Estadística de datos ómicos

Guillermo Ayala Gallego

2024-04-02

Table of contents

Estadística y datos ómicos

- Aplicar técnicas estadísticas a datos ómicos.
- Nos centraremos en lo que tienen en común.
- El énfasis de este texto es en **métodos estadísticos**.
- Una frase (que subscribo) de un gran estadístico inglés

```
fortunes::fortune(50)
```

```
To paraphrase provocatively, 'machine learning is statistics minus any checking
of models and assumptions'.
   -- Brian D. Ripley (about the difference between machine learning and
      statistics)
      useR! 2004, Vienna (May 2004)
```

Estructura de los datos

- Observaremos un gran número de características sobre un pequeño número de muestras.
- Las características serán de distinto carácter:
  - Fluorescencia cuando trabajemos con distintos tipos de microarrays y que se asocia con abundancia de transcritos o presencia de metilación,

  - número de lecturas alineadas cuando hablamos de procedimientos de secuenciación,
  - presencia o ausencia de una mutación cuando estudiemos asociación.

- Esta característica puede estar asociada a:

  - una sonda o a un grupo de sondas en un microarray.
  - O bien la información corresponde a un gen o a un exon,
  - o a un péptido, a una proteína,
  - o a una región genómica,
  - o ...

El mundo al revés

- La característica con la que trabajamos en cada momento se cuantifica con distintos procedimientos.

- El número lo denotamos por N donde este valor es grande (miles).

- El número de muestras es n (decenas con suerte).

- N es mucho mayor que n: $n \ll N$.
- Estadística de alta dimensión (High dimensional statistics).

Matriz de ...

- Las características las recogemos en una matriz.
- Matriz de expresión o de metilación o de conteos o de mutaciones.

$$\mathbf{y} = [y_{ij}]_{i,j=1,\ldots,n}$$

- y_{ij} nos cuantifica la característica i en la muestra j.
- Es la traspuesta de una matriz de datos en Estadística.

- Las distintas muestras son **independientes**.

- Son **condicionalmente** independientes.

- Las filas de $\mathbf{y}$ son realizaciones de vectores dependientes.

Preprocesado

- Habitualmente los datos de las columnas de la matriz **y** no son directamente comparables.
- Hay muchos artefactos técnicos así como ruido en la observación de la característica de interés.
- Se han desarrollado técnicas para corregirlo: el preprocesado de la información.

Metadatos o variables fenotípicas

- De cada muestra tendremos información.

- Si es una muestra control o de tratamiento.

- Las variables que describen las muestras son los **metadatos** o **variables fenotípicas**.
- Denotaremos por $x = (x_1, \dots, x_n)$ los valores observados de una variable en las n muestras.
- Casos y controles: $x_i = 1$ si es un caso e $x_i = 0$ si es un control.

Problemas estadísticos

- **Expresión diferencial** (marginal o en grupos)
- **Metilación diferencial** (lo mismo que en el punto anterior)
- **Estudios de asociación** (en todo el genoma o localizados)
- **Clasificación** con o sin supervisión.