Guía vizualización y análisis de datos

Diego H Halabi, PhD, Assist. Prof.

Resumen

El análisis de la varianza, Anova, es una herramienta estadística que nos permite identificar si existen diferencias entre 3 o más grupos, cuando en ellos medimos una variable cuantitativa. Un ejemplo clásico en Medicina, es cuando tenemos que comparar los resultados de 3 tratamientos. En este ejercicio, trabajaremos con los datos simulados de un Ensayo Clínico Aleatorio, pasando por las diferentes etapas que nos permitirán obtener un modelo para tomar la mejor decición de tratamiento.

Índice

Descripción de los datos

Estos datos provienen de un Ensayo Clínico Aleatorio de diseño paralelo (simulados) en el que se trató durante 2 meses a 60 pacientes diabéticos tipo II con metformina, ejercicio o antocianinas. El outcome fue el nivel de hemoglobina glicosilada al inicio del tratamiento (basal) y la misma medición una vez finalizado el tratamiento. En la figura 1 se puede ver el flujograma del estudio, para su mejor comprensión.

En este documento veremos cómo realizar un análisis de los datos para verificar si alguno de los tratamientos es más efectivo.

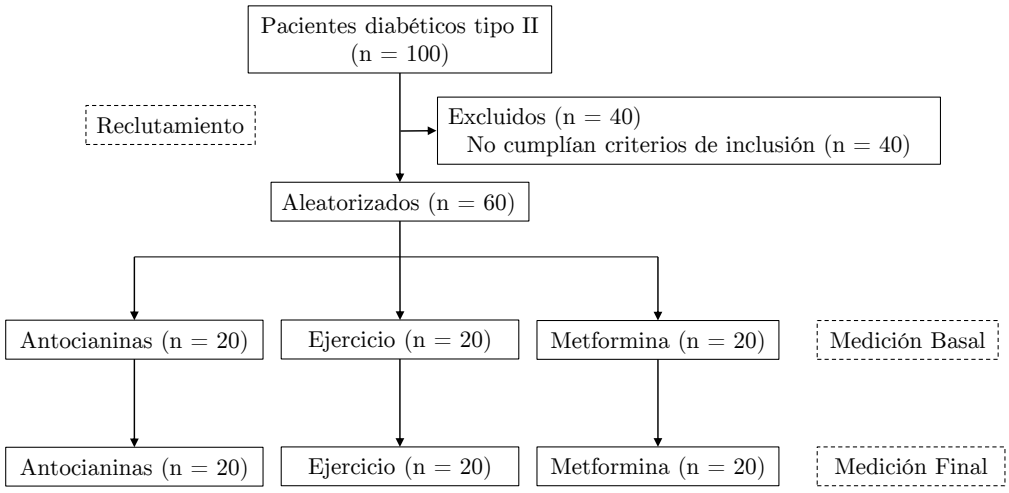


Figura 1: Flujograma del Ensayo Clínico Aleatorio del que provienen los datos (simulados).

1. Importar los datos.

En R, podemos ingresar los datos directamente en la línea de comandos. Sin embargo, muchas veces tendremos extensas bases de datos que ya están tabulados, y resulta más simple y seguro importarlos. Es importante conocer los formatos en que se puede encontrar una base de datos. El formato más simple es csv, el cual puede ser exportado desde programas como Excel, Google Docs, Calc, etc.

Para este ejemplo, utilizaremos un archivo con los datos tabulados que se llama `analisis-estadistico.csv`.

```
dataR <- read.csv("data.csv", sep=";")
attach(dataR)
str(dataR)
```

```
## 'data.frame':    60 obs. of  4 variables:
##  $ Grupo : chr  "Metformina" "Metformina" "Metformina" "Metformina" ...
##  $ Sexo  : chr  "masculino" "femenino" "masculino" "masculino" ...
##  $ HbA1cA: num  9.9 10.2 9.6 9.1 9.9 9.9 9.8 9.2 11 9.6 ...
##  $ HbA1cB: num  7.2 7.6 7.1 7.3 7.4 8.5 7.8 7.3 8.1 6.6 ...
```

Las variables en nuestro set de datos son:

- Grupo: variable independiente, categorizada en los 3 posibles tratamientos.
- HbA1cA: covariable, corresponde a los niveles de hemoglobina glicosilada basal.
- Hba1cB: variable dependiente, que corresponde a la hemoglobina glicosilada al finalizar los tratamientos.

