Estadística II
Examen final enero - 14/01/16
Curso 2015/16 – Soluciones

Duración del examen: 2 h. y 45 min.

---

1. (3 puntos) Queremos analizar el tiempo que los niños españoles dedican a ver la televisión durante la semana. En un muestra aleatoria simple (m.a.s) de 16 niños se ha obtenido que el tiempo medio dedicado a ver la televisión es de 18,36 horas a la semana, con una cuasidesviación típica de 3,92 horas. Suponemos que la variable aleatoria "número de horas dedicadas semanalmente a ver la televisión" sigue una distribución normal.

   a) (0,5 puntos) Calcule un intervalo de confianza al 90 % para el tiempo medio que los niños españoles dedican a ver la televisión.

   b) (0,5 puntos) Se ha obtenido la siguiente salida de Excel, correspondiente a un contraste bilateral realizado sobre la muestra anterior. Indique las hipótesis nula y alternativa de dicho contraste, y la conclusión a la que llegaría (para un nivel de significación del 10 %).

Prueba t para dos muestras suponiendo varianzas desiguales		
	Variable 1	*Variable 2*
Media	18,3625	17
Varianza	15,3665	0
Observaciones	16	16
Diferencia hipotética de las medias	0	
Grados de libertad	15	
Estadístico t	1,3903016	
P(T<=t) una cola	0,09236181	
Valor crítico de t (una cola)	1,75305036	
P(T<=t) dos colas	0,18472363	
Valor crítico de t (dos colas)	2,13144955	

   c) (1 punto) En una investigación similar realizada hace 5 años, se concluyó que el tiempo medio dedicado a ver la televisión por los niños españoles era de 16 horas a la semana. Con un nivel de significación del 1 %, ¿se puede afirmar que el tiempo medio dedicado a ver la televisión ha aumentado en estos 5 años? Explique los supuestos que ha empleado, los diferentes pasos del proceso y sus conclusiones.

   d) (1 punto) Indique si las afirmaciones siguientes son verdaderas o falsas, y justifique en caso de que sean falsas cuál sería la respuesta verdadera:

      i. La probabilidad de un error de Tipo II es igual a uno menos el valor de la función de potencia obtenido cuando la hipótesis alternativa es cierta.

      ii. Si se rechaza una hipótesis nula frente a su alternativa al nivel del 1 %, entonces no se rechaza frente a esa alternativa al nivel del 5 %.

      iii. El p-valor de un contraste se define como la probabilidad de que la hipótesis nula sea correcta.

      iv. Se hace un contraste y no se rechaza la hipótesis nula, aunque ésta realmente sea falsa. Entonces, en este caso se ha cometido un error de Tipo I.

Solución.

a) Sea $X \equiv$ "número de horas dedicadas semanalmente a ver la televisión", $X \sim N(\mu, \sigma^2)$. Datos: $\alpha = 0{,}1$, $n = 16$, $\bar{x} = 18{,}36$ y $s_x = 3{,}92$. La distribución pivotal es una t_{n-1} y el cuantil de interés es $t_{15,0,05} = 1{,}753$. El intervalo de confianza viene dado por:

$$\mathrm{IC}_\alpha(\mu) = [\bar{x} \mp t_{n-1,\alpha/2}\frac{s_x}{\sqrt{n}}] = [16{,}642; 20{,}078]$$

b) El contraste es

$$\begin{aligned} H_0 &: \mu = \mu_0 = 17 \\ H_1 &: \mu \neq 17 \end{aligned}$$

Como el p-valor del contraste es 0,185 > 0,1, no rechazamos H_0. *También es posible responder a esta pregunta a partir de la región crítica, pero es necesario utilizar la tabla de la t de Student, ya que el valor crítico indicado en la salida de Excel corresponde a un nivel de significación del 5 %.*

c) Asumiremos que el valor 16 indicado en el enunciado es una aproximación razonable de la media poblacional en el periodo anterior. El contraste es

