instituto de cálculo
UBA - CONICET

Trabajo práctico - IECD

8 de diciembre de 2025 Intro a la Estadística y Ciencia de Datos

Integrante	LU	Correo electrónico
Allami, Florencia	484/23	allami.florencia@gmail.com
Torres, Emiliano	80/23	emilianomtorres1@gmail.com
Vanotti, Franco	464/23	fvanotti15@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
http://www.exactas.uba.ar

0. 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

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

y la *especificidad Sp* como

$$\mathbb{P}(T = 0 | Y = 0) = Sp$$

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.

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 $T_{\text{per}} \sim \text{Bi}(n, \theta)$

   Sea $Y_i = 1$ si la i-esima persona en la muestra está enferma y $Y_i = 0$ caso contrario, con esta definicion obtenemos que $Y_i \sim Be(\theta)$. Además como tenemos una muestra aleatoria tenemos que $\forall i, j \in [1, ...n]$ con i distinto de j, Y_i es independiente de Y_j. Por esto mismo, podemos ver que $T_{per} = \sum_{i=1}^{n} Y_i \sim Bin(n, \theta)$.

2. Calcule el estimador de máxima verosimilitud (EMV) de θ. Llame a ese estimador $\hat{\theta}_{\text{per}}$.

   Sea $L(\theta)$ la verosimilitud, tenemos que $L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$.

   Tomamos logaritmo natural.

   $$ln(L(\theta)) = ln(\theta) \sum_{i=1}^{n} y_i + (ln(1-\theta)) \left(n - \sum_{i=1}^{n} y_i \right)$$

   Llamamos $\ell(\theta) = ln(L(\theta))$ y tomamos ℓ'.

   $$\ell'(\theta) = \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n - \sum_{i=1}^{n} y_i}{1-\theta}$$

   Ahora igualamos a $\ell'(\theta) = 0$.

   $$\frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n - \sum_{i=1}^{n} y_i}{1-\theta} = 0$$

   $$\frac{\sum_{i=1}^{n} y_i}{\theta} = \frac{n - \sum_{i=1}^{n} y_i}{1-\theta}$$

   $$(1-\theta) \sum_{i=1}^{n} y_i = \theta \left(n - \sum_{i=1}^{n} y_i \right)$$

   $$\sum_{i=1}^{n} y_i - \theta \sum_{i=1}^{n} y_i = n\theta - \theta \sum_{i=1}^{n} y_i$$

   $$\theta = \frac{\sum_{i=1}^{n} y_i}{n}$$

   Encontramos un punto crítico. Veamos que es un máximo, para ello tomamos $\ell''\left(\frac{\sum_{i=1}^{n} y_i}{n} \right)$ y veamos que es negativo.

Calculemos $\ell''(\theta)$.

$$\ell'(\theta) = \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n - \sum_{i=1}^{n} y_i}{1-\theta}$$

$$\ell''(\theta) = -\frac{\sum_{i=1}^{n} y_i}{\theta^2} - \frac{n - \sum_{i=1}^{n} y_i}{(1-\theta)^2}$$

Ahora remplazamos $\theta = \frac{\sum_{i=1}^{n} y_i}{n}$.

$$-\frac{\sum_{i=1}^{n} y_i}{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2} - \frac{n - \sum_{i=1}^{n} y_i}{\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2}$$

$$-\frac{\sum_{i=1}^{n} y_i}{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2} - \frac{n}{\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2} + \frac{\sum_{i=1}^{n} y_i}{\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2}$$

Notar que el único término positivo es el tercero, y que $n \geq \sum_{i=1}^{n} y_i$. Luego,

$$\frac{n}{\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2} \geq \frac{\sum_{i=1}^{n} y_i}{\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right)^2}$$

Entonces $\ell''\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \leq 0$. Llamamos $\hat{\theta}_{per} = \frac{\sum_{i=1}^{n} y_i}{n}$.

3. Analice sesgo, varianza, error cuadrático medio (ECM), consistencia y distribución asintótica del estimador $\hat{\theta}_{\text{per}}$.

Calculemos el sesgo de $\hat{\theta}_{per} = \frac{\sum_{i=1}^{n} y_i}{n}$. Primero calculemos la $\mathbb{E}(\hat{\theta}_{per})$.

$$\mathbb{E}(\hat{\theta}_{per}) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)$$

Por linealidad de la esperanza tenemos que:

$$\mathbb{E}(\hat{\theta}_{per}) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{\sum_{i=1}^{n} \mathbb{E}(Y_i)}{n}$$

Recordemos que Y_i es una variable aleatoria Bernuli. Entonces $\mathbb{E}(Y_i) = \mathbb{P}(Y_i = 1) = \theta$, remplazando:

$$\mathbb{E}(\hat{\theta}_{per}) = \frac{\sum_{i=1}^{n} \theta}{n} = \theta$$

Luego es insesgado. Calculemos ahora la varianza de $\hat{\theta}_{per}$.

$$Var(\hat{\theta}_{per}) = Var\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)$$

Por propiedades de la varianza.

$$Var(\hat{\theta}_{per}) = \frac{\sum_{i=1}^{n} Var(Y_i)}{n^2}$$

Usando que Y_i es una variable aleatoria Bernuli. Entonces $Var(Y_i) = \mathbb{P}(Y_i = 1)P(Y_i = 0) = \theta(1-\theta)$, remplazando:

$$Var(\hat{\theta}_{per}) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}$$

Usamos que: Sea X una v.a $ECM(X) = Sesgo(X)^2 + Var(X)$

$$ECM(\hat{\theta}_{per}) = Sesgo(\hat{\theta}_{per})^2 + Var(\hat{\theta}_{per})$$

Remplazando con los valores ya calculados:

$$ECM(\hat{\theta}_{per}) = \frac{\theta(1-\theta)}{n}$$

Encontremos W tal que $c_n(\hat{\theta} - \theta) \xrightarrow{d} W$ con c_n una sucesión de números naturales, con $cn \to \infty$. Comenzamos usando el Teorema Central del Limite, estamos bajo las hipótesis varianza finita y esperanza finita.

$$\sqrt{n}\frac{\bar{Y} - \mathbb{E}(Y)}{\sqrt{Var(Y)}} \xrightarrow{d} N(0,1)$$

Despejando la varianza por propiedades del limite en distribución.

$$\sqrt{n}(\bar{Y} - \mathbb{E}(Y)) \xrightarrow{d} N(0, Var(Y))$$

Remplazando $\hat{\theta} = \bar{Y}$, $\mathbb{E}(Y) = \theta$ y $Var(Y) = \theta(1-\theta)$ como habíamos calculado previamente.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \theta(1-\theta))$$

Luego $W \sim N(0, \theta(1-\theta))$.

