Universidad de La Coruña

Análisis de la varianza

Modelización en Biomedicina

Miriam Gutiérrez Serrano

Índice

1. Introducción **2**

2. Diferencias en el nivel de expresión **2**

3. Modelo matemático propuesto **4**

4. Tabla ANOVA **5**

5. 10000 genes **6**

1. Introducción

En esta práctica se estudia el análisis de la varianza (ANOVA) aplicado a datos reales de expresión génica obtenidos mediante secuencias de ADN. El objetivo es evaluar si existen diferencias significativas en el nivel de expresión de ciertas secuencias cortas de ADN (genes) en función de dos factores: el tipo de tejido del que proceden las muestras celulares y su estado (sano o canceroso).

Se nos proporciona el fichero `'datos.10000randomGenes.2025.rda'` que contiene los niveles de expresión genética de 10000 secuencias cortas de ADN, observados en cinco tipos de tejidos (óseo, cerebral, mamario, adiposo y renal). Para cada tejido disponemos de 20 muestras de células sanas y 20 muestras de células cancerosas, según esta indicado en las cabeceras de las columnas del archivo.

Como caso de estudio inicial, se analiza la secuencia `204489_s_at`, correspondiente al gen CD44, conocido por estar implicado en diversos procesos celulares relacionados con el cáncer, como la adhesión celular, la migración y la metástasis. A través de herramientas gráficas y modelos estadísticos, se pretende comprobar si este gen presenta diferencias de expresión significativas entre tejidos y/o estados, y posteriormente extender el análisis al conjunto completo de genes.

2. Diferencias en el nivel de expresión

El primer objetivo es representar gráficamente los niveles de expresión del gen CD44 (fila 50), con el fin de explorar si existen diferencias apreciables en función del tejido y del estado (sano o canceroso) de las muestras.

Para ello se extrajo la fila correspondiente a dicho gen de la matriz de expresión y se construyeron dos variables:

- tejido: óseo, cerebral, mamario, adiposo y renal;
- estado: sano y canceroso.

Estas variables se organizaron en un data.frame y representamos un gráfico de cajas entre el tipo de tejido y el nivel de expresión. El color muestra el estado de la muestra (sano o cáncer). El código empleado es el siguiente:

```
    load("ruta/a/tu/archivo/datos.10000randomGenes.2025.rda")
ls()

dim(datosPractica.2025)
rownames(datosPractica.2025)[50]
colnames(datosPractica.2025)[1:200]

# Extraemos la fila 50 (CD44)
expresion <- as.numeric(datosPractica.2025[50, ])

# Construimos los factores "tejido" y "estado"
estado <- rep(c("Sano", "Cancer"), each = 20, times = 5)
```

```
tejido <- rep(c("Oseo", "Cerebral", "Mamario", "Adiposo", "Renal"), each = 40)

# Creamos un data.frame
df <- data.frame(
  expresion = expresion,
  tejido = factor(tejido),
  estado = factor(estado)
)

# Grafica
library(ggplot2)

ggplot(df, aes(x = tejido, y = expresion, fill = estado)) +
  geom_boxplot(position = position_dodge(0.8)) +
  labs(title = "Expresión del gen CD44 (204489_s_at)",
       x = "Tejido",
       y = "Nivel de expresión",
       fill = "Estado") +
  theme_minimal()
```

Y el resultado gráfico:

