Técnicas de Análisis Cuantitativas y Cualitativas
Resolución del ejercicio de evaluación 1

Marcos Rial Docampo

16 de enero de 2016

En el ejercicio propuesto se presentan dos tipos de resultados de forma gráfica aparecidos en prensa. Los primeros resultados son los referentes a la información extraída del resultado de las elecciones municipales de Madrid del 24 de mayo de 2015 en la que se muestra en forma de gráfico el porcentaje de votos recibidos por las diferentes formaciones políticas. Los segundos resultados son extraídos de una encuesta de estimación de voto en el municipio de Madrid hecha entre los días 4 y 10 de septiembre de 2015.

El primer error, en lo referente a la presentación de los datos, es la de no incluir en los datos de la encuesta los intervalos de confianza. En el primer caso no sería correcto incluir el intervalo de confianza puesto que se trata de datos obtenidos de unos resultados electorales donde se convocaba a participar al total de la población con derecho a voto del municipio de Madrid (censo electoral de 2.597.411 personas[1]). No se trata de una encuesta, como en el segundo caso, donde el tamaño muestral es de 502 personas consultadas.

Para empezar, podemos tomar el ejercicio como un conjunto de cuatro apartados individuales en los que cada uno trata un modelo discreto distinto bajo una distribución binomial. Por ejemplo, tomar para el caso del PP que un 33,7 % de la población lo votaría frente al 66,7 % que no lo haría (tomamos, por tanto, un éxito de 0,337).

Para obtener los datos de límites superior e inferior del intervalo de confianza mostrados en el cuadro 1 utilizamos el comando *prop.test()* de R como se especifica en la figura 1. En él simplemente necesitamos especificar el número de elementos de éxito en la muestra frente al tamaño de la misma así como el nivel de confianza (para este ejercicio será del 95 %). Estos datos podrían verse reflejados en una gráfica similar a la de la figura 1 del enunciado del ejercicio pero, en lugar de los porcentajes y número de concejales, se pondrían los intervalos de confianza calculados.

Se indica en la noticia que el error de muestreo es de $\pm 4,5\,\%$. Esto quiere decir que la diferencia entre los resultados de la encuesta y los reales (preguntando a toda la población) sería de un 4,5 %. Una forma de reducir este error podría ser la de ampliar el tamaño de la muestra pero es un valor que no debería generalizarse a toda la encuesta puesto que tenemos resultados porcentuales de lo que se espera obtener en la encuesta

[1] Fuente: Datos electorales del Instituto Nacional de Estadística para las elecciones del 24 de mayo de 2015.

Partido	Elecciones (mayo)	Encuesta (septiembre)		
	Valor	Valor	LI*	LS*
PP	34,6 %	33,7 %	29,6	38,0
Cs	11,4 %	13,8 %	10,9	17,1
AM	31,9 %	29,1 %	25,2	33,3
PSOE	15,3 %	17,4 %	14,2	21,0

* LI y LS límites inferior y superior del intervalo de confianza al 95 %

Cuadro 1: Resumen de los resultados de las elecciones de mayo y la encuesta de septiembre.

para cada uno de los partidos políticos (reflejados en los resultados electorales de mayo). Es decir, que cada dato para cada partido político obtendría su propio dato de error de muestreo.

Un error muestral del $\pm 4,5\,\%$ se antoja bastante alto para afirmar con rotundidad, como dice el titular del artículo, que la candidata del Partido Popular (PP) recuperaría la alcaldía de haber un pacto con Ciudadanos (C's). Si a esto añadimos el hecho de que los intervalos de confianza son bastante amplios refuerza la incorrecta inferencia hecha por el periodista. Puede deberse a que en encuestas electorales publicadas en prensa este valor se toma como si fuera el intervalo de confianza haciendo pensar que se trata de un margen de error[2]. El error de muestreo correcto para esta encuesta, siguiendo la ecuación 1 sería de un $\pm 2,2\,\%$.

$$E = \sqrt{\frac{p \cdot (1-p)}{n}} \tag{1}$$

siendo p el valor tomado de 0,5 y n el tamaño de la muestra de 502 elementos.

Para asegurar lo dicho en el párrafo anterior podemos tomar los votos conjuntos a PP y C's para evaluarlos separadamente. Obtenemos así para un 47,5 % de los votos un intervalo de confianza al 95 % de 43,0 - 51,9.

Por extensión al ejercicio podemos realizar un contraste de hipótesis también con el comando *prop.test()* utilizando como hipótesis nula que el valor obtenido en la encuesta es igual al de los resultados electorales del 24 de mayo. Por lo tanto, aplicaríamos el comando como se muestra en la figura 2 obtenemos los resultados siguientes:

Vemos un valor bajo de *p-value* de C's, pero que no indica que sea correcto rechazar la hipótesis nula. En general ninguno lo indica, no pudiendo rechazar en ningún caso dicha hipótesis nula, así que podemos aceptar que los valores de la encuesta se adaptan a los resultados obtenidos en las elecciones.