4. Construya un intervalo de confianza para θ de nivel asintótico 0.95.

Sea $Z = \frac{\sqrt{n}(\hat{\theta}_{per} - \theta)}{\sqrt{\theta(1-\theta)}}$, entonces $Z \sim N(0,1)$

Luego, para calcular el intervalo de confianza de nivel 0.95 tenemos:

$$\mathbb{P}(|Z| \leq z_{0,025}) = \mathbb{P}\left(\frac{\sqrt{n}|\hat{\theta}_{per} - \theta|}{\sqrt{\theta(1-\theta)}} \leq z_{0,025}\right) = \mathbb{P}\left(-z_{0,025} \leq \frac{\sqrt{n}(\hat{\theta}_{per} - \theta)}{\sqrt{\theta(1-\theta)}} \leq z_{0,025}\right) = 0.95$$

Seguidamente, como $\hat{\theta}_{per}$ es fuertemente consistente, se obtiene:

$$= \mathbb{P}\left(-z_{0,025} \leq \frac{\sqrt{n}(\hat{\theta}_{per} - \theta)}{\sqrt{\hat{\theta}_{per}(1-\hat{\theta}_{per})}} \leq z_{0,025}\right) = \mathbb{P}\left(\hat{\theta}_{per} - z_{0,025}\frac{\sqrt{\hat{\theta}_{per}(1-\hat{\theta}_{per})}}{\sqrt{n}} \leq \theta \leq \hat{\theta}_{per} + z_{0,025}\frac{\sqrt{\hat{\theta}_{per}(1-\hat{\theta}_{per})}}{\sqrt{n}}\right) = 0.95$$

Finalmente, se consigue que el intervalo de confianza:

$$\left[\hat{\theta}_{per} - z_{0,025}\frac{\sqrt{\hat{\theta}_{per}(1-\hat{\theta}_{per})}}{\sqrt{n}} \quad , \quad \hat{\theta}_{per} + z_{0,025}\frac{\sqrt{\hat{\theta}_{per}(1-\hat{\theta}_{per})}}{\sqrt{n}}\right]$$

5. Supongamos $\theta = 0.25$. Evalúe coberturas empíricas mediante simulación Monte Carlo.
Observe qué ocurre para distintos valores de n.

Para analizar el desempeño del intervalo de confianza asintótico construido en el punto anterior, se realizaron simulaciones de Monte Carlo. Para distintos tamaños muestrales $n \in \{10, 50, 100, 200, 500\}$, se generaron 500 muestras independientes y, para cada una, se calculó el intervalo de confianza del 95 % y se verificó si contenía o no el valor verdadero de θ (siendo este 0.25). Se obtuvieron los siguientes resultados:

Número de Muestras	Porcentaje de Cobertura
10	92 %
50	94 %
100	93.4 %
200	93 %
500	93.6 %

Tabla 1: Porcentaje de Cobertura x Número de Muestras

A continuación se puede observar con mayor precisión los resultados obtenidos para $n = 10$ y $n = 200$. En los gráficos se puede visualizar tanto la variabilidad de los estimadores como la amplitud de los intervalos y la frecuencia con la que incluyen la prevalencia real. Mostramos solo los resultados de las primeras 100 visualizaciones para no sobrecargar el gráfico.

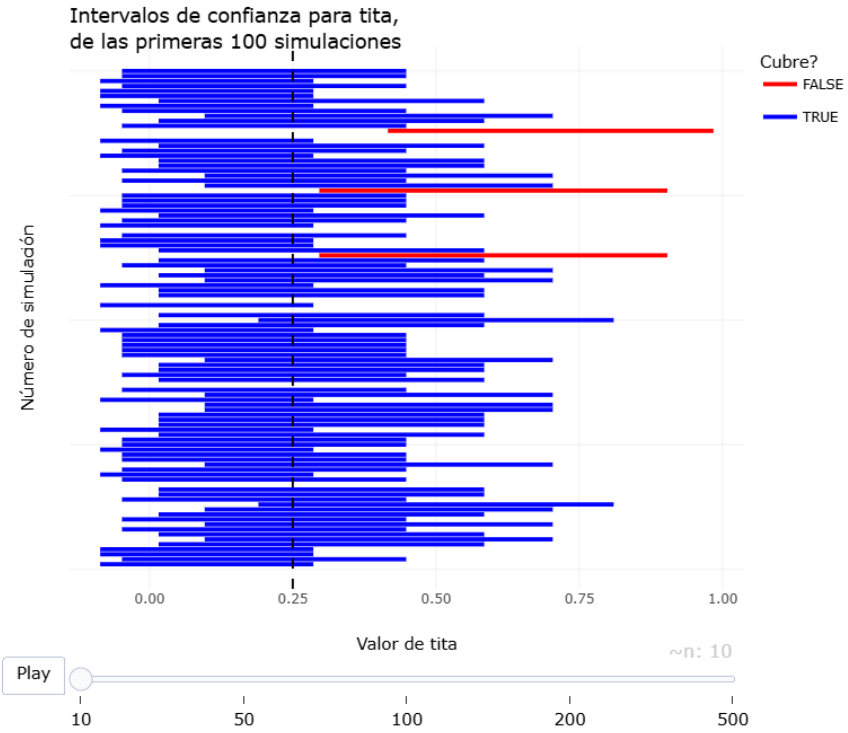

Figura 1: Intervalos de Confianza de 100 simulaciones para $n = 10$

En esta figura se representan 100 intervalos de confianza construidos para un número pequeño de muestras (n = 10).

Se observa que:

- Los intervalos presentan una pequeña variabilidad en sus longitudes, como consecuencia de usar la estimación de la varianza en su cálculo.
- Algunos intervalos (marcados en rojo) no contienen el valor verdadero $\theta = 0.25$, lo que refleja una cobertura empírica imperfecta. La proporción de intervalos que no contienen al verdadero valor es cercana al 5 %.

Figura 2: Intervalos de Confianza obtenidos con 200 simulaciones para n=200

En cambio, al considerar un número mucho más grande de muestras (n = 200):

- Los intervalos son significativamente más estrechos, reflejando la menor varianza del estimador $\hat{\theta}_{per}$ cuando aumenta n.
- La cobertura empírica se mantiene alrededor del 95 %.

