EDIP

Examen I

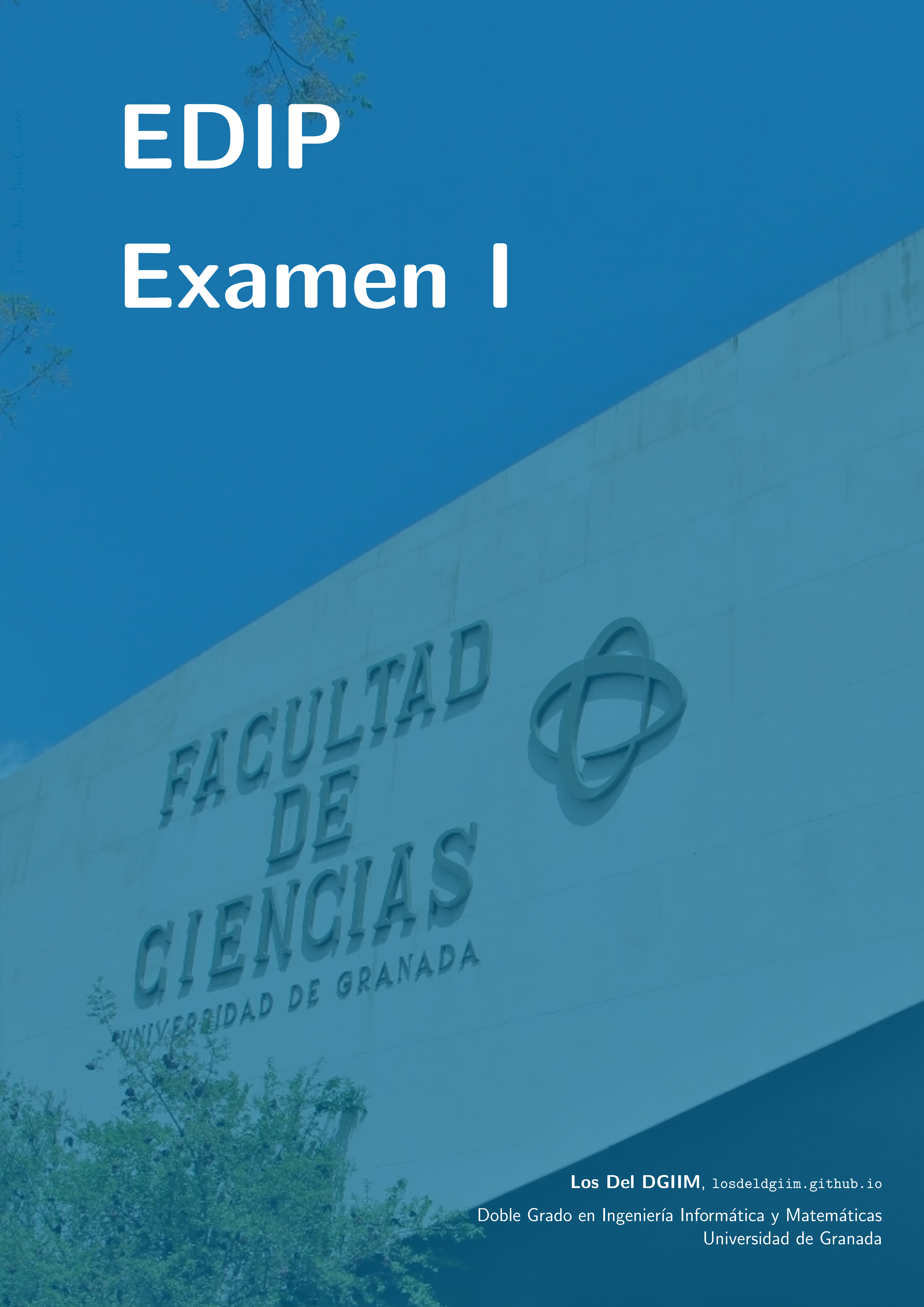

Foto: José Juan Castro

Los Del DGIIM, `losdeldgiim.github.io`

Doble Grado en Ingeniería Informática y Matemáticas
Universidad de Granada

EDIP
Examen I

Los Del DGIIM, `losdeldgiim.github.io`

Arturo Olivares Martos

Granada, 2023

Asignatura Estadística Descriptiva e Introducción a la Probabilidad.

