Estadística Aplicada. Lab5: Comparaciones múltiples y ANOVAs complejos

Edlin Guerra Castro

07/10/2021

Parte A. Aplicación de comparaciones múltiples luego del Análisis de Varianza (ANOVA)

Continuamos con el caso de la planta procesadora de celulosa. Ya hicimos la primera evaluación estadística a los valores de DBO, y conseguimos evidencias estadísticas para indicar que la DBO cambia a lo largo del río. Ahora se requiere evaluar el patrón de diferencia, ya que solo algunas combinaciones de diferencias pueden o no reflejar impacto ambiental. Es importante destacar que el resultado del ANOVA que realizamos no implica autonámticamente que podemos hablar de un impacto ambiental, se debe detallar más en los resultados para poder señalar esto. Inspeccione los datos (archivo dats5a.csv) y responda:

1. Identifica todas las posibles hipótesis alternativas y qué significaría cada una en el contexto del estudio.
2. En caso no tengas desplegado las gráficas y resultados del ANOVA, repite el ANOVA, tablas y gráficos.
3. Para aplicar comparaciones pareadas entre todos los pares de medias de un factor, se puede usar la función `pairwise.t.test`. Esta función toma en cuenta un valor de error común ("pooled"), y realiza todas las comparaciones posibles (como si no hubieras aplicado un ANOVA). Conviene utilizar un método de ajuste a alfa por el alto número de pruebas aplicadas ("family wise error rate"). Resuelto esto, responda:
   a) ¿Cuántas pruebas de t habría que aplicar?
   b) ¿Qué representan los números de esta matriz triangular?
4. Para aplicar el procedimiento Tukey's HSD (Honest Smallest Difference) es necesario que el objeto sea el resultado de una función `aov` (una función distinta de aplicar un ANOVA e R). Toma los mismos datos de DBO analizados, y aplica la función `aov` en sustitución de la función `lm` usada antes. Al objeto resultante aplícale la función `TukeyHSD`.
   a) ¿Qué representan los datos de la columna con el nombre diff? Usa la función 'aggregate' recien aprendida para ayudarte en los cálculos.
   b) ¿Qué crees que sean los que están bajo lwr y upr?
   c) Aplica la función `plot` al objeto que resultó de aplicar la de Tukey. Estudia e interpreta el gráfico que produce.
5. Para aplicar el procedimiento SNK de la librería `GAD` es necesario establecer cuáles son factores fijos y cuáles aleatorios. Primero tienes que instalar el paquete `GAD`, y luego llamarlo para hacerla disponible en esta sesión de R. Luego tienes que volver explícito que el factor dietas es un factor fijo. Ajustas el modelo, y después aplicas el procedimiento. Identifica estos pasos con las líneas de código a continuación e interpreta la salida.

```
library(GAD)
datos$localidades <- as.fixed(datos$localidades)
```

```
mod.lm<-lm(datos$DBO~datos$localidades)
snk.test(mod.lm, term="datos$localidades")
```

6. Compara los resultados de los 3 métodos usados.
    a) ¿Cuál produce un mayor número de resultados significativos?
    b) ¿Cuál involucra menor número de pruebas?
    c) ¿Cuál prefieres en este caso? ¿Por qué?

Parte B. ANOVA multifactorial

Asumamos que se efectuó un Estudio de impacto ambiental con un diseño BACI (**B**efore-**A**fter x **C**ontrol-**I**mpact), en el que se midió la DBO tres meses antes de operaciones de la planta y luego a los tres meses de iniciadas las operaciones. El modelo acá es que la puesta en operaciones de la planta promueve el incremento de la DBO del agua. No obstante, las mediciones se efectuaron en *dos* localidades: Loc1 (río arriba) y Loc2 (justo en las inmediaciones de la descarga de la planta). El modelo espacial implica que la DBO del agua en Loc2 debe ser mayor a Loc1 debido a las cercanía con la descargas de la planta. Note, sin embargo, que esto debe ocurrir sólo despues de iniciadas las operaciones de la planta, ya que si antes la loc2 ya presentaba mayores niveles de DBO que loc1, no se puede asumir que la planta incrementó la DBO del agua en loc2. Inspeccione los datos (archivo dats5b.csv) y responda:

1. Identifica todas las fuentes de variación y define las hipótesis estadísticas de cada una. Para ello defina la naturaleza del factor (fijo o aleatorio)
2. Elabora dibujos de los posibles resultados combinando cambios solo temporales, solo espaciales, e interacciones. Discuta qué significaría cada una en el contexto del estudio. ¿Cuál de los posibles resultados es un indicador de impacto ambiental?
3. Efectue un ANOVA bifactorial usando el paquete `GAD`. Este es uno de los pocos paquetes en R que efectúa adecuada descomposición de la variación considerando la naturaleza de los factores (use la función `as.fixed` o `as.random`). Primero cargue el paquete y luego defina la naturaleza del factor. Luego ajuste un modelo lineal (llámelo mod1) que considere cada uno de los términos principales, asi como la interacción de ambos, escribiendo alguno de los siguientes códigos.

```
library(GAD)
BA <- as.fixed(datos$BA)
CI <- as.fixed(datos$CI)
DBO <-datos$DBO
mod1<-lm(DBO~BA*CI)
```

4. Calcula los grados de libertad de los distintos términos en el modelo.
5. Obtén los valores de los MS estimados (EMS) de la tabla de ANOVA usando la función `estimates` aplicada a mod1. ¿Cuáles son los términos del modelo que deberán ser usados en el numerador y denominador para probar cada una de las Hipótesis propuestas?
6. Obtén los resultados del análisis usando la función `gad` sobre el objeto mod1. Corrobora que los g.l. del numeral 4 fueron correctamente calculados.
7. Examina la tabla de ANOVA obtenida y responde a las siguientes preguntas:
    a) ¿Es significativa la interacción? ¿Cómo se interpreta este resultado?
    b) ¿Es significativo alguno de los términos principales en el modelo?¿qué significaría esto?
    c) ¿Con esta información es suficiente para sustentar impacto ambiental, o faltan más procedimientos estadísticos? Si piensas que falta algo, ¿qué faltaría?
8. Realice un gráfico exploratorio de la interacción usando `boxplot`

9. Elabora un gráfico de interacción **BACI** con promedios y desviaciones estándar usando `ggplot2` (vas a requerir estimar promedios y desviación estándar para cada grupo para generar este gráfico)

10. Imagina que loc1 y loc2 son en realidad localidades bien extensas, por lo que las muestras de agua (25 en cada muestreo) en realidad provienen de combinar cinco muestras de agua de cinco zonas aleatorias de cada localidad. El objetivo de ello era tener una mejor representatividad de cada localidad.

    a) ¿Cuáles serian los pasos que se verían modificados en el procedimiento que acabas de aplicar?
    b) ¿Cuántas fuentes de variación hay ahora? Lleva a cabo los cambios que has propuesto para el modelo y grábalo con el nombre mod2.

```
BA<-as.fixed(datos$BA)
CI<-as.fixed(datos$CI)
Zona<-as.random(datos$Zona)
DBO <- datos$DBO
mod2<-lm(DBO ~ BA + CI + BA*CI + Zona%in%(BA*CI))
```

c) Obtén los MS estimados y la tabla de ANOVA bajo este supuesto, y responde a las siguientes preguntas
d) ¿Qué cambió en la tabla de ANOVA?
e) ¿Cuál es la interpretación estadística del nuevo resultado? ¿Cómo sería el gráfico ideal?
f) ¿Sería este resultado evidencia de impacto ambiental?