A nivel general, los resultados muestran un comportamiento interesante: el intervalo asintótico mejora su precisión (se hace más angosto) cuando aumenta n, como predice la teoría. En conjunto, los resultados indican que la aproximación asintótica funciona razonablemente. La cobertura empírica puede fluctuar, pero se mantiene menor y cercana al 95 %. Por ser un intervalo asintótico, esperaríamos que su cobertura empírica se acerque más al 95 % cuando aumenta n, pero no fue lo que observamos. Quizá se necesita un n aún mayor para ver esto, o quizá el hecho de que usamos una varianza estimada para crear el intervalo disminuye su efectividad de forma que no se ve un aumento significativo en su porcentaje de cobertura.

2. Parte II: Test imperfecto con *Se* y *Sp* conocidos

Definamos la probabilidad p como

$$p = \mathbb{P}(T = 1)$$

2. Escribir a p como en función de la prevalencia θ, de las sensibilidad Se y de la especificidad Sp.

$$p = \mathbb{P}(T = 1)$$

Usando el Teorema de Probabilidad Total.

$$\mathbb{P}(T = 1) = P(T = 1|Y = 1)\,\mathbb{P}(Y = 1) + P(T = 1|Y = 0)\,\mathbb{P}(Y = 0)$$

Remplazamos con $\mathbb{P}(Y = 1) = \theta$, $\mathbb{P}(Y = 0) = 1 - \theta$, y notamos que $\mathbb{P}(T = 1|Y = 0) = 1 - \mathbb{P}(T = 0|Y = 0)$. Luego obtenemos la siguiente expresión.

$$\mathbb{P}(T = 1) = \mathbb{P}(T = 1|Y = 1)\,\theta + (1 - \mathbb{P}(T = 0|Y = 0))(1 - \theta)$$

Recordando que $Se = \mathbb{P}(T = 1|Y = 1)$ y $Sp = \mathbb{P}(T = 0|Y = 0)$. Luego

$$\mathbb{P}(T = 1) = Se\,\theta + (1 - Sp)(1 - \theta)$$

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.

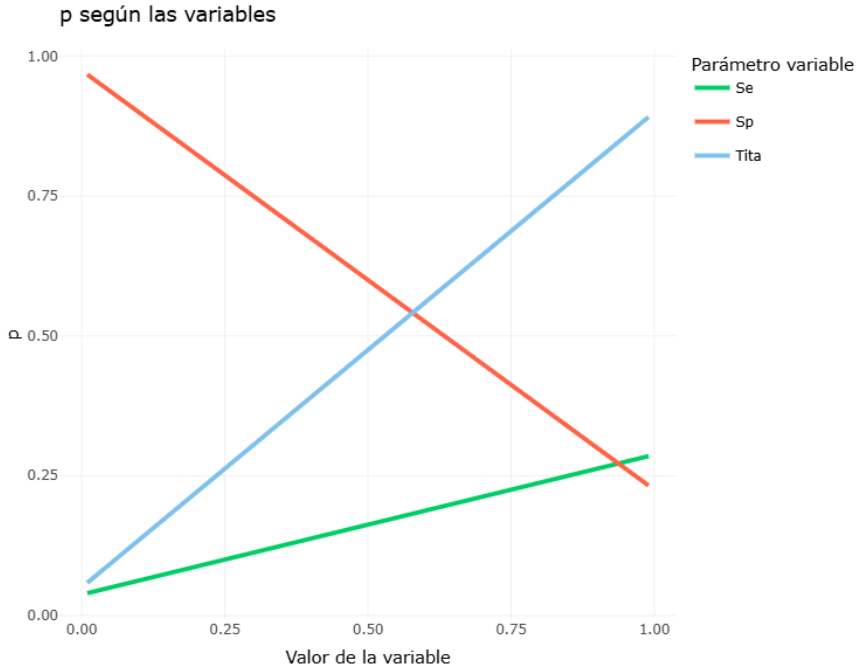

Figura 3: Valor real de p en función de distintos valores de Se, Sp y θ

En el gráfico podemos ver la relación entre p y las distitnas variables que la definen. Como era de esperarse por la relación calculada en el punto anterior, la función es lineal para cada una de las variables. En particular para Sp, están inversamente relacionados.

En la variablidad según Se, notamos que mientras más aumenta Se, más se acerca el valor de p al de θ. Cuando Se vale 1, p vale más de 0.25, lo que muestra que como Sp no es 1, hay falsos positivos que diferencian el valor de p respecto de θ.

El caso de Sp es muy similar, solo que va disminuyendo hasta quedar por debajo de 0.25 cuando $Sp = 1$ y $Se < 1$.

El caso de θ es distinto. Mientras aumenta θ, p se mantiene como una bastante buena aproximación (por Se y Sp cercanos a 1) y su valor aumenta linealmente. Cuando $\theta = 0$, p vale más que θ, y cuando $\theta = 1$, p vale menos que θ.

2.1. Estimador de momentos (MoM)

4. Teniendo en cuenta la relación entre p y θ, calcular el estimador de momentos de θ. Llamarlo $\hat{\theta}_{\text{MoM}}$.

   La información que tengo es el resultado de mi test, representado por la variable aleatoria $T \sim Be(p)$. Para calcular el estimador de momentos $\hat{\theta}_{MoM}$ a partir de p, el parámetro de T, necesito una función g que dependa de mi variable aleatoria T tal que $\mathbb{E}(g(T)) = \theta$.

   La esperanza de T es:

$$\mathbb{E}(T) = p = Se\,\theta + (1 - Sp)(1 - \theta)$$

   distribuyendo y reordenando,

$$\mathbb{E}(T) = Se\,\theta + 1 - \theta - Sp + Sp\,\theta$$
$$\mathbb{E}(T) = 1 + \theta(Se + Sp - 1) - Sp$$
$$\frac{\mathbb{E}(T) - 1 + Sp}{Se + Sp - 1} = \theta$$

   Ahora, por linealidad de la esperanza,

$$\mathbb{E}\left(\frac{T - 1 + Sp}{Se + Sp - 1}\right) = \theta$$

   y luego defino

$$g(T) = \frac{T - 1 + Sp}{Se + Sp - 1}$$

   Entonces ahora puedo estimar θ como $\overline{g(T)}$

$$\hat{\theta}_{MoM} = \overline{g(T)} = \frac{1}{n}\sum_{i=0}^{n} \frac{t_i - 1 + Sp}{Se + Sp - 1} = \frac{\bar{T} - 1 + Sp}{Se + Sp - 1}$$

5. Analice sesgo, varianza y ECM. ¿Qué se puede decir de la consistencia?

   Busquemos la $\mathbb{E}(\hat{\theta}_{MoM})$

$$\mathbb{E}(\hat{\theta}_{MoM}) = \mathbb{E}\left(\frac{\bar{T} + Sp - 1}{Se + Sp - 1}\right) = \frac{\mathbb{E}(\bar{T}) + Sp - 1}{Se + Sp - 1}$$