Curso Académico 2022-23.

Grado Doble Grado en Ingeniería Informática y Matemáticas.

Grupo Único.

Profesor Fernando Jesús Navas Gómez.

Descripción Parcial. Parte de Estadística Descriptiva.

Fecha 27 de abril de 2023.

Ejercicio 1. [2 puntos] Sea (X, Y) una variable estadística bidimensional con valores (x_i, y_j), $i = 1, \dots, k$, $j = 1, \dots, p$. Contestar razonadamente a las siguientes cuestiones:

1. **[0.75 puntos]** Ajustar por el método de mínimos cuadrados un modelo del tipo $Y = ax^2 + 3x$. Comprobar que el valor de a es un mínimo.

   En el ajuste mediante mínimos cuadrados, buscamos minimizar el error cuadrático medio ECM. Determinamos en primer lugar su expresión.

$$\Psi(a) = ECM(a, x) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} e_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} (y_j - f(x_i))^2$$

   Como en nuestro caso $f(x) = ax^2 + 3x$, tenemos que:

$$\Psi(a) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} [y_j - (ax_i^2 + 3x_i)]^2$$

   Para hallar el mínimo del ECM, derivamos parcialmente respecto de a.

$$\frac{\partial \Psi}{\partial a} = -2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} [y_j - (ax_i^2 + 3x_i)] x_i^2$$

   Como, al ser el ECM derivable, el mínimo anula la primera derivada, buscamos los valores que anulan la primera derivada:

$$\begin{aligned}
\frac{\partial \Psi}{\partial a} = 0 &\iff \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} [y_j - (ax_i^2 + 3x_i)] x_i^2 = 0 \\
&\iff \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} y_j x_i^2 - f_{ij} x_i^2 (ax_i^2 + 3x_i) = 0 \\
&\iff \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} y_j x_i^2 - f_{ij} x_i^4 a - 3 f_{ij} x_i^3 = 0 \\
&\iff m_{21} - a m_{40} - 3 m_{30} = 0 \\
&\iff a = \frac{m_{21} - 3 m_{30}}{m_{40}}
\end{aligned}$$

   Por tanto, ya tenemos realizado el ajuste. Para comprobar que es un mínimo, simplemente hay que demostrar que el candidato a extremo relativo es un mínimo. Para ello, se puede proceder de diversas formas. Por ejemplo, se puede optar por que el coeficiente líder de $\Psi(a)$ es $\sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i^4 > 0$, por lo que se trata de una parábola convexa, y por tanto su extremo relativo es un mínimo absoluto. Otra opción es determinar la segunda derivada:

$$\frac{\partial^2 \Psi}{\partial a^2} = -2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i^2 (-x_i^2) = 2 \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} x_i^4 > 0$$

Por tanto, como la segunda derivada es positiva, tenemos que efectivamente se trata de un mínimo relativo. Como el Ψ es continua y solo tiene un extremo relativo, dicho valor de a es mínimo absoluto.

2. **[0.4 puntos]** Si $\bar{x} = 1$ e $\bar{y} = 3$, determinar las condiciones para que la varianza de los residuos coincida con la media de los residuos.

Calculamos en primer lugar la media de los residuos. Sea $\bar{y}_i = f(x_i)$.

$$\bar{e} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}e_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - f(x_i)) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[y_j - (ax_i^2 + 3x_i)] =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[y_j - (ax_i^2 + 3x_i)] = m_{01} - \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}ax_i^2 + f_{ij}3x_i = m_{01} - am_{20} - 3m_{10}$$

