Estadística II
Examen final enero 23/1/19
Curso 2018/19 – Soluciones

Duración del examen: 2 h y 15 min

---

1. (3,5 puntos) En la fabricación de un producto se hacen controles diarios para garantizar su calidad. El sistema de control consiste en seleccionar una muestra aleatoria de la producción diaria, y estudiar el porcentaje de piezas cuyo nivel de defectos supera un cierto umbral.

   a) (1 punto) Se ha analizado una muestra de 50 piezas correspondientes a la producción del último día, siendo el resultado del análisis que 7 piezas superaban el valor umbral de defectos. Construya un intervalo de confianza al 95 % para el porcentaje de piezas defectuosas en la producción de ese día.

   b) (1 punto) La empresa ha decidido que descartará la producción de un día determinado si el porcentaje de piezas defectuosas ese día supera el 8 %. Utilizando los datos de la muestra del apartado anterior, realice el contraste oportuno (usando $\alpha = 0{,}05$) para saber si la empresa debe descartar la producción de ese día.

   c) (1 punto) Para el contraste descrito en el apartado $1b$, si el porcentaje real de defectos en la producció de ese día fuese del 15 %, ¿cuál sería la potencia del contraste?

   d) (0,5 puntos) Discuta, justificando la respuesta, la veracidad de las siguientes afirmaciones:

      1) En un contraste bilateral cuyo resultado fuese aceptar H_0, podríamos estar cometiendo un error de tipo I.
      2) Cuanto mayor es el nivel de confianza $1-\alpha$, mayor es la probabilidad de un error de tipo II, β.
      3) Si en un contraste bilateral para la proporción obtenemos un estadístico de contraste $z = 1{,}82$ y no rechazamos H_0, siempre se cumple que $P(Z > 1{,}82) > \alpha$, donde $Z \sim \mathcal{N}(0,1)$.

   Solución:

   a) Teniendo en cuenta que $\hat{p} = 7/50 = 0{,}14$, el intervalo de confianza pedido es:

   $$\text{IC}(p) = \hat{p} \pm z_{0{,}025}\sqrt{\hat{p}(1-\hat{p})/n} = 0{,}14 \pm 1{,}96\sqrt{0{,}14(1-0{,}14)/50} = [0{,}0438; 0{,}2361].$$

   b) El contraste a realizar es:

   $$\begin{cases} H_0: & p \leq p_0 = 0{,}08 \\ H_1: & p > p_0 \end{cases}$$

   El estadístico de contraste (bajo H_0) y su distribución aproximada son:

   $$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathcal{N}(0,1)$$

   Para los datos del enunciado, este estadístico toma el valor:

   $$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{7/50 - 0{,}08}{\sqrt{0{,}08(1-0{,}08)/50}} = 1{,}564$$

   Como la región de rechazo es $\text{RR}_\alpha = \{Z > z_\alpha\}$ y $z_\alpha = z_{0{,}05} = 1{,}645$, no rechazamos la hipótesis nula. Esto implica que para la muestra empleada y el nivel de significación establecido, no tenemos suficiente evidencia estadística para descartar la producción del día.

c) Para obtener la potencia de este contraste para el caso en que $p = p_1 = 0{,}15$ debemos calcular (de manera aproximada)

$$\begin{aligned}
\text{potencia}(0{,}15) &= P(\text{rechazar } H_0 \mid p = 0{,}15) = P\left(\frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} > 1{,}645 \mid p = 0{,}15\right) \\
&= P\left(\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} - \frac{p_0 - p}{\sqrt{p(1-p)/n}} > 1{,}645\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} \mid p = 0{,}15\right) \\
&= P\left(Z > 1{,}645\sqrt{\frac{0{,}08(1-0{,}08)}{0{,}15(1-0{,}15)}} + \frac{0{,}08 - 0{,}15}{\sqrt{0{,}15(1-0{,}15)/50}}\right) \\
&= P(Z > -0{,}1364) = 0{,}554,
\end{aligned}$$

donde hemos hecho uso de la propiedad de que si $p = p_1$ (aproximadamente)

$$\frac{\hat{P} - p_1}{\sqrt{p_1(1-p_1)/n}} \sim N(0,1).$$

d) Las respuestas son:

1) Falso: El error de tipo I solo se puede cometer al rechazar H_0, por lo que para el caso propuesto al indicarnos que se acepta H_0 no sería posible cometer un error de tipo I.
2) Verdadero: Cuanto mayor es el nivel de confianza $1-\alpha$, menor es el nivel de significación α. También se tiene que la probabilidad de un error de tipo II, β, es mayor cuanto menor es la de un error de tipo I, α. Por tanto, cuanto mayor es $1-\alpha$, menor es α y mayor es β.
3) Falso: Al no haber rechazado H_0, sabemos que el nivel de significación α es menor que el p-valor, igual a $2P(Z > 1{,}82)$. Pero con esta información no podemos asegurar que $P(Z > 1{,}82)$ (la mitad del p-valor) vaya a ser siempre mayor que α.