   Usando y remplanzando que $\mathbb{E}(\bar{T}) = \mathbb{E}(T) = Se\,\theta + (1 - Sp)(1 - \theta)$

$$\mathbb{E}(\hat{\theta}_{MoM}) = \frac{Se\,\theta + (1 - Sp)(1 - \theta) + Sp - 1}{Se + Sp + 1}$$

$$\mathbb{E}(\hat{\theta}_{MoM}) = \frac{Se\,\theta + 1 - \theta - Sp + Sp\,\theta + Sp - 1}{Se + Sp - 1}$$

$$\mathbb{E}(\hat{\theta}_{MoM}) = \frac{Se\,\theta - \theta + Sp\,\theta}{Se + Sp - 1}$$

$$\mathbb{E}(\hat{\theta}_{MoM}) = \frac{Se + Sp - 1}{Se + Sp - 1}\theta$$

$$\mathbb{E}(\hat{\theta}_{MoM}) = \theta$$

   Luego $\hat{\theta}_{MoM}$ es un estimador insesgado de θ.

   Ahora calculemos la varianza, para ello recordemos que $T \sim Be(p)$.

$$Var(\hat{\theta}_{MoM}) = Var\left(\frac{\bar{T} + Sp - 1}{Se + Sp - 1}\right)$$

Usando propiedades de la varianza llegamos a que:

$$Var(\hat{\theta}_{MoM}) = \frac{Var(\bar{T})}{(Se + Sp - 1)^2} = \frac{Var(T)}{n(Se + Sp - 1)^2}$$

$$Var(\hat{\theta}_{MoM}) = \frac{p\,(1-p)}{n\,(Se + Sp - 1)^2}$$

Ahora escribimos al ECM en función del sesgo y la varianza.

$$ECM(\hat{\theta}_{MoM}) = Sesgo(\hat{\theta}_{MoM})^2 + Var(\hat{\theta}_{MoM})$$

$$ECM(\hat{\theta}_{MoM}) = \frac{p(1-p)}{n(Se + Sp - 1)^2}$$

Como $ECM(\hat{\theta}_{MoM}) \xrightarrow[n\to\infty]{} 0$ tenemos garantizada la consistencia débil. La consistencia fuerte no es difícil de demostrar.

Por Ley Fuerte de los Grandes Números sabemos que:

$$\bar{T} \xrightarrow{cs} p$$

Por propiedades de límite casi seguro, podemos sumar $Sp - 1$ y dividir por $Se + Sp - 1$ a ambos lados, siempre que sea distinto de 0:

$$\frac{\bar{T} + Sp - 1}{Se + Sp - 1} \xrightarrow{cs} \frac{p + Sp - 1}{Se + Sp - 1}$$

el lado izquierdo es $\hat{\theta}_{MoM}$, y si remplazamos p en el lado derecho por $Se\,\theta + 1 - Sp - \theta + Sp\,\theta$, obtenemos que

$$\hat{\theta}_{MoM} \xrightarrow{cs} \frac{Se\,\theta + 1 - Sp - \theta + Sp\,\theta + Sp - 1}{Se + Sp - 1}$$

$$\hat{\theta}_{MoM} \xrightarrow{cs} \frac{Se + Sp - 1}{Se + Sp - 1}\theta$$

$$\hat{\theta}_{MoM} \xrightarrow{cs} \theta$$

6. Grafique el ECM en función de n y compárelo con el ECM del test perfecto.

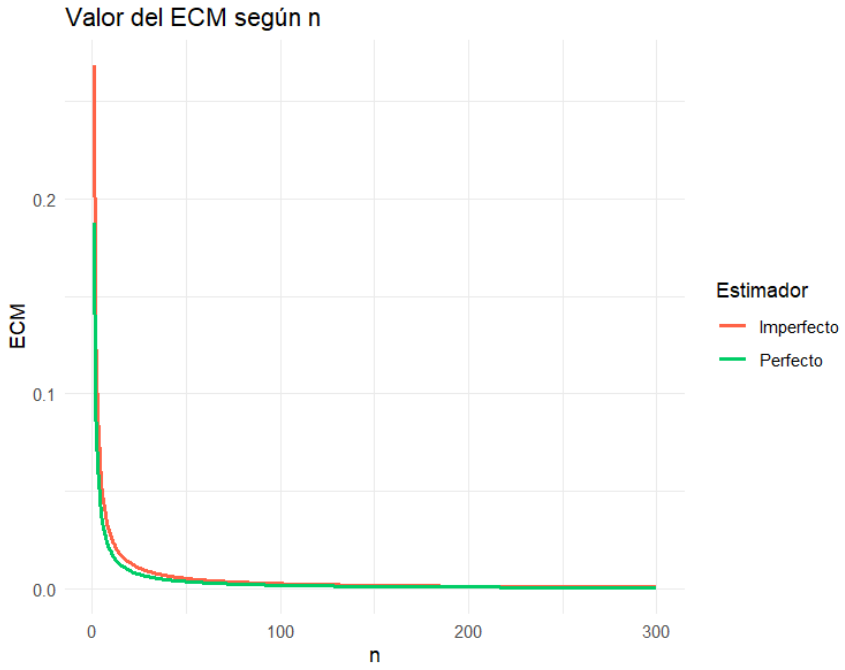

Figura 4: ECM del Test Imperfecto vs ECM del Test Perfecto

En la figura se observa el error cuadrático medio (ECM) del estimador de momentos en el caso del test imperfecto, comparado con el ECM correspondiente al test perfecto. En ambos escenarios el ECM decrece rápidamente a medida que aumenta el tamaño muestral, lo cual es consistente con la convergencia de los estimadores al verdadero valor de la prevalencia. Sin embargo, se aprecia que el ECM del test imperfecto es mayor que el del test perfecto para todos los valores de n. Esta diferencia refleja la pérdida de información introducida por la presencia de sensibilidad y especificidad menores que uno, lo que genera un aumento en la varianza del estimador ajustado. A pesar de esto, para tamaños muestrales grandes, ambas curvas se vuelven muy similares y tienden a cero, lo cual indica que el estimador de momentos sigue siendo consistente aún en presencia del error de medición.

Probando para distintos valores de Se y Sp, y viendo la fórmula de $\hat{\theta}_{MoM}$, vemos que el ECM cae más lentamente mientras menor sea la diferencia entre $Se + Sp$ y 1, y cuando $Se + Sp = 1$ el ECM es constante $\forall n$.