Concluimos que una propuesta para un posible titular apto para la noticia teniendo en cuenta lo anteriormente expuesto sería "Según la encuesta de GAD3 para ABC existiría

[2]Entrada en Explorable.com donde se trata el tema del error de muestreo en encuestas de opinión de periódicos https://explorable.com/es/error-de-muestreo-aleatorio

Partido	*p-value*
PP	0,6941
Cs	0,1134
AM	0,1916
PSOE	0,2294

Cuadro 2: Datos obtenidos del contraste de hipótesis.

```r
#Aplicacion de prop.test para definir los intervalos de confianza con n
    = 502 y nivel de confianza del 95%
#PP con 33.7% de los votos
prop.test(round(0.337*502, digits = 0),502,
          conf.level = 0.95)

#Cs con 13.8% de los votos
prop.test(round(0.138*502, digits = 0),502,
          conf.level = 0.95)

#AM con 29.1% de los votos
prop.test(round(0.291*502, digits = 0),502,
          conf.level = 0.95)

#PSOE con 17.4% de los votos
prop.test(round(0.174*502, digits = 0),502,
          conf.level = 0.95)
```

Figura 1: Script en R con la aplicación de prom.test a los datos de la encuesta. Se aplica redondeo para ofrecer un número entero de votantes por cada partido.

la posibilidad de que el PP recuperara Madrid tras 90 días".

```r
#Contraste de hipotesis
#H0 = porcentajes de votos de la encuesta iguales al resultado electoral
#PP con 33.7% de los votos frente al 34.6% obtenido en las elecciones
prop.test(round(0.337*502, digits = 0),502,
          p = 0.346, alternative = "two.sided",
          conf.level = 0.95)

#Cs con 13.8% de los votos frente al 11.4% obtenido en las elecciones
prop.test(round(0.138*502, digits = 0),502,
          p = 0.114, alternative = "two.sided",
          conf.level = 0.95)

#AM con 29.1% de los votos frente al 31.9% obtenido en las elecciones
prop.test(round(0.291*502, digits = 0),502,
          p = 0.319, alternative = "two.sided",
          conf.level = 0.95)

#PSOE con 17.4% de los votos frente al 15.3% obtenido en las elecciones
prop.test(round(0.174*502, digits = 0),502,
          p = 0.153, alternative = "two.sided",
          conf.level = 0.95)
```

Figura 2: Script en R con la aplicación de prom.test para realizar un contraste de hipótesis. Se aplica redondeo para ofrecer un número entero de votantes por cada partido.

Técnicas de Análisis Cuantitativas y Cualitativas
Resolución del ejercicio de evaluación 2

Marcos Rial Docampo

7 de noviembre de 2015

En este ejercicio de evaluación se nos presentan datos de superficie agrícola abandonada entre 1980 y 2010 (porcentaje de la superficie agrícola que ha cambiado su uso en el periodo estudiado), densidad de población (hab/km^2 con datos de 1991) y altitud media (metros sobre el nivel del mar) de una serie de 50 observaciones tomadas en otros tantos municipios gallegos. Se destaca la importancia que tiene la densidad de población o la elevación sobre los cambios de uso de suelo que afectan a superficie agrícola. Asignamos la variable dependiente al abandono de superficie agrícola y las variables dependientes a la elevación y a la densidad de población.

En los gráficos de la figura 1 se presenta la relación existente entre el abandono de la superficie agrícola y las otras dos variables: elevación y densidad de población. A primera vista podemos observar como aparentemente la relación entre el abandono y la densidad de población no ofrece indicios de correlación entre ambas. Esto no ocurre con la segunda relación entre el abandono y la elevación, donde sí parece haber correlación. Esta es una valoración fundada en la simple observación del aspecto de las relaciones de variables en una gráfica. Podría ser que alguna de las variables necesitara ser transformada para que en la gráfica se ofreciera una visión más fidedigna, pero no se ve necesario.

Comprobamos la existencia de correlación entre las variables abandono y elevación y densidad de población. Partimos de la hipótesis nula (H_0) de que no existe correlación entre las dos variables estudiadas en cada caso. Para ello empleamos el comando de R *cor.test()* como se muestra en la figura 2. Vemos en el cuadro 1 que los resultados nos arrojan una alta correlación entre la variable abandono y la de elevación. Mientras que para la otra relación, entre el abandono y la densidad de población, la correlación es mucho más baja. Se confirma lo expuesto en el párrafo anterior.

Variable	p-valor	Coef. Correlación
Dens. Población	0,9951	-0,3620322
Elevación	2,2e^{-16}	0,8717672

Cuadro 1: Resultados del test de correlación en relación con la variable abandono de superficie agraria.

