Paula Vargas Pellicer

15/03/2022

ANOVA

ANOVA es uno de los análisis estadísticos más utilizados en el dominio de las ciencias ecológicas y ambientales.

Para entender ANOVA, recordemos algunas definiciones importantes:

Las variables categóricas contienen un número finito de categorías o grupos distintos, (tratamientos, tipo de material, forma de pago)

Las variables continuas son medidas a lo largo de una escala continua, variables numéricas que tienen un número infinito de valores entre dos valores, por ejemplo. tiempo, altura.

Una variable explicativa (también llamada variable independiente, factor, tratamiento o variable predictora) es una variable que se manipula en un experimento para observar su efecto en una variable de respuesta (también llamada variable dependiente o variable de resultado).

En esta clase, nos centraremos en un diseño de un solo factor (es decir, el más simple) y aprenderemos cómo ejecutar e interpretar un ANOVA de una vía. La lógica básica de ANOVA es simple: compara la variación entre grupos con la variación dentro de los grupos para determinar si las diferencias observadas se deben al azar o no. Un ANOVA unidireccional solo considera UN factor.

1. Establecer una pregunta de investigación

Siempre establece tu pregunta de investigación antes de comenzar a pensar cuál es la prueba estadística más apropiada para usar en tus datos.

Debes ser clar@ y concis@ y debes contener tanto tu respuesta como tus variables explicativas.

En esta clase, nuestra pregunta de investigación es: ¿Cómo varía el tiempo de eclosión de las crías de rana con la temperatura?

Imagina que hicimos un experimento de manipulación.

Un estudio manipulativo es aquel en el que el experimentador cambia algo sobre el sistema de estudio experimental y estudia el efecto de este cambio.

Recolectamos huevos de rana recién puestos de un estanque y los llevamos al laboratorio, donde los dividimos en 60 recipientes de agua. La temperatura del agua de 20 de los contenedores se mantuvo a 13°C, 20 contenedores se mantuvieron a 18°C y los 20 contenedores restantes se mantuvieron a 25°C. (Tener una gran cantidad de repeticiones

aumenta nuestra confianza en que la diferencia esperada entre los grupos se debe al factor que nos interesa, en este caso, la temperatura).

Supervisamos cada contenedor de agua y registramos los tiempos de eclosión (días hasta la eclosión de los huevos) en una hoja de cálculo (aquí llamada `frogs_messy_data.csv`).

Nuestra variable de respuesta es `Hatching_time`. Nuestra variable explicativa es `Temperature`, con 3 niveles: 13°C, 18°C y 25°C.

Queremos comparar las medias de 3 grupos independientes (grupos de temperatura de 13 °C, 18 °C y 25 °C) y tenemos una variable de respuesta continua (Tiempo de eclosión) y una variable explicativa categórica (Temperatura). ¡ANOVA unidireccional es el análisis apropiado!

2. Formulación de una hipótesis

Siempre haz una hipótesis y predicción, antes de profundizar en el análisis de datos.

Una hipótesis es una respuesta tentativa a una pregunta bien formulada, que se refiere a una explicación mecanicista del patrón esperado. Se puede verificar a través de predicciones, que se pueden probar haciendo observaciones adicionales y realizando experimentos.

Esto debe estar respaldado por cierto nivel de conocimiento sobre tu sistema de estudio.

En nuestro caso, sabiendo que las crías de rana tardan entre 2 y 3 semanas en eclosionar a temperaturas óptimas (15-20 °C), podemos suponer que cuanto más baja sea la temperatura, más tiempo tardará en eclosionar. Por lo tanto, nuestra hipótesis puede ser: el tiempo medio de eclosión de los huevos de rana variará con el nivel de temperatura. Podemos predecir que dado nuestro rango de temperatura, a la temperatura más alta (25°C) se reducirá el tiempo de eclosión.

3. Manipulación de datos

Importación de datos

```
# carga paquetes
library(tidyverse)

library(ggplot2)

# Carga la base de datos
ranas_data <- read_csv("~/Documents/Posdoc/Curso_R/ANOVA/frogs_messy_data.csv
")

head(ranas_data)
```

Echemos un vistazo más de cerca a nuestro conjunto de datos. Como puedes ver a primera vista, este marco de datos tiene Temperatura13, Temperatura18 y Temperatura25 (los 3 niveles de nuestra variable explicativa) como columnas separadas dentro de las cuales se

ha registrado el tiempo de eclosión para cada muestra de desove de rana. Este es nuestro conjunto de datos en formato ancho.