7. Realice simulaciones para comparar los valores teóricos hallados en el item 5 con los simulados.

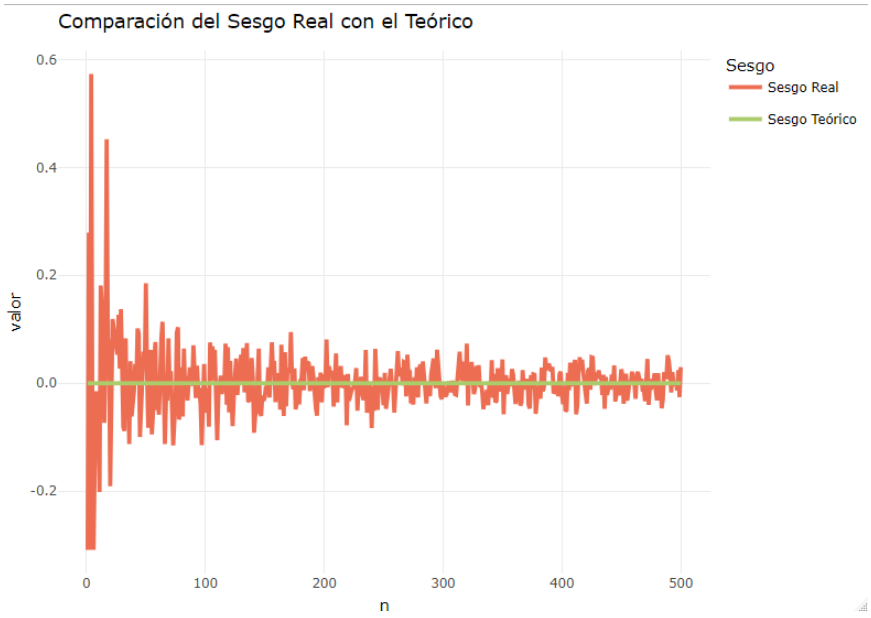

Figura 5: Comparación del sesgo real y el teórico a medida que aumenta el n del estimador $\hat{\theta}_{MoM}$

Naturalmente, el sesgo real es muy alto con pocas muestras, pero a medida que aumenta el valor de n, podemos ver que el sesgo disminuye y tiende a 0, como predice la teoría.

Figura 6: Comparación del varianza real y el teórica a medida que aumenta el n del estimador $\hat{\theta}_{MoM}$

Notemos que la varianza real es mayor que la teorica, lo que tiene sentido porque la varianza real depende de $\bar{T}$. Sin embargo, converge rápidamente a medida que crece el n. Esto es esperable dado la consistencia fuerte del estimador $\hat{\theta}_{MoM}$.

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?

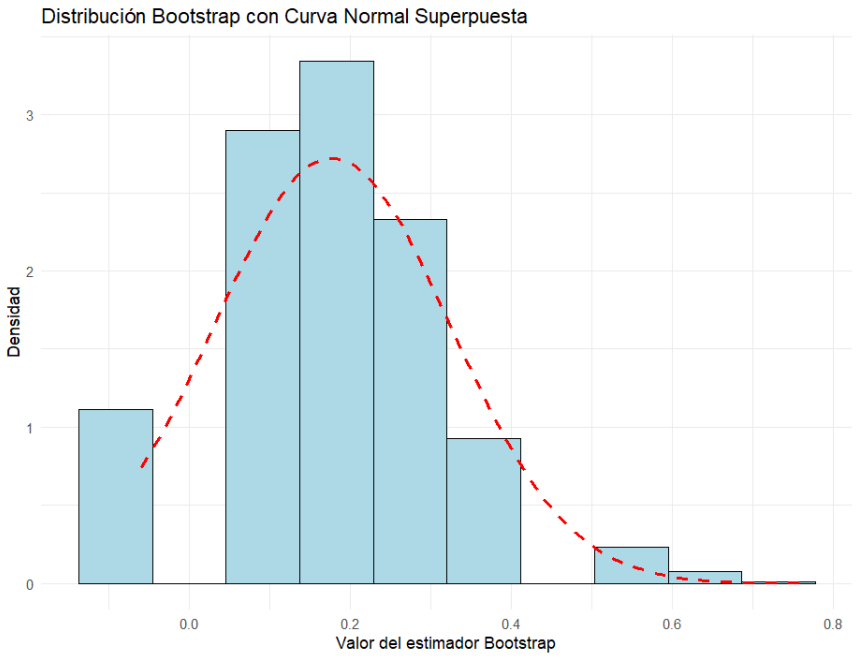

Figura 7: Histograma de muestras bootstrap para $\hat{\theta}_{MoM}$, con una densidad de una normal superpuesta

En el histograma se puede ver que sigue una distribución aproximadamente normal. Esto tiene sentido, ya que $\hat{\theta}_{MoM} = \frac{\bar{T}-1+Sp}{Se+Sp-1}$. Notemos que $\bar{T}$ ya se distribuye aproximadamente como una normal por ser un promedio, entonces sumarle una constante y dividir por otras constantes no modifican demasiado su distribución. Los espacios vacíos en el histograma se deben al bajo valor de n. Con $n = 10$ el promedio solo puede tomar 11 valores distintos.

2.2. Intervalos de confianza

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

   Primero usamos una muestra de tamaño n para estimar $\hat{\theta}$. Luego, usando $\hat{\theta}$ podemos hacer muestras bootstraps. El intervalo de confianza bootstrap percentil nivel $1-\alpha$ se calcula tomando como límites el cuantil $\alpha/2$ y $1-\alpha/2$ de entre los resultados de los estimadores calculados a partir de las muestras bootstrap.

   En el ejercicio 11 analizamos los resultados.

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

    Para construir un intervalo de confianza nivel asintótico 0.95, aprovechamos el hecho de que la distribución de los estimadores bootstrap es aproximadamente normal. Luego, tomamos el intervalo $[\hat{\theta} - z_{\alpha/2}\, \hat{se}_{boot}\, ,\, \hat{\theta} + z_{\alpha/2}\, \hat{se}_{boot}]$, siendo $\hat{se}_{boot}$ el estimador del desvío estándar calculado a partir de los valores del estimador obtenidos de las muestras bootstrap.

    En el ejercicio 11 analizamos los resultados.