Para analizar la relación entre la variable dependiente y las independientes de forma conjunta podemos calcular un modelo de regresión lineal múltiple del que obtenemos el

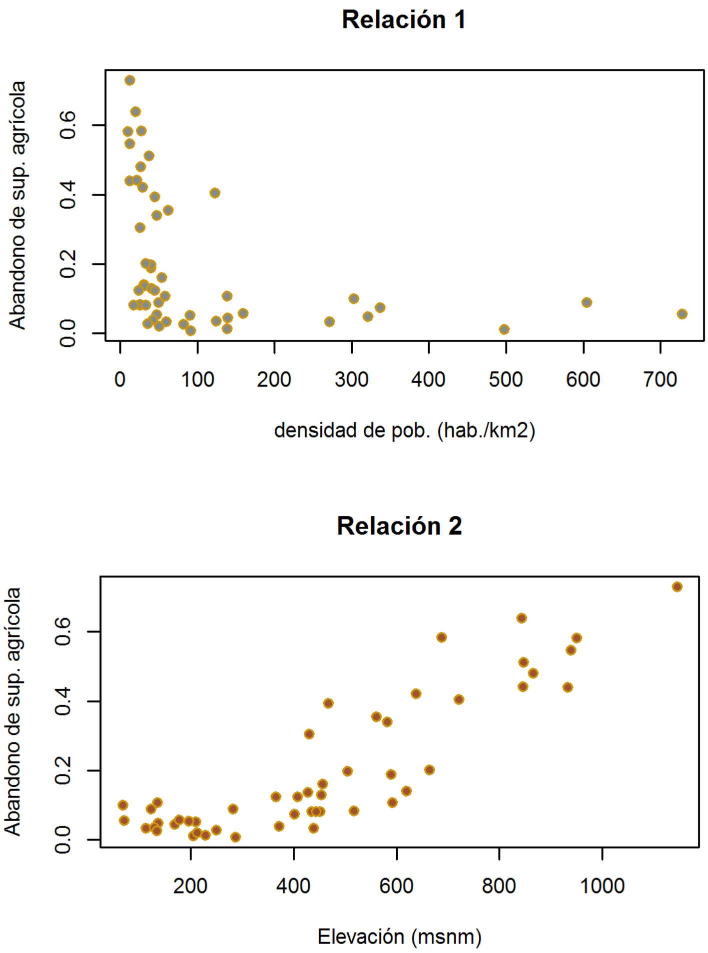

Figura 1: Relación entre el abandono de superficie agrícola y las variables densidad de población y elevación.

```
# Estudio de correlacion entre variables
# H0 = las variables no estan correlacionadas
# Abandono vs densidad de poblacion
cor.test(datos$abandon.uaa, datos$pop.dens, alternative = "greater",
         method = "pearson", conf.level = 0.95)
# Abandono vs elevacion
cor.test(datos$abandon.uaa, datos$elevation, alternative = "greater",
         method = "pearson", conf.level = 0.95)
```

Figura 2: Empleo del comando cor.test con el método de Pearson y nivel de confianza al 95 %.

plano de regresión siguiente:

$$y = -0{,}1375 + 0{,}0002x_1 + 0{,}0007x_2 \tag{1}$$

siendo y la variable dependiente y x_1 y x_2 las variables independientes densidad de población y elevación respectivamente.

De dicho modelo de regresión múltiple extraemos la información que nos permitirá saber qué variable es significativa y cuál no, en el caso de haberlas, respecto de la variable abandono de superficie agraria. Partiendo de la H_0 de que no hay correlación entre las variables, observamos los coeficientes de regresión ofrecidos de el cuadro 2 donde vemos que existe una fuerte relación del abandono con la variable elevación, mientras que con la variable densidad de población es menor. Por lo tanto, con un p-valor de $1{,}94e^{-15}$ (muy cercano a cero) rechazamos la hipótesis nula y aceptamos que la variable elevación es significativa en relación con la variable abandono.

	Estimate	Std. Error	t value	$\Pr(> \|t\|)$
(Intercept)	-1,375e^{-01}	3,732e^{-02}	-3,684	0,000592
pop.dens	1,967e^{-04}	1,071e^{-04}	1,837	0,072527
elevation	6,987e^{-04}	6,005e^{-05}	11,635	1,94e^{-15}

Cuadro 2: Resultados del análisis de significación de la regresión múltiple.

Analizamos los supuestos de partida de la figura 3. En la primera gráfica vemos, representados con línea continua roja, los valores medios de los residuos comparados con la regresión o ajuste. Tenemos que el valor medio de los residuos es próximo a cero a lo largo del ajuste, lo que nos indicaría la independencia del residuo con los valores ajustados.

En la segunda gráfica vemos el diagrama de cuantiles para los residuos del ajuste que nos permite comprobar que los residuos siguen una distribución normal al estar situados cerca de la diagonal.

El tercer gráfico muestra la variabilidad de los residuos en función de los valores ajustados lo que nos indica una baja homocedasticidad al ser la varianza de los residuos independientes del ajuste.

Y en el cuarto y último gráfico de supuestos de partida observamos que no hay puntos cercanos a la distancia de Cook, lo que nos indica la ausencia de apalancamiento del modelo, lo que es un buen indicativo.

