Estadística II
Examen final junio 26/6/18
Curso 2017/18 – Soluciones

Duración del examen: 2 h y 30 min

---

1. (3 puntos) Un Ayuntamiento está interesado en comprobar si es cierto que la comida que anualmente se desaprovecha en el sector de restauración de su ciudad no supera el millar de kilos por restaurante. Ha contratado a una consultora para que realice un seguimiento a una muestra aleatoria simple de 10 restaurantes, obteniendo un promedio de 1,15 miles de kilos de comida desaprovechada anualmente, con una cuasi-desviación típica de 150 kilos. Suponga que los kilos de comida desaprovechada siguen una distribución normal.

   a) (1 punto) Ayude al Ayuntamiento a contrastar la afirmación sobre los restaurantes con un nivel de significación del 5 %, construyendo la región de rechazo del mismo e indicando su conclusión.

   b) (0,5 puntos) Calcule el p-valor del contraste y comente el resultado.

   c) (1 punto) Calcule la potencia del contraste cuando la cantidad de comida desaprovechada realmente es de 1,1 miles de kilos.

   d) (0,5 puntos) Indique si las siguientes afirmaciones son verdaderas o falsas; justifique solo las afirmaciones falsas:

   - Para un valor de la cantidad real de comida desaprovechada dado, y manteniendo todo lo demás constante, si aumenta el número de restaurantes de la muestra la potencia mejora.
   - Si el contraste se realiza al 1 % de significación, en lugar de hacerlo para el 5 %, el p-valor resultante cambia.
   - La masa de probabilidad de la región de rechazo o crítica es $1 - \alpha$.

   Solución:

   a) Sea $X =$ "comida desaprovechada anualmente por un restaurante," y $\mu = E[X]$ su promedio. El contraste a llevar a cabo es

   $$H_0 : \mu \leq 1$$
   $$H_1 : \mu > 1.$$

   Suponemos que X sigue una distribución normal y que la muestra es i.i.d. El estadístico del contraste será:

   $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

   donde S denota la quasi-desviación típica muestral y $n = 10$.
   La región de rechazo corresponderá a

   $$\mathrm{RR}_\alpha = \{(x_1, \ldots, x_n) : t > t_{n-1;\alpha}\}.$$

   Como nos indican que $\alpha = 0{,}05$, tenemos $t_{n-1;\alpha} = t_{9;0,05} = 1{,}833$.
   El valor del estadístico para nuestra muestra es

   $$t = \frac{1{,}15 - 1}{0{,}15/\sqrt{10}} = 3{,}162.$$

   Este valor está en la región de rechazo, por lo que concluimos que existe evidencia suficiente para rechazar la hipótesis nula, y para sostener que el promedio de comida desaprovechado anualmente por un restaurante es superior a mil kilos.

   b) El p-valor del contraste vendrá dado por

   $$\text{p-valor} = P(T > t) = P(t_9 > 3{,}162) \in [0{,}005; 0{,}01]$$

   (un valor mas preciso es 0,0058). Este p-valor es muy bajo (menor que el 1 %), y ofrece una indicación clara para rechazar H_0.

c) Para obtener la potencia del contraste, potencia(μ), cuando $\mu = 1{,}1$, partimos de la definición de potencia como la probabilidad de rechazar H_0 cuando no es cierta.

$$\begin{aligned}\text{potencia}(1{,}1) &= P(\text{rechazar } H_0 \mid \mu = 1{,}1) = P(T > 1{,}833 \mid \mu = 1{,}1)\\ &= P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > 1{,}833 \mid \mu = 1{,}1\right) = P\left(\bar{X} > 1{,}087 \mid \mu = 1{,}1\right)\\ &= P\left(t_9 > \frac{1{,}087 - \mu}{0{,}15/\sqrt{10}} \mid \mu = 1{,}1\right) = P\left(t_9 > \frac{1{,}087 - 1{,}1}{0{,}0474}\right)\\ &= P(t_9 > -0{,}274) \in [0{,}5; 0{,}75],\end{aligned}$$

o siendo más precisos, la potencia vale 0,605.

En el desarrollo anterior hemos utilizado que, bajo la hipótesis de que el valor correcto de μ es 1,1, se tiene

$$\frac{\bar{X} - 1{,}1}{S/\sqrt{n}} \sim t_9.$$

d) Las respuestas son:

- Verdadera. La potencia aumenta con el tamaño de la muestra.
- Falsa. El p-valor es independiente del nivel de significación utilizado, ya que corresponde al valor en probabilidad de la muestra bajo H_0. Una vez calculado, se contrasta con el α que se quiera utilizar.
- Falsa. La masa de probabilidad de la región de rechazo o crítica es α.