Sin embargo, para analizar datos, necesitamos reordenar la hoja de datos en formato largo: esto significa ordenar los datos para que cada variable sea una columna y cada observación sea una fila.

```
### Formatea tu base de datos

# Reúne los tiempos de eclosión por cada temperatura

# Selecciona las dos columnas que ocuparemos "Hatching_time", "Temperature"

# Elimina las NAs

# Guarda la nueva base como un objeto nuevo
```

Hatching_time y Temperature son variables numéricas. ¿Es esto correcto?

Ahora es un buen momento para pensar en lo que queremos lograr con el conjunto de datos. Recuerda la pregunta de investigación: ¿Cómo varía el tiempo de eclosión de los huevos de rana con la temperatura?

Queremos modelar el tiempo de eclosión en función de la temperatura.

La temperatura es nuestra variable explicativa y aquí se codifica como variable numérica, cuando debería codificarse como factor (variable categórica) con 3 niveles (“13”, “18”, “25”). Los números representan las diferentes categorías de nuestra variable explicativa, no los datos de conteo reales. Por lo tanto, necesitamos transformar la temperatura de variable numérica a factorial.

```
# Haz temperatura a una variable de factores
```

4. Visualización de la distribución con un histograma

Siempre echa un vistazo a la distribución de tu variable de respuesta antes de profundizar en el análisis estadístico. Esto se debe a que muchas pruebas estadísticas paramétricas (dentro de las cuales ANOVA) asumen que las variables dependientes continuas se distribuyen normalmente, por lo que debemos verificar que se cumplan los supuestos para confiar en el resultado de nuestro modelo.

Ten en cuenta que los datos se pueden transformar logarítmicamente para cumplir con los supuestos de normalidad. Alternativamente, las pruebas no paramétricas están disponibles para datos que no se distribuyen normalmente.

👇

```
# Crea un histograma con ggplot
```

5. Visualización de medias con un diagrama de caja

Sigamos explorando nuestro conjunto de datos, usando un diagrama de caja.

Un diagrama de caja nos permite ver la variación en una variable continua entre categorías, la dispersión de los datos y nos da una idea de lo que podríamos encontrar con ANOVA en términos de diferencias entre grupos. Si los recuadros no se superponen, probablemente tengamos diferencias significativas entre los grupos, pero debemos verificar esto mediante un análisis estadístico.

👇

```
# Crea un diagrama de cajas con ggplot
```

6. Ejecutar un ANOVA unidireccional simple

👁*OJO. Mi base de datos se llama, `frogs_tidy_data`, sin embargo, tú probablemente le pusiste otro nombre

```
ranas_anova <- aov(Hatching_time ~ Temperature, data = frogs_tidy_data)
```

Puedes leer tu código de modelado como si fuera una oración: el código anterior ejecuta la prueba ANOVA (aov), analizando el tiempo de eclosión (Hatching_time) en función del (~) nivel de temperatura (Temperature), obteniendo datos (data = .. .) del marco de datos frogs_tidy_data.

Visualización de la tabla de resultados del modelo e interpretación

La función summary() muestra el resumen de tu ANOVA, también conocida como su tabla ANOVA, con grados de libertad, valor F y valor p.

```
summary(ranas_anova)
```

ANOVA divide la varianza total en: a) Un componente que puede ser explicado por la variable predictora (varianza entre niveles del tratamiento, es decir, grupos de temperatura): la primera fila de su tabla. b) Un componente que no puede ser explicado por la variable predictora (varianza dentro de los niveles, la varianza residual): la segunda fila de su tabla.

La estadística de prueba, F, es la relación de estas dos fuentes de variación. La probabilidad de obtener el valor observado de F se calcula a partir de la distribución de probabilidad

conocida de F, con dos grados de libertad: uno para el numerador (el número de niveles -1) y otro para el denominador (número de repeticiones - 1 x número de niveles). Por lo tanto, en nuestro caso Df entre niveles = 3-1 = 2 y Df dentro de niveles = 60 - 3 = 57. Esto representa cuántos valores involucrados en el cálculo tienen la libertad de variar.

El ANOVA muestra el valor p asociado al estadístico F. El valor p es la probabilidad del valor F observado de la distribución F (con los grados de libertad dados). El valor p es nuestro umbral de significancia. Un valor p es la probabilidad de ver una estadística de prueba tan grande o más grande que la que realmente observamos si la hipótesis nula es verdadera. Si $p < 0.05$ rechazamos la hipótesis nula. Sin embargo, la prueba debe repetirse varias veces para poder aceptar o rechazar con confianza la hipótesis nula.

Aquí, p es altamente significativo ($p < 2e\text{-}16$ ***). Esto significa que hay una diferencia significativa entre los tiempos de eclosión bajo diferentes niveles de temperatura. Nuestra variable predictora ha tenido un efecto significativo en su variable de respuesta.