Una vez aplicado y analizado el modelo de regresión múltiple y extraída la información obtendremos la recta de regresión entre la variable abandono y elevación, puesto que es la única relación que presenta correlación. Obtenemos la gráfica de la figura 4 y observamos gracias a invocar el valor "modelo2", creado para aplicar la función de regresión lineal *lm()* como en el caso anterior pero esta vez simple (a una sola variable independiente), que devuelve los coeficientes de significación que nos permiten obtener la recta de regresión siguiente y el cuadro de datos 3:

$$y = -0{,}0896 + 0{,}0006x \tag{2}$$

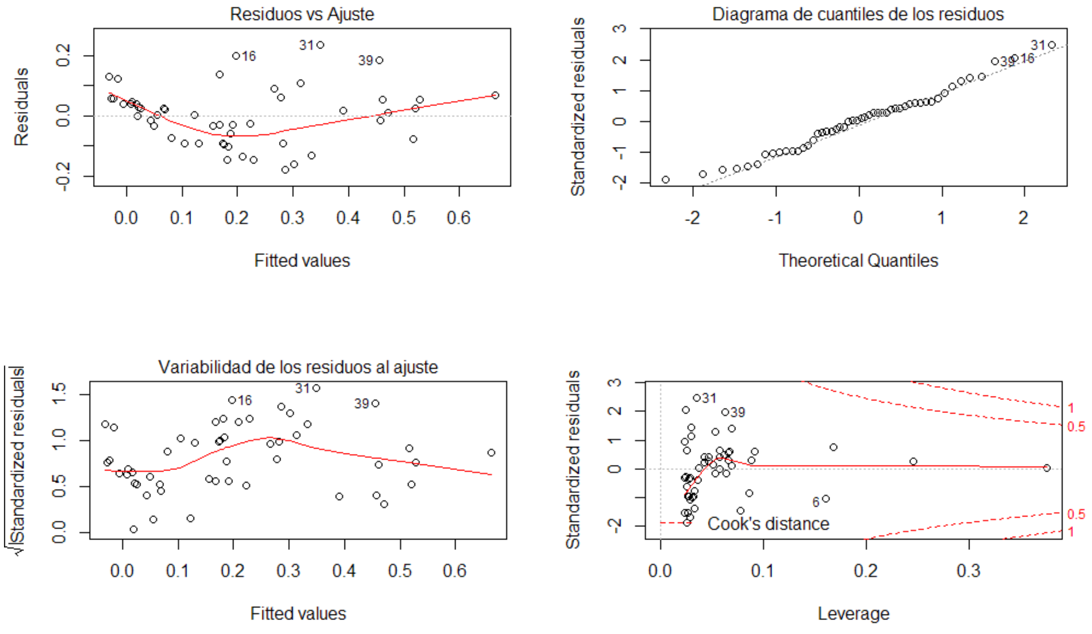

Figura 3: Supuestos de partida de la regresión múltiple.

siendo y la variable dependiente y x la variable independiente. Lo que nos va a permitir estimar el porcentaje de superficie agraria abandonada con la aplicación de un dato de elevación de la zona de estudio.

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	-1,375^{e-01}	3,732^{e-02}	-3,684	0,000592
elevation	6,393e^{-04}	5,186e^{-05}	12,328	<2e^{-16}

Cuadro 3: Resultados del análisis de significación de la regresión simple.

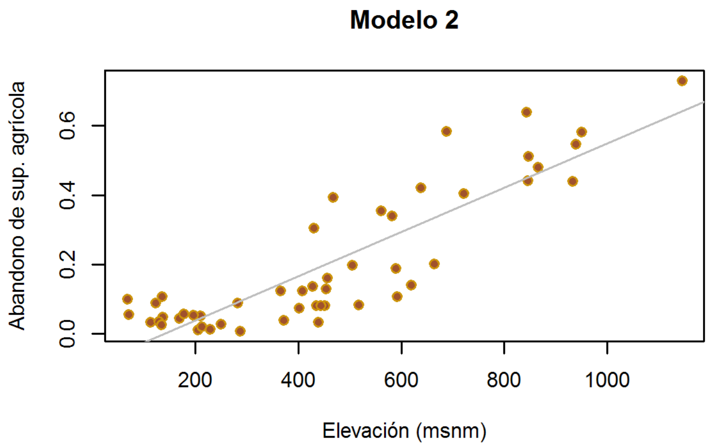

Figura 4: Modelo de regresión lineal para la relación entre abandono y elevación.

Técnicas de Análisis Cuantitativas y Cualitativas
Resolución del ejercicio de evaluación 3

Marcos Rial Docampo

16 de enero de 2016

En este ejercicio de evaluación se nos presentan datos de superficie agrícola abandonada (porcentaje de la superficie agrícola al inicio del periodo), densidad de población (hab/km^2) y altitud media (metros sobre el nivel del mar) de una serie de 50 observaciones tomadas en otros tantos municipios gallegos.

