Estadística II
Examen final enero 11/1/18
Curso 2017/18 – Soluciones

Duración del examen: 2 h y 15 min

---

1. (3 puntos) El Barómetro del CIS (Centro de Investigaciones Sociológicas) en su pregunta 4 recoge información sobre la percepción de la situación política en España. En el barómetro de septiembre de 2017 se han obtenido los siguientes resultados:

   Pregunta 4:
   Y refiriéndonos ahora a la situación política general en España, ¿cómo la calificaría usted?

	Sept 2017
Muy buena	5
Buena	66
Regular	576
Mala	893
Muy mala	865
N.S.	75
N.C.	10
(Total)	2490

   En base a la tabla anterior responda a los siguientes apartados:

   a) (1 punto) Obtenga un intervalo de confianza al 90 % para la proporción de españoles que considera la situación política en España mala o muy mala en 2017.

   b) (1 punto) El Gobierno afirma que la imagen que los españoles tienen de la situación política no ha empeorado en los últimos años. Más concretamente, considere la afirmación de que la proporción de españoles que valoran la situación política en 2017 como mala o muy mala no es mayor del 68 % (valor que se obtuvo en 2016). Contraste si esta afirmación es cierta utilizando el p-valor para alcanzar una conclusión.

   c) (1 punto) Calcule la potencia del contraste anterior si la verdadera proporción de españoles que valoran la situación política como mala o muy mala fuera de un 70 %, utilizando un nivel de significación del 5 %.

   Solución:

   a) Sea X una variable aleatoria que vale 1 si el individuo piensa que la situación política en España es mala o muy mala y 0 en otro caso.

   Supuestos:

   1) $X \sim Ber(p)$.
   2) $\underline{X}(2490)$. Tenemos una muestra aleatoria grande donde cada $X_i \sim iidBer(p)$.
   3) $\alpha = 0{,}1$

   El estimador muestral de la proporción es:

$$\hat{p}_x = \frac{893 + 865}{2490} = 0{,}706.$$

   Y por lo tanto el intervalo pedido será:

$$\mathrm{IC}_{0,9}(p) = \hat{p}_x \pm z_{0,05}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}} = 0{,}706 \pm 1{,}645\sqrt{\frac{0{,}706 \cdot 0{,}294}{2490}} = [0{,}6926; 0{,}7193]$$

   b) El contraste pedido es

$$\begin{aligned} H_0 &: \quad p \leq 0{,}68 \\ H_1 &: \quad p > 0{,}68 \end{aligned}$$

   Supuestos:

1) $X \sim Ber(p)$.

2) $\underline{X}(2490)$. Tenemos una muestra aleatoria grande donde cada $X_i \sim iidBer(p)$.

El estadístico de contraste será:

$$Z = \frac{\hat{p}_x - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0{,}706 - 0{,}68}{\sqrt{\frac{0{,}68 \cdot 0{,}32}{2490}}} = 2{,}7813$$

Tenemos que p-valor $= P(Z > 2{,}7813) \backsimeq 0{,}0027$. Solo con niveles de significación menores al $0{,}27\,\%$ podría mantenerse la hipótesis nula.

Conclusión: Existe evidencia muestral para afirmar que la proporción de españoles que piensan que la situación política en España es mala o muy mala es superior al $68\,\%$, para niveles de significación razonables.

c) Sustituyendo en la región de rechazo, obtenemos que:

$$\text{RR} = \{Z > Z_{0{,}05}\} = \left\{ \frac{\hat{p}_x - 0{,}68}{\sqrt{\frac{0{,}68 \cdot 0{,}32}{2490}}} > 1{,}645 \right\} \quad \Rightarrow \quad \text{RR} = \{\hat{p}_x > 0{,}6953\}.$$