Nótese que p es un valor arbitrario. ¡Así que ten cuidado! No es una medida universal y puede ser engañosa, dando como resultado falsos positivos. Te recomiendo esta publicación de blog sobre Métodos en Ecología y Evolución: "Hay locura en nuestros métodos".

7. Comprobación de supuestos

ANOVA hace 3 suposiciones fundamentales:

A. Los datos se distribuyen normalmente.

B. Las varianzas son homogéneas.

C. Las observaciones son independientes.

Necesitamos verificar que se cumplan los supuestos del modelo para confiar en los resultados de ANOVA. Revisemos estos uno por uno con tramas específicas:

A. Histograma de residuos y gráfico Q-Q normal:

la normalidad se puede verificar a través de un histograma de frecuencia de los residuos y un gráfico de cuantiles donde los residuos se representan frente a los valores esperados de una distribución normal.

Los residuos son la desviación de muestras medidas individualmente de la media.

Qué buscar: el histograma de residuos debe seguir una distribución normal (gaussiana) y los puntos en el gráfico Q-Q deben estar en su mayoría en línea recta.

```
#
par(mfrow = c(1,2))  # esto pone ls dos graficas en la misma ventana
hist(ranas_anova$residuals)   # histograma de residuoss
plot(ranas_anova, which = 2)   # hace la grafica Q-Q
```

Si no se cumple la suposición de normalidad, puedes transformar tus datos en logaritmo para que se distribuyan normalmente o ejecutar la alternativa no paramétrica a ANOVA: la prueba H de Kruskal-Wallis.

B. Residuos VS Gráfica ajustada:

para verificar que la variación en los residuos sea aproximadamente igual en todo el rango de la variable predictora (es decir, verificar la homocedasticidad), podemos graficar los residuos contra los valores ajustados del objeto del modelo aov.

Los valores ajustados son lo que predice el modelo para la variable de respuesta.

Qué buscar: ¡Queremos ver una línea roja recta centrada alrededor de cero! Esto significa que los residuos NO difieren sistemáticamente entre diferentes grupos.

```
# checando homoscedasticidad (Homogeneidad de varianzas)
plot(ranas_anova, which = 1)  # residuos VS datos
```

Si se viola la suposición de homogeneidad de varianzas, ejecute una prueba Welch F

C. ANOVA asume que todas las medidas replicadas son independientes entre sí:

Dos medidas son independientes si la medida de un individuo no indica qué valor producirá la medida de otro individuo.

Las medidas replicadas deben ser igualmente probables de ser muestreadas de la población de valores posibles para cada nivel. Este problema debe ser considerado en la etapa de diseño experimental. Si los datos se agrupan de alguna manera, se necesitan diseños más complejos para tener en cuenta factores adicionales. Se recomienda un enfoque de modelo mixto para datos jerárquicos.

Nuestros datos no violan ninguna de las suposiciones de ANOVA: ¡por lo tanto, podemos confiar en el resultado de nuestro modelo! Si las suposiciones no se cumplen al 100%, no hay problema, la mayoría de las veces es suficiente para que las suposiciones se cumplan aproximadamente.

8. Comunicar los resultados del modelo con un diagrama de barras

Podemos comunicar nuestros hallazgos de varias maneras:

```
Verbalmente: “El tiempo medio de eclosión de los huevos de rana varió signifi
cativamente con la temperatura (ANOVA, F = 385,9, df = 2, 57, p = 2,2e-16)” O
“El nivel de temperatura tuvo un efecto estadísticamente significativo en el
tiempo medio de eclosión de los huevos de rana (ANOVA, F = 385,9, gl = 2, 57,
p = 2,2e-16)”.
```

Después de ejecutar un ANOVA, siempre debe de informarse al menos su valor F, los grados de libertad y valor p.

```
Visualmente: podemos visualizar nuestros resultados con un gráfico de caja, como hicimos anteriormente, y con un gráfico de barras de medias de grupo con barras de error estándar.
```

En primer lugar, creemos un nuevo marco de datos con la función summarise(), que permite calcular estadísticas de resumen, incluido nuestro tamaño de muestra (n), tiempo medio de eclosión por nivel de temperatura, desviación estándar y valores de error estándar.

```
resumen_stats <- frogs_tidy_data %>%
                 group_by(Temperature) %>%
                 summarise(n = n(),  # calcular el tamaño de muestra (n)
                           average_hatch = mean(Hatching_time),
                           # calcular la media de tiempo de eclosión
                           SD = sd(Hatching_time))%>%  # calcular la desviaci
ón estándar
                 mutate(SE = SD / sqrt(n))  # Calcular el error estándar
```

La desviación estándar es una medida de la dispersión de valores alrededor de la media. El error estándar es una medida de la precisión estadística de una estimación.

Ahora, tracemos nuestro gráfico.

```
# Haz un gráfico de barras con la base de datos anterior
```