Usando los valores dados por el enunciado, tenemos que:

$$\bar{e} = 3 - am_{20} - 3 = -am_{20}$$

Calculamos ahora la varianza de los residuos:

$$\sigma_r^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}e_{ij}^2 - \bar{e}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - f(x_i))^2 - \bar{e}^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}y_j^2 + f_{ij}(ax_i^2 + 3x_i)^2 - 2f_{ij}y_j(ax_i^2 + 3x_i) - \bar{e}^2 =$$

$$= m_{02} + \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(a^2x_i^4 + 9x_i^2 + 6ax_i^3) - 2am_{21} - 6m_{11} - \bar{e}^2 =$$

$$= m_{02} + a^2m_{40} + 9m_{20} + 6am_{30} - 2am_{21} - 6m_{11} - a^2m_{20}^2$$

$$\sigma_r^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(e_{ij} - \bar{e})^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - (ax_i^2 + 3x_i) + am_{20})^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[y_j^2 + (ax_i^2+3x_i)^2 + a^2m_{20}^2 + 2y_jam_{20} - 2y_j(ax_i^2+3x_i) - 2am_{20}(ax_i^2+3x_i)] =$$

$$= m_{02} + a^2m_{40} + 9m_{20} + 6am_{30} + a^2m_{20}^2 + 2am_{10}m_{20} - 2am_{21} - 6m_{11} - 2a^2m_{20}^2 + 6am_{20}m_{10} =$$

$$= m_{02} + a^2m_{40} + 9m_{20} + 6am_{30} + 8am_{10}m_{20} - 2am_{21} - 6m_{11} - a^2m_{20}^2$$

Por tanto, es necesario que $8am_{10}m_{20} = 0$. Como $m_{10} = \bar{x} = 1$, tenemos que es necesario que:

$$8am_{20} = 0 \Longleftrightarrow (m_{21} - 3m_{30})m_{20} = 0 \Longleftrightarrow \left\{ \begin{array}{l} m_{21} = 3m_{30} \\ \qquad \vee \\ m_{20} = 0 \Longleftrightarrow x_i = \bar{x} = 1 \quad \forall i \end{array} \right.$$

3. [**0.85 puntos**] Consideramos la variable $Z = 3X - 2Y$:

   a) [**0.65 puntos**] Deducir la covarianza entre las variables Z y X en términos de σ_{xy}.

   Tenemos que $z_{ij} = 3x_i - 2y_j$. Calculamos $\bar{z}$:

$$\bar{z} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} z_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(3x_i - 2y_j) = 3\bar{x} - 2\bar{y}$$

   Calculamos por tanto la covarianza buscada:

$$\sigma_{xz}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(z_{ij} - \bar{z})(x_i - \bar{x}) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(3x_i - 2y_j - 3\bar{x} + 2\bar{y})(x_i - \bar{x}) =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[3(x_i - \bar{x}) - 2(y_j - \bar{y})](x_i - \bar{x}) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[3(x_i - \bar{x})^2 - 2(y_j - \bar{y})(x_i - \bar{x})] =$$

$$= 3\mu_{20} - 2\mu_{11} = 3\sigma_x^2 - 2\sigma_{xy}$$

   b) [**0.2 puntos**] Determinar la covarianza entre las variables Z y X si se sabe que las variables X y Y son independientes.

   Como X e y son independientes, tenemos que $\sigma_{xy} = 0$. Por tanto, se tiene que $\sigma_{zx} = 2\sigma_x^2$.

Ejercicio 2. Indica la opción correcta:

(a) El coeficiente de variación de una variable tipificada es nulo. Una variable Z es tipificada si $Z = \frac{X - \bar{x}}{\sigma_x}$.

Se le está aplicando una transformación lineal, por lo que:

$$\bar{z} = \frac{\bar{x} - \bar{x}}{\sigma_x} = 0 \qquad\qquad \sigma_z^2 = \frac{1}{\sigma_x^2} \cdot \sigma_x^2 = 1$$

Por tanto, tenemos que:

$$C.V.(Z) = \frac{\sigma_z}{|\bar{z}|} = \frac{1}{0}$$

Por tanto, para variables tipificadas no está definido. Es **falso**.

(b) El coeficiente de determinación, en el caso de regresión lineal, coincide con el coeficiente de correlación lineal.

Falso, ya que $r = \pm\sqrt{r^2} \iff r = 0, 1$. Por tanto, por norma general no se da.

(c) Si el valor de la vivienda se ha incrementado un $2\,\%, 3\,\%, 10\,\%$ y $9\,\%$, respectivamente durante los últimos 4 años, el incremento medio anual del valor de la vivienda durante dicho periodo ha sido de un $8\,\%$.