Por lo tanto, si la verdadera proporción es $p = 0{,}7$,

$$\begin{aligned} \text{Potencia} &= 1 - \beta = P(\hat{p}_x > 0{,}6953 \mid p = 0{,}7) = P\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} > \frac{0{,}6953 - p}{\sqrt{p(1-p)/n}} \;\middle|\; p = 0{,}7 \right) \\ &= P\left(Z > \frac{0{,}6953 - 0{,}7}{\sqrt{\frac{0{,}7 \cdot 0{,}3}{2490}}} \right) = P(Z > -0{,}5121) = 0{,}6959. \end{aligned}$$

2. (3 puntos) Una consultora española quiere realizar un estudio sobre los salarios mensuales pagados por una entidad financiera a sus empleados. Para ello, se seleccionan aleatoriamente dos muestras de diez observaciones, una de hombres y otra de mujeres. De dichas muestras se obtienen los siguientes resultados a partir de los salarios expresados en miles de euros:

Muestra de hombres (x_i)	Muestra de mujeres (y_i)
$\sum_{i=1}^{10} x_i = 17{,}1$; $\sum_{i=1}^{10} x_i^2 = 29{,}75$	$\sum_{i=1}^{10} y_i = 13{,}5$; $\sum_{i=1}^{10} y_i^2 = 18{,}45$

a) (1 punto) ¿Se podría concluir que el salario medio pagado por la entidad a los hombres es superior al salario medio pagado a las mujeres? (utilice $\alpha = 5\,\%$). Especifique claramente:

- La hipótesis nula y la alternativa.
- Los supuestos necesarios para realizar el contraste.
- El estadístico de contraste y su distribución.
- La región de rechazo y el p-valor.
- La conclusión del contraste.

b) (1 punto) ¿Se podría afirmar que la variabilidad de los salarios entre los hombres es la misma que entre las mujeres? Realice el contraste de hipótesis con un nivel del significación del $10\,\%$.

c) (1 punto) Se ha aumentado el tamaño de ambas muestras a 20 observaciones en cada una de ellas y con las nuevas muestras se ha llevado a cabo en Excel el contraste del apartado $2a$), obteniendo la salida que se indica en la siguiente tabla.

Indique y justifique la conclusión a la que llegaría para el contraste con las nuevas muestras (utilice $\alpha = 5\,\%$).

Prueba t para dos muestras suponiendo varianzas iguales

	Variable 1	*Variable 2*
Media	1,63	1,44
Varianza	0,055894737	0,035157895
Observaciones	20	20
Varianza agrupada	0,045526316	
Diferencia hipotética de las medias	0	
Grados de libertad	38	
Estadístico t	2,815933197	
P(T<=t) una cola	0,003834542	
Valor crítico de t (una cola)	1,68595446	
P(T<=t) dos colas	0,007669084	
Valor crítico de t (dos colas)	2,024394164	

Solución:

a) *Supuestos necesarios:* Como los tamaños muestrales son menores a 30, necesitamos suponer normalidad y varianzas poblacionales iguales.
El contraste a realizar es:

$$H_0 : \mu_x \leq \mu_y \quad \text{y} \quad H_1 : \mu_x > \mu_y$$

El estadístico de contraste es:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \backsim t_{n_x+n_y-2} \equiv t_{18}$$

Para un nivel de significación del 5 %, la región de rechazo corresponde a todos aquellos valores del estadístico de contraste superiores a $t_{18;0,05} = 1{,}734$.
De los datos muestrales proporcionados se tiene que

$$\begin{aligned} \bar{x} &= 17{,}1/10 = 1{,}71, \quad \bar{y} = 13{,}5/10 = 1{,}35 \\ s_x^2 &= \frac{29{,}75 - 17{,}1^2/10}{9} = 0{,}0566, \quad s_y^2 = \frac{18{,}45 - 13{,}5^2/10}{9} = 0{,}025 \\ s_p^2 &= \frac{0{,}0566 + 0{,}025}{2} = 0{,}04078 \Rightarrow s_p = 0{,}2019 \end{aligned}$$

El estadístico de contraste evaluado bajo H_0 toma el valor:

$$T|_{H_0} = \frac{1{,}71 - 1{,}35}{0{,}2019\sqrt{\frac{1}{10} + \frac{1}{10}}} = 3{,}986$$

