Estadística Avanzada - Actividad 4: Análisis de varianza y repaso del curso

Enunciado

Índice

INTRODUCCIÓN

En una investigación médica se propuso estudiar la capacidad pulmonar de los fumadores y no fumadores. Se recogieron datos de una muestra de la población fumadora, no fumadora y fumadores pasivos. A cada persona se le realizó un test de capacidad pulmonar consistente en evaluar la cantidad de aire expulsado (AE). La muestra de n individuos se categorizó en 6 tipos:

- No fumadores (NF)
- Fumadores pasivos (FP)
- Fumadores que no inhalan (NI): personas que fuman pero no inhalan el humo.
- Fumadores "light" (FL): personas que fuman e inhalan de uno a 10 cigarrillos al día durante 20 años o más.
- Fumadores moderados (FM): personas que fuman e inhalan entre 11 y 39 cigarrillos por día durante 20 años o más.
- Fumadores intensivos (FI): personas que fuman e inhalan 40 cigarrillos o más durante 20 años o más.

En esta actividad se realizará un análisis de si la capacidad pulmonar está influida por el tipo de fumador y si los fumadores pasivos presentan diferencias en relación al resto de grupos. Para ello, se aplicarán distintos tipos de análisis, revisando los contrastes de hipótesis de una y dos muestras, vistos en la actividad A2, y luego realizando análisis más complejos como ANOVA.

En la segunda parte de la actividad (sección 7), aplicaremos el análisis de varianza a un problema de analítica de datos. Concretamente, aplicaremos un algoritmo de minería de datos (k-NN) para un problema de clasificación. El algoritmo se ejecutará en diferentes configuraciones y a partir del análisis de varianza se investigará si existen diferencias significativas entre las diferentes configuraciones del algoritmo y en caso afirmativo, cuál de las versiones obtiene los mejores resultados.

Debido a la extensión y relativa complejidad de la actividad, esta segunda parte de la actividad es **opcional** y sirve para subir nota. Sólo es imprescindible entregar la primera parte (apartados 1-6). Si se entrega la primera parte, podéis optar a la nota máxima igualmente. La parte opcional mejora la nota global de la actividad.

Notas importantes a tener en cuenta para la entrega de la actividad:

- Es necesario entregar el fichero Rmd y el fichero de salida (PDF o html). El fichero de salida debe incluir el código y el resultado de su ejecución (paso a paso). Se debe incluir un índice o tabla de contenidos. Y se debe respetar la numeración de los apartados del enunciado.
- No realizar listados de los conjuntos de datos, puesto que estos pueden ocupar varias páginas. Si queréis comprobar el efecto de una instrucción sobre un conjunto de datos podéis usar la función head que muestra las primeras 10 filas del conjunto de datos.

1. Preprocesado

1.1. Carga de datos

Cargar el fichero de datos "Fumadores.csv".

1.2. Tipos de datos

Consultar los tipos de datos de las variables y si es necesario, aplicar las conversiones apropiadas. Averiguar posibles inconsistencias en los valores de Tipo o AE. En caso de que existan inconsitencias, corregidlas.

1.3. Realizar un análisis descriptivo de la muestra en su totalidad en relación a la variable AE

1.4. Analizar los datos según el tipo de fumador

Mostrar el número de personas en cada tipo de fumador, la media de AE de cada tipo de fumador y un gráfico que muestre esta media. Se recomienda ordenar el gráfico de menos a más AE.

Luego, se debe representar un boxplot donde se muestre la distribución de AE por cada tipo. Para calcular la media o otras variables para cada tipo de fumador, podéis usar las funciones **summarize** y **group_by** de la libreria **dplyr** que os serán de gran utilidad. Para realizar la visualización de los datos, podéis usar la función **ggplot** de la librería **ggplot2**.

---

2. Intervalo de confianza

Calcular el intervalo de confianza de la capacidad pulmonar de toda la muestra. El nivel de confianza es 95 %. Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ... En cuanto a la elección del método para el cálculo del intervalo de confianza, debéis justificar vuestra elección.