Lo primero que haremos será suprimir la variable categórica X puesto que no ofrece información alguna al tratarse de un código referente a la zona donde se tomaron las observaciones. Una vez hecho esto extraemos gráficamente la relación directa entre las tres variables restantes como se muestra en la figura 1. La estimación numérica de la correlación entre estas variables sería la mostrada en el cuadro 1 mediante el método de Pearson.

	Abandono	Densidad pob.	Elevación
Abandono	1		
Densidad pob.	-0,3620	1	
Elevación	0,8718	-0,5379	1

Cuadro 1: Matriz de correlaciones de las variables.

El siguiente paso previo a la generación de agrupamientos sería el de la estandarización de las variables. Tenemos tres variables cuyos datos no se encuentran entre los mismos intervalos donde, por ejemplo, la variable abandono *abandon.uaa* muestra valores entre 0 y 0,73 mientras que la densidad de población *pop.dens* por ejemplo está entre 9 y 728. Lo que se busca con este estandarizado es el de aproximar los intervalos de los valores de las tres variables y realizar un agrupamiento lo más satisfactorio posible. El procedimiento seguido es el de restar a los valores de cada observación la media y dividir por la desviación típica (1). Una vez realizada esta estandarización se obtienen los datos del cuadro 2 donde a cada valor se le resta la media y divide por la desviación típica.

$$x_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

donde x_i es el valor estandar de la observación i cuyo valor original es X_i, μ es la media de la observación y σ la desviación típica.

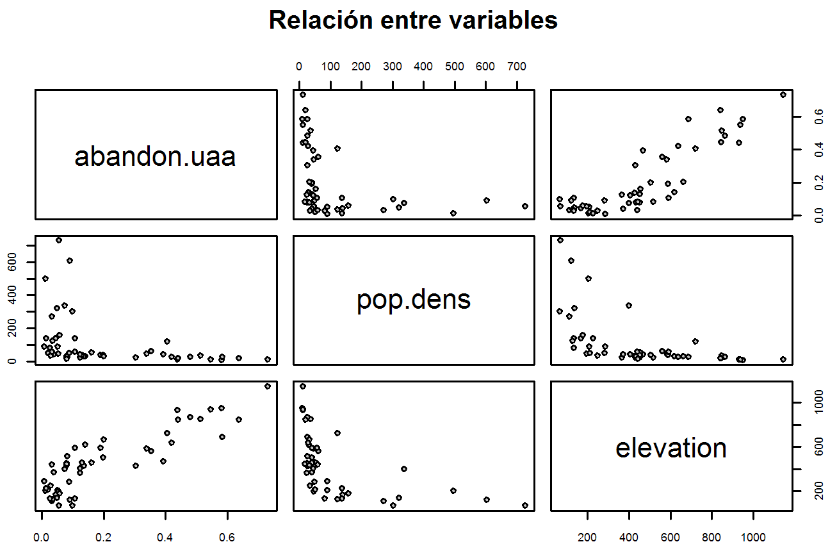

Figura 1: Relación entre variables.

	abandon.uaa		pop.dens		elevation	
	ant.	desp.	ant.	desp.	ant.	desp.
Mínimo	0,01	-0,97	9,38	-0,63	67,00	-1,42
1^{er} cuartil	0,05	-0,74	27,47	-0,52	209,80	-0,90
Mediana	0,11	-0,47	44,50	-0,40	437,00	-0,06
Media	0,20	0,00	106,22	0,00	454,10	0,00
3^{er} cuartil	0,35	0,75	114,55	0,05	612,00	0,58
Máximo	0,73	2,66	727,83	4,07	1145,00	2,54

Cuadro 2: Valores de las variables antes y después del estandarizado.

Pasamos a realizar el agrupamiento jerárquico. Para empezar calculamos la matriz de distancias entre cada observación en tantas dimensiones como variables tenemos, en este caso, tres. Para este ejercicio aplicaremos el cálculo de una distancia euclídea (cálculo habitual para la distancia entre dos puntos).

Para realizar el agrupamiento se aplica el algoritmo de Ward y obtenemos el dendrograma de la figura 2. En él podemos apreciar los agrupamientos en base a la distancia y el peso de la similitud entre grupos (cuanto mayor es este, menor es la similitud entre grupos). Decidimos realizar los grupos cortando el dendrograma por el peso 5 (línea roja a trazos de la figura 2) resultando así un total de 5 grupos.