$$\begin{aligned} H_0 &: \mu \leq \mu_0 = 16 \\ H_1 &: \mu > 16 \end{aligned}$$

En nuestro caso, el estadístico del contraste y su distribución vienen dados por

$$\frac{\bar{X} - \mu_0}{\sqrt{s_x^2/n}} \sim t_{n-1}.$$

La región de rechazo es $\mathrm{RR}_{0,01} = \{t : t > t_{15,0,01}\} = \{t : t > 2{,}602\}$. El valor del estadístico para nuestra muestra es

$$t = \frac{18{,}36 - 16}{\sqrt{3{,}92^2/16}} = 2{,}408$$

Como el valor no está en la región crítica, no rechazamos H_0 y a un nivel de significación del 1 % concluimos que no tenemos suficiente evidencia para afirmar que el tiempo dedicado a ver la televisión haya aumentado respecto al observado hace 5 años.

d) Las respuestas son:

i. Verdadero.
ii. Falso. Si se rechaza rechaza una hipótesis nula frente a su alternativa al nivel del 1 % entonces el estadístico pertenece a la región crítica correspondiente que está incluida en la región crítica al 5 %.
iii. Falso. El p-valor es el menor valor de α para el que la hipótesis nula puede ser rechazada.
iv. Falso. Se comete un error de tipo II.

2. (2,25 puntos) Para una muestra aleatoria simple de 125 empresarios de una nación A (35 mujeres y 90 hombres) se ha obtenido que el número medio de cambios de trabajo en un periodo de 5 años fue de 1,91, con una cuasidesviación típica de 1,32. Se ha tomado otra muestra aleatoria simple de 86 empresarios de otra nación B (38 mujeres y 48 hombres), en la que el número medio de cambios de empleo en el mismo periodo fue de 0,21, con una cuasidesviación típica de 0,53.

a) (1 punto) ¿Se puede afirmar que, en media, los empresarios de A cambian un mayor número de veces de empleo que los empresarios de B? (Se asume un nivel de significación $\alpha = 0{,}05$.) Indique los supuestos necesarios para realizar el contraste.

b) (0,25 puntos) Indique una cota (lo más ajustada posible) para la potencia del test cuando las medias de ambas poblaciones sean iguales.

c) (1 punto) ¿Se puede afirmar que la proporción de hombres empresarios es igual para ambas nacionalidades? (Se asume un nivel de significación $\alpha = 0{,}05$.) Indique los supuestos necesarios para realizar este contraste.

Solución.

a) El contraste es:

$$\begin{aligned} H_0 &: \mu_A = \mu_B \\ H_1 &: \mu_A > \mu_B \end{aligned}$$

Los datos son: $n_A = 125$, $\bar{x}_A = 1{,}91$, $s_A = 1{,}32$, $\hat{p}_A = 0{,}72$ y $n_B = 86$, $\bar{x}_B = 0{,}21$, $s_B = 0{,}53$, $\hat{p}_B = 0{,}56$. El estadístico del contraste y su valor para estas muestras es:

$$\frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{1{,}91 - 0{,}21}{\sqrt{\frac{1{,}32^2}{125} + \frac{0{,}53^2}{86}}} = 12{,}96$$

Por el TCL, este estadístico sigue una distribución aproximadamente normal. La correspondiente región crítica es $\mathrm{RR}_{0{,}05} = \{z : z > 1{,}645\}$. Comparando el valor del estadístico con la región crítica, se rechaza H_0, esto es, concluimos que existe una diferencia significativa en los números medios de cambios de trabajo en ambas poblaciones, al nivel de significación indicado.

No hace falta asumir normalidad dado que las muestras son grandes ($n > 30$). Además, no hace falta asumir igualdad de las varianzas poblacionales y éstas se estiman a partir de las cuasivarianzas muestrales.

b) La potencia del test cuando $\mu_A = \mu_B$ es $\leq 0{,}05$, porque coincide con el valor de α cuando se está bajo la H_0.

c) Se tiene el contraste:

$$\begin{array}{ll} H_0: & p_A = p_B \\ H_1: & p_A \neq p_B \end{array}$$

El estadístico y su valor son

$$\frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{0{,}72 - 0{,}56}{\sqrt{0{,}6548(1-0{,}6548)\left(\frac{1}{125} + \frac{1}{86}\right)}} = 2{,}40,$$

siendo

$$\hat{p}_0 = \frac{n_A\hat{p}_A + n_B\hat{p}_B}{n_A + n_B} = 0{,}6548$$

La correspondiente región crítica es $\mathrm{RR}_{0{,}05} = \{z : z < -1{,}96, z > 1{,}96\}$. Por tanto se rechaza H_0.