En este caso, al tratarse de incrementos tenemos que se trata de media geométrica. Por tanto,

$$G = \sqrt[4]{1{,}02 \cdot 1{,}03 \cdot 1{,}1 \cdot 1{,}09} = 1{,}059 \Longrightarrow 5{,}9\,\%$$

Por tanto, tenemos que es falso.

(d) Todas las afirmaciones anteriores son falsas.

Ejercicio 3. El cambio de origen y escala, $Y = \frac{X-x_0}{a}$, afecta a los momentos centrales de la siguiente forma:

(a) $\mu_3^3(X) = a^3\mu_3^3(Y)$

(b) $\mu_3(Y) = a^3\mu_3(X)$

(c) $\boldsymbol{\mu_3(X) = a^3\mu_3(Y)}$

(d) $\mu_3(Y) = a^3\mu_3(X)$

Calculamos el siguiente momento central:

$$\mu_3(Y) = \sum_{i=1}^{k} f_i(y_i - \bar{y})^3 = \sum_{i=1}^{k} f_i\left(\frac{x_i}{a} - \cancel{\frac{x_0}{a}} - \frac{\bar{x}}{a} + \cancel{\frac{x_0}{a}}\right)^3 = \sum_{i=1}^{k} f_i\frac{(x_i-\bar{x})^3}{a^3} = \frac{\mu_3(X)}{a^3} \Longrightarrow$$
$$\Longrightarrow \mu_3(X) = a^3\mu_3(Y)$$

donde he aplicado que, por ser una transformación afín, tenemos que $\bar{y} = \frac{\bar{x}-x_0}{a}$. Por tanto, tenemos que la opción correcta es la (c).

Ejercicio 4. La recta de regresión de Y sobre X es $y = 5$, y $\sigma_Y^2 = 2$. Entonces:

(a) $\boldsymbol{\eta^2_{Y/X} = 0}$

Tenemos que la recta de regresión es:

$$Y = \frac{\sigma_{xy}}{\sigma_x^2}x + \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x} = 5$$

Por tanto, deducimos que $\sigma_{xy} = 0$, $\sigma_x^2 \in \mathbb{R}^*$. Por tanto,

$$\eta^2_{Y/X} = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} = 0$$

(b) $\eta^2_{Y/X} = 1$

(c) Los residuos de la recta son todos nulos.

Para que todos los residuos de la recta fuesen nulos, tendrían que depender linealmente y tener $\eta_{Y/X} = 1$. No obstante, esto no se da, por lo que no son nulos.

(d) La varianza de los residuos de la recta es 2.

Por ser un ajuste lineal en los parámetros, tenemos que:

$$\sigma_y^2 = \sigma_{ey}^2 + \sigma_{ry}^2 \Longrightarrow \sigma_{ry}^2 = \sigma_y^2 - \sigma_{ey}^2 = \sigma_y^2 - \sigma_y^2\eta^2_{Y/X} = \sigma_y^2 = 2$$

Ejercicio 5. Indica la afirmación correecta:

(a) Dos variables estadísticas son independientes si y solo si su covarianza es nula.

Falso, ya que la implicación hacia la izquierda no se da.

(b) Dos variables estadísticas son independientes si su coeficiente de correlación es nulo.

Falso, la implicación va en sentido contrario.

(c) Los coeficientes de determinación lineal de Y/X y X/Y pueden no coincidir.

Falso, ya que por definición coinciden.

(d) Sean X y Y dos variables estadísticas con $\sigma_{\mathbf{xy}} = \mathbf{0}$. Entonces, podemos afirmar que $\mathbf{m_{11} = m_{10}m_{01}}$.