2. (2,5 puntos) Para comprobar la utilidad de un nuevo modelo de formación en una empresa, se ha llevado a cabo un experimento consistente en realizar unas pruebas de mejora de capacitación a dos muestras aleatorias independientes compuestas por 32 empleados (diferentes) cada una de ellas. Una muestra ha seguido un proceso de formación basado en el nuevo modelo y la otra lo ha hecho con el modelo tradicional. Un resumen de los resultados obtenidos, medidos en una escala de 0 a 25, es el siguiente:

$$\sum_{i=1}^{32} x_i = 408, \quad \sum_{i=1}^{32} x_i^2 = 5602, \quad \sum_{i=1}^{32} y_i = 480, \quad \sum_{i=1}^{32} y_i^2 = 7648.$$

Para los datos anteriores se pide lo siguiente:

a) (1 punto) Plantee un contraste para determinar si se puede rechazar que la mejoría en la formación es la misma empleando los dos modelos. Obtenga la región de rechazo del contraste para un nivel de significación del 1 %, e indique sus conclusiones.

b) (0,5 puntos) Calcule el p-valor del contraste anterior y comente el resultado obtenido.

c) (1 punto) Realice el contraste indicado en el apartado 2*a*) con el mismo nivel de significación para contrastar la creencia de que al aplicar el nuevo modelo de formación la mejora en el rendimiento es de al menos dos puntos. Indique la región de rechazo y la conclusión para este contraste.

Solución:

a) Teniendo en cuenta que los sujetos son distintos en ambas muestras, se trata de un contraste de igualdad de medias con muestras independientes. Denotaremos como X = "resultado en las pruebas de mejora obtenido con la formación tradicional," y como Y = "resultado en las pruebas de mejora obtenido con el nuevo modelo de formación."

Denotamos como μ_X y μ_Y la medias poblacionales de las pruebas de mejora con la formación tradicional y el nuevo modelo de formación, respectivamente. El contraste que debemos plantear es:

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

Como el tamaño de ambas muestras es suficientemente elevado, el estadístico del contraste y su distribución vendrán dados por:

$$T = \frac{X - Y - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n_x + S_Y^2/n_Y}} \sim_{\text{aprox.}} N(0,1),$$

donde S_X^2 y S_Y^2 denotan las cuasi-varianzas muestrales.
La región de rechazo (aproximada) vendrá dada por

$$\text{RR}_\alpha = \{(d_1, \ldots, d_n) \mid |T| > z_{\alpha/2}\},$$

donde $z_{\alpha/2} = z_{0,005} = 2{,}576$.
Para obtener el valor del estadístico a partir de los datos de las dos muestras usaremos

$$\bar{x} = \frac{1}{n}\sum_i x_i = 408/32 = 12{,}75, \qquad \bar{y} = \frac{1}{n}\sum_i y_i = 480/32 = 15$$

$$s_X^2 = \frac{1}{n-1}\left(\sum_i x_i^2 - n\bar{x}^2\right) = 12{,}903, \qquad s_Y^2 = \frac{1}{n-1}\left(\sum_i y_i^2 - n\bar{y}^2\right) = 14{,}452.$$

El valor buscado es

$$t = \frac{12{,}75 - 15 - 0}{\sqrt{12{,}903/32 + 14{,}452/32}} = -2{,}434.$$

Este valor no está en la región crítica. Por tanto, no rechazamos H_0 para el nivel de significación indicado, esto es, no tenemos evidencia suficiente para pensar que el nuevo modelo de formación ofrece un aumento en los resultados de las pruebas de mejora mayor que el modelo tradicional.

b) El p-valor de este contraste vendrá dado por

$$\text{p-valor} = 2P(Z < -2{,}434) = 0{,}015.$$

Este valor implica que rechazaríamos la hipótesis nula para cualquier nivel de significación superior a 0,015, y no lo haríamos para valores inferiores a este. En particular para $\alpha = 0{,}01$ no rechazaríamos H_0.