Se tienen muestras aleatorias simples y los tamaños muestrales son grandes, por lo tanto la suma de distribuciones de Bernoulli se puede aproximar a una normal aplicando el TCL.

3. (4,75 puntos) Se ha obtenido información sobre el gasto realizado por cinco partidos políticos en sus respectivas campañas electorales x_i (en millones de euros) y el número de escaños obtenidos en las elecciones y_i. De la muestra correspondiente se tienen los valores siguientes:

$$\sum_{i=1}^{5} x_i = 20, \quad \sum_{i=1}^{5} y_i = 350, \quad \sum_{i=1}^{5} x_i y_i = 1710, \quad \sum_{i=1}^{5} x_i^2 = 106, \quad \sum_{i=1}^{5} y_i^2 = 33100.$$

Se pide:

a) (0,75 puntos) Construir un modelo lineal, usando Mínimos Cuadrados Ordinarios, para prever el número de escaños obtenidos en función del gasto realizado en la campaña electoral,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

Interpretar el modelo obtenido.

b) (0,25 puntos) Calcular una estimación puntual para la varianza de los errores, empleando el estimador insesgado. *(Si este valor fuese necesario en apartados posteriores y no ha sido capaz de calcularlo, puede emplear $s_R^2 = 1634{,}62$.)*

c) (1 punto) ¿Aportan los datos evidencia significativa al 10 % para concluir que el número de escaños obtenidos depende linealmente del gasto en campaña electoral?

d) (0,75 puntos) Construir un intervalo de confianza al 95 % para la predicción del número de escaños obtenidos para un partido que ha tenido un gasto en campaña de 3 millones de euros (independientemente de la respuesta obtenida en el apartado anterior). ¿Tendría la misma longitud si construyésemos el intervalo de confianza para la respuesta promedio del número de escaños? ¿Por qué?

e) (0,5 puntos) ¿Qué proporción de la variabilidad del número de escaños se explica por el gasto en campaña? ¿El modelo estimado tiene poder explicativo? ¿Por qué? (Los cálculos deben basarse en los datos anteriores.)

f) (1 punto) Para el modelo anterior se ha obtenido la siguiente salida de Excel:

Resumen					
Estadísticas de la regresión					
Coeficiente de correlación múltiple	XXX				
Coeficiente de determinación R^2	XXX				
R^2 ajustado	0,239713775				
Error típico	40,430377				
Observaciones	5				
ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	3696,153846	3696,153846	XXX	0,229692624
Residuos	3	XXX	XXX		
Total	4	8600			
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*
Intercepción	XXX	XXX	0,611034312	0,584360403	-93,87732645
Variable X 1	XXX	7,92904928	XXX	0,229692624	-13,31069666

Completar los valores indicados con "XXX" en las tablas anteriores.

g) (0,5 puntos) Para el modelo de regresión lineal simple $Y_i = \beta_0 + \beta_1 X_i + U_i$ y su estimación por MCO, $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$, indica si las afirmaciones siguientes son verdaderas o falsas y justifica tu respuesta:

1) $U_i \sim N(0, \sigma_i^2)$
2) $0 \leq R^2 \leq 1$
3) $\sum_{i=1}^{N} e_i = 0$
4) $\hat{\beta}_1$ sigue una distribución normal.