Cierto, ya que:

$$\sigma_{xy} = \mu_{11} = m_{11} - m_{10}m_{01} = 0 \Longleftrightarrow m_{11} = m_{10}m_{01}$$

Ejercicio 6. Se han tomado 50 mediciones de láminas de acero de distintos grosores, en mm, (Y) y la temperatura en $°C$, (X), que éstas pueden alcanzar hasta su fundición. La siguiente tabla muestra los resultados obtenidos:

X/Y	$(1-3]$	$(3-6]$	$(6-10]$	$n_{i.}$	$h_{i.}$
$(0-20]$	8	5	6	19	19/20
$(20-35]$	2	8	3	13	13/15
$(35-40]$	3	7	8	18	18/5
$n_{.j}$	13	20	17	50	

Contestar a las siguientes cuestiones:

1. **[0.75 puntos]** Determina el valor medio más representativo.

Calculamos el coeficiente de variación de Pearson marginal en cada caso. Para ello, calculamos previamente la media marginal de cada variable y la desviación típica.

$$\bar{x} = \frac{1}{50}\sum_{i=1}^{3} n_{i.}c_i = \frac{1222{,}5}{50} = 24{,}45$$

$$\bar{y} = \frac{1}{50}\sum_{j=1}^{3} n_{.j}c_j = \frac{252}{50} = 5{,}04$$

$$\sigma_x^2 = \frac{1}{50}\sum_{i=1}^{3} n_{i.}c_i^2 - \bar{x}^2 = \frac{37043{,}75}{50} - \bar{x}^2 = 143{,}0725$$

$$\sigma_y^2 = \frac{1}{50}\sum_{j=1}^{3} n_{.j}c_j^2 - \bar{y}^2 = \frac{1545}{50} - \bar{y}^2 = 5{,}4984$$

Por tanto, tenemos que los coeficientes son:

$$CV(X) = \frac{\sigma_x}{\bar{x}} \approx 0{,}48921 \qquad CV(Y) = \frac{\sigma_y}{\bar{y}} \approx 0{,}46525$$

Por tanto, tenemos que la media de Y es más representativa.

2. **[1 punto]** Determina la temperatura más frecuente para fundir láminas cuyo grosor es como máximo 6 mm.

   Tenemos que se condicona a que $y \leqslant 6$, por lo que la tabla de la distribucón es:

X/Y	$(1-3]$	$(3-6]$		$n_{i.}^{j=1,2}$	$h_{i.}^{j=1,2}$
$(0-20]$	8	5		13	13/20
$(20-35]$	2	8		10	10/15
$(35-40]$	3	7		10	10/5
$n_{.j}$	13	20		23	

   Buscamos en primer lugar el intervalo modal. Este es el que tiene la densidad de frecuencia h_i mayor, que como podemos ver es el último, I_3. Entonces, interpolamos el valor de la moda en dicho intervalo.

$$\frac{Mo_x - e_i}{e_{i+1} - Mo_x} = \frac{h_i - h_{i-1}}{h_i - h_{i+1}} \Longrightarrow \frac{Mo_x - 35}{40 - Mo_x} = \frac{2 - \frac{2}{3}}{2 - 0} \Longrightarrow$$
$$\Longrightarrow 2Mo_x - 70 = 80 - \frac{80}{3} - 2Mo_x + \frac{2}{3}Mo_x \Longrightarrow \frac{10}{3}Mo_x = \frac{370}{3} \Longrightarrow Mo_x = 37$$

   donde hay que tener en cuenta que $h_{i+1} = 0$, ya que no hay más intervalos en la distribución.

3. **[1 punto]** Determina el porcentaje de láminas de acero en las que la temperatura es superior a $25°C$, si el grosor es superior a 3 mm.

   Tomamos la distribución condicionada a un grosor mayor a 3 mm, es decir, $j = 2, 3$.

X/Y	$(3-6]$	$(6-10]$		$n_{i.}^{j=2,3}$	$N_{i.}^{j=2,3}$
$(0-20]$	5	6		11	11
$(20-35]$	8	3		11	22
$(35-40]$	7	8		15	37
$n_{.j}$	20	17		37	

   Buscamos $P_\alpha = 25$. Como no hay ningún intervalo que comience en el 25, y tenemos que $25 \in (20, 35]$, entonces:

$$25 = P_\alpha = e_i + \frac{\frac{n^{j=2,3}r}{100} - N_{i-1}}{N_i - N_{i-1}} \cdot a_i = 20 + \frac{\frac{37r}{100} - 11}{11} \cdot 15 \Longleftrightarrow 5 = \frac{\frac{37r}{100} - 11}{11} \cdot 15 \Longleftrightarrow$$
$$\Longleftrightarrow \frac{11}{3} = \frac{37r}{100} - 11 \Longleftrightarrow r = 39.\overline{639}$$

