Trabajo Práctico – IECD

- La entrega consiste en **un archivo PDF** con las respuestas y **un script de R** que permita reproducir todos los resultados obtenidos.
- La fecha límite de entrega es el **8 de diciembre**. No se aceptarán entregas fuera de este plazo.
- Los grupos deben ser de **2 o 3 estudiantes** (idealmente 3). Cada grupo debe inscribirse en el Drive grupal correspondiente.
- El trabajo debe ser completamente **reproducible**. Para ello, fije la semilla utilizando el **mínimo número de libreta** entre los integrantes del grupo.
- Las consultas que se responderán serán únicamente sobre el **enunciado**. No se resolverán dudas sobre implementación de código o desarrollo de las soluciones.
- Las **exposiciones orales** se realizarán por **Zoom** en la fecha que indicará la cátedra.

Contexto del problema

Se desea estimar la **prevalencia** θ de una condición en una población. Para ello, se dispone de un **test diagnóstico imperfecto** T con dos resultados posibles: 1 si el resultado del test es positivo y 0 si es negativo. Sea Y la variable aleatoria que representa el estado verdadero de una persona y que vale 1 si la persona presenta la enfermedad, 0 si no.

Se definen así la *sensibilidad* Se como

$$\mathbb{P}(T = 1 \mid Y = 1) = Se\,,$$

y la *especificidad* Sp como

$$\mathbb{P}(T = 0 \mid Y = 0) = Sp\,.$$

El objetivo de este trabajo será estimar la prevalencia $\theta = \mathbb{P}(Y = 1)$ de la enfermedad y cuantificar la incertidumbre, comparando casos con y sin error de medición.

Importante: Para resolver el TP, puede definir variables auxiliares y agregar las hipótesis que considere necesarias.

1 Parte I: Test perfecto (baseline)

Supongamos primero que contamos con un test **perfecto**. Obsérvese que este caso correspondería a considerar $Se = Sp = 1$.

Se seleccionan n personas al azar y se define T_{per} la variable aleatoria que cuenta la cantidad de personas enfermas en la muestra.

1. Observe que $T_{\text{per}} \sim Bi(n, \theta)$.
2. Calcule el estimador de máxima verosimilitud (EMV) de θ. Llame a ese estimador $\hat{\theta}_{\text{per}}$.
3. Analice sesgo, varianza, error cuadrático medio (ECM), consistencia y distribución asintótica del estimador $\hat{\theta}_{\text{per}}$.
4. Construya un intervalo de confianza para θ de nivel asintótico 0.95.
5. Suponga $\theta = 0.25$. Evalúe coberturas empíricas mediante simulación Monte Carlo. Observe qué ocurre para distintos valores de n.

2 Parte II: Test imperfecto con Se y Sp conocidos

Definamos la probabilidad p como

$$p = P(T = 1)\,.$$

1. ¿Cómo estimaría p con T_{per}? Llame a ese estimador $\widehat{p}$.
2. Escribir a p como en función de la prevalencia θ, de la sensibilidad Se y de la especificidad Sp.
3. Mostrar con gráficos, para $Se = 0.9$, $Sp = 0.95$ y $\theta = 0.25$, cómo cambia p en función de:
   (a) θ dejando fijos Se y Sp,
   (b) Se dejando fijos θ y Sp, y
   (c) Sp dejando fijos θ y Se.

2.1 Estimador de momentos (MoM)

4. Teniendo en cuenta la relación entre p y θ, calcular el estimador de momentos de θ. Llamarlo $\widehat{\theta}_{\text{MoM}}$.
5. Analice sesgo, varianza y ECM. ¿Qué se puede decir de la consistencia?
6. Grafique el ECM en función de n y compárelo con el ECM del test perfecto.
7. Realice simulaciones para comparar los valores teóricos hallados en el item 5 con los simulados.
8. Para $\theta = 0.25$, $Se = 0.9$ y $Sp = 0.95$, construya muestras bootstrap para observar la distribución del estimador de momentos de θ cuando $n = 10$. ¿Qué observa?