---

3. Comparación de dos muestras: Fumadores vs No Fumadores

¿Podemos afirmar que la capacidad pulmonar de los fumadores es inferior a la de no fumadores? Incluid dentro de la categoría de no fumadores los fumadores pasivos. Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ... Seguid los pasos que se indican a continuación.

3.1. Escribir la hipótesis nula y alternativa

3.2. Preparar los datos para realizar el contraste

3.3. Especificar qué tipo de contraste usaréis

Podéis consultar Gibergans Baguena (2009).

3.4. Realizar los cálculos del valor p

3.5. Interpretar el resultado del contraste

---

4. ANOVA

A continuación se realizará un análisis de varianza, donde se desea comparar la capacidad pulmonar entre los seis tipos de fumadores/no fumadores clasificados previamente. El análisis de varianza consiste en evaluar si la variabilidad de una variable dependiente puede explicarse a partir de una o varias variables independientes, denominadas factores. En el caso que nos ocupa, nos interesa evaluar si la variabilidad de la variable AE puede explicarse por el factor tipo de fumador. Hay dos preguntas básicas a responder:

- ¿Existen diferencias entre la capacidad pulmonar (AE) entre los distintos tipos de fumadores/no fumadores?
- Si existen diferencias, ¿entre qué grupos están estas diferencias?

Para la resolución de esta sección, se seguirán los apuntes de López-Roldán y Fachelli (2015).

4.1. Verificar la asunción de normalidad

Atendiendo al material, gráfico III.8.6 y página 25, evaluar si el conjunto de datos cumple las condiciones de aplicación de ANOVA. Seguid los pasos que se indican a continuación.

4.1.1. Representar gráficamente la normalidad de la muestra de datos AE con la función qqnorm

4.1.2. Escribir la hipótesis nula e hipótesis alternativa

4.1.3. Aplicar un test de normalidad

Aplicar un test de normalidad, siguiendo las recomendaciones del material mencionado López-Roldán y Fachelli (2015). Justificar la elección.

4.1.4. Interpretar los resultados a partir del gráfico qqnorm y de los valores que devuelve el test.

4.2. Homoscedasticidad: Homogeneidad de varianzas

Otra de las condiciones de aplicación de ANOVA es la igualdad de varianzas (homoscedasticidad). Aplicar un test para validar si los grupos presentan igual varianza. Seguid las indicaciones de los apuntes de López-Roldán y Fachelli (2015).

4.2.1. Escribir hipótesis nula y alternativa

4.2.2. Realizar los cálculos

4.2.3. Interpretar los resultados

4.3. ANOVA unifactorial (One Way ANOVA)

Calcular ANOVA de un factor (one-way ANOVA o independent samples ANOVA) para investigar si existen diferencias en el nivel de aire expulsado (AE) entre los distintos tipos de fumadores.

4.3.1. Escribir la hipótesis nula y alternativa

4.3.2. Realizar los cálculos

Podéis usar la función **aov**.

4.3.3. Interpretar los resultados de la prueba ANOVA y relacionarlos con el resultado gráfico del boxplot

4.4. Cálculos ANOVA

4.4.1. Identificar las variables

A partir de los resultados del modelo devuelto por **aov**, identificar las variables SST (Total Sum of Squares), SSW (Within Sum of Squares), SSB (Between Sum of Squares) y los grados de libertad. A partir de estos valores, calcular manualmente el valor F, el valor crítico (a un nivel de confianza del 95 %), y el valor p.Interpretar los resultados.

4.4.2. Calcular manualmente las variables

Calcular manualmente SSB, SSW y SST a partir de los datos de la muestra. Comprobad que el cálculo coincide con los valores que devuelve la función anova. Las funciones **summarize** y **group_by** os pueden ser útiles para simplificar los cálculos.

4.5. Calcular la fuerza de la relación e interpretar el resultado

Podéis consultar López-Roldán y Fachelli (2015), p.37.