   Por tanto, tenemos que el porcentaje que se encuentra por encima es

$$100 - r = 60.\overline{360}\,\%$$

En la siguiente tabla de observan las variables X e Y para 5 láminas de acero distintas:

X	10,5	16,8	27,5	32,7	37,5
Y	2	4	5	8	9

4. **[2.5 puntos]** Ajustar mediante un modelo hiperbólico. ¿Es este ajuste mejor que un ajuste lineal para Y? Interprete los resultados.

Buscamos ajustarla de la forma $y = az + b$, donde $z = \frac{1}{x}$. Tenemos que:

$$y - \bar{y} = \frac{\sigma_{zy}}{\sigma_z^2}(z - \bar{z})$$

$$\bar{y} = \frac{1}{5}\sum_{j=1}^{5} n_{.j} y_j = 5{,}6 \qquad \bar{z} = \frac{1}{5}\sum_{i=1}^{5} n_{i.} z_i = \frac{1}{5}\sum_{i=1}^{5} \frac{n_{i.}}{x_i} = 0{,}04967$$

$$\sigma_y^2 = \frac{1}{5}\sum_{j=1}^{5} n_{.j} y_j^2 - \bar{y}^2 = \frac{190}{5} - \bar{y}^2 = 6{,}64$$

$$\sigma_z^2 = \frac{1}{5}\sum_{i=1}^{5} n_{i.} z_i^2 - \bar{z}^2 = \frac{1}{5}\sum_{i=1}^{5} \frac{n_{i.}}{x_i^2} - \bar{z}^2 = \frac{0{,}015582}{5} - \bar{z}^2 = 0{,}648829 \cdot 10^{-3}$$

$$\sigma_{zy} = \frac{1}{5}\sum_{i,j=1}^{5} n_{ij} z_i y_j - \bar{z}\bar{y} = \frac{1}{5}\sum_{i,j=1}^{5} \frac{y_j}{x_i} - \bar{z}\bar{y} = \frac{1{,}095}{5} - \bar{z}\bar{y} = -0{,}059144$$

Por tanto, tenemos que el ajuste hiperbólico es:

$$y = -91{,}155627z + 10{,}1277 \Longrightarrow y = -91{,}155627 \cdot \frac{1}{x} + 10{,}1277$$

Para estudiar la bondad de los ajustes calculamos r^2. En el caso hiperbólico,

$$r^2 = \frac{\sigma_{zy}^2}{\sigma_z^2 \sigma_y^2} = 0{,}8119$$

Para el caso lineal, calculamos los siguientes resultados previos:

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} n_{i.} x_i = \frac{125}{5} = 25$$

$$\sigma_x^2 = \frac{1}{5}\sum_{i=1}^{5} n_{i.} x_i^2 - \bar{x}^2 = \frac{3624{,}28}{5} - \bar{x}^2 = 99{,}856$$

$$\sigma_{xy} = \frac{1}{5}\sum_{i,j=1}^{5} n_{ij} x_i y_j - \bar{x}\bar{y} = \frac{824{,}8}{5} - \bar{x}\bar{y} = 24{,}96$$

Por tanto, calculamos r^2 en el caso lineal:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = 0{,}9396$$

Por tanto, como r^2 en el caso lineal es mayor, tenemos que el ajuste lineal es mejor. Explica el 93,96 % de los casos.

5. [**0.75 puntos**] Estudia la interdependencia lineal.

   Estamos estudiando la interdependencia entre X e Y. Tenemos que:

$$r = \ +\sqrt{r^2} = \ 0{,}9693$$

   donde he elegido el valor positivo ya que la covarianza es positiva. Por tanto, tenemos que están muy relacionadas linealmente, ya que $r \approx 1$. Por tanto, se ajustan prácticamente a una recta. Además, como $r > 0$, tenemos que la correlación es positiva.