2. Transformar.

Como nos interesa conocer la efectividad de cada tratamiento en la reducción de la hemoglobina glicosilada, debemos sustraer el valor final (HbA1cB) a la medición basal (HbA1cA). En otras palabras, crearemos una nueva variable; reducción de la hemoglobina glicosilada, y la llamaremos HbA1cR.

```
HbA1cR <- HbA1cA-HbA1cB
str(HbA1cR)
```

```
##  num [1:60] 2.7 2.6 2.5 1.8 2.5 1.4 2 1.9 2.9 3 ...
```

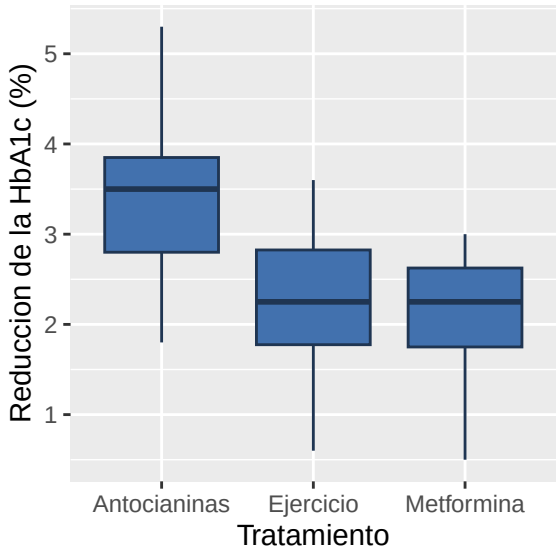
3. Visualizar.

Visualizaremos nuestros datos mediante gráficos y números.

3.1 Gráficos.

Una forma sencilla de observar los datos, es mediante los gráficos. En este caso, graficaremos los niveles de hemoglobina glicosilada al inicio y al final del tratamiento. Para realizar un gráfico de cajas y bigotes, cargaremos la librería `ggplot2`.

```
library(ggplot2)
ggplot(dataR,aes(x=Grupo,y=HbA1cR))+
  geom_boxplot(fill="#4271AE",colour="#1F3552")+ scale_x_discrete(name="Tratamiento")+
  scale_y_continuous(name="Reduccion de la HbA1c (%)")
```



Al finalizar el tratamiento, observamos que el grupo tratado con antocianinas reduce los niveles de hemoglobina glicosilada en comparación a los grupos tratados con ejercicio o metformina. Sin embargo, alguien podría considerar que esta diferencia no es suficiente para tomar decisiones clínicas, por lo que debemos profundizar más nuestro análisis.

3.2 Números.

Aquí nos interesa obtener medidas de tendencia central y dispersión, principalmente. Comenzamos por cargar la librería `FSA` que nos permitirá esto.

```
library(FSA)
```

```
## ## FSA v0.9.5. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

Ahora utilizamos la función `Summarize` para obtener los resultados que nos interesan.

```
Summarize(HbA1cA~Grupo) #medición basal, por grupo
```

```
##          Grupo  n   mean        sd min    Q1 median     Q3  max
## 1 Antocianinas 20 10.210 0.5485867 9.1 9.800  10.35 10.700 10.9
## 2    Ejercicio 20 10.045 0.6210983 9.1 9.575   9.95 10.525 11.0
## 3   Metformina 20 10.075 0.6163347 9.1 9.600   9.90 10.575 11.0
```

```
Summarize(HbA1cB~Grupo) #final del tratamiento, por grupo
```

```
##          Grupo  n  mean        sd min   Q1 median    Q3  max
## 1 Antocianinas 20 6.770 0.9674165 4.6 6.20   7.00 7.500  8.1
## 2    Ejercicio 20 7.770 1.0301252 6.1 7.05   7.45 8.750  9.6
## 3   Metformina 20 7.955 0.8432238 6.6 7.30   7.95 8.325 10.3
```

```
Summarize(HbA1cR~Grupo) #reducción de la hemoglobina glicosilada
```

```
##          Grupo  n  mean        sd min    Q1 median    Q3 max
## 1 Antocianinas 20 3.440 0.8964257 1.8 2.800   3.50 3.850 5.3
## 2    Ejercicio 20 2.275 0.7047769 0.6 1.775   2.25 2.825 3.6
## 3   Metformina 20 2.120 0.7097813 0.5 1.750   2.25 2.625 3.0
```

Resulta muy útil observar si al inicio del tratamiento, los grupos eran similares. También se puede observar la misma tendencia visualizada en el gráfico; el tratamiento con antocianinas parece reducir más la hemoglobina glicosilada, cuando lo comparamos a la metformina y el ejercicio.

4. Modelamiento.

Evaluaremos si la reducción de la hemoglobina glicosilada en el grupo tratado con antocianinas es significativa; es decir, si se debe al azar o el tratamiento realmente es efectivo. Para esto, ajustaremos nuestros datos a un modelo de Análisis de la Varianza; Anova.

Antes de ajustar a Anova, debemos verificar que los datos cumplan con ciertos requisitos, o *assumptions*. En primer lugar, los datos deben ajustarse a la distribución normal, y en segundo lugar, debe existir homogeneidad en las varianzas.

4.1. Ajuste a la distribución normal.

Verificaremos si nuestros resultados se ajustan a una distribución normal, mediante el test de Shapiro Wilk. Por lo tanto, la hipótesis nula será que nuestros datos tienen una distribución normal.