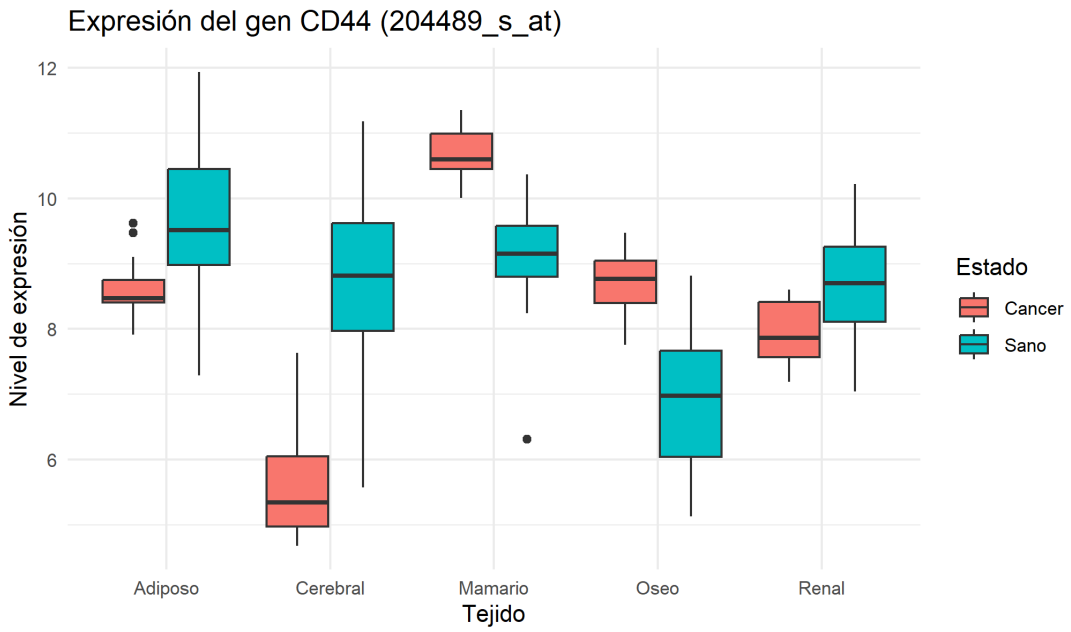

Figura 1: Gen 204489_s_at

Esta representación nos permite detectar diferencias aparentes en la expresión del gen entre distintos tejidos. En alguno de ellos, como el mamario o el cerebral, también observamos variaciones notables entre los grupos sano y canceroso. Nos sugiere que tanto el tejido como el estado podrían tener un efecto significativo sobre la expresión del gen. Por lo tanto, tiene sentido hacer un análisis estadístico mediante un modelo ANOVA con dos factores.

3. Modelo matemático propuesto

Tras la comparación gráfica de los datos, planteamos un análisis através de un modelo matemático. Para estudiar si el nivel de expresión del gen `CD44` difiere significativamente entre tejidos y estados (sano/cáncer), se plantea un análisis de la varianza de dos factores con interacción. Este modelo permite evaluar no solo los efectos individuales de cada factor, sino también si existe interacción entre ambos, es decir, si el efecto del estado depende del tejido.

El modelo considerado es el siguiente

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \quad 1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K \tag{1}$$

donde:

- y_{ijk} es la observación correspondiente al k-ésimo individuo del grupo formado por el i-ésimo nivel del factor A (tejido) y el j-ésimo nivel del factor B (estado);
- μ es la media general;
- α_i es el efecto del tejido i;
- β_j es el efecto del estado j;
- $(\alpha\beta)_{ij}$ es la interacción entre el tejido i y el estado j;
- $e_{ijk} \sim \mathcal{N}(0, \sigma^2)$ son errores aleatorios independientes e idénticamente distribuidos.

El modelo se somete a las siguientes restricciones, que aseguran su identificabilidad:

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \sum_{i=1}^{I} (\alpha\beta)_{ij} = 0 \ \forall j, \quad \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0 \ \text{ para todo } i$$

Y podemos establecer las siguientes hipótesis:

- $H_0^A : \alpha_1 = \cdots = \alpha_I = 0$ (no hay efecto del tejido),
- $H_0^B : \beta_1 = \cdots = \beta_J = 0$ (no hay efecto del estado),
- $H_0^{AB} : (\alpha\beta)_{ij} = 0 \quad \forall i, j$ (no hay interacción entre los factores).

Estas hipótesis se contrastan mediante la descomposición de la variabilidad total en la tabla ANOVA, que haremos más adelante. Con los datos proporcionamos podemos definir el modelo con $I = 5$ (tipos de tejidos), $J = 2$ (estados) y $K = 20$ (réplicas por combinación).

Bajo las condiciones de ortogonalidad impuestas, los estimadores de los parámetros del modelo son:

$$\hat{\mu} = \bar{y}_{\cdots}, \qquad \hat{\alpha}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}, \qquad \hat{\beta}_j = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}, \qquad \widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}$$

donde las medias están calculadas sobre los índices correspondientes.

A continuación, en la siguiente figura se muestra la media de expresión del gen para combinación de tipo de tejido y estado. Cada línea representa uno de los dos niveles de estado.