c) En este caso el contraste a llevar a cabo sería:

$$H_0 : \mu_Y - \mu_X \geq 2$$
$$H_1 : \mu_Y - \mu_X < 2.$$

Mantenemos el mismo estadístico del contraste T definido en el primer apartado, y la región de rechazo pasa a ser

$$\text{RR}_\alpha = \{(d_1, \ldots, d_n) \mid T > z_\alpha\},$$

teniendo en cuenta que T está definido en términos de $X - Y$.
El valor del estadístico para la muestra dada (y bajo H_0) pasa a ser

$$t = \frac{12{,}75 - 15 - (-2)}{\sqrt{12{,}903/32 + 14{,}452/32}} = -0{,}270.$$

Este valor no está en la región crítica, y por tanto no rechazamos H_0, esto es, no tenemos suficiente evidencia para creer que la mejora con la nueva formación sea de menos de dos puntos.

3. (4,5 puntos) Se quiere analizar la influencia de los años de experiencia en el sector bancario (variable X) sobre los salarios anuales de los empleados (variable Y), medidos en miles de euros, en uno de los principales bancos nacionales. Para ello se ha tomado una muestra aleatoria de doce empleados, obteniendo los siguientes datos:

$$\sum_{i=1}^{12} x_i = 149, \quad \sum_{i=1}^{12} x_i^2 = 2611, \quad \sum_{i=1}^{12} y_i = 312, \quad \sum_{i=1}^{12} y_i^2 = 8776,$$
$$\sum_{i=1}^{12} x_i y_i = 4484, \quad \sum_{i=1}^{12} e_i^2 = 174{,}98.$$

a) (0,5 puntos) Estime el modelo de regresión lineal $Y = \beta_0 + \beta_1 X + u$.

b) (0,75 puntos) Realice un contraste que determine si existe relación lineal entre los años de experiencia en banca y el salario a percibir por los trabajadores de esta entidad, con un nivel de significación del 5 %.

c) (0,5 puntos) Complete los valores de la Tabla ANOVA para este modelo.

d) (0,5 puntos) Calcule el coeficiente de determinación del modelo e interprete su resultado.

e) (0,5 puntos) Obtenga una predicción para el salario de un empleado recién contratado con una experiencia (previa) en el sector bancario de 13 años. Proporcione una medida de la fiabilidad de dicha predicción mediante un intervalo de confianza al 90 %.

Para mejorar este estudio sobre salarios se ha incluido una nueva variable que mide el nivel de formación del empleado (X_2). Los valores obtenidos para el modelo de regresión múltiple son los dados en la salida de Excel siguiente:

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	15,20389506	1,6595637	9,161380824	7,38103E-06
Variable X 1	0,230479612	0,214655309	1,073719598	0,310891603
Variable X 2	1,763181429	0,567148689	3,108852161	0,012542058

f) (0,75 puntos) Calcule un intervalo de confianza al 99 % para el coeficiente correspondiente a la variable X_2. A partir de este intervalo, comente si esta variable sería significativa.

g) (0,5 puntos) Complete la tabla ANOVA mostrada a continuación, calculando los valores indicados como "`XXXX`":

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	579,62	289,81	XXXX	9,29E-05
Residuos	9	XXXX	XXXX		
Total	XXXX	664			

h) (0,5 puntos) Indique si son significativos cada uno de los coeficientes del modelo. Razone su respuesta. ¿Es globalmente significativo el modelo?

Solución:

a) Para estimar el modelo partimos de los siguientes valores obtenidos de los datos de las muestras

$$\bar{x} = 149/12 = 12{,}42, \quad \bar{y} = 312/12 = 26$$
$$s_x^2 = (2611 - 12 \times 12{,}42^2)/11 = 69{,}17, \quad s_y^2 = (8776 - 12 \times 26^2)/11 = 60{,}36$$
$$s_{xy} = (4484 - 12 \times 12{,}42 \times 26)/11 = 55{,}45.$$

Empleamos las fórmulas de los estimadores de mínimos cuadrados de los dos parámetros para obtener

$$\begin{aligned} \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = 0{,}802 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 16{,}05 \end{aligned}$$

El modelo estimado es

$$\hat{y} = 16{,}05 + 0{,}802x.$$

b) El contraste a realizar es

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

El estadístico del contraste es

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s_R^2}{(n-1)s_x^2}}} \sim t_{n-2}.$$

