ANOVA

Paula

17/03/2022

Otros ejemplos

Análisis descriptivo

```
# Cargar el archivo
diet <- read.csv("~/Downloads/stcp-Rdataset-Diet.csv",row.names=1)
head(diet)

# Arreglar las columnas a mis necesidades
## generar una columna de peso perdido
class(diet$pre.weight)
diet$pre.weight <- as.numeric(diet$pre.weight)
class(diet$weight6weeks)
diet$weight.loss <- diet$pre.weight - diet$weight6weeks

##cambiar los factores de las columnas "dieta" y "genero"
library(dplyr)

class(diet$Diet)

diet$Diet <- as.factor(diet$Diet)
diet$diet.type <- recode_factor(diet$Diet, "1"="A","2"="B","3"="C")
diet$gender <- as.factor(diet$gender)
diet$Gender <- recode_factor(diet$gender, "0"="Female","1"="Male")

# Graficarlos datos
boxplot(weight.loss ~ diet.type, data= diet, col="white",
        ylab = "Perdida de peso (kg)", xlab = "Tipo de dieta")
abline(h=0,col="blue")
```

ANOVA o Kruskal Wallis?

```
# para decidir cual análisis correr, debemos de revisar la distribución de lo
s datos
hist(diet$weight.loss,col="green",main="")
# Si estos datos están distribuidos de manera normal, entonces podemos proced
er a hacer un ANOVA, de lo contrario, realizamos KW
```

```
lines(seq(-4,10,0.1),length(diet$weight.loss)*dnorm(seq(-4,10,0.1),mean(diet$
weight.loss),sqrt(var(diet$weight.loss))))

# Prueba de Shapiro. La prueba de Shapiro nos indica si los valores están dis
tribuidos de manera normal.
# OJO: La hipótesis nula indica que los valores ESTAN normalmente distribuido
s
shapiro.test(diet$weight.loss)
```

OJO! Un valor *p* no es la probabilidad de que la hipótesis nula sea cierta (esto es un malentendido común). Por el contrario, el valor *p* se basa en el supuesto de que la hipótesis nula es verdad. Un valor *p* es una estimación de la probabilidad de que un resultado particular (W = 0.98991 en este caso), o un resultado más extremo que el resultado observado, podría haber ocurrido por casualidad, si la hipótesis nula fuera cierta.

Un valor *p* grande (digamos, p = 0.802) significa que no hay evidencia convincente sobre la cual rechazar la hipótesis nula. Por supuesto, al decir "No rechazamos la hipótesis nula" y "la hipótesis nula es verdadera" son dos cosas distintas. Por ejemplo, es posible que no hayamos rechazado una hipótesis nula falsamente porque nuestro tamaño de muestra era demasiado bajo, o porque nuestro error de medición fue demasiado grande. Por lo tanto, los valores *p* son interesantes, pero no cuentan toda la historia: Los tamaños de efecto y tamaños de las muestras son igualmente importantes para sacar conclusiones.

```
diet.fisher<- aov(weight.loss~diet.type,data=diet)
summary(diet.fisher)
```

Revisión del Modelo

Para esto hay que definir los residuos del modelo restando la media de cada grupo a la pérdida de peso de los participantes correspondientes

```
# La media o mediana de cada grupo se pueden obtener mediante la función
tapply() que permite aplicar cualquier función a los datos

mean_group<- tapply(diet$weight.loss,diet$diet.type,mean)
mean_group

median_group<- tapply(diet$weight.loss,diet$diet.type,median)
median_group

## Ahora generemos dos columnas para los residuos de la media y mediana
diet$resid.mean<- (diet$weight.loss - mean_group[as.numeric(diet$diet.type)])
diet$resid.median<- (diet$weight.loss - median_group[as.numeric(diet$diet.typ
e)])
diet[1:10]

#Visualizar los residuos (estos deben de tener una distribución normal)
par(mfrow=c(1,2),mar=c(4.5,4.5,2,0)) #par() me permite poner dos gráficos jun
```

```
tos y definir los márgenes
boxplot(resid.mean~diet.type,data=diet,main="Residual boxplot per group",col="light gray",xlab="Diet type",ylab="Residuals")
abline(h=0,col="blue")
col_group = rainbow(nlevels(diet$diet.type))
qqnorm(diet$resid.mean,col=col_group[as.numeric(diet$diet.type)])
qqline(diet$resid.mean)
legend("top",legend=levels(diet$diet.type),col=col_group,pch=21,ncol=3,box.lwd=NA)
```

