ESTADÍSTICA II (2022/23). Grado en Matemáticas
Examen final, 16 de enero de 2023

APELLIDOS: ______________________________ **Nombre:** ______________

En todos los contrastes de hipótesis se deberán plantear claramente las hipótesis nula H_0 y alternativa H_1.
La nota máxima de este examen son 10 puntos.

Problema 1: (2.5 puntos) Sea $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ un vector aleatorio $N_3(\mathbf{0}_3, \mathbf{I}_3)$.

a) **(1 punto)** Determinar la distribución de X_2 condicionada por X_1, siendo $\mathbf{X} = (X_1, X_2)' = (0, 1)' + \mathbf{B}\mathbf{Y}$, con

$$\mathbf{B} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & -1 \end{pmatrix}.$$

b) **(1.5 puntos)** Determinar la matriz $\mathbf{A}$ simétrica e idempotente que permite expresar la forma cuadrática

$$Q = \frac{(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + (Y_3 - Y_1)^2}{3}$$

en la forma $Q = \mathbf{Y}'\mathbf{A}\mathbf{Y}$. ¿Cuál es la distribución de Q?

Problema 2: (2.5 puntos) Se han obtenido de la web del INE el número de defunciones en España desagregadas según la causa de la muerte y según el sexo:

	Hombre	Mujer
Enfermedades infecciosas y parasitarias	25728	20273
Tumores	67844	45818
Enfermedades del sistema circulatorio	55905	63291
Causas externas de mortalidad	10689	6142

a) **(2 puntos)** A nivel de significación 0.01, ¿hay evidencia de que la causa de defunción y el sexo sean dependientes? ¿Qué puedes decir acerca del p-valor del contraste?

b) **(0.5 puntos)** Debajo se han escrito dos posibles maneras de hacer el contraste de (a) en `R` y sus correspondientes resultados. Razona cuál de los dos es el código adecuado para resolver el test de (a) y qué contraste resuelve el otro código. La respuesta debe ser breve: no es necesario hacer cálculos.

```
Datos <- c(25728,67844,55905,10689,20273,45818,63291,6142)
chisq.test(Datos,correct=F)

        Chi-squared test for given probabilities

data:  Datos
X-squared = 111713, df = 7

Datos <- matrix(c(25728,67844,55905,10689,20273,45818,63291,6142),nrow=4)
chisq.test(Datos,correct=F)

        Pearson's Chi-squared test

data:  Datos
X-squared = 4579.5, df = 3
```

Problema 3: (2.5 puntos) Consideremos un modelo de regresión lineal en el que se quiere estudiar el peso Y de un individuo en términos de su estatura x_3, pero teniendo en cuenta el sexo del individuo. Se plantea el siguiente modelo:

$$Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad i = 1, \ldots, n,$$

donde la variable x_1 **toma el valor 1 si el individuo es mujer y 0 si es hombre**, y la variable x_2 **toma el valor 0 si es mujer y 1 si es hombre**. Suponemos que $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$ y que los individuos son independientes entre sí. Denotamos por n_1 el número de mujeres y por $n_2 = n - n_1$ el número de hombres. Suponemos que la muestra está ordenada de tal manera que los primeros n_1 individuos son las mujeres y los últimos son los hombres.

a) **(0.5 puntos)** Obtener la expresión de la suma de cuadrados residual RSS para el modelo del enunciado.

b) **(2 puntos)** Determinar los estimadores de mínimos cuadrados de los parámetros β_j, $j = 1, 2, 3$. ¿Qué diferencias observas entre estos estimadores y los obtenidos en el modelo clásico de regresión simple, $Y = \beta_0 + \beta_3 x_3 + \epsilon$?

Problema 4: (2.5 puntos) En un estudio se analizó la asociación entre la remisión (`Y`=1=Sí, `Y`=0=No) del cáncer en unos pacientes y un índice, `LI`, que mide la actividad proliferativa de las células después de que el paciente recibe una inyección de timidina tritiada. La tabla siguiente resume la información muestral:

LI	Nº de casos	Nº de remisiones	LI	Nº de casos	Nº de remisiones
8	2	0	22	2	1
10	2	0	24	1	0
12	3	0	26	1	1
14	3	0	28	1	1
16	3	0	32	1	0
18	1	1	34	1	1
20	3	2	38	3	2

Al ajustar con `R` un modelo de regresión logística a estos datos se obtiene:

```
RegLog <- glm(Y~LI,family=binomial)
summary(RegLog)

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -3.77714    1.37862  -2.740  0.00615
  LI           0.14486    0.05934     A    #######

      Null deviance: 34.372  on 26  degrees of freedom
  Residual deviance: 26.073  on 25  degrees of freedom
  AIC: 30.073
```

a) (1 punto) Plantea el modelo logístico ajustado. Estima la probabilidad de remisión en un paciente con `LI = 28`. En base a esta estimación ¿en cuál de los dos grupos se le clasificaría?

b) (1 punto) Calcula el valor de `A`. A nivel $\alpha = 0.05$, ¿hay suficiente evidencia de que un **aumento** del índice `LI` **aumenta** la probabilidad de remisión del cáncer?

c) (0.5 puntos) Estima la disparidad O (*odds*) para un paciente con `LI = 34`. Interpreta el valor obtenido.