Estadística II
Examen final junio - 17/06/16
Curso 2015/16 – Soluciones

Duración del examen: 2 h. y 45 min.

1. (3,5 puntos) La publicidad de un fondo de inversión afirma que la rentabilidad media anual del mismo es superior al 7%. Para apoyar tal afirmación, la publicidad indica que las rentabilidades anuales del fondo en los últimos 5 años fueron del 8%, 6.1%, 9%, 12% y 10%.

   a) (0,5 puntos) Suponiendo que las rentabilidades anuales del fondo son i.i.d. con distribución normal, plantea el contraste de hipótesis apropiado para contrastar la afirmación de la publicidad, identificando las hipótesis nula y alternativa.

   b) (0,75 puntos) Calcula el estadístico de contraste. Resuelve el contraste con un nivel de significación del 5% e interpreta el resultado.

   c) (0,5 puntos) Calcula aproximadamente (acota) el p-valor. Indica para qué niveles de significación puedes afirmar que no se rechaza la hipótesis nula con los datos obtenidos. Interpreta el resultado.

   d) (0,75 puntos) Supongamos ahora que la desviación típica de la rentabilidad anual del fondo es conocida e igual a $\sigma = 0{,}022$. Bajo este supuesto, resuelve de nuevo el apartado 1*b* y comenta el resultado.

   e) (1 punto) Bajo esta última hipótesis, calcula la potencia del contraste para una rentabilidad media anual del 9% y un nivel de significación del 5%. Interpreta el resultado.

   Solución.

   a) Se trata de un contraste para la media μ de una población normal con varianza desconocida partir de una muestra pequeña de tamaño $n = 5$. El contraste de hipótesis es:

$$\begin{aligned} &\mathrm{H}_0 : \mu \leqslant 0{,}07 \quad \text{vs.} \\ &\mathrm{H}_1 : \mu > 0{,}07. \end{aligned}$$

   b) El estadístico de contraste es

$$T = \frac{\bar{X} - 0{,}07}{s_X/\sqrt{n}} = t \approx \frac{0{,}0902 - 0{,}07}{0{,}022/\sqrt{5}} \approx 2{,}05.$$

   Con un nivel de significación del 5%, rechazamos H_0 si $T > t_{4,0{,}05} \approx 2{,}132$. Por tanto, con los datos dados no rechazamos H_0: no aceptamos la afirmación de la publicidad sobre la rentabilidad media del fondo.

   c) A partir de la tabla de valores críticos para la distribución t de Student acotamos el p-valor del contraste como $p = P\{T_4 > 2{,}05\} \in (0{,}05, 0{,}1)$, es decir, el p-valor se sitúa entre el 5% y el 10% (no incluidos).

   Con la información proporcionada podemos afirmar que no se rechaza H_0 para niveles de significación $\alpha \leqslant 0{,}05$. Por tanto, los datos mostrados en la publicidad no representan una evidencia sólida para avalar la afirmación de que la rentabilidad media es superior al 7%.

   d) El estadístico de contraste es

$$Z = \frac{\bar{X} - 0{,}07}{\sigma/\sqrt{n}} = z \approx \frac{0{,}0902 - 0{,}07}{0{,}022/\sqrt{5}} \approx 2{,}05.$$

   Con un nivel de significación del 5%, rechazamos H_0 si $Z > z_{0{,}05} \approx 1{,}645$. Por tanto, con los datos dados rechazamos H_0: aceptamos la afirmación de la publicidad sobre la rentabilidad media del fondo. La discrepancia con el apartado (b) se debe a que ahora los datos nos proporcionan más información, al ser la varianza conocida. La potencia del contraste cuando $\mu = 0{,}09$ y $\alpha = 0{,}05$ vale

$$\begin{aligned}
\text{potencia}(\mu = 0{,}09) &= P\left\{ \frac{\bar{X} - 0{,}07}{\sigma/\sqrt{5}} > 0{,}0119 \,\Big|\, \mu = 0{,}09 \right\} \\
&= P\left\{ \frac{\bar{X} - 0{,}09 + 0{,}09 - 0{,}07}{0{,}022/\sqrt{5}} > 1{,}645 \,\Big|\, \mu = 0{,}09 \right\} \\
&= P\left\{ \frac{\bar{X} - 0{,}09}{0{,}022/\sqrt{5}} > 1{,}645 - \frac{0{,}02}{0{,}022/\sqrt{5}} \,\Big|\, \mu = 0{,}09 \right\} \\
&= P\{Z > -0{,}39\} = 1 - P\{Z > 0{,}39\} = 1 - 0{,}3483 = 0{,}6517.
\end{aligned}$$

Por tanto, en el caso $\mu = 0{,}09$ el contraste rechaza H_0 con una probabilidad del 65.17%.