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

    Para cada n, hicimos 500 simulaciones. En cada simulación creamos una "muestra real"de tamaño n, de la cual estimamos θ. Luego usamos esa $\hat{\theta}_{MoM}$ para hacer 1000 muestras bootstrap, con las cuales generamos un intervalo de confianza percentil y uno normal asintótico. Obtuvimos los siguientes resultados:

n	Cobertura Bootstrap (%)	Cobertura $\hat{\theta}_{\text{MoM}}$ (%)	Longitud Promedio Percentil	Longitud Promedio Asinstótico
10	93 %	93 %	0.568	0.586
50	90.8 %	90.8 %	0.280	0.283
100	94.8 %	95 %	0.200	0.201
200	93.8 %	93.2 %	0.142	0.143
500	95.8 %	95.4 %	0.090	0.091

Tabla 2: Comparación Bootstrap vs $\hat{\theta}_{\text{MoM}}$ x Número de Muestras

En la Tabla 2 se comparan las coberturas y longitudes promedio de los intervalos de confianza obtenidos mediante el método bootstrap percentil y mediante el intervalo asintótico basado en $\hat{\theta}_{\text{MoM}}$, para distintos tamaños muestrales n.

En primer lugar, ambos métodos presentan coberturas cercanas al 95 %. Aun así, la cobertura bootstrap resulta levemente superior a la cobertura del intervalo asintótico para los tamaños muestrales más elevados, aunque siempre la diferencia entre coberturas es mínima.

Respecto a la longitud de los intervalos, se observa que ambos métodos muestran la disminución esperada al aumentar n, lo que refleja la convergencia del estimador. Para tamaños muestrales grandes (por ejemplo, $n = 200$ y $n = 500$), las longitudes del intervalo bootstrap y del intervalo asintótico prácticamente coinciden, indicando que en el régimen asintótico ambos métodos se comportan de forma similar. Sin embargo, para muestras pequeñas (particularmente $n = 10$), el intervalo bootstrap resulta ligeramente más corto que el intervalo asintótico, aunque conservando una cobertura comparable. En general, los datos sugieren que el intervalo asintótico funciona ligeramente mejor.

A continuación se puede observar con mayor precisión los resultados obtenidos para $n = 10$ y $n = 200$ para ambos métodos. En los gráficos se puede visualizar tanto la variabilidad de los estimadores como la amplitud de los intervalos y la frecuencia con la que incluyen la prevalencia real. Solo mostramos los resultados de 100 simulaciones. Nótese que para cada simulación, la $\hat{\theta}_{MoM}$ en la que se basaron las muestras bootstrap es la misma para ambos métodos. Esto se evidencia en que los intervalos que no cubren el valor real se dan en las mismas simulaciones, porque las muestras bootstrap se basaron en el mismo $\hat{\theta}_{MoM}$ poco representativo.

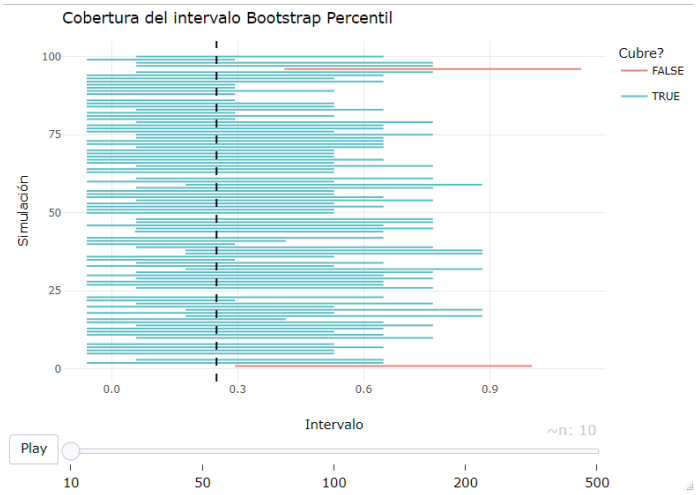

Figura 8: Intervalos de Confianza Bootstrap Percentil con 500 simulaciones para $n = 10$

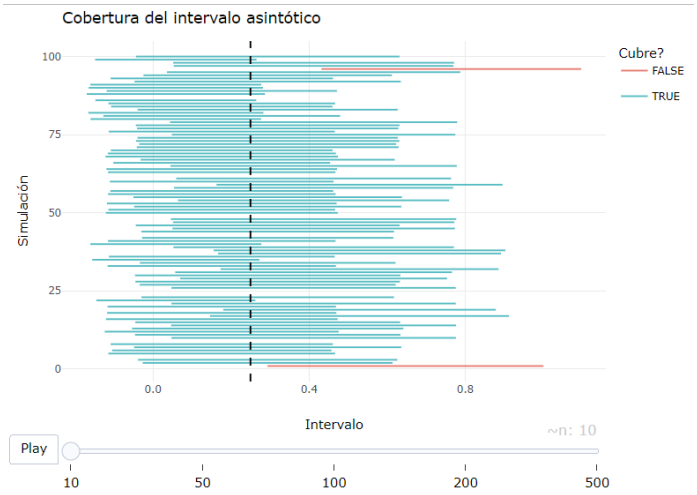

Figura 9: Intervalos de Confianza Bootstrap Asintótico con 500 simulaciones para $n = 10$

En ambos gráficos se ve que, para $n = 10$, los intervalos presentan una dispersión muy grande, lo que produce coberturas algo inestables.

En el bootstrap percentil, la mayoría de los intervalos cubren el valor verdadero (línea punteada), pero aparece un pequeño número de intervalos rojos (no cobertura), generalmente asociados una estimación de θ inicial muy mala.

En el bootstrap asintótico, la cobertura visualmente es muy similar, pero los intervalos tienden a ser más centrados y ligeramente más amplios, lo que explica que la cobertura también sea parecida.

En conjunto, las figuras muestran que ambos métodos funcionan razonablemente bien incluso con $n = 10$, pero también evidencian que, con muestras pequeñas, la variabilidad de los intervalos es grande y aparecen fallas de cobertura en ambos enfoques, aunque sin un ganador claro.