Solución.

a) Tenemos los siguientes valores:

$$\begin{aligned} \bar{x} &= 4, \quad \bar{y} = 70, \quad s_x^2 = \frac{\sum_i x_i^2 - n\bar{x}^2}{n-1} = 6{,}5, \quad s_y^2 = \frac{\sum_i y_i^2 - n\bar{y}^2}{n-1} = 2150 \\ \operatorname{cov}(x,y) &= \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{n-1} = 77{,}5 \\ \hat{\beta}_1 &= \frac{\operatorname{cov}(x,y)}{s_x^2} = 11{,}92, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 22{,}3 \end{aligned}$$

El modelo resultante es:

$$\hat{Y} = 22{,}3 + 11{,}92X,$$

esto es, en promedio, por cada millón de euros adicionales gastados se consiguen 11,92 escaños más. También, si no se gasta ningún dinero en campaña el número medio de escaños obtenidos es 22,3.

b) Tenemos que

$$\begin{aligned} s_R^2 &= \frac{\sum_i e_i^2}{n-2} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} \\ &= \frac{1}{n-2}\left(\sum_i y_i^2 + \hat{\beta}_0^2 + \hat{\beta}_1^2 \sum_i x_i^2 - 2\hat{\beta}_0 \sum_i y_i - 2\hat{\beta}_1 \sum_i x_i y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum_i x_i \right). \end{aligned}$$

Todos estos valores son conocidos, y sustituyendo se obtiene el valor indicado.

Alternativamente, de la tabla ANOVA al final del ejercicio se tiene que $\text{SCR} = 8600 - 3696{,}15 = \sum_i e_i^2$, y sustituyendo en la definición de s_R^2 se obtiene el mismo valor.

Se considerarán también como válidas las respuestas que se hayan calculado en los apartados siguientes utilizando el valor $s_R^2 = 1364{,}62$.

c) El contraste a aplicar es

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

Tenemos para el estadístico de este contraste que

$$t = \frac{\hat{\beta}_1}{\sqrt{s_R^2/(n-1)s_x^2}} = \frac{11{,}92}{\sqrt{1364{,}62/4 \times 6{,}5}} = 1{,}504,$$

y como $t_{n-2;\alpha/2} = t_{3;0,05} = 2{,}35$, no podemos rechazar la hipótesis nula, y por tanto no tenemos suficiente evidencia para concluir que existe una relación lineal entre las variables.

d) En este caso, la fórmula a emplear (intervalo de confianza para una predicción) es

$$\mathrm{IC}_\alpha(y_0) = \hat{y}_0 \mp t_{n-2;\alpha/2}\sqrt{s_R^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)},$$

donde $\hat{y}_0 = 22{,}3 + 11{,}92x_0 = 81{,}92$, and $t_{3;0,025} = 3{,}182$. Obtenemos

$$\mathrm{IC}_\alpha(y_0) = [-85{,}08; 201{,}22]$$

El intervalo no tiene la misma longitud para la respuesta promedio, porque la variabilidad del estimador correspondiente es menor, al tratarse de un promedio.

e) Nos piden el valor del coeficiente de determinación del modelo, R^2. Obtenemos dicho valor como

$$R^2 = \mathrm{cor}(x,y)^2 = \frac{\mathrm{cov}(x,y)^2}{s_x^2 s_y^2} = \frac{77{,}5^2}{6{,}5 \times 2150} = 0{,}4298.$$

El poder explicativo es por tanto reducido, al ser el valor de R^2 relativamente pequeño (mucho menor de 1).

f) El cuadro completo se indica a continuación:

Resumen					
Estadísticas de la regresión					
Coeficiente de correlación múltiple	0,655580148				
Coeficiente de determinación R^2	0,429785331				
R^2 ajustado	0,239713775				
Error típico	40,430377				
Observaciones	5				
ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	3696,153846	3696,153846	2,261176471	0,229692624
Residuos	3	4903,846154	1634,615385		
Total	4	8600			
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*
Intercepción	22,30769231	36,50808454	0,611034312	0,584360403	-93,87732645
Variable X 1	11,92307692	7,92904928	1,503720875	0,229692624	-13,31069666

g) Las respuestas son:

1) Falso. Se supone $\sigma_i = \sigma \ \forall i$
2) Verdadero. De $R^2 = \mathrm{cor}(x,y)^2$ y $-1 \leq \mathrm{cor}(x,y) \leq 1$.
3) Verdadero. Consecuencia de las fórmulas obtenidas de MCO.
4) Verdadero. $\hat{\beta}_1 \sim N(\beta_1, \sqrt{\sigma^2/((n-1)s_x^2)})$