Análisis de datos ómicos (M0-157)

Primera prueba de evaluación continua

Isabel Terrero Aragón

El data set de metabolómica utilizado para este estudio es el que se encuentra en el repositorio de GitHub (https://github.com/nutrimetabolomics/metaboData). Esta base de datos ha sido seleccionada ya que contiene una gran cantidad de datos lo que hace que el análisis sea más veraz y realista. Además, tiene una amplia variedad de variables lo que hace que se puedan hacer diferentes estudios y, por tanto, reduce las limitaciones del análisis.

El análisis de estos datos se realiza con Rstudio, donde primero se han cargado los datos y metadatos de este data set. Para evitar problemas en los diferentes análisis, se han reemplazado los valores nulos por 0, ya que hay bastantes valores nulos y si se eliminan las muestras con datos nulos se acaban eliminando todos los datos (Imagen 1). En el caso de los valores nulos en las variables categóricas se han sustituido por NA. Una vez que se han preprocesado los datos correctamente se pasa a crear el objeto *SummarizedExperiment* (Imagen 2). Este se utiliza para almacenar matrices rectangulares de resultados experimentales, pudiendo gestionar simultáneamente varios resultados experimentales o ensayos que tengan las mismas dimensiones. Lo más importante de esta clase es que puede coordinar los metadatos y los ensayos al crear subconjuntos. Esta clase es muy similar al *ExpressionSet* pero se diferencia principalmente en que *SummarizedExperiment* es más flexible en la información de sus filas, permitiendo tanto los basados en *GRanges* como los descritos por *DataFrames*.

```r
{r}
# Cargar las librerías necesarias
library(SummarizedExperiment)

# Definir la ruta de los archivos
ruta <- "/Users/isabelterrero/Documents/Datos_omicos/"

# Cargar los datos
data_values <- read.csv(paste0(ruta, "DataValues_S013.csv"), row
.names = 1)
data_info <- read.csv(paste0(ruta, "DataInfo_S013.csv"), row.names =
1)

# Sustituir los valores nulos en data_values (variables numéricas)
por 0
num_cols_values <- sapply(data_values, is.numeric)

# Reemplazar NAs por 0 en las columnas numéricas
data_values[, num_cols_values] <- lapply(data_values[,
num_cols_values], function(x) {
  x[is.na(x)] <- 0
  return(x)
})

# Sustituir los valores nulos en data_info (variables categóricas)
por NA
cat_cols_info <- sapply(data_info, is.factor)

# Reemplazar NAs por NA en las columnas categóricas
data_info[, !cat_cols_info] <- lapply(data_info[, !cat_cols_info],
function(x) {
  x[is.na(x)] <- NA
  return(x)
})
```

Imagen 1: Script de la carga de librerías y datos en Rstudio y sustitución de los valores nulos.

```r
{r}
# Crear el objeto SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = as.matrix(data_values)),
  colData = data_info
)
# Verificar el objeto SummarizedExperiment creado
se
```

Imagen 2: Script de la creación del objeto SummarizedExperiment

Una vez visualizada las variables y el contenido de este data set se procede a un análisis estadístico de los datos. Primero se hace un análisis univariante para ver cómo se comporta cada variable por separado. La media de edad de los participantes en este estudio es de 40.7 (Figura 1). Además, la mayoría de los participantes son mujeres (Figura 2). En cuanto al índice de masa corporal (ICM) la media se encuentra en 50.50, aunque el valor de ICM con más frecuencia se encuentra entre 40 y 50 (Figura 3). El valor de glucosa normal en un adulto es entre 70 y 100 mg/dL y la mayoría de los participantes de este estudio se encuentran dentro del rango normal. Sólo algunos sobre pasan los 125 mg/dL que es cuando empieza a ser grave (Figura 4). Por último, el peso de los participantes es demasiado elevado en general, se podría decir que en su mayoría tienen sobrepeso (Figura 5).

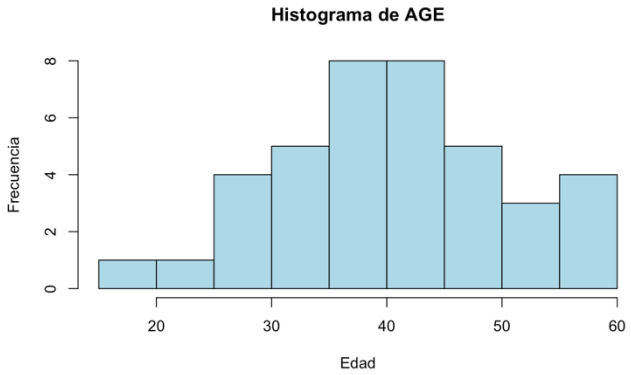

Figura 1: Histograma con la frecuencia de la distribución de las edades.