El grupo 1 se caracteriza por una baja densidad de población en una zona de elevación media respecto del resto de grupos, al igual que el grupo 5 pero presentando menos porcentaje de abandono que este, en torno a un 50 % menos. El grupo 2 presenta niveles

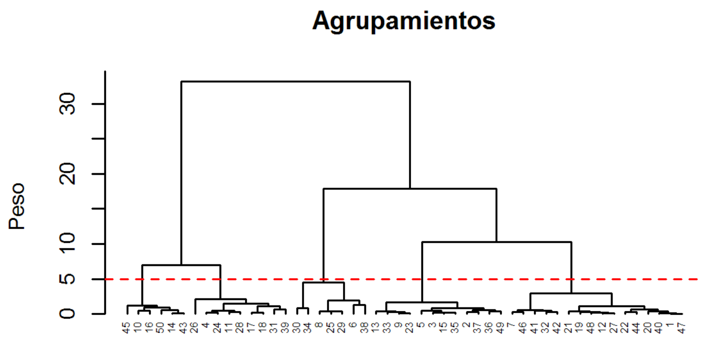

Figura 2: Dendrograma resultante del agrupamiento de observaciones.

bajos de las tres variables. En el grupo 3 claramente se agrupan observaciones hechas en municipios rurales de interior con mucha elevación, mucho abandono y una densidad de población baja. Al contrario que con el grupo 4 en el que su densidad alta y su elevación baja nos indica que se trata de observaciones de municipios costeros. Estos últimos son los grupos más claramente identificables de entre los cinco propuestos, dejando en duda a los otros tres.

Nos ayudamos de los gráficos de cajas de la figura 3 para analizar los datos de las observaciones por grupos. Extraemos la siguiente información:

- En la variable abandono los grupos 2 y 4 presentan mucha similaridad y podríamos decir que el grupo 1 se les parece levemente con una media próxima a la de estos otros grupos. Destaca el grupo 3 por ser el más diferente.
- En cuanto a la variable densidad de población los grupos 1, 3 y 5 se parecen notablemente. Se destaca la gran amplitud de los datos de las observaciones del grupo 4.
- En la variable elevación hay diferencias en todos los grupos. Los datos están notablemente bien agrupados.

Para realizar un agrupamiento no jerárquico aplicamos el algoritmo *k-means*. Utilizaremos, al aplicar dicho algoritmo, los datos escalados empleados en el agrupamiento jerárquico así como un total de 5 grupos. Adicionalmente, al comparar la coincidencia entre ambos métodos de clasificación resulta el cuadro 3 En él vemos que las dos agrupaciones son bastante similares (obviando el hecho de que los nombres de los grupos no coinciden) sobre todo en el caso del grupo 4 jerárquico y 3 no jerárquico donde solo se distinguen por una observación agrupada en otro grupo.

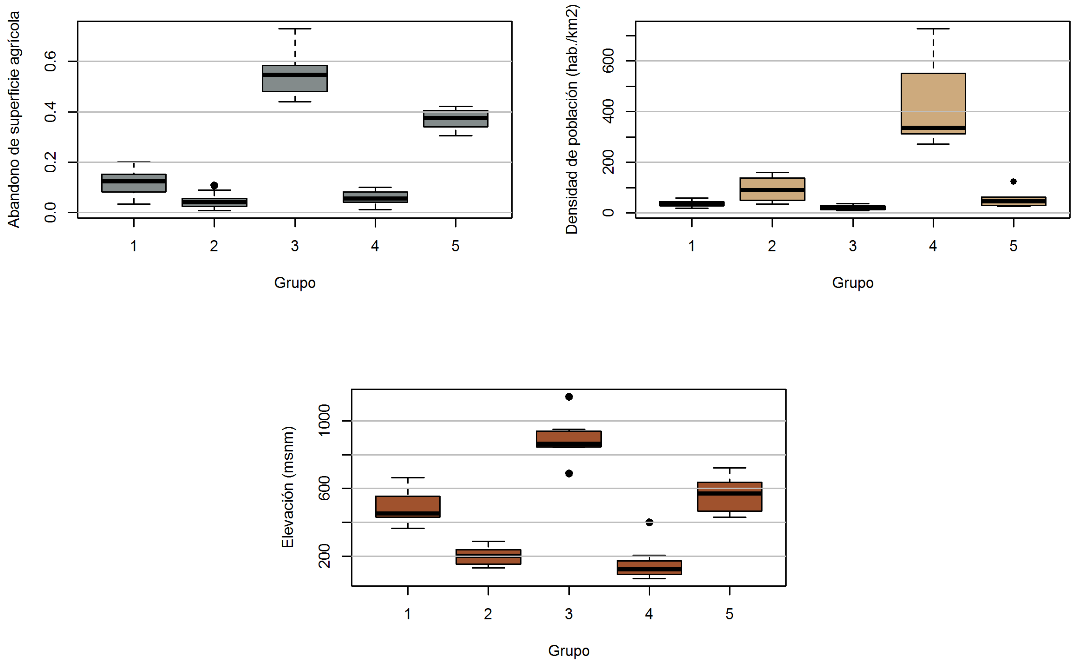

Figura 3: Diagramas de cajas de las observaciones agrupadas.

Grupos	1	2	3	4	5
1	0	1 0	0	15	
2	0	12	0	0	0
3	8	0	0	1	0
4	0	1	6	0	0
5	2	0	0	0	4

Cuadro 3: Tabla de comparación entre los dos agrupamientos

El algoritmo *k-means* crea un número k de centroides iniciales, tantos como grupos, originalmente repartidos de forma aleatoria entre las observaciones. Los grupos se generan asignando las observaciones al centroide cuyo valor se aproxime más al valor del dato. El valor del centroide se actualiza al de la media del grupo al que representa y vuelve a realizarse el paso anterior. El proceso es iterativo hasta conseguir convergencia o un número de iteraciones predeterminado.