El p-valor vendrá dado por

$$\text{p-valor} = P(T > 3{,}986) \leq 0{,}001$$

Como el valor del estadístico de contraste pertenece a la región de rechazo, rechazamos la hipótesis nula, es decir, *podemos afirmar que el salario medio pagado por la entidad a los hombres es superior al de las mujeres.*

b) El contraste de hipótesis a realizar es el definido por

$$H_0 : \frac{\sigma_x^2}{\sigma_y^2} = 1 \quad \text{y} \quad H_1 : \frac{\sigma_x^2}{\sigma_y^2} \neq 1$$

Bajo el supuesto de normalidad, el estadístico de contraste es:

$$F = \frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \sim F_{n_x-1;n_y-1},$$

y bajo la hipótesis nula toma el valor

$$F = \frac{0{,}0566}{0{,}025} = 2{,}2622.$$

La región de rechazo vendrá definida por:

$$\text{RR} = \{F < F_{9;9}^{0,95} = 0{,}3144\} \cup \{F > F_{9;9}^{0,05} = 3{,}18\}$$

Al no pertenecer el valor de nuestro estadístico a la región de rechazo, concluimos que *no tenemos suficiente evidencia para rechazar la hipótesis de que la variabilidad entre los salarios de los hombres es la misma que la variabilidad de salarios entre las mujeres.*

c) Dado que el p-valor del contraste unilateral, 0,003834542, es menor al 5 %, rechazamos la hipótesis nula y concluimos (de nuevo) que *el salario medio pagado a los hombres es superior al salario medio pagado a las mujeres.*

3. (4 puntos) En una empresa se pretende analizar el absentismo laboral y las causas que explican este fenómeno. Para ello y como punto de partida se realizó un estudio en el que se ajustó un modelo de regresión lineal simple, donde el índice de absentismo laboral (variable Y) se intentó explicar en función de la remuneración mensual de los trabajadores (variable X, medida en miles de euros).

   Se trabajó con una muestra de 20 empleados elegidos aleatoriamente, obteniéndose los resultados siguientes:

   $$\sum_{i=1}^{20} y_i = 71{,}68, \quad \sum_{i=1}^{20} x_i = 43{,}9, \quad \sum_{i=1}^{20} y_i^2 = 322{,}57, \quad \sum_{i=1}^{20} x_i^2 = 109{,}11$$

   $$\sum_{i=1}^{20} x_i y_i = 138{,}11, \quad \sum_{i=1}^{20} e_i^2 = 36{,}69$$

   a) (1 punto) Estime la recta de regresión e interprete sus coeficientes.

   b) (0,5 puntos) Realize el contraste adecuado para determinar si entre las dos variables existe una relación lineal significativa. Utilice un nivel de significación del 5 %.

   c) (1 punto) ¿Cuál sería el índice previsto de absentismo laboral para un empleado que tuviese una remuneración mensual de 3000 euros? Calcule e interprete un intervalo de confianza al 95 % para dicha predicción.

   Para mejorar el modelo anterior, y utilizando la misma muestra, se añadieron dos nuevas variables explicativas: Años de antigüedad en la empresa (variable X_2) y un índice de satisfacción laboral (variable X_3) que mide aspectos no remunerativos como: adecuación al puesto de trabajo, horarios, ambiente laboral, etc.

   Los resultados de la nueva regresión lineal múltiple se muestran en la siguiente salida de Excel:

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	7,7687	0,9314	8,3409	0,0000	5,79420073	9,74314282
Variable X 1	-1,0260	0,4245	-2,4168	0,0280	-1,9258914	-0,1260499
Variable X 2	0,0070	0,0611	0,1143	0,9104	-0,1226383	0,13661636
Variable X 3	-0,6279	0,2486	-2,5251	0,0225	-1,1549577	-0,1007537

   d) (0,5 puntos) Calcule e interprete un intervalo de confianza al 99 % para el coeficiente asociado a la variable X_3.

   e) (0,5 puntos) Sabiendo que la varianza residual en este nuevo modelo es 1,627, construya la tabla ANOVA.

   f) (0,5 puntos) Calcule e interprete el valor del coeficiente de determinación.