2. (2,5 puntos) Para comprobar la utilidad de una técnica de gestión de procesos un investigador ha pasado una prueba de rendimiento a una muestra de 16 sujetos. A continuación se lleva a cabo una formación en su técnica de gestión y tras ello, se vuelve a pasar la prueba de rendimiento. Los resultados de rendimiento obtenidos fueron los siguientes:

	Pre	Post
1	8	9
2	12	16
3	14	23
4	11	21
5	16	17
6	6	10
7	11	14
8	9	8
9	10	11
10	10	12
11	19	19
12	12	16
13	17	16
14	8	13
15	13	17
16	12	11

De los datos anteriores se tiene que $\bar{X}_A = 11{,}75$ y $\bar{X}_B = 14{,}56$, donde X_A denota la puntuación de rendimiento antes de la formación y X_B la puntuación posterior.

a) (1 punto) Plantea un contraste unilateral para determinar si los rendimientos han mejorado tras llevar a cabo la formación. Indica las hipótesis nula y alternativa, e identifica el estadístico del contraste y su distribución.

b) (1 punto) Calcula el p-valor del contraste anterior, y comenta tu conclusión para un nivel de significación del 1 %.

c) (0,5 puntos) Para una variante del contraste anterior has obtenido la siguiente salida de Excel (en la que se han eliminado algunos valores). Se pide que indiques el valor de la casilla indicada como "XXX". Razona tu respuesta.

	Variable 1	*Variable 2*
Media	14,5625	11,75
Varianza	18,6625	12,06666667
Observaciones	16	16
Coeficiente de correlación de Pearson	0,663047302	
Diferencia hipotética de las medias	-	
Grados de libertad	15	
Estadístico t	2,203131843	
P(T<=t) una cola	0,02181697	
Valor crítico de t (una cola)	1,340605608	
P(T<=t) dos colas	XXX	
Valor crítico de t (dos colas)	-	

Solución.

a) Teniendo en cuenta que los sujetos son los mismos en ambas muestras, se trata de un contraste de igualdad de medias con datos emparejados. Denotamos por μ_A y μ_B la medias poblacionales del rendimiento antes y despuÃ©s de usar la técnica, respectivamente. De este modo el contraste que debemos plantear es:

$$H_0: \ \mu_A \geq \mu_B$$
$$H_1: \ \mu_A < \mu_B.$$

Calculamos la diferencia de ambas muestras: $D = X_B - X_A$, y obtenemos:

X_A	X_B	D
8	9	1
12	16	4
14	23	9
11	21	10
16	17	1
6	10	4
11	14	3
9	8	-1
10	11	1
10	12	2
19	19	0
12	16	4
17	16	-1
8	13	5
13	17	4
12	11	-1

Tenemos $n = 16$, $\bar{D} = 2{,}81$ y $s_D = 3{,}19$. El contraste unilateral en términos de D tendrá la forma

$$H_0: \ \mu_D \leq 0 \tag{1}$$
$$H_1: \ \mu_D > 0. \tag{2}$$

Suponiendo normalidad en los datos, ya que el tamaño de la muestra es reducido, el estadÃstico del contraste vendrá dado por

$$T = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} \sim t_{n-1}.$$

La regla de decisión supone rechazar H_0 si y solo si

$$\frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} > t_{n-1;\alpha}.$$

b) De los valores precedentes tenemos que

$$t = \frac{2{,}81}{3{,}19/\sqrt{16}} = 3{,}52,$$

y p-valor $= P(t_{15} > 3{,}52)$. Ese valor en las tablas está en el intervalo $(0{,}0005; 0{,}005)$. Este p-valor es menor que el nivel de significación indicado del $1\,\%$, y por tanto podemos rechazar la hipótesis nula y concluir que la formación aumenta el rendimiento.

c) De los valores anteriores tenemos que

$$t = \frac{2{,}81}{3{,}19/\sqrt{16}} = 3{,}52,$$

y p-valor $= P(t_{15} > 3{,}52)$ y ese valor en las tablas está en el intervalo $(0{,}0005; 0{,}005)$. Para un nivel de significación del $1\,\%$ este p-valor es menor, y por tanto podemos rechazar la hipótesis nula y concluir que la formación aumenta el rendimiento.

d) El valor que nos piden es el p-valor del contraste bilateral. Como conocemos el p-valor del contraste unilateral, el valor pedido es el doble del p-valor unilateral, esto es,

$$P(|t_{15}| > t) = 2P(t_{15} > t) \Rightarrow XXX \equiv 2 \times 0{,}021817 = 0{,}043634$$

3. (4 puntos) De acuerdo con la teoría de disponibilidad de recursos (TDR), las ventajas de una compañía están asociadas al nivel de sus recursos tangibles e intangibles. Deseas analizar la ventaja competitiva de un hotel, a partir de la relación existente entre sus ingresos (variable Y en millones de euros), su distancia del centro de la ciudad (variable X_1 en kilómetros) y su gasto en publicidad (variable X_2 en miles de euros). Para este estudio dispones de una muestra aleatoria simple de 17 hoteles.

