# Ejercicios Prácticos: Simulaciones en Filiaciones Complejas

Análisis de Distribuciones de LR

Dr. Franco Marsico
Taller de filiaciones complejas, SAGF
19 de noviembre, 2025

## Introducción

Los siguientes ejercicios están diseñados para trabajar con las distribuciones de LR obtenidas mediante simulaciones. Se realizaron **1,000 simulaciones** para cada escenario, evaluando tanto la hipótesis H1 (relacionado) como H2 (no relacionado). Trabajaremos con dos pedigríes diferentes:

- **sim1.txt** corresponde al pedigrí 1 (Figura 1)
- **sim2.txt** corresponde al pedigrí 2 (Figura 2)

En ambos casos, el parentesco evaluado es entre los **individuos de referencia (rayados) y el individuo 7**.

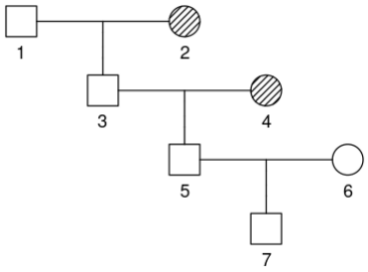


Figura 1: Escenario sim1

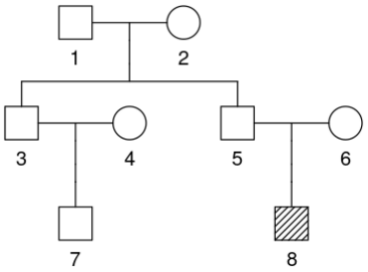


Figura 2: Escenario sim2

## Ejercicio 1: Caracterización de las Distribuciones

Acá buscamos familiarizarnos con las distribuciones de LR bajo ambas hipótesis y calcular sus medidas de tendencia central.

a) Abrir los datos de ambos archivos (`sim1.txt` y `sim2.txt`) en alguna hoja de cálculo o R.

b) Para cada archivo, identifique las columnas correspondientes a LR bajo H1 (related, hipótesis de parentesco verdadera) y LR bajo H2 (unrelated, hipótesis de no parentesco verdadera).

c) Calcule las siguientes medidas de tendencia central para ambas distribuciones:

- Media, Mediana, Desviación estándar
- Mínimo y máximo
- Percentiles 5 %, 25 %, 75 % y 95 %

d) Genere histogramas para visualizar ambas distribuciones. Utilice escala logarítmica en el eje x: $\log_{10}(LR)$.

e) ¿Qué diferencia observa entre la media y la mediana en cada distribución? ¿Qué indica esto sobre la forma de las distribuciones?

# Ejercicio 2: Selección de Umbral y Tasas de Error

Este ejercicio sirve para comprender la relación entre el umbral seleccionado y las tasas de falsos positivos (TFP) y falsos negativos (TFN).

**Marco Teórico:** Para un umbral $T$ dado (ver abajo):

- **TFN:** Proporción de casos bajo H1 (related) donde $LR < T$
- **TFP:** Proporción de casos bajo H2 (unrelated) donde $LR > T$

a) Seleccione los siguientes umbrales: $T = 1, 10, 100, 1000, 10000$

b) Para cada umbral, calcule TFN y TFP.

c) Complete la siguiente tabla:

| Umbral (T) | TFN | TFP |
|---|---|---|
| 1 | | |
| 10 | | |
| 100 | | |
| 1000 | | |
| 10000 | | |

d) Grafique la curva TFP vs TFN.

e) Pregunta:

- ¿Qué umbral consideraría adecuado si su prioridad es minimizar falsos positivos?
- ¿Y si su prioridad es minimizar falsos negativos?
- ¿Algún umbral representa un balance razonable entre ambos tipos de error?

# Ejercicio 3: Falsos Positivos Esperados en Búsquedas a Gran Escala

Que pasa si queremos estimar el número esperado de falsos positivos al realizar búsquedas en bases de datos de distintos tamaños. En búsquedas mediante bases de datos de ADN, la TFP puede parecer pequeña (ej: 0.001), pero al comparar contra una base de datos grande, puede generar muchos candidatos.

a) Fije un umbral $T = 10$.

b) Calcule la TFP para este umbral.

c) Estime el número esperado de falsos positivos para bases de datos de los siguientes tamaños: $N = 100, 1000, 10000, 100000$ individuos. Esto te realiza multiplicando el TFP obtenido por el N.

d) Complete la tabla:

| **Tamaño BD (N)** | **Falsos Positivos Esperados** |
|---|---|
| 100 | |
| 1,000 | |
| 10,000 | |
| 100,000 | |

e) **Preguntas:**

- ¿Cómo cambia el número de falsos positivos esperados con el tamaño de la base de datos?
- ¿Qué estrategias podría implementar para reducir la carga de trabajo (casos que superan el umbral) sin perder verdaderos positivos?

# Ejercicio 4: Estabilidad de Medidas con Tamaño Muestral

Evaluar cómo las medidas de tendencia central se estabilizan al aumentar el número de simulaciones.

Las simulaciones son aproximaciones estocásticas. Con pocas simulaciones, las estimaciones pueden variar considerablemente. Este ejercicio explora cuántas simulaciones son necesarias para obtener estimaciones estables.

a) Tome submuestras aleatorias de tamaños: 10, 50, 100, 300, 500 simulaciones de los datos bajo H1.

b) Para cada submuestra, calcule: Media de LR, Mediana de LR y Desviación estándar de LR.

c) Repita el muestreo 10 veces para cada tamaño muestral.

d) Para cada tamaño muestral, calcule:

- Media de las 10 medias
- Desviación estándar de las 10 medias
- Coeficiente de variación (CV = SD/Media)

e) Grafique la evolución de la media y su CV en función del tamaño muestral.

f) **Preguntas:**

- ¿A partir de qué número de simulaciones las medidas comienzan a estabilizarse? ¿Se estabiliza?
- ¿Qué medida (media o mediana) se estabiliza más rápidamente?
- ¿Cuántas simulaciones recomendaría realizar en la práctica?

**Referencias:**


- Marsico, F., Sibilla, G., Escobar, M. S., and Chernomoretz, A. (2024). The missing person problem through the lens of information theory. *Forensic Science International: Genetics*, 70, 103025.
- Slooten, K. (2020). Likelihood ratio distributions and the (ir)relevance of error rates. *Forensic Science International: Genetics*, 44, 102173.
- Marsico, F., and Amigo, M. (2025). Ethical and security challenges in AI for forensic genetics: From bias to adversarial attacks. *Forensic Science International: Genetics*, 76, 103225.
- Marsico, F. L., and Caridi, I. (2023). Incorporating non-genetic evidence in large scale missing person searches: A general approach beyond filtering. *Forensic Science International: Genetics*, 66, 102891.
- Marsico, F. L., Vigeland, M. D., et al. (2021). Making decisions in missing person identification cases with low statistical power. *Forensic Science International: Genetics*, 54, 102519.
- Vigeland, M. D., Marsico, F. et al. (2020). Prioritising family members for genotyping in missing person cases: a general approach combining the statistical power of exclusion and inclusion. *Forensic Science International: Genetics*, 49, 102376.


**Software Disponible:**

- **Familias:** `https://familias.name`
- **App Vínculos:** `https://vinculosgen.com`
- **Mispitools (R):** `https://github.com/MarsicoFL/mispitools`