Finalmente, vamos a realizar una prueba de Shapiro para evaluar si hay suficiente evidencia de que los residuos no están distribuidos normalmente (mediante la función shapiro.test ()) y a realizar una prueba de Bartlett para evaluar si existe suficiente evidencia de que los residuos por grupo no tienen varianza diferente (mediante la función bartlett.test ().)

```
shapiro.test(diet$resid.mean)


# y una prueba para la homogeneidad de varianzas
bartlett.test(diet$resid.mean~as.numeric(diet$diet.type))
```

Comparaciones múltiples

Vamos a realizar una prueba Tukey HSD para definir qué pares de grupos tienen diferentes medias (mediante la función TukeyHSD ()) y a comparar el tamaño del intervalo de confianza de Tukey HSD para la diferencia de medias entre las pérdidas de peso de la Dieta A y la Dieta B con la obtenida mediante una prueba t de Student (función t.test () con argumento var.equal establecido en TRUE)

```
plot(TukeyHSD(diet.fisher))

t.test(weight.loss~diet.type,data=diet[diet$diet.type!="C",],var.equal = TRUE)
```

OTRO EJEMPLO

Primero carga los paquetes necesarios

```
library(tidyverse)
library(ggpubr)
library(rstatix)
```

ANOVA de un solo factor

Preparación de los datos

Usaremos una base de datos de R llamada `r PlantGrowth'. Contiene el peso de plantas obtenidas bajo dos tratamientos y un control

```r
data("PlantGrowth")
Plantas<- PlantGrowth
head(Plantas)

levels(Plantas)

Plantas$group<-factor(Plantas$group, levels=c("ctrl", "trt1", "trt2"))
```

Visualizacion

```r
library("ggpubr")
ggboxplot(Plantas, x = "group", y = "weight",
          color = "group", palette = c("#00AFBB", "#E7B800", "#FC4E07"),
          order = c("ctrl", "trt1", "trt2"),
          ylab = "Weight", xlab = "Treatment")
```

Checar los supuestos

Valores atípicos

Los valores atipicos pueden verse con graficas o corriendo la funcion

```r
Plantas %>%
  group_by(group) %>%
  identify_outliers(weight)
```

Sin valores atípicos. En la situación en la que se tienen valores atípicos extremos, puede deberse a errores de entrada de datos, errores de medición o valores inusuales. Puedes incluir el valor atípico en el análisis de todos modos si no crees que el resultado se verá afectado sustancialmente. Esto se puede evaluar comparando el resultado de la prueba ANOVA con y sin el valor atípico.

Normalidad

Se puede revisar usando uno de los siguientes metodos: 1. Analizar los residuos del modelo de ANOVA 2. Revisar la normalidad de cada grupo por separado

1. Usando los residuos del modelo Se usan las graficas tipo QQ y Shapiro-Wilk

```r
#Crea un modelo lineal
modelo<- lm(weight ~ group, data=Plantas)
# Crea la grafica tipo QQ de residuos
ggqqplot(residuals(modelo))
```

Este tipo de grafica muestra la correlación entre ciertos datos y la distribución normal

```r
# Haz Shapiro-Wilk
shapiro_test(residuals(modelo))
```

Tanto la grafica como la prueba indican normalidad de los datos

2. Usando la normalidad por grupo

```
Plantas %>%
  group_by(group) %>%
  shapiro_test(weight)
```

Distribución normal de todos los grupos. !OJO! si tu muestra es >50, la grafica QQ es preferida pues cuando las muestras son muy grandes, Shapiro-Wilk se vuelve muy sensible a la menor desviacion de la normalidad

```
ggqqplot(Plantas, x="weight", facet.by = "group")
```

NOTA: Si tus datos resultan no estar normalmente distribuidos, checa esta pagina para transformarlos de manera correcta: https://rcompanion.org/handbook/I_12.html

Homogeneidad de varianza

1. Con una grafica

```
plot(modelo, 1)
```

No hay evidencia de que haya una relación entre los residuos y la media de cada grupo, lo cual es bueno, por lo tanto, asumimos homogeneidad de las varianzas

2. Con una prueba

```
Plantas %>% levene_test(weight ~ group)
```

Con base en los resultados, podemos ver que el valor de p>0.05, lo cual lo hace no-significante. Esto significa que no hay diferencias significativas entre las varianzas de los grupos.

