesis nula vendrá dada por

$$\mathrm{RR}_{0{,}05} = \{t : t > t_{19;0{,}05}\} = \{t : t > 1{,}73\}.$$

- El p-valor del contraste será

$$\text{p-valor} = P(t_{19} > t = 4{,}05) \approx 0.$$

- Decisión: Como el p-valor es tan bajo, o alternativamente dado que el valor observado del estadístico del contraste está claramente dentro de la región de rechazo, rechazamos la hipótesis nula de igualdad de medias para este caso.
- Interpretación: Concluimos que, a partir de la evidencia disponible, las puntuaciones promedio de mujeres directivas en la prueba de habilidades de negociación son claramente superiores a la media de las puntuaciones obtenidas por los hombres en dicha prueba.

c) Si el cero estuviese incluido en el intervalo para la diferencia de medias, podríamos afirmar que no existen diferencias significativas entre las puntuaciones promedio de las mujeres y los hombres en la prueba. Por el apartado anterior, sabemos con una confianza del 95 % que las mujeres puntúan más alto de los hombres. Por tanto, existen diferencias significativas en las medias y podemos concluir que, si calculásemos el intervalo de confianza para la diferencia de medias para un nivel de confianza del 95 %, el cero no estaría contenido en dicho intervalo.

Como un intervalo al 90 % tiene el mismo centro y una longitud menor que uno al 95 %, el intervalo al 90 % tampoco contendría al cero.

3. (4 puntos) En una localidad se ha intentado explicar la demanda anual de alarmas de seguridad para uso doméstico (variable Y) en función de la venta de viviendas unifamiliares (variable X), en términos también anuales. Para ello, se quiere realizar en esta localidad una regresión lineal simple con la evolución de las dos variables en los últimos 16 años—el periodo 2000-2015—, medidas ambas en miles de unidades.

Se tienen los siguientes datos:

$$\sum_{i=1}^{16} y_i = 4{,}305, \quad \sum_{i=1}^{16} x_i = 6{,}195, \quad \sum_{i=1}^{16} x_i y_i = 1{,}782, \quad \sum_{i=1}^{16} y_i^2 = 1{,}290,$$
$$\sum_{i=1}^{16} x_i^2 = 2{,}508, \quad \sum_{i=1}^{16} e_i^2 = 0{,}0103.$$

a) (1 punto) Estima y especifica el modelo de regresión lineal $Y = \beta_0 + \beta_1 X + u$.

b) (1 punto) Realiza el contraste adecuado para determinar si la demanda de alarmas está relacionada linealmente con la venta de viviendas unifamiliares con un nivel de significación del 5 %. Interpreta el resultado obtenido.

En un estudio posterior se realizó una nueva regresión incorporando dos nuevas variables explicativas: Un índice de siniestralidad doméstica (robos, vandalismo, incendios, etc) (variable X_2) y el índice de precios de las alarmas (variable X_3). Denominaremos como X_1 en lo que sigue a nuestra variable X anterior. Los resultados en Excel de la nueva regresión son los siguientes:

ANÁLISIS DE VARIANZA					
	Grados de libertad	Suma de cuadrados	romedio de los cuadrado	F	Valor crítico de F
Regresión	3	0,1314712	0,0438237	711,2175327	0,0000000
Residuos	12	0,0007394	0,0000616		
Total	15	0,1322106			

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95,0%	Superior 95,0%
Intercepción	-0,08190	0,07228	-1,13318	0,27927	-0,23938	0,07558	-0,23938	0,07558
Variable X 1	0,29433	0,09700	3,03443	0,01038	0,08299	0,50567	0,08299	0,50567
Variable X 2	0,41376	0,03323	12,45198	0,00000	0,34136	0,48616	0,34136	0,48616
Variable X 3	-0,06942	0,05286	-1,31336	0,21362	-0,18460	0,04575	-0,18460	0,04575

c) (0,5 puntos) Realiza el contraste adecuado para determinar si la variable X_2 (índice de siniestralidad) está relacionada linealmente con la demanda anual de alarmas (variable Y). Especifica la hipótesis nula y alternativa, el estadístico de contraste, así como tu conclusión. Utiliza el concepto de p-valor.

d) (0,5 puntos) Obtén un intervalo de confianza al 95 % para el parámetro de la regresión asociado a la variable X_3 (índice de precios de las alarmas) y decide si con un nivel de significación del 5 % podríamos mantener que este parámetro es distinto de cero.

e) (0,5 puntos) Calcula e interpreta el coeficiente de determinación múltiple del modelo.

f) (0,5 puntos) El p-valor asociado al estadístico F es próximo a cero, ¿qué interpretación tiene este resultado?