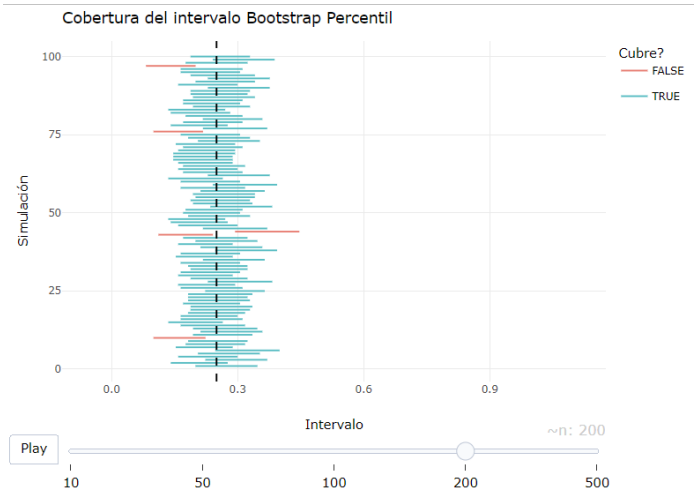

Figura 10: Intervalos de Confianza Bootstrap Percentil con 500 simulaciones para $n = 200$

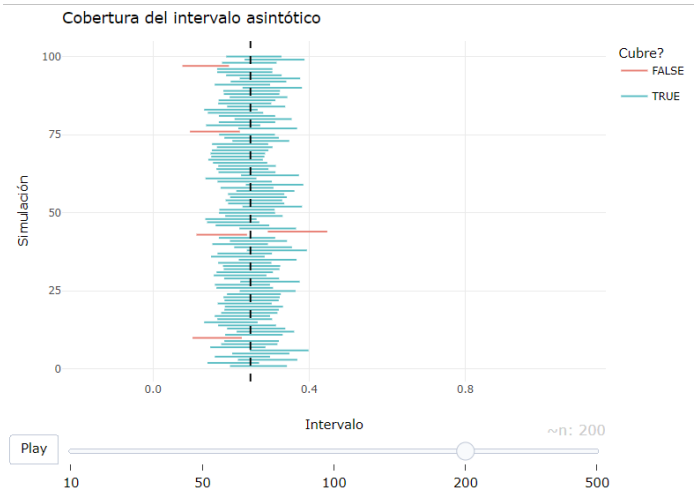

Figura 11: Intervalos de Confianza Bootstrap Asintótico con 500 simulaciones para $n = 200$

En ambos gráficos se observa que casi todos los intervalos cubren el valor real de $\theta = 0.25$, marcado con la línea vertical punteada. Solo aparecen unos pocos casos en rojo (intervalos que no cubren), y están distribuidos de manera aislada.

En resumen, ambos métodos funcionan bien para valores grandes de n, pero el bootstrap percentil parece mostrar una cobertura ligeramente más estable con menos intervalos que fallan.

2.3. Estimador truncado

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

Debido a que el estimador de momentos realiza una corrección basada en la sensibilidad y especificidad del test, la transformación aplicada a $\bar{T}$ no garantiza que el resultado permanezca dentro del intervalo [0,1]. En particular, cuando $Se + Sp$ - 1 es pequeño o la muestra presenta valores extremos de $\bar{T}$, el estimador puede tomar valores negativos o mayores que uno.

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

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

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

Para cada valor de n, hicimos 100 simulaciones. En cada simulación creamos 100 muestras de tamaño n, y para cada muestra calculamos $\hat{\theta}_{trunc}$. Luego, para cada simulación tenemos el sesgo, la varianza, y el ECM del estimador, calculados a partir de esas 100 muestras.

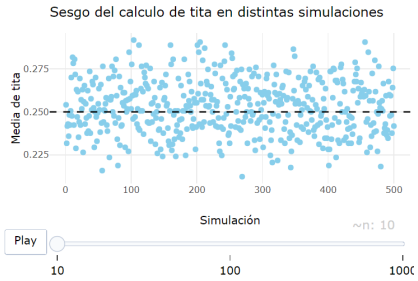

Figura 12: Sesgo calculado en cada una de las 100 simulaciones con $n = 10$

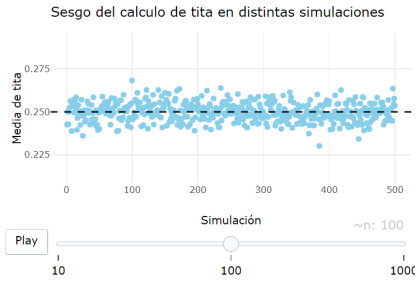

Figura 13: Sesgo calculado en cada una de las 100 simulaciones con $n = 100$

En los gráficos, los puntos representan la media de los $\hat{\theta}_{trunc}$ calculados en cada simulación, es decir, la esperanza empírica, y la línea punteada el valor real de θ. Con $n = 10$ hay un sesgo menor a 0.05 $(0{,}3 - 0{,}25)$, y rápidamente este número disminuye, como se puede ver en el gráfico de $n = 100$. Esto sugiere que el estimador es insesgado.

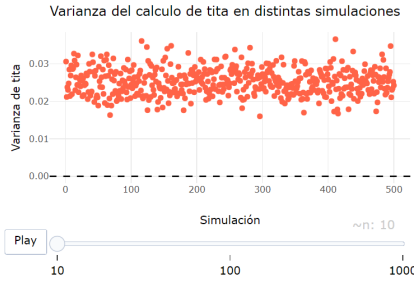

Figura 14: Varianza calculada en cada una de las 100 simulaciones con $n = 10$

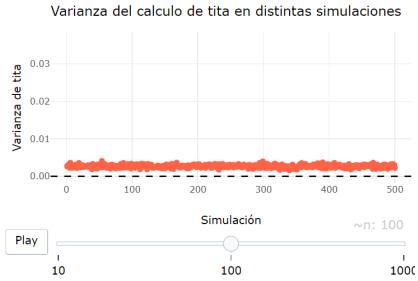

Figura 15: Varianza calculada en cada una de las 100 simulaciones con $n = 100$

Como se puede ver, la varianza también es baja, y disminuye significativamente al aumentar n.

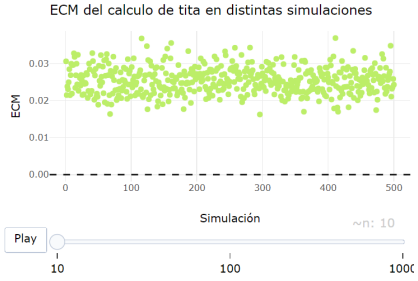

Figura 16: ECM calculado en cada una de las 100 simulaciones con $n = 10$

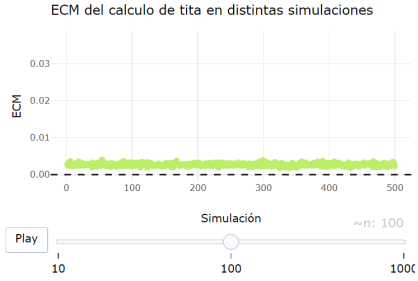

Figura 17: ECM calculado en cada una de las 100 simulaciones con $n = 100$

Luego, el ECM tiende a 0. Se parece mucho a la varianza, porque el sesgo, de a lo sumo 0.05, aparece en la formula al cuadrado, por lo que es a lo sumo 0.0025, y comparado a una varianza de a lo sumo 0.03, tiene poco impacto.