5. Comparaciones múltiples

Independientemente del resultado obtenido en el apartado anterior, realizamos un test de comparación múltiple entre los grupos. Este test se aplica cuando el test ANOVA devuelve rechazar la hipótesis nula de igualdad de medias. Por tanto, procederemos como si el test ANOVA hubiera dado como resultado el rechazo de la hipótesis nula.

5.1. Sin corrección

Calcular las comparaciones entre grupos sin ningún tipo de corrección. Podéis usar la función **pairwise.t.test**.

5.2. Corrección de Bonferroni

Aplicar la corrección de Bonferroni en la comparación múltiple. Interpretar el resultado y contrastar el resultado con el obtenido sin corrección.

5.3. Prueba de Scheffé

La prueba de Scheffé es una de las más usadas cuando se cumplen condiciones de igualdad de varianza. Según López, está basada en la distribución F. Permite la comparación binaria entre pares de medias y también la comparación múltiple. En las comparaciones de pares es conservadora. Se puede aplicar con muestras de tamaño desigual y es bastante robusta ante el incumplimiento de la condición de homoscedasticidad.

Aplicad la prueba de Scheffé e interpretar el resultado.

6. ANOVA multifactorial

En una segunda fase de la investigación se evaluó el efecto del sexo como variable independiente. Con este objetivo, se recolectó un segundo conjunto de datos con las variables independientes sexo y nivel de fumador y con la variable dependiente capacidad pulmonar medida según el aire expulsado, al igual que con el primer conjunto de datos. Este conjunto de datos se encuentra en el fichero `Fumadores2.csv`

6.1. Estudiar los efectos principales y posibles interacciones

Examinar las características del conjunto de datos y realizar un estudio visual de los datos. A continuación, aplicad ANOVA. Seguid los pasos que se indican a continuación.

6.1.1. Análisis visual

Se realizará un primer estudio visual para determinar si sólo existen efectos principales o hay efectos de interacción entre sexo y tipo de fumador. Para ello, seguir los pasos que se indican a continuación:

1. Leer el conjunto de datos
2. Agrupar el conjunto de datos por tipo de fumador y sexo y calcular la media de AE en cada grupo. Podéis usar las instrucciones **group_by** y **summarise** de la librería **dplyr** para realizar este proceso. Mostrar el conjunto de datos en forma de tabla, donde se muestre la media de cada grupo según el sexo y tipo de fumador.

3. Mostrar en un plot el valor de AE medio para cada tipo de fumador y sexo. Podéis inspiraros en los gráficos de López-Roldán y Fachelli (2015), p.38. Podéis realizar este tipo de gráfico usando la función **ggplot** de la librería **ggplot2**.
4. Interpretar el resultado sobre si existen sólo efectos principales o existe interacción. Si existe interacción, explicar cómo se observa y qué efectos produce esta interacción.

6.1.2. ANOVA multifactorial

Calcular ANOVA multifactorial para evaluar si la variable dependiente AE se puede explicar a partir de las variables independientes sexo y tipo de fumador. Incluid el efecto de la interacción sólo si se ha observado dicha interacción en el análisis visual del apartado anterior. Interpretad el resultado.

7. Aplicación de ANOVA a la investigación en minería de datos

Aplicaremos ahora el análisis de varianza en un caso de analítica de datos. En analítica de datos, es frecuente aplicar uno o varios algoritmos sobre un conjunto de datos con el objetivo de elegir el mejor algoritmo para el problema. En este ejercicio, aplicaremos diferentes algoritmos en un conjunto de datos de tipo clasificación, con el objetivo de elegir el algoritmo que clasifica mejor el conjunto de datos. El conjunto de datos es el **BreastCancer**, que está integrado en la librería "mlbench". Se debe instalar este paquete con **install.packages** y posteriormente, cargar la librería (función **library**). Una vez se ha hecho, se puede usar la instrucción **data(BreastCancer)** para comenzar a usar el conjunto de datos. El nombre que recibe el data frame es BreastCancer.

Aplicaremos el algoritmo de classificación K-NN para varios valores de k (número de vecinos). La pregunta a responder es:

- ¿Existen diferencias entre los resultados del algoritmo K-NN para distintos valores de K?
- Si hay diferencias, cuál es el valor o valores de K mejores?