NOTA: En caso de no cumplirse este supuesto, puedes correr una prueba de Welch usando la siguiente función `r welch_anova_test()' en el paquete *rstatix*

Ahora si, ANOVA:

```
res.aov<- aov(weight ~ group, data = Plantas)
summary(res.aov)
```

Df = Es el número de piezas de información independientes que se utilizaron para calcular la estimación. En otras palabras, son el número de valores que pueden variar libremente en un conjunto de datos. POR ESO SIEMPRE ES EL NUMERO DE OBSERVAIONES QUE TIENES MENOS 1 (N-1) Sum Sq = Esta es la suma de la diferencia al cuadrado de cada valor real con respecto al valor previsto. Mean sq = Es en realidad la varianza Valor F = Te dice si las medias entre los dos grupos de datos son significativamente diferentes. Pr(>F) = El valor p

Prueba Post-hoc

Cuando nuestro modelo de ANOVA indica que hay diferencias significativas entre los grupos, podemos entonces investigar qué grupo(s) es(son) diferente(s) de los demás, para ellos necesitamos comparar los grupos de 2 en 2, y en ese caso se usa precisamente una prueba *t*.

Las pruebas post-hoc son una familia de pruebas estadísticas, y hay un montón. Las más utilizadas son las pruebas de Tukey HSD y Dunnett:

Tukey HSD se utiliza para comparar todos los grupos entre sí. Dunnett se utiliza para hacer comparaciones con un grupo de referencia. Por ejemplo, considere 2 grupos de tratamiento y un grupo de control.

OJO, Ten en cuenta que se supone que las varianzas son iguales para ambas pruebas. Si las variaciones no son iguales, puede utilizar la prueba de Games-Howell, entre otras.

Prueba de Tukey HSD

En nuestro caso, hay un tratamiento "de referencia" o control, por lo que nos interesa comparar los tratamientos 1 y 2 contra el control, entonces vamos a utilizar la prueba de Dunnett.

En R, la prueba de Dunnett se realiza de la siguiente manera. Aquí es donde el segundo método para realizar el ANOVA resulta útil porque los resultados (`r res.aov') se reutilizan para la prueba post-hoc:

```
library(DescTools)
DunnettTest(Plantas$weight, Plantas$group)
```

Y entonces, ¿por quá ANOVA es significativo?

```
TukeyHSD(res.aov)
plot(TukeyHSD(res.aov, conf.level=.95), las = 2)

#Caclular la media de cada grupo
mean<- aggregate(Plantas$weight, by=list(Plantas$group),mean)
sd<- aggregate(Plantas$weight, by=list(Plantas$group),sd)
resutados<-cbind(mean, sd)
names(resutados)<- c("Tratamiento","mean_distance","sd")
```

Reporte

Ahora hay que informar los resultados de ANOVA unidireccional de la siguiente manera:

Se realizó un ANOVA de una vía para evaluar si el crecimiento de la planta era diferente para los 3 grupos de tratamiento diferentes: ctr (n = 10), trt1 (n = 10) y trt2 (n = 10).

El crecimiento de las plantas fue estadísticamente significativo entre los diferentes grupos de tratamiento, F(2,27) = 4.846, p = 0.0159.

El crecimiento de la planta disminuyó en el grupo trt1 (4.66 +/- 0.79) en comparación con el grupo ctr (5.03 +/- 0.58). Aumentó en el grupo trt2 (5.53 +/- 0.44) en comparación con el grupo trt1 y ctr.

Los análisis post-hoc de Tukey revelaron que el aumento de trt1 a trt2 (0.865, IC del 95% (0.17 a 1.56)) fue estadísticamente significativo (p = 0.012), pero ninguna otra diferencia de grupo fue estadísticamente significativa.

Visualizacion

```
#Vamos a calcular ANOVa y Tukey con el paquete [rstatix] para hacer la grafic
a linda
aovTest<- Plantas %>% anova_test(weight ~ group)
#> Coefficient covariances computed by hccm()
pwc<- PlantGrowth %>% tukey_hsd(weight ~ group)

# Grafica con valores-p
pwc <- pwc %>% add_xy_position(x = "group")
ggboxplot(Plantas, x = "group", y = "weight") +
  stat_pvalue_manual(pwc, hide.ns = TRUE) +
  labs(
    subtitle = get_test_label(aovTest, detailed = TRUE),
    caption = get_pwc_label(pwc)
    )
```