Solución:

a) Las estimaciones de los parámetros del modelo se pueden obtener de las fórmulas

$$\begin{aligned}
\bar{x} &= 6{,}195/16 = 0{,}387 \quad \bar{y} = 4{,}305/16 = 0{,}269 \\
s_x^2 &= (2{,}508 - 16 \times 0{,}387^2)/15 = 0{,}00728 \quad s_y^2 = (1{,}290 - 16 \times 0{,}269^2)/15 = 0{,}00881 \\
\text{cov}(x,y) &= (1{,}782 - 16 \times 0{,}387 \times 0{,}269)/15 = 0{,}00769 \\
\hat{\beta}_1 &= \frac{\text{cov}(x,y)}{s_x^2} = \frac{0{,}00769}{0{,}00728} = 1{,}057 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 0{,}269 - 1{,}057 \times 0{,}387 = -0{,}140.
\end{aligned}$$

El modelo de regresión lineal simple resultante es:

$$\hat{y} = -0{,}140 + 1{,}057x.$$

b) El contraste que nos piden es el siguiente:

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

El estadístico del contraste y su distribución son

$$T = \frac{\hat{\beta}_1}{\sqrt{\dfrac{s_R^2}{(n-1)s_x^2}}} \sim t_{n-2}$$

Para obtener el valor del estadístico necesitamos calcular la varianza residual,

$$s_R^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{0{,}0103}{14} = 0{,}000736$$

El valor del estadístico será por tanto

$$t = \frac{1{,}057}{\sqrt{\dfrac{0{,}000736}{15 \times 0{,}00728}}} = 12{,}868.$$

La región de rechazo del contraste es

$$\text{RR}_{0,05} = \{t : |t| > t_{14;0,025}\} = \{t : |t| > 2{,}145\},$$

y el valor de $t = 12{,}868$ que hemos obtenido está claramente en esta región de rechazo.

Por tanto, rechazamos H_0 para un nivel de significación del 5 %, esto es, concluimos que tenemos evidencia suficiente de que existe una relación lineal entre la demanda anual de alarmas de seguridad para uso doméstico y la venta anual de viviendas unifamiliares.

c) Para contrastar si la variable X_2 está relacionada linealmente con la demanda anual de alarmas, debemos llevar a cabo el contraste

$$\begin{aligned}
H_0 &: \beta_2 = 0 \\
H_1 &: \beta_2 \neq 0
\end{aligned}$$

El estadístico del contraste y su distribución son

$$T = \frac{\hat{\beta}_2}{s(\hat{\beta}_2)} \sim t_{n-k-1}.$$

En nuestro caso, como $n = 16$ y $k = 3$, la distribución es una t_{12}.

De los resultados de Excel tenemos el p-valor de este contraste bajo la columna “Probabilidad”, que es 0,00000. Al ser tan reducido, rechazamos H_0 para cualquier nivel de significación razonable (superior a 10^{-5}). Y concluimos que la variable “índice de siniestralidad” está relacionada linealmente con la demanda anual de alarmas.

d) Para obtener un intervalo de confianza al 95 % para el parámetro asociado a la variable X_3, β_3, utilizamos los datos de la salida de Excel bajo las columnas “Inferior 95 %” y “Superior 95 %”. Los valores correspondientes son

$$\mathrm{IC}_{0,95}(\beta_3) = [-0,185; 0{,}046].$$

Como este intervalo contiene el valor 0, no podemos rechazar la hipótesis de que el valor del parámetro sea igual a cero, esto es, no tenemos suficiente evidencia para concluir que la variable “índice de precios de las alarmas” esté relacionada linealmente con la demanda anual de alarmas.

e) El coeficiente de determinación múltiple del modelo, R^2, se puede calcular (utilizando los datos de la salida de Excel) como

$$R^2 = \frac{\mathrm{SCM}}{\mathrm{SCT}} = \frac{0{,}13147}{0{,}13221} = 0{,}994$$

Este valor implica que la proporción de la variabilidad de la variable respuesta “demanda anual de alarmas de seguridad para uso doméstico” explicada por las tres variables explicativas es el 99,4 %.

f) El p-valor asociado al estadístico F aparece en la salida de Excel como 0,0000000. Este p-valor corresponde al contraste de significación global del modelo de regresión múltiple, esto es, al contraste

$$H_0 : \beta_1 = \beta_1 = \beta_3 = 0$$
$$H_1 : \beta_i \neq 0 \text{ para algún } i = 1, 2, 3$$

Un p-valor tan reducido implica que para cualquier valor razonable del nivel de significación rechazamos H_0 y concluimos que (al menos) alguna de las variables explicativas del modelo tiene una relación lineal significativa con la variable respuesta.