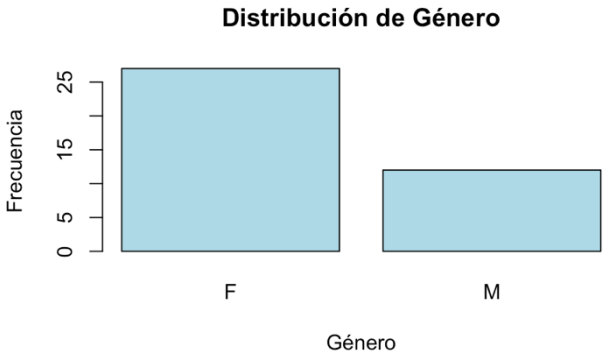

Figura 2: Histograma con la frecuencia de los géneros.

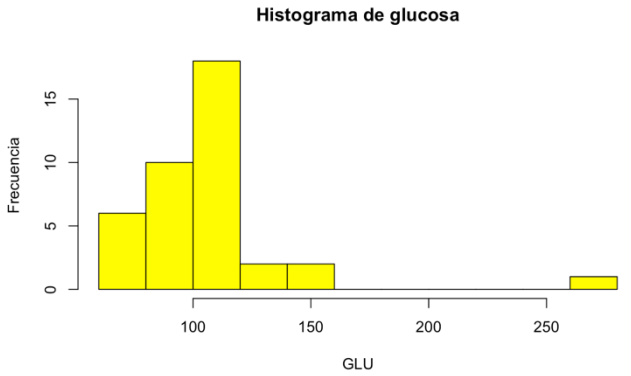

Figura 3: Histograma con la frecuencia de los niveles de glucosa.

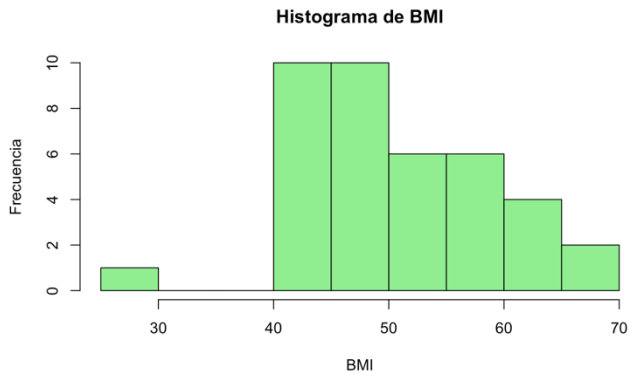

Figura 4: Histograma con la frecuencia del índice de masa corporal (ICM) (BMI en inglés).

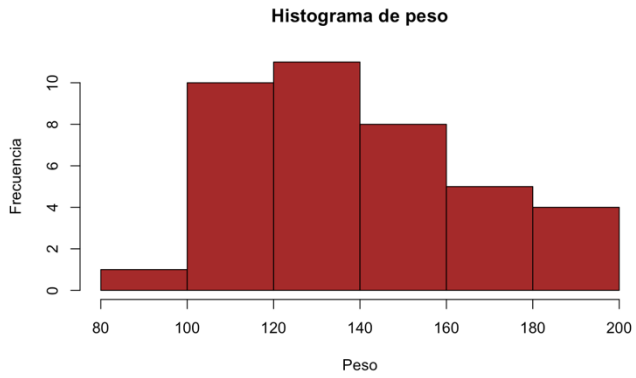

Figura 5: Histograma con la frecuencia del peso.

Ahora que se tiene una visión general de cómo se distribuyen estás variables, se pasa a un análisis multivariante de las diferentes variables. Primero se hace un análisis general para ver qué tipo de personas participan, por ejemplo, como se distribuye la edad según el género (Figura 6). En este caso, las mujeres participantes son mayores que los hombres. Además, se estudiará que variables pueden estar relacionadas significativamente o no. Para ello, se hace una matriz de correlación (Figura 7) en la que se puede destacar la relación del peso con el ICM y una pequeña relación del peso con los niveles de glucosa.

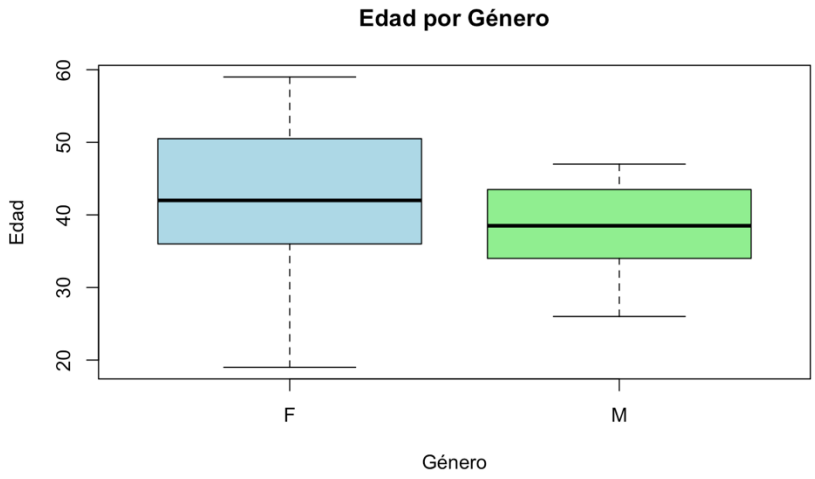

Figura 6: Boxplot para comparar la edad por género.

