ESTADÍSTICA II (2021/22). Grado en Matemáticas
Examen final extraordinario, 20 de junio de 2022

Nombre: _____

En todos los contrastes de hipótesis se deberán plantear claramente las hipótesis nula H_0 y alternativa H_1. Todas las respuestas deberán ser razonadas.

Problema 1: (2 puntos) La Encuesta Nacional de Salud (ENSE), realizada por el Instituto Nacional de Estadística (INE), tiene como objetivo proporcionar información sobre la salud de la población española para poder planificar actuaciones en materia sanitaria. En la ENSE de 2017, en el apartado relativo a determinantes de salud, una de las preguntas fue el nivel de actividad física que realizaba el encuestado. La Tabla 1 muestra los resultados obtenidos (número de individuos por nivel de actividad física), desagregados por sexos.

	Nivel ejercicio físico		
Sexo	Alto	Moderado	Bajo
Hombres	4955	5922	5482
Mujeres	3027	7327	6091

Tabla 1

A un nivel de significación del 5 % contrastar si el nivel de ejercicio físico se distribuye de manera diferente según el sexo del individuo. Decir todo lo que se puede acerca del p-valor del contraste.

Problema 2: (2 puntos) Sean $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ y $\mathbf{X}_4$ vectores aleatorios independientes, cada uno de ellos con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Determinar la distribución del vector

$$\mathbf{V} = \frac{1}{4}\mathbf{X}_1 - \frac{1}{4}\mathbf{X}_2 + \frac{1}{4}\mathbf{X}_3 - \frac{1}{4}\mathbf{X}_4.$$

Problema 3: **(3 puntos)** Consideramos un conjunto de pacientes en cada uno de los cuales se han observado las siguientes variables biomecánicas relacionadas con la forma y orientación de la pelvis y de la columna vertebral[1]:

$$
\begin{array}{rcl}
X_1 & = & \text{Incidencia pélvica} \\
X_2 & = & \text{Basculación pélvica} \\
X_3 & = & \text{Ángulo de lordosis lumbar} \\
X_4 & = & \text{Inclinación sacra} \\
X_5 & = & \text{Radio pélvico} \\
X_6 & = & \text{Grado de espondilolistesis}
\end{array}
$$

a) **(1 p.)** Los datos de las variables X_i, $i = 1, \ldots, 6$, de un grupo de 100 pacientes normales (sin patologías; grupo de control `NO`) se almacenan en `R` en la matriz `column_3CNO`. Explica línea a línea qué hace el siguiente código e interpreta el resultado obtenido en términos de la distribución del vector $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)'$:

```
m <- colMeans(column_3CNO)
S = cov(column_3CNO)
D22 <- mahalanobis(column_3CNO,m,S)
ks.test(D22,"pchisq",6)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  D22
## D = 0.092788, p-value = 0.3554
## alternative hypothesis: two-sided
```

b) **(2 p.)** Ahora también consideramos un grupo de 150 pacientes con espondilolistesis (grupo `SL`). Los datos de las variables X_i, $i = 1, \ldots, 6$ de este grupo se almacenan en `R` en la matriz `column_3CSL`. Hacemos el siguiente cálculo:

```
colMeans(column_3CNO)
```

```
##       V1       V2       V3       V4       V5       V6
##  51.6856  12.8218  43.5423  38.8638 123.8912   2.1870
```

```
colMeans(column_3CSL)
```

```
##        V1        V2        V3        V4        V5        V6
##  71.51367  20.74800  64.10987  50.76613 114.51833  51.89687
```

Juntamos las observaciones de $\mathbf{X}$ en una matriz `X` de dimensión 250×6 y el grupo al que pertenece cada paciente (`NO` o `SL`) en un vector `Y` de 250 componentes. Explica qué hace el siguiente código de `R` y qué resultado se obtiene con `L$scaling`:

```
library(MASS)
L <- lda(X,Y)
L$scaling
```

```
##            LD1
## V1 -9.62322674
## V2  9.62873437
## V3  0.03466479
## V4  9.62008282
## V5 -0.03986606
## V6  0.02267635
```

Utilizando la anterior información, clasificar razonadamente a un nuevo paciente con $\mathbf{x} = (55, 15, 45, 40, 130, 4)'$ en uno de los dos grupos.

[1] Fuente de los datos: `http://archive.ics.uci.edu/ml/datasets/Vertebral+Column`

Problema 4: (3 puntos) Utilizando datos de varios países correspondientes a los años 50 del siglo XX se han ajustado dos modelos de regresión lineal. El objetivo es estudiar la tasa de natalidad (nat) en función de algunas variables socioeconómicas: x_1 = renta per cápita (`renta`), x_2 = proporción de la población que vive en granjas (`granja`) y x_3 = tasa de mortalidad infantil (`mortinf`). Reproducimos algunos de los resultados obtenidos:

```
Call:
lm(formula = nat ~ renta + granja + mortinf, data = birth)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.553868   7.898060   0.703  0.48818
renta        0.006566   0.006239   1.052  0.30225
granja       9.104755          A   0.710  0.48420
mortinf      0.242690   0.072845   3.332  !!!!!!!


Residual standard error: B on 26 degrees of freedom
Multiple R-squared:  D, Adjusted R-squared:  0.403
F-statistic: 7.525 on  and  DF,  p-value: !!!!!!!

----------

Call:
lm(formula = nat ~ mortinf, data = birth)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.43237    2.75145   4.882 3.83e-05
mortinf      0.21574    0.04587   4.703 6.25e-05

----------

Analysis of Variance Table

Model 1: nat ~ mortinf
Model 2: nat ~ renta + granja + mortinf
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 1486.2
2     26 1423.8  2    62.296      C 0.5731
```

A partir de estos resultados, contesta a las siguientes preguntas:

a) **(1 p.)** En el modelo completo, lleva a cabo los contrastes ($\alpha = 0.05$) de las dos hipótesis nulas siguientes: (i) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, y (ii) $H_0 : \beta_3 = 0$.

b) **(1 p.)** Calcula el valor de `A`, `B` y `D` en los resultados anteriores.

c) **(1 p.)** Calcula el valor de `C` en la última línea de los resultados anteriores. Indica a qué contraste de hipótesis corresponde ese estadístico. Interpreta el valor 0.5731 que aparece en la última línea de los resultados anteriores.