   Solución:

   a) La estimación de los coeficientes del modelo da como resultado:

   $$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{138{,}11 - 20 \cdot 3{,}584 \cdot 2{,}195}{109{,}11 - 20 \cdot 2{,}195^2} = -1{,}5078 \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 6{,}8936.$$

   Por lo que la recta de regresión estimada es:

   $$\hat{y}_i = 6{,}8936 - 1{,}5078 x_i.$$

   La constante $\hat{\beta}_0$ no tiene interpretación económica puesto que el salario bruto mensual no puede ser igual a 0. Por cada 1000 euros más de salario bruto mensual, el índice de absentismo laboral disminuye en media 1,5078 puntos.

b) Para saber si entre las dos variables existe una relación lineal significativa debemos realizar el contraste:

$$H_0 : \beta_1 = 0 \quad \text{y} \quad H_1 : \beta_1 \neq 0.$$

El valor del estadístico de contraste es:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{s_R^2}{(n-1)s_x^2}}} = \frac{-1{,}5078}{\sqrt{\frac{2{,}0383}{19 \cdot 0{,}6710}}} = -3{,}771$$

donde hemos utilizado el valor $s_R^2 = \sum_i e_i^2/(n-2) = 36{,}69/18 = 2{,}0383$.
Como se tiene que $t_{18;0{,}025} = 2{,}101$, el valor del estadístico está en la región de rechazo, se rechaza la hipótesis nula y se concluye que existe una relación lineal significativa entre las dos variables.

c) Si consideramos un salario bruto mensual de 3000 euros ($x_0 = 3$), la predicción del índice de absentismo laboral sería:

$$\hat{y}_0 = 6{,}8936 - 1{,}5078 \cdot 3 = 2{,}3702.$$

El intervalo de confianza pedido es entonces,

$$\begin{aligned} \text{IC}_{0,95}(y_0) &= \hat{y}_0 \pm t_{n-2}^{\alpha/2} \sqrt{s_R^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)} \\ &= 2{,}3702 \pm 2{,}101 \sqrt{2{,}0383 \left(1 + \frac{1}{20} + \frac{(3 - 2{,}195)^2}{(19)0{,}67102}\right)} = [-0{,}7767; 5{,}5172]. \end{aligned}$$

Por tanto, con una confianza del 95 % el valor del índice de absentismo observado en un trabajador con una remuneración mensual de 3000 euros estaría entre $-0{,}7767$ y $5{,}5172$.

d) El intervalo para el coeficiente asociado a la variable x_3 puede obtenerse como,

$$\text{IC}_{0,99}(\beta_3) = \hat{\beta}_3 \pm t_{n-4}^{0,005} \cdot \text{se}(\hat{\beta}_3) = -0{,}6279 \pm 2{,}921 \cdot 0{,}2486 = [-1{,}3541; 0{,}0984],$$

donde hemos obtenido el valor estimado ($\hat{\beta}_3$) y el error estándar ($\text{se}(\hat{\beta}_3)$) de la salida de Excel.
Con una confianza del 99 %, la variable x_3 no es significativa, puesto que el valor 0 está incluido en el intervalo.

e) Nos basamos en los valores siguientes:

$$\begin{aligned} \text{SCR} &= (n-k-1)s_R^2 = 16 \cdot 1{,}627 = 26{,}032 \\ \text{SCT} &= (n-1)s_y^2 = 322{,}57 - 71{,}68^2/20 = 65{,}669 \end{aligned}$$

La tabla ANOVA completa sería:

ANÁLISIS DE VARIANZA				
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*
Regresión	3	39,637	13,212	8,121
Residuos	16	26,032	1,627	
Total	19	65,669		

f) El valor del coeficiente de determinación de este modelo es:

$$R^2 = \frac{\text{SCM}}{\text{SCT}} = \frac{39{,}637}{65{,}669} = 0{,}6036.$$

El modelo explica un 60,36 % de la variabilidad en el índice de absentismo laboral.