2.2 Intervalos de confianza

9. Construya intervalos de confianza bootstrap percentil de θ basado en $\widehat{\theta}_{\text{MoM}}$. Para ello, realice simulaciones para $\theta = 0.25$, $Se = 0.9$ y $Sp = 0.95$ y distintos valores de n.

10. Construya intervalos de confianza de nivel asintótico 0.95 para θ basado en $\hat{\theta}_{\text{MoM}}$.

11. Con simulaciones, compare coberturas y longitudes promedio de los intervalos de confianza de los items anteriores.

2.3 Estimador truncado

12. Observe que, para ciertas muestras, el $\widehat{\theta}_{\text{MoM}}$ puede encontrarse fuera del intervalo $[0, 1]$.

13. Definamos así el estimador de momentos *truncado* como

$$\widehat{\theta}_{\text{trunc}} = \begin{cases} \widehat{\theta}_{\text{MoM}} & \text{si } 0 \leq \widehat{\theta}_{\text{MoM}} \leq 1 \\ 0 & \text{si } \widehat{\theta}_{\text{MoM}} < 0 \\ 1 & \text{si } \widehat{\theta}_{\text{MoM}} > 1 \end{cases}$$

Para este estimador, aproxime, utilizando simulaciones para valores de $n = 10, 100$ y 1000, si es insesgado y/o asintóticamente insesgado, su varianza, el ECM y su distribución asintótica.

3 Parte III: Dos muestras (pre–post intervención)

Se realizó una campaña de vacunación contra la enfermedad. Lo que se desea ahora es evaluar si cambió la prevalencia de la enfermedad tras dicha campaña, comparando las prevalencias antes y después de la intervención. A estas prevalencias las llamaremos θ_{pre} y θ_{post}.

Supogamos que se muestrearon n_{pre} personas antes de llevar adelante la campaña y n_{post} después de llevar adelante la campaña y que las muestras son **independientes**. En ambos casos aplicaremos el mismo test diagnóstico imperfecto, caracterizado por su sensibilidad Se y su especificidad Sp.

Definimos X_{pre} a la cantidad de personas a las que el test les dio positivo en la etapa previa a la campaña y X_{post} a la cantidad de personas a las que el test les dio positivo después de llevar adelante la campaña. Luego,

$$X_{\text{pre}} \sim \text{Bi}(n_{\text{pre}}, p_{\text{pre}}) \quad \text{y} \quad X_{\text{post}} \sim \text{Bi}(n_{\text{post}}, p_{\text{post}}).$$

con $p_A = (Se + Sp - 1)\,\theta_A + (1 - Sp)$ siendo A igual a pre o post, según corresponda.

El parámetro de interés es la diferencia de prevalencias verdaderas:

$$\Delta = \theta_{\text{post}} - \theta_{\text{pre}}.$$

3.1 Test de Hipótesis

1. Utilizando el estimador $\widehat{\theta}_{\mathrm{MoM}}$ (el que no está truncado), plantee un test de nivel aproximado 0.05 para las hipótesis

$$H_0 \,:\, \Delta = 0 \quad \text{vs} \quad H_1 \,:\, \Delta \neq 0\,.$$

2. Aplique a un caso ficticio con $n_{\mathrm{pre}} = n_{\mathrm{post}} = 100$, $Se = 0.9$, $Sp = 0.95$, $\theta_{\mathrm{pre}} = 0.2$ y $\theta_{\mathrm{post}} = 0.15$ y $\alpha = 0.05$. ¿Qué ocurre si achicamos los tamaños de muestra?

3. Construya intervalos de confianza de nivel asintótico 0.95 para Δ.

4. Fijado el tamaño de muestras n_{pre} y n_{post}, calcule el nivel empírico del test, es decir, genere $Nrep$ muestras con los tamaños establecidos y calcule la proporción de rechazos a nivel 0.05. Utilice tamaños de muestras pequeños y grandes.