TP1 2026 (WIP)

Introducción

La ciudad costera de Puerto Bruma vive su temporada de fútbol con estadios llenos y un clásico que paraliza todo: los Albatros del Muelle contra los Búhos de la Ribera.

Luego de incidentes aislados y mucha presión mediática, el Ministerio de Convivencia Pública implementó un sistema de IA llamado Centinela Delta para gestionar el derecho de admisión en los estadios: el Templo de la Almeja (localía de los Albatros) y el Fortín de la Bruma (localía de los Búhos).

Para cada persona que intenta ingresar, Centinela Delta asigna un Score de Riesgo $S \in (0, 100)$ y se instruyó a la policía y al gobierno que consideren que S es la probabilidad estimada de que la persona cometa un incidente en la cancha.

La regla de decisión es simple:

```
- Si $S>T$ (umbral fijo $T$), se prohíbe el acceso del simpatizante, es decir, predicción: "
- SI $S\leq T$, se permite el acceso, es decir, predicción: "pacífico"
```

La empresa proveedora, NebulaGuard Analytics, sostiene que su score está perfectamente calibrado: "un *score* del 80% es una probabilidad del 80%, sin importar el color de la camiseta".

Tras unos meses del sistema funcionando, la ONG Gradas en Paz presenta una denuncia ante la Defensoría del Pueblo. Su argumento es:

> "Históricamente hubo mayor vigilancia y controles alrededor del Fortín de la Bruma. Eso hizo que, incluso entre gente pacífica, los hinchas de Lobos tengan más registros, más 'señales' y más fricción con el sistema."

La denuncia es contra el sistema, en términos estadísticos

> "A nuestros hinchas pacíficos (inocentes) de los Búhos les prohíben entrar más seguido que a los pacíficos del Albatros."

Nos pondremos en el rol de consultores de ciencia de datos contratados por la Defensoría del Pueblo para auditar el sistema. Nos entregan un dataset anonimizado de la temporada anterior para realizar el análisis.

Los objetivos son:

```
- Estimar con estadística Bayesiana si hay evidencia de disparidad en los errores (más allá
- Explicar el resultado en términos comprensibles para un organismo público.
- Conectar el resultado con el "trade-off" de fairness: calibración vs igualdad de tasas de
- Estudiar mediante simulación la posibilidad de generar un sistema más justo.
```

Actividades

Para todo lo que sigue, tenga en consideración lo siguiente:

- Se toma una muestra al azar con reemplazo de tamaño $n = 100$ de una población de tamaño $N = 1000$.
- Para nuestros estudios comparativos, supondremos que el porcentaje de alumnos que apuesta (lo que se quisiera estimar) es 40%.

- Cuando se les pregunta directamente si han hecho alguna vez apuestas, los estudiantes que sí han apostado alguna vez mienten con probabilidad μ.
- Si se utilizan técnicas de respuesta aleatorizada, los estudiantes no mienten.

Comenzaremos realizando un estudio de simulación para estudiar el efecto de la mentira en las estimaciones. En esta primera aproximación, consideraremos que se realiza la pregunta directa.

1. Proponga un modelo bayesiano que, a partir de encuestar a n estudiantes, permita estimar π_a. Explique la elección de la función de verosimilitud y el *prior* e indique cómo se obtiene el *posterior*.
2. Utilizando R, simule la obtención de una muestra para el caso en que los estudiantes no mienten y para el caso en que los estudiantes mienten con tres niveles de mentira μ bajo, medio y alto. Compare los resultados de la inferencia.
3. Realice ahora 1000 simulaciones y compare los resultados de las inferencias.

Considere ahora el caso del método propuesto por Warner:

4. Según este método, ¿cuál es la probabilidad (llamémosla λ_W) de que un estudiante responda afirmativamente? ¿cuál es la probabilidad de que un estudiante responda por la contraria?
5. A partir de lo anterior, proponga un modelo razonable sobre cómo se generan los datos.
6. Considere un *prior* uniforme y halle el *posterior* exacto.

> **Ayuda**
>
> Escriba el *posterior* dejando la constante de normalización expresada como Z. Es decir, escriba
>
> $$p(\pi_a|y) = \frac{N(\pi_a)}{Z}$$
>
> Naturalmente, Z es una integral. Muestre que el resultado de esa integral es:
>
> $$Z = \frac{B(1-p; y+1, n-y+1) - B(p; y+1, n-y+1)}{1-2p}$$
>
> donde $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ es la función beta incompleta, p es la probabilidad con la que se le pregunta si apuesta e y es la cantidad de personas que responden afirmativamente.
> Para resolver la integral de Z, utilice el método de sustitución y hágalo en términos de λ_W. No es tan terrible como parece.

7. Grafique el *posterior* para diferentes valores de p y concluya. Si quiere evaluar el comportamiento a través de múltiples muestras, puede hacerlo replicando el proceso generador de los datos que planteó en el ítem 5 (no hace falta obtener la muestra aleatoria simple cada vez).
8. ¿Qué pasaría si el porcentaje de alumnos que apuesta fuera diferente al 40%? Analice los resultados en función de diferentes niveles de p y π_a.
9. Escriba una función de R que le permita realizar la inferencia (en forma aproximada) con un *prior* beta no necesariamente uniforme. Para eso, utilice una grilla de valores de π_a

Para el método propuesto por Greenberg:

10. ¿Cuál es la probabilidad λ_G de que un estudiante responda que sí? ¿y de que responda que no?

11. Escriba una función de R que le permita realizar la inferencia (en forma aproximada) con un *prior* beta no necesariamente uniforme.

Para terminar, es hora de comparar todos los escenarios.

12. Utilizando R, simule la obtención de una muestra para el caso en que los estudiantes no mienten, mienten con tres niveles de mentira μ (bajo, medio y alto), se utiliza el método de Werner ($p = 0.3$) y se utiliza el método de Greenberg. Compare los resultados de las inferencias en cada caso.

13. Realice ahora 1000 simulaciones y analice los resultados.