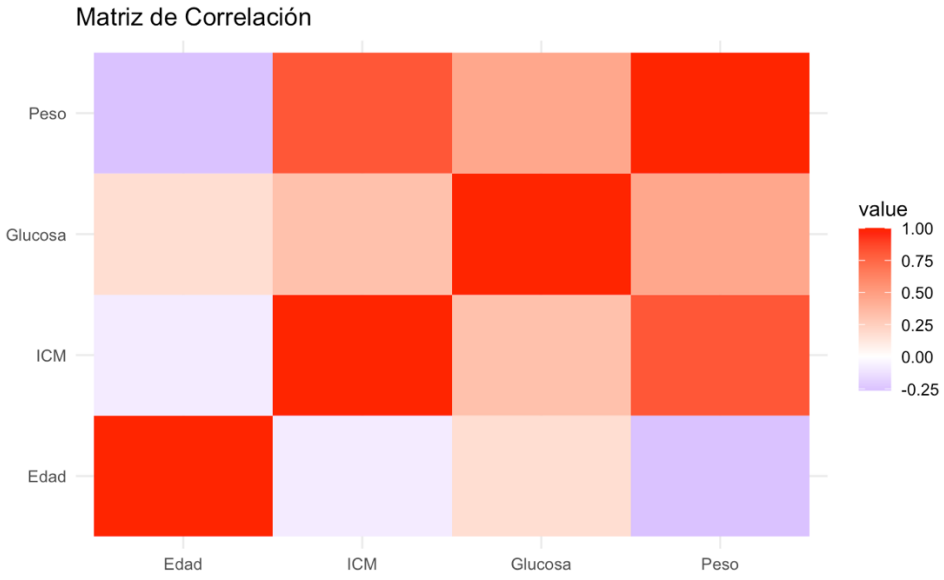

Figura 7: Matriz de correlación entre las diferentes variables.

Ahora que se sabe que el peso puede estar relacionado con el ICM, se hace una prueba ANOVA para comprobarlo. En esta prueba se obtiene un valor de p-value de 2.69e-10 lo que indica que el ICM tiene una relación estadísticamente significativa con el peso (Figura 8).

```
            Df Sum Sq Mean Sq F value   Pr(>F)
bmi          1  18829   18829   73.19 2.69e-10 ***
Residuals   37   9518     257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 8: Resultados de la prueba ANOVA entre peso e ICM.

También se vio que el peso podría estar relacionado también con la glucosa, por lo que se hace otra prueba ANOVA entre estas dos variables. Obteniendo un valor de p-value muy grande por lo que indica que la glucosa no tiene un efecto significativo sobre el peso (Figura 9).

```
            Df Sum Sq Mean Sq F value Pr(>F)
glu_data     1    114  113.81    1.17  0.286
Residuals   37   3599   97.26
```

Figura 9: Resultados de la prueba ANOVA entre peso y glucosa.

Por último, se hará una regresión múltiple para ver si la glucosa tiene efecto sobre el peso mientras se controlan los efectos de las otras variables. Con este análisis se puede confirmar la prueba ANOVA realizada anteriormente entre el peso y el ICM, ya que el ICM tiene un fuerte efecto positivo sobre el peso (Figura 10). En cambio, se obtiene que la glucosa tiene un efecto positivo sobre el peso, lo que sugiere que los niveles más altos de glucosa están relacionados con un mayor peso, aunque su efecto es más pequeño en comparación con el ICM (Figura 10).

```
Call:
lm(formula = peso_data ~ age_data + bmi_data + glu_data, data = counts_df)

Residuals:
    Min      1Q  Median      3Q     Max
-27.819  -7.879  -2.119   8.270  35.828

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.24202   17.24835   1.579  0.12324
age_data    -0.69331    0.23300  -2.976  0.00527 **
bmi_data     2.30645    0.28639   8.054 1.76e-09 ***
glu_data     0.23049    0.07609   3.029  0.00459 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.82 on 35 degrees of freedom
Multiple R-squared:  0.7642,    Adjusted R-squared:  0.744
F-statistic: 37.82 on 3 and 35 DF,  p-value: 4.434e-11
```

Figura 10: Resultados de la regresión múltiple entre el peso y la glucosa, edad y el ICM.

Con este último análisis se puede ver como la glucosa tiene un efecto positivo sobre el peso cuando se tienen en cuenta otras variables, por lo que nos proporciona un resultado más preciso sobre como la glucosa impacta en el peso.

Todo el data set y el código utilizado para hacer este análisis se encuentre en el repositorio de GitHub: https://github.com/ita98/Terrero-Aragon-Isabel-PEC1.git