2. (2 puntos) Para estudiar el impacto de una campaña de concienciación medioambiental, se recogió una muestra de 20 mediciones de concentraciones de contaminantes antes del inicio de la campaña, con una media muestral de 3,10 y cuasidesviación típica de 0,415. Otra muestra de 20 mediciones de esta concentración se ha recogido dos meses después de llevar a cabo la campaña, con media 2,90 y cuasidesviación típica 0,521.

a) (1,5 puntos) Suponemos que las muestras son independientes. Supondremos también que las mediciones siguen una distribución normal y las varianzas de las dos poblaciones son iguales. Se quiere contrastar si la concentración de contaminantes se ha reducido significativamente, para un nivel de significación del 5 %. Lleve a cabo dicho contraste unilateral comparando las medias de las dos poblaciones. Justifique los pasos del procedimiento aplicado y comente su conclusión.

b) (0,5 puntos) Suponemos ahora que estas dos muestras corresponden a mediciones emparejadas antes y después de la campaña. Se quiere contrastar de nuevo si se ha producido una reducción significativa en la concentración de contaminantes, para un nivel de significación del 1 %. Indique su conclusión para este contraste basándose en los resultados de Excel que se muestran a continuación.

Prueba t para medias de dos muestras emparejadas		
	Muestra antes	*Muestra después*
Media	3,1	2,9
Varianza	0,172631579	0,271578947
Observaciones	20	20
Coeficiente de correlación de Pearson	0,768113344	
Diferencia hipotética de las medias	0	
Grados de libertad	19	
Estadístico t	2,677650336	
P(T<=t) una cola	0,007444662	
Valor crítico de t (una cola)	1,729132812	
P(T<=t) dos colas	0,014889324	
Valor crítico de t (dos colas)	2,093024054	

Solución:

a) Deseamos llevar a cabo el siguiente contraste:

$$\begin{cases} H_0: & \mu_a \leq \mu_d \\ H_1: & \mu_a > \mu_d \end{cases}$$

donde μ_a y μ_d denotan la concentración promedio de contaminantes correspondiente a las medidas "antes" y "después", respectivamente.
Para las hipótesis indicadas, el estadístico a emplear y su distribución son

$$T = \frac{\hat{X}_a - \hat{X}_d - (\mu_a - \mu_d)}{s_P\sqrt{1/n_a + 1/n_d}} \sim t_{n_a+n_d-2},$$

donde

$$s_P^2 = \frac{(n_a-1)s_a^2 + (n_d-1)s_d^2}{n_a+n_d-2} = \frac{19\times 0{,}415^2 + 19\times 0{,}521^2}{38} = 0{,}2218 \Rightarrow s_P = 0{,}4710$$

La región de rechazo vendrá dada por $\mathrm{RR}_\alpha = \{T > t_{38;0,05} = 1{,}686\}$. Y el valor del estadístico para las muestras recogidas es

$$t = \frac{3{,}1 - 2{,}9 - 0}{0{,}471\sqrt{1/19 + 1/19}} = 1{,}3088$$

Este valor no está en la región de rechazo, por lo que no podemos rechazar la hipótesis nula al nivel de significación indicado. Esto es, no podemos rechazar la posibilidad de que no se haya producido un descenso en el nivel de contaminación, para un nivel de significación del 5 %.

b) De la salida de Excel tenemos que el p-valor correspondiente a este contraste unilateral es 0,007444. Como este valor es menor que el nivel de significación indicado, 0,01, rechazamos la hipótesis nula y concluimos que se ha producido una reducción significativa en el nivel de concentración tras la campaña.

3. (4,5 puntos) Una empresa de investigación de mercados analiza periódicamente datos de consumo de un producto de primera necesidad. Las variables de interés son el consumo del producto (Y en kg por mes) y su precio (X_1 en euros por kg). Su interés es obtener una ecuación de regresión lineal (Y en función de X) a fin de determinar si existe una relación lineal significativa entre consumo y precio. Con este fin, ha recogido 16 observaciones de ambas variables en distintas zonas, obteniéndo los siguientes resultados:

$$\sum_{i=1}^{16} x_i = 54{,}51; \quad \sum_{i=1}^{16} y_i = 34{,}81; \quad \sum_{i=1}^{16} x_i^2 = 187{,}207; \quad \sum_{i=1}^{16} y_i^2 = 76{,}040$$
$$\sum_{i=1}^{16} x_i y_i = 118{,}113; \quad \sum_{i=1}^{16} e_i^2 = 0{,}1522.$$

a) (1 punto) Obtenga la tabla ANOVA correspondiente a la variable Y.

b) (0,5 puntos) Lleve a cabo un contraste al 5 % de significación para analizar la influencia del precio del producto en la demanda del mismo.

c) (0,5 puntos) Calcule el coeficiente de determinación e interprételo.

d) (0,5 puntos) Determine las estimaciones de mínimos cuadrados de los parámetros de la recta de regresión.

e) (1 punto) Estime el consumo para un caso en el que el precio del producto sea de 3,4 euros/kg. Proporcione, con un nivel de confianza del 95 %, un intervalo de confianza para dicha predicción.