Técnicas de Análisis Cuantitativas y Cualitativas
Resolución del ejercicio de evaluación 4

Marcos Rial Docampo

16 de enero de 2016

Disponemos de datos de los tiempos obtenidos por los participantes de una carrera popular. Se trata de 80 muestras de un total de 1140 que están estratificadas por sexo y categoría de edad.

Mediante un test de Shapiro-Wilk comprobamos si la variable *total.minutos* tiene una distribución normal. Para ello podemos aplicar el test a todas las observaciones de la variable o dividirla en grupos. En este caso dividiremos la variable en los cuatro grupos correspondientes a la categoría de la prueba (infantil-cadete, junior, senior y veteranos), pero podríamos haberlo hecho por sexo o por subcategoría. Obtenemos los resultados mostrados en el cuadro 1 y los gráficos de la figura 1 donde se indica que la variable *total.minutos* tiene una distribución normal en cada uno de los grupos de edad. En el caso del cuadro vemos que el p-valor supera el 5 % en todos los casos, lo que nos permite admitir la hipótesis nula H_0 de que existe una distribución normal de los datos. Lo mismo que se observa en las gráficas donde los valores de las observaciones de adaptan a la recta. En las gráficas no se observa desviaciones de linealidad respecto a la recta, sobre todo en el centro, lo que demuestra la normalidad de las observaciones para cada grupo.

	Test Shapiro-Wilk	
	W	p-valor
Infantil-Cadete	0,93455	0,1888
Junior	0,93301	0,1765
Senior	0,96890	0,7315
Veterano	0,96757	0,7029

Cuadro 1: Resultado del test de Shapiro-Wilk para la variable "total.minutos".

Es interesante saber si la variable sexo tiene influencia sobre los tiempos obtenidos por los participantes de la prueba. Al ser el sexo de los participantes un factor de dos niveles, podemos realizar un contraste de hipótesis mediante un test de Welch. Obtenemos los resultados del cuadro 2 y la figura 2 donde vemos que la media del tiempo empleado por las mujeres es casi 9 minutos superior al de los hombres. Además, un p-valor de

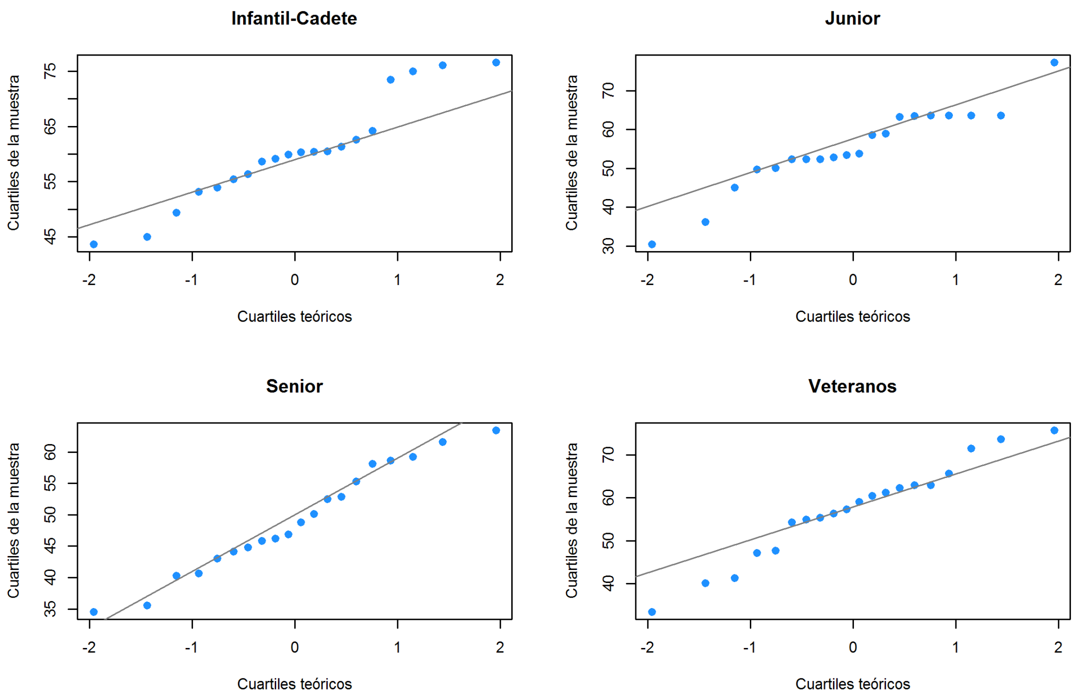

Figura 1: Diagramas de cuartiles para las categorías de la muestra.