Los resultados obtenidos utilizando Excel se muestran en la tabla siguiente:

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	XXX	258,9235217	XXX	XXX	9,5134E-09
Residuos	14	XXX	XXX		
Total	XXX	278,8494118			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	2,05536248	0,88185499	2,330726144	0,035233054
Variable X 1	-0,350936953	0,152105551	-2,3071936	0,036844084
Variable X 2	0,042916883	0,003189592	13,45528761	2,12413E-09

a) (0,5 puntos) Indica la forma del modelo de regresión obtenido, e interpreta los valores de los coeficientes de regresión.

b) (1 punto) Completa los valores indicados como “XXX” en la tabla ANOVA.

c) (0,5 puntos) Calcula el coeficiente de determinación e interpreta su valor.

d) (0,5 puntos) Contrasta la significación global de los parámetros asociados a las variables explicativas.

e) (0,5 puntos) Obtén una estimación de la varianza de los errores empleando un estimador insesgado.

f) (0,5 puntos) Calcula un intervalo de confianza al 90 % para el coeficiente de X_1 (distancia al centro de la ciudad) e interpreta tu resultado.

g) (0,5 puntos) Indica si las afirmaciones siguientes son verdaderas o falsas para el modelo de regresión lineal múltiple anterior, y justifica tus respuestas:

1) La variabilidad del estimador del intercepto $\hat{\beta}_0$ decrece cuando aumenta el tamaño muestral.
2) La variable X_1 no es significativa para un nivel de significación del 1 %.
3) El valor de $(X^T X)^{-1}_{22}$ es 0,107

Solución.

a) El modelo será

$$\hat{y} = 2{,}055 - 0{,}351x_1 + 0{,}043x_2.$$

Un hotel en el centro de la ciudad que no gastase dinero en publicidad tendría unos ingresos promedio de 2,055 millones de euros ($\hat{\beta}_0$). Un incremento de una unidad en la distancia al centro de la ciudad cambiaría los ingresos promedio en $-0{,}351$ ($\hat{\beta}_1$), suponiendo que el gasto en publicidad no variase. Un aumento de una unidad en el gasto en publicidad cambiaría los ingresos promedio en 0,043 ($\hat{\beta}_2$), suponiendo que no cambiase la distancia al centro.

b) Los valores pedidos se indican en la tabla ANOVA siguiente:

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	258,9235217	129,4617608	90,96028558	9,5134E-09
Residuos	14	19,92589008	1,423277863		
Total	16	278,8494118			

c) El coeficiente de determinación se define como

$$R^2 = \frac{\text{SCM}}{\text{SCT}} = \frac{258{,}92}{278{,}85} = 0{,}9285$$

El modelo lineal explica casi el 93 % de la variabilidad en los valores de la variable dependiente Y (ingresos).

d) Para llevar a cabo el contraste global de significación emplearemos el valor del estadístico F dado en la tabla ANOVA. Este valor es menor que 10^{-8}, por lo que concluimos que el modelo es globalmente significativo.

e) Un estimador insesgado de la varianza de los errores es la varianza residual, y su valor en este caso es

$$s_R^2 = \frac{\text{SCR}}{n-k-1} = \frac{19{,}93}{14} = 1{,}423.$$

Este valor también puede encontrarse directamente en la tabla ANOVA.

f) El intervalo pedido se puede obtener como:

$$\text{IC}_\alpha = \hat{\beta}_1 \mp t_{n-k-1;\alpha/2} s(\hat{\beta}_2) = -0{,}351 \mp 1{,}761 \times 0{,}152 = [-0{,}619; -0{,}083].$$

Como el valor 0 no pertenece al intervalo, concluimos que el coeficiente es significativo a un nivel del 10 %.

g) Las respuestas son:

1) Cuando aumenta el tamaño muestral, la variabilidad de los estimadores de los parámetros del modelo decrece. VERDADERO
2) El p-valor del contraste de significación para la variable X_1 es 0,0368 (de la tabla ANOVA), mayor que 0,01. Por tanto, la variable X_1 no es significativa para un nivel de significación del 1 %. VERDADERO
3) El valor de $(X^T X)^{-1}_{22}$ se puede obtener como $s(\hat{\beta}_1)/s_R^2 = 0{,}152/1{,}423 = 0{,}107$. VERDADERO