Revisión Práctica Segundo Parcial

Ejercicio 1. Un fabricante sospecha que los lotes de materia prima recibidos de un proveedor difieren significativamente de su contenido en calcio. Elige al azar 5 lotes diferentes y un químico hace cinco determinaciones del contenido en calcio de cada lote. Los resultados obtenidos han sido:

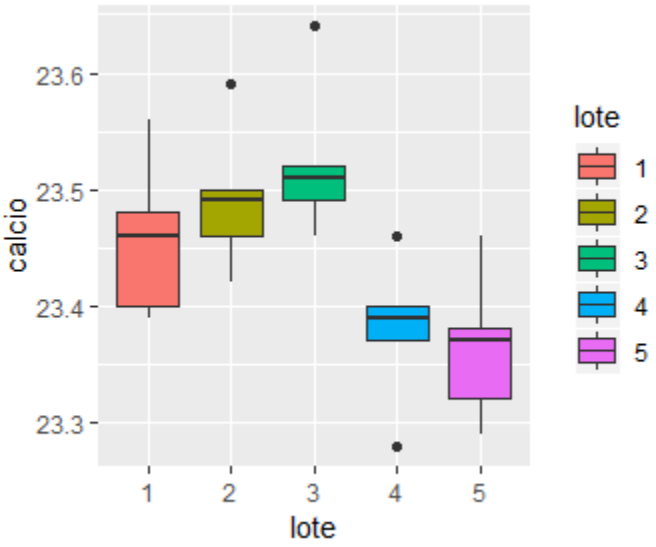
Table 1: Contenido de Calcio

lote1	lote2	lote3	lote4	lote5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.4	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.4	23.5	23.49	23.29	23.38

1. Establecer las hipótesis de interés.
2. Señalar los supuestos del modelo a testear.
3. Analizar el cumplimiento de los supuestos del modelo, detallando en cada caso las hipótesis testeadas.
4. Realizar un test con un nivel de significación del 0.05.
5. Si el fabricante tiene dos líneas de productos, cuál de los lotes seleccionaría para la línea con alto contenido de calcio y cuál para la línea de bajo contenido de calcio. Fundamente su respuesta.

1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_5$ versus $H_1 : \exists(i; j)/\mu_i \neq \mu_j$
2. a) Independencia de las observaciones, b) normalidad de los residuos y c) homocedasticidad de los residuos.
3. Visualizamos las distribuciones del calcio en los lotes:

Figure 1: Calcio por Lote



Testeamos la normalidad de los residuos
shapiro.test(residuals(calcio.aov))

Shapiro-Wilk normality test

data: residuals(calcio.aov)
W = 0.97728, p-value = 0.8264

No se rechaza la hipótesis de normalidad.

Testeamos la homocedasticidad de los residuos

Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group 4 0.0322 0.9978
20

No se rechaza la hipótesis de homocedasticidad.

Dados estos dos resultados no es necesario utilizar una transformación de Box& Cox de la variable objetivo ni aplicar una prueba no paramétrica.

4. Realizar un test con un nivel de significación del 0.05.

   calcio.aov=aov(Calcio$calcio Calcio$Lote,data=Calcio)
   summary(calcio.aov)
   Df Sum Sq Mean Sq F value Pr(>F)
   Lote 1 0.0450 0.04500 7.415 0.0121 *
   Residuals 23 0.1396 0.00607
   — Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   Se rechaza la hipótesis de igualdad de medias de anova.

5. Si el fabricante tiene dos líneas de productos, cuál de los lotes seleccionaría para la línea con alto contenido de calcio y cuál para la línea de bajo contenido de calcio. Fundamente su respuesta.

   library(stats)
   tukey.test <- TukeyHSD(calcio.aov)
   tukey.test

Table 2: Comparaciones Post Hoc

	diff	lwr	upr	padj
1-Feb	0.034	-0.091	0.159	0.924
1-Mar	0.066	-0.059	0.191	0.528
1-Apr	-0.078	-0.203	0.047	0.368
1-May	-0.094	-0.219	0.031	0.204
2-Mar	0.032	-0.093	0.157	0.938
2-Apr	-0.112	-0.237	0.013	0.094
2-May	-0.128	-0.253	-0.003	0.044
3-Apr	-0.144	-0.269	-0.019	0.019
3-May	-0.160	-0.285	-0.035	0.008
4-May	-0.016	-0.141	0.109	0.995