```
tapply(HbA1cR,Grupo,shapiro.test)
```

```
## $Antocianinas
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97937, p-value = 0.9257
##
##
## $Ejercicio
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97576, p-value = 0.8685
##
##
## $Metformina
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.93067, p-value = 0.159
```

Observamos que el valor W de los 3 grupos es cercano a 1, y su p-value correspondiente es mayor a 0.05, por lo tanto, podemos aceptar la hipótesis nula y asumir una distribución normal.

Si aun existieran dudas, podríamos corroborar el ajuste mediante un gráfico Q-Q, con la función `qqnorm`.

4.2 Verificar la homogeneidad de las varianzas.

Mediante el test de Levene, verificaremos si las varianzas son homogéneas en nuestros 3 grupos. Para esto, tenemos que cargar la librería `car`. La hipótesis nula es que no existen diferencias en las varianzas de los 3 grupos.

```
library(car)
```

```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method       from
##   hist.boot    FSA
##   confint.boot FSA
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:FSA':
##
##     bootCase
```

```
leveneTest(HbA1cR~Grupo)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.3411 0.7125
##       57
```

Al igual que el test anterior, al encontrar un p-value elevado, podemos aceptar la hipótesis nula y asumir una similitud en las varianzas de los 3 grupos.

4.3 Análisis de la varianza.

Ya hemos verificado que nuestros datos cumplen con los requisitos o *assumptions* para realizar un test de Anova. Lo primero que haremos, será ajustar el modelo:

```
Anova <- aov(HbA1cR~Grupo)
summary(Anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Grupo        2  20.82  10.412   17.31 1.33e-06 ***
## Residuals   57  34.28   0.601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(Anova,~.,test="F")
```

```
## Single term deletions
##
## Model:
## HbA1cR ~ Grupo
```

```
##        Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>              34.277 -27.5913
## Grupo   2    20.824 55.102  -3.1097  17.314 1.332e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos observar que el p-value es menor a 0.001. Esto significa que existen diferencias significativas entre al menos un grupo con el resto, ergo rechazamos la hipótesis nula. Podemos presumir que se trata del grupo tratado con antocianinas, pero desconocemos si hay diferencias entre los otros 2 grupos.

Para dilucidar esto, haremos un test post-hoc, que básicamente corresponde a hacer 3 t-test con un ajuste para las observaciones múltiples. En este caso, ajustaremos nuestro modelo al test de Tukey, ya que se encuentra en un sano equilibrio entre test muy conservadores como Bonferroni, pero sin el alto riesgo de cometer el error de tipo I por no ajustar los p-value acumulados, como el caso de Dunn.

```
TukeyHSD(Anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = HbA1cR ~ Grupo)
##
## $Grupo
##                             diff        lwr        upr     p adj
## Ejercicio-Antocianinas  -1.165 -1.7551172 -0.5748828 0.0000416
## Metformina-Antocianinas -1.320 -1.9101172 -0.7298828 0.0000043
## Metformina-Ejercicio    -0.155 -0.7451172  0.4351172 0.8031395
```

Aquí podemos ver en los p-value ajustados, que las diferencias solo son significativas entre las antocianinas y los demás grupos. El ejercicio y la metformina no difieren entre ellos. Además, podemos obtener datos respecto a la magnitud del efecto; la diferencias de medias de las antocianinas fue 1.17 (IC 95% 1.76 - 0.57) con el ejercicio, y 1.32 (IC95% 1.91 - 0.73) con la metformina.

5. Conclusión.

Basado en los resultados obtenidos, he confeccionado la tabla1, procurando que se mantenga lo más simple posible, pero sin perder información relevante para la práctica clínica.

Cuadro 1: Reducción de la hemoglobina glicosilada (HbA1c) después de 2 meses de tratamiento con antocianinas, ejercicio o metformina. Los resultados están representados como promedio ± desviación estándar (DE).

Tratamiento	n[a]	Reducción HbA1c(%)	dm[b] (95% IC)
Antocianinas	20	3.44 ± 0.90[c]	Ref
Metformina	20	2.28 ± 0.70	1.32 (0.73 - 1.91)
Ejercicio	20	2.12 ± 0.71	1.17 (0.57 - 1.75)

[a]número de pacientes
[b]diferencia de medias
[c]diferente de metformina y ejercicio ($p < 0.001$; Anova, post-hoc Tukey)

El tratamiento con antocianinas es más efectivo que la metformina o el ejercicio para el manejo de pacientes diabéticos, reduciendo el porcentaje de hemoglobina glicosilada entre 0.57% a 1.91% más.

Es necesario mencionar que otra posibilidad de abordar este conjunto de datos, hubiese sido analizando los niveles de hemoglobina glicosilada basal y final mediante Anova de 2 factores.