Se ha incorporado información sobre el nivel de ingresos en cada zona como una variable adicional, X_2, para mejorar el modelo anterior.

Con los datos adicionales se ha ajustado un modelo de regresión múltiple en Excel, obteniendo la siguiente salida:

ANÁLISIS DE VARIANZA						
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
Regresión	2	0,212784577	0,106392288	14,80721126	0,000445158	
Residuos	13	0,093407173	0,007185167			
Total	15	0,30619175				
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	3,247319109	0,237013013	13,7010161	4,19761E-09	2,735283624	3,759354594
Variable X 1	-0,551627223	0,106383966	-5,18524776	0,000175541	-0,78145581	-0,321798637
Variable X 2	0,050496656	0,017651721	2,860721332	0,013379094	0,01236243	0,088630881

f) (0,5 puntos) ¿Es significativa la nueva variable X_2 a un nivel de significación del 5 %? ¿Y del 1 %? Justifique su respuesta.

g) (0,5 puntos) ¿Es globalmente significativo el modelo de regresión múltiple a un nivel de significación del 1 %? Justifique su respuesta.

Solución:

a) De los datos en el enunciado tenemos que $\text{SCR} = \sum_i e_i^2 = 0{,}1522$.
También tenemos (obsérvese que este valor también aparece en la tabla ANOVA del modelo de regresión múltiple):

$$\text{SCT} = \sum_{i=1}^{16}(y_i - \bar{y})^2 = (16-1)s_y^2 = \sum_{i=1}^{16} y_i^2 - 16\bar{y}^2 = 0{,}3062.$$

Con estos datos, la tabla ANOVA pedida queda como sigue:

Fuente	Suma cuadrados	G.L.	Promedio cuadrados	Razón-F
Modelo	0,1540	1	0,1540	14,163
Residuos	0,1522	14	0,01087	
Total	0,3062	15		

b) El valor crítico para el contraste de significación es $F_{1;14;0,05} = 4{,}60$.
Como se cumple que Razón-F $= 14{,}163 > F_{1;14;0,05} = 4{,}60$, podemos rechazar la hipótesis nula de este contraste (no influencia del precio en la demanda) a un nivel de significación del 5 %, y por tanto concluimos que existe una relación lineal significativa entre el precio y la demanda de este producto.

c) El coeficiente de determinación viene dado por:

$$R^2 = \frac{\text{SCM}}{\text{SCT}} = \frac{0{,}1540}{0{,}3062} = 0{,}5029$$

Por tanto, el precio explica un 50 % de la variabilidad en la demanda.

d) Las estimaciones de los parámetros de la recta de regresión se obtienen de las fórmulas de mínimos cuadrados como:

$$\hat{\beta}_1 = \frac{\text{cov}(X,Y)}{s_x^2} = \frac{\sum_{i=1}^{16} x_i y_i - 16\bar{x}\bar{y}}{\sum_{i=1}^{16} x_i^2 - 16\bar{x}^2} = -0{,}3206, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3{,}268.$$

e) La estimación puntual pedida para $x_0 = 3{,}4$ será:

$$\hat{y}_0 = 3{,}268 - 0{,}3206 \times 3{,}4 = 2{,}178.$$

Para obtener el intervalo de confianza, recordamos que la varianza residual (que también aparece en la tabla ANOVA) es

$$s_R^2 = \frac{\sum_{i=1}^{16} e_i^2}{16-2} = 0{,}01087.$$

Empleamos la fórmula del intervalo de confianza correspondiente a una predicción,

$$\begin{aligned} \text{IC}_{0,05} &= \hat{y}_0 \pm t_{n-2;0,025} \sqrt{s_R^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)} \\ &= 2{,}178 \pm 2{,}145 \sqrt{0{,}01087 \left(1 + \frac{1}{16} + \frac{(3{,}4 - 3{,}407)^2}{(16-1) \times 0{,}0999}\right)} \\ &= [1{,}947; 2{,}408]. \end{aligned}$$

f) El p-valor del contraste de significación individual de la variable X_2 aparece en la salida bajo la columna "Probabilidad". Dicho valor es 0,0134, y por tanto la variable es significativa a un nivel de significación del 5 %, pero no al 1 %.

g) Para llevar a cabo el contraste de significación global podemos utilizar el p-valor indicado en la tabla ANOVA bajo la columna "Valor crítico de F", en nuestro caso 0,00045. Como este valor es muy inferior a $\alpha = 0{,}01$, concluimos que el modelo es globalmente significativo.