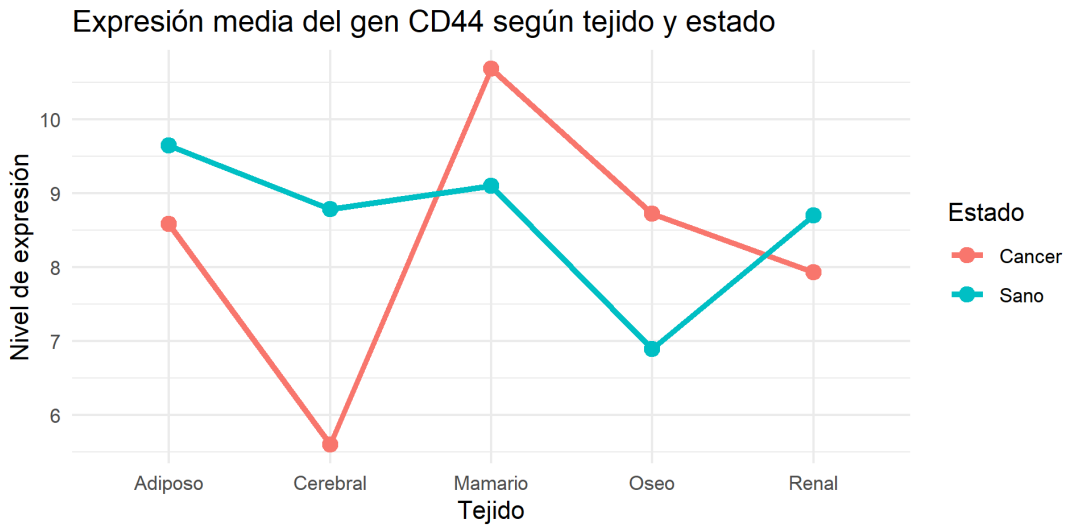

Figura 2

Como vemos las líneas no son paralelas, es decir, el efecto del estado depende del tejido y no es constante. Por ejemplo, en el tejido mamario la expresión es notablemente mayor en muestras cancerosas. Mientras que en el tejido cerebral ocurre lo contrario: las muestras sanas tienen una expresión superior. De esta forma, podemos justificar la inclusión del término de interacción en el modelo matemático.

El código utlizado para la representación gráfica ha sido el siguiente:

```
  ggplot(df, aes(x = tejido, y = expresion, color = estado, group = estado)) +
stat_summary(fun = mean, geom = "line", size = 1.2) +
stat_summary(fun = mean, geom = "point", size = 3) +
labs(title = "Expresión media del gen CD44 según tejido y estado",
     x = "Tejido",
     y = "Nivel de expresión",
     color = "Estado") +
theme_minimal()
```

4. Tabla ANOVA

Para evaluar los efectos de los factores tejido, estado y su interacción sobre la expresión del gen CD44, se ha ajustado un modelo ANOVA bifactorial con interacción y se ha generado la tabla ANOVA correspondiente utilizando la función `anova()`.

La tabla siguiente recoge los grados de libertad, sumas de cuadrados, medias cuádraticas, estadísticos F y los correspondientes p valores para cada uno de los términos del modelo.

Fuente	Df	Sum	Mean	F value	Pr(>F)
tejido	4	181.355	45.339	53.830	$< 2 \times 10^{-16}$
estado	1	5.063	5.063	6.011	0.01512
tejido:estado	4	171.836	42.959	51.005	$< 2 \times 10^{-16}$
Residuals	190	160.028	0.842		

Tabla 1: Tabla ANOVA para el gen CD44 (modelo bifactorial con interacción).

Los resultados muestran:

- El factor tejido es altamente significativo ($p < 2e - 16$), indicando que existen diferencias claras entre los cinco tipos de tejidos.
- El factor estado también influye en el nicel de expresión del gen ($p = 0,015$).
- La interacción entre tejido y estado es igualmente muy significativo ($p < 2e - 16$), lo que confirma que el efecto canceroso no es homogéneo entre tejidos.

Estos resultados respaldan la elección del modelo del ejercicio anterior, y permite concluir que el gen presenta un perfil de expresión dependiente tanto del tejido como del estado y su combinación.

5. 10000 genes

En este apartado extendemos el ánalisis realizado sobre el gen CD44 al conjunto completo de 10000 secuencias de ADN. Queremos determinar cuántas de ellas muestran evidencia de estar relacionadas con el estado de las muestras (sano o canceroso).

Se ha ajustaado el modelo para cada gen, y se ha extraído el valor p correspondiente al efecto del estado. A continuación, se ha contado cuántos de estos valores son menotes que $\alpha = 0,05$. El código utilizado es el siguiente:

```
pvalores_estado <- numeric(nrow(datosPractica.2025))  # 10,000

for (i in 1:nrow(datosPractica.2025)) {
  expresion <- as.numeric(datosPractica.2025[i, ])
  df <- data.frame(expresion = expresion,
                   tejido = factor(tejido),
                   estado = factor(estado))

  modelo <- aov(expresion ~ tejido * estado, data = df)
  resumen <- summary(modelo)

  pvalores_estado[i] <- resumen[[1]]["estado", "Pr(>F)"]
}
# Contamos cuantos genes tienen p < 0.05
sum(pvalores_estado < 0.05)
```

Hemos obtenido 5744 genes con un p valor menor que 0,05 para el efecto del estado. Esto sugiere que más de la mitad de las secuencias presentan diferencias de expresión asociadas al cáncer. Cabe destacar que al estar realizando 10000 contrastes simultáneamente,

el uso de un umbral fijo como α puede afectar y obtener un número importante de falsos positivos.

Todo el código en R utilizado para el tratamiento de los datos, el ajuste de modelos y la generación de gráficos se adjunta al presente informe para su consulta y posible reproducción de los resultados.