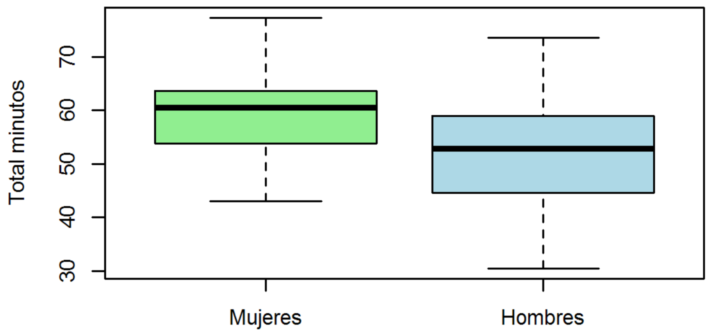

Figura 2: Representación gráfica de la diferencia en tiempos respecto al sexo de los participantes.

7,405e^{-05} sugiere que la interpretación anterior de la diferencia de las medias es correcta. Realizando un análisis de varianza de un factor habríamos obtenido un resultado similar.

Sexo	Tiempo empleado (minutos)
Femenino	59,93
Masculino	50,96

Cuadro 2: Resultados de las medias de tiempo por sexo de los participantes en la prueba.

Analizamos ahora la influencia que tienen los factores sexo y categoría sobre la variable tiempo en minutos. Primero analizamos la influencia de los dos factores por separado con una hipótesis nula (H_0) de que no afectan estos factores a los tiempos obtenidos por los participantes. Obtenemos los resultados del cuadro 3 donde vemos que ambos factores influyen por separado en los tiempos de la prueba deportiva.

Factor	Pr(>F)
Sexo	2,09e^{-05}
Categoría	0,0016

Cuadro 3: Resultado del análisis de varianza para los factores sexo y categoría.

Una vez analizados por separado procedemos a hacerlo conjuntamente teniendo en cuenta la interacción de ambos factores. En este caso el resultado que nos da es que conjuntamente no afectan al tiempo total de los participantes. Con un Pr(>F) obtenido de 0,2478 aceptamos la H_0. Observamos en la gráfica de interacción de la figura 3 que efectivamente los tiempos obtenidos por participantes femeninos están por encima de los

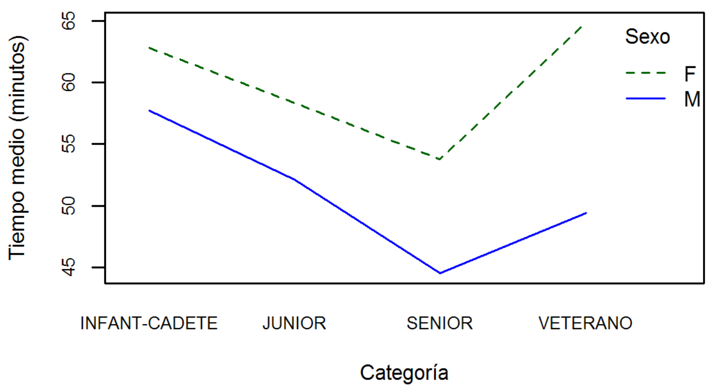

Figura 3: Gráfica de interacción entre los factores sexo y categoría con el tiempo de los participantes.

masculinos en todas las categorías. En cambio en las categorías los tiempos se mezclan en cuanto al sexo como se muestra en la figura 4 en la categoría de veteranos.

Mediante un test de varianzas comprobamos la igualdad de estas para cada factor. Es decir, que gracias a este test y con una H_0 igual a que los factores estudiados presentan varianzas iguales comprobaremos la condición de homocedasticidad de estos con la variable *total.minutos*.

Para el caso del factor sexo esta condición de homocedasticidad se cumple ya que el test devuelve un p-palor de 0,8756. En el caso de las categorías y aplicando un test de Bartlett (debido a que este factor tiene más de dos niveles) el resultado también nos dice que existe homocedasticidad con un p-valor de 0,6638.

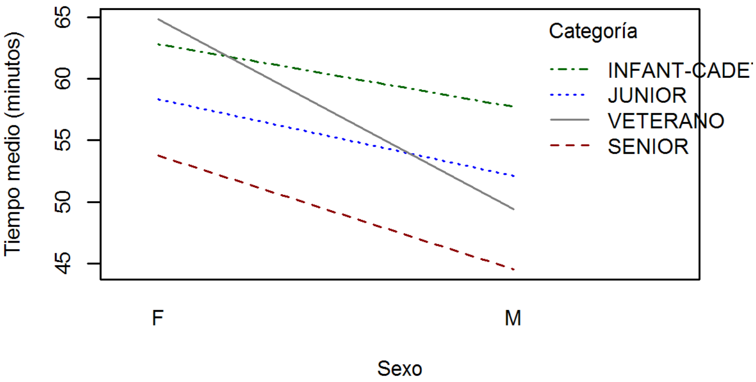

Figura 4: Gráfica de interacción entre los factores categoría y sexo con el tiempo de los participantes.