Para resolver este ejercicio, es necesario leer el módulo 4 de la asignatura: "Diseño experimental de analítica de datos."

A continuación, se indican los pasos a seguir.

7.1. Carga de datos

En primer lugar, cargamos el conjunto de datos BreastCancer. Para hacer el manejo de datos más fácil, podéis cambiar el nombre "BreastCancer" por "BC".

7.2. Crear el conjunto de datos

Se desea aplicar un algoritmo de clasificación para predecir si el tipo de cáncer es `maligno` o `benigno` a partir de las variables explicativas siguientes:

Cl.thickness, Cell.size, Cell.shape, Epith.c.size.

Crear un conjunto de datos que contenga estas variables y la variable Class.

7.3. Validación cruzada

A continuación, se aplica validación cruzada a los datos (leer el módulo 4, apartado 3 "Evaluación del modelo"). Hay que aplicar validación cruzada con 5 divisiones (folds=5). Usar la librería `caret` y la función `createFolds` de la misma.

7.4. Cálculo de precisión

Para cada fold, se obtiene un conjunto de train y un conjunto de test. Por lo tanto, para 5 folds, habrá 5 conjuntos emparejados de entrenamiento y test. Se debe aplicar el K-NN para valores de k (3, 5, 7). Para cada valor de k del K-NN y para cada par de conjuntos train-test, hay que calcular el valor de precisión (accuracy).

Para entrenar el algoritmo, se debe usar la función `knn3` y para obtener los valores de precisión, la función `predict`.

7.5. Preparación de datos para el análisis de varianza

Una vez obtenida la matriz de valores de precisión (accuracy) para cada configuración del algoritmo y para cada fold, aplicaremos análisis de varianza para averiguar si existen diferencias significativas entre las diferentes configuraciones del algoritmo. Es decir, si hay diferencias cuando ejecutamos el algoritmo para k = (3,5,7).

Si tenemos guardados los datos en la disposición de data frame, con columnas 3, 5, 7 que corresponden a las diferentes configuraciones del algoritmo K-NN, y donde cada fila corresponde al "fold", hay que transformar estos datos a un formato "desplegado" o vertical para poder aplicar el análisis de varianza con anova. Se puede realizar con la función `melt` de la librería `reshape2`.

7.6. Asunciones ANOVA

Una vez preparados los datos, comprobamos si se cumplen las asunciones de ANOVA. Testear normalidad y homoscedasticidad.

7.7. Aplicación ANOVA

A continuación, aplicad ANOVA para identificar si existen diferencias significativas entre las tres configuraciones del algoritmo. Aplicad el test más apropiado, en función del cumplimiento de las condiciones de normalidad y homoscedasticidad. Justificar la elección. Finalmente, interpertad el resultado del análisis de varianza en relación a la pregunta sobre si existen diferencias entre las distintas configuraciones del algoritmo K-NN.

7.8. Comparación múltiple

Si el test ANOVA da como resultado diferencias entre las medias, aplicad un test a posteriori de comparación múltiple para determinar qué configuración del algoritmo da mejores valores de precisión.

8. Puntuaciones de los apartados

- Pregunta 1: 10 %
- Pregunta 2: 10 %

- Pregunta 3: 10 %
- Pregunta 4: 30 %
- Pregunta 5: 10 %
- Pregunta 6: 20 %
- Calidad del informe dinámico: 10 %
- Opcional: Pregunta 7 (sube nota)

Referencias

Gibergans Baguena, Josep. 2009. «Contraste de dos muestras». En *Estadística*, editado por Josep Gibergans Baguena, Angel J. Gil Estallo, y Carles Rovira Escofet. Barcelona: FUOC.

López-Roldán, P., y S. Fachelli. 2015. «Capítulo III.8 Análisis de varianza». En *Metodología de la investigación social cuantitativa*, editado por P. López-Roldán y S. Fachelli. Barcelona: UAB.