Ejercicio 2. En este ejercicio utilizaremos la base de datos Wisconsin cáncer de mama. La base consiste en 683 casos de potenciales tumores cancerígenos en Wisconsin, de los cuales 283 resultaron malignos. La determinación de la malignidad del tumor es generalmente posterior a un procedimiento quirúrgico. El objetivo de este estudio es determinar si la extracción con una pequeña

aguja de tejido corresponde a células malignas o no en función de alguna/s variable/s predictora/s adecuada/s.
Se han registrado para cada caso las siguientes 10 variables:

- **Class**: 0 si es maligno, 1 si es benigno.
- **Adhes**: adhesión marginal.
- **BNucl**: núcleos desnudos.
- **Chrom**: cromatina suave.
- **Epith**: tamaño de las células epiteliales.
- **Mitos**: mitosis.
- **NNucl**: núcleos normales.
- **Thick**: grosor del borde.
- **UShap**: uniformidad celular.
- **USize**: tamaño celular.

(a) Hallar el vector medio total y el vector medio por grupos.

(b) Comparar las variables de a una entre los grupos mediante análisis univariado. ¿Cuál/es le parece que pueden aportar a discriminar entre los grupos?

(c) Utilice toda la información disponible para construir una función discriminante lineal o cuadrática, según corresponda, entre los grupos.

(d) Es adecuado el análisis discriminante lineal en este caso?

(e) Estime en forma ingenua la capacidad discriminante de esta función.

(f) Seleccione aleatoriamente el 70% de la base. Construya una regla discriminante lineal y utilice el30% restante de la base para cuantificar el poder discriminante de la regla construida.

(a) **vector medio del grupo total**

Adhes	BNucl	Chrom	Epith	Mitos	NNucl	Thick	UShap	USize
2.816	3.542	3.433	3.231	1.604	2.859	4.436	3.204	3.140

vector medio clase 0 (maligno)

Adhes	BNucl	Chrom	Epith	Mitos	NNucl	Thick	UShap	USize
5.567	7.651	5.966	5.328	2.609	5.861	7.197	6.546	6.563

vector medio clase 1 (benigno)

Adhes	BNucl	Chrom	Epith	Mitos	NNucl	Thick	UShap	USize
1.339	1.334	2.072	2.104	1.063	1.246	2.953	1.409	1.300

- **Adhes**: adhesión marginal → p valor « 0.0001
- **BNucl**: núcleos desnudos → p valor « 0.0001
- **Chrom**: cromatina suave → p valor « 0.0001
- **Epith**: tamaño de las células epiteliales →p valor « 0.0001
- **Mitos**: mitosis → p valor « 0.0001
- **NNucl**: núcleos normales → p valor « 0.0001
- **Thick**: grosor del borde → p valor « 0.0001
- **UShap**: uniformidad celular → p valor « 0.0001
- **USize**: tamaño celular → p valor « 0.0001

Todas parecen discriminar!!!

(b) can.lda=lda(Class Adhes+BNucl+Chrom+Epith+NNucl+Mitos+UShap+USize, can.dat)

(c) Prueba de normalidad y homocedasticidad

(d) Estimacion ingenua.

(e) cross validation.

Ejercicio 3. Se intenta estudiar la contaminación atmosférica en ciudades de USA. Los datos incluyen una variable de contaminación atmosférica, cuatro variables climáticas y dos indicadores de ecología humana en 41 ciudades de Estados Unidos.(**city.xls**)

- **SO2** contenido de SO2 en aire, en mg/m3.
- **TEMP** Temperatura media anual, en ^{0} F
- **MANUF** Número de empresas manufactureras con 20 empleados o más
- **POP** Tamaño de la población, en miles.

- **WIND** Velocidad media del viento, en millas por hora
- **PRECI** Precipitación media anual en pulgadas
- **DAYS** Número medio de días con precipitación al año

(a) Realizar un análisis de cluster jerárquico explicando la clasificación obtenida y el método elegido.

(b) Decidir el número de clusters y justificar.

(c) Realizar un análisis de cluster no jerárquico utilizando esta información.

(d) Elegir una de las dos clusterizaciones y caracterizar los grupos formados.