La región de rechazo viene dada por

$$\mathrm{RR}_\alpha = \{(x_i, y_i) \mid |t| > t_{n-2;\alpha/2}\},$$

donde $t_{n-2;\alpha/2} = t_{10;0,025} = 2{,}228$.
El valor del estadístico para nuestra muestra se puede obtener de

$$s_R^2 = \frac{1}{n-2}\sum_i e_i^2 = \frac{174{,}98}{10} = 17{,}50, \quad t = \frac{0{,}802}{\sqrt{\frac{17{,}50}{11\times 69{,}17}}} = 5{,}286.$$

Este valor está en la región de rechazo. Por tanto, concluimos que para el nivel de significación indicado, existe una relación lineal significativa entre las variables.

c) Necesitamos realizar los siguientes cálculos:

$$\mathrm{SCR} = \sum_{i=1}^{n} e_i^2 = 174{,}98, \quad \mathrm{SCT} = (n-1)s_y^2 = 664, \quad \mathrm{SCM} = \mathrm{SCT} - \mathrm{SCR} = 489{,}02$$
$$s_R^2 = \mathrm{SCR}/(n-2) = 17{,}50, \quad F = \mathrm{SCM}/s_R^2 = 27{,}95.$$

La tabla ANOVA será:

Fuente de variación	SC	GL	Media	Cociente F
Modelo	489,02	1	489,02	27,95
Residuos/errores	174,98	10	17,50	
Total	664	11		

d) El coeficiente de determinación se obtiene como

$$R^2 = \frac{\mathrm{SCM}}{\mathrm{SCT}} = \frac{489{,}02}{664} = 0{,}736.$$

Su interpretación es que el 73,6 % de la variabilidad en la variable dependiente (salarios anuales) se puede explicar a partir de los valores de la variable independiente (experiencia en el sector), a través del modelo de regresión.

e) Para predecir el salario de un empleado recién contratado con una experiencia previa en el sector bancario de $x_0 = 13$ años, empleamos el modelo de regresión

$$\hat{y}_0 = 16{,}05 + 0{,}802x_0 = 26{,}47.$$

Para obtener el intervalo de confianza pedido, aplicamos la fórmula

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

En nuestro caso tenemos que

$$\hat{y}_0 = 26{,}47, \qquad t_{n-2;\alpha/2} = t_{10;0,05} = 1{,}812, \qquad s_R^2 = 17{,}50,$$
$$x_0 - \bar{x} = 13 - 12{,}42 = 0{,}58, \qquad s_x^2 = 69{,}17.$$

Sustituyendo estos valores, obtenemos el intervalo de confianza

$$\mathrm{IC}_{0,9}(y_0) = 26{,}47 \pm 1{,}812\sqrt{17{,}50\left(1 + \frac{1}{12} + \frac{0{,}58^2}{11\times 69{,}17}\right)} = [18{,}57; 34{,}36].$$

f) Para obtener el intervalo pedido, empleamos la información de la tabla de salida de Excel y el cuantil de la distribución t de Student correspondiente, $t_{n-3;\alpha/2} = t_{9;0,005} = 3{,}250$,

$$\mathrm{IC}_\alpha(\beta_2) = \hat{\beta}_2 \pm t_{n-3;\alpha/2} \times \text{error típico} = 1{,}763 \pm 3{,}250 \times 0{,}567 = [-0{,}080; 3{,}606]$$

Como el intervalo contiene el valor 0, la variable no sería significativa en este modelo para un nivel de significación del 1 %.

g) Para completar la tabla ANOVA calculamos

$$\text{GL Total} = 9 + 2 = 11, \qquad \mathrm{SCR} = \mathrm{SCT} - \mathrm{SCM} = 664 - 579{,}62 = 84{,}38$$
$$s_R^2 = \mathrm{SCR}/(n-3) = 9{,}38, \qquad F = \mathrm{SCM}/s_R^2 = 30{,}91.$$

La tabla ANOVA queda como

Fuente de variación	GL	SC	Media	Cociente F
Modelo	2	579,62	289,81	30,91
Residuos/errores	9	84,38	9,38	
Total	11	664		

h) De los p-valores indicados en la salida de Excel para los tres parámetros del modelo, podemos concluir que la variable X_1 no será significativa en general (tan solo para niveles de significación superiores al 31 %) y que la variable X_2 solo será significativa para niveles de significación superiores al 1,25 %. La constante del modelo (el coeficiente β_0) es siempre significativa, ya que su p-valor es muy reducido ($7{,}38\,10^{-6}$).

El modelo es globalmente significativo, dado que el p-valor del cociente F es muy reducido ($9{,}29\,10^{-5}$).