Por último, analicemos la distribución asintótica de nuestra estimación, es decir, dado $c_n \to \infty$, decimos que $c_n\,(\hat{\theta}_{trunc} - \theta) \xrightarrow{d} W$, y analizamos la distribución de W. Sospechamos que se distribuye como una normal ya que $\hat{\theta}_{MoM}$ se distribuía

aproximadamente como una normal. Por su relación con la normal, y por ser el factor más utilizado, decidimos usar $c_n = \sqrt{n}$. Luego, analizamos el límite en distribución de $\sqrt{n}\,(\hat{\theta}_{trunc} - \theta)$. Planteamos la distribución asintotica para un n suficientemente grande ($n = 1000$), y realizamos un qqplot para comparar la distribución asintótica contra la distribución Normal(0,1).

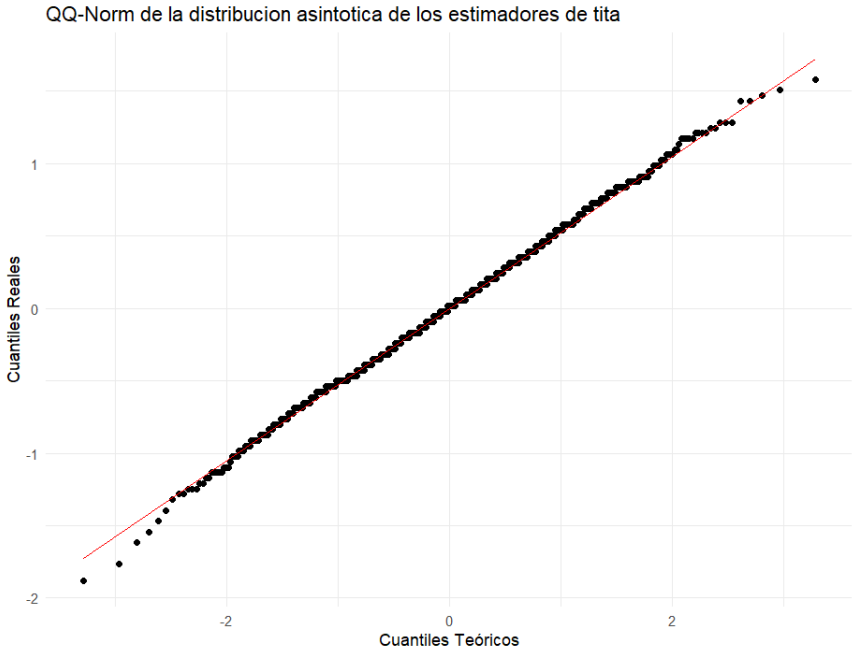

Figura 18: QQ-plot de la distribución asintótica de nuestro estimador vs $N(0,1)$

Podemos asumir que sigue teniendo una distribución normal, ya que los puntos se acercan consistentemente a los cuantiles teóricos de la normal con media 0 y desvio 1.

Graficamos $\sqrt{n}\,(\hat{\theta}_{trunc} - \theta)$, y superpusimos una normal con la media y varianza empíricas de la expresión, es decir, una normal $N(0{,}0054, 0{,}5329501)$. Adjuntamos un gráfico para ver la bondad de este ajuste.

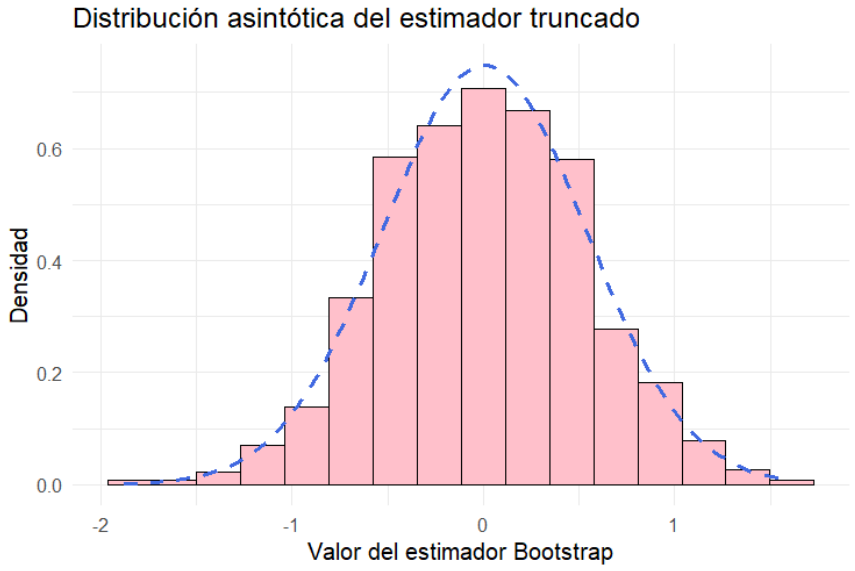

Figura 19: Ajuste de la Normal propuesta al histograma de frecuencias de la distribución asintótica.

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}.

Supongamos 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,

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

con $p_A = (Se + Sp\ \text{-}1)\ \theta_A + (1\ \text{-}\ Sp)$ 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 $\hat{\theta}_{\text{MoM}}$ (el que no está truncado), plantee un test de nivel aproximado 0.05 para las hipótesis:

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

Empezamos viendo Δ

$$\Delta = \theta_{post} - \theta_{prev}$$

Usando la relación entre p y θ tenemos que:

$$\Delta = \frac{p_{post} + Sp - 1}{Se + Sp - 1} - \frac{p_{prev} + Sp - 1}{Se + Sp - 1}$$

$$\Delta = \frac{p_{post} - p_{prev}}{Se + Sp - 1}$$

Entonces, definimos la nueva variable aleatoria $X_i = \hat{\theta}_{\text{i,post}} - \hat{\theta}_{\text{i,prev}} \Rightarrow X_i = \frac{T_{\text{i,post}} - T_{\text{i,prev}}}{Se+Sp-1}$

Este análisis requiere entonces que para todo $T_{i,prev}$ exista un $T_{i,post}$. Por lo tanto, agregamos el supuesto de que $n = n_{prev} = n_{post}$.

Vemos que $\mathbb{E}(X_i) = \frac{p_{post} - p_{prev}}{Se+Sp-1} = \Delta$ y $\mathbb{V}ar(X_i) = \frac{\mathbb{V}ar(T_{\text{i,post}}) + \mathbb{V}ar(T_{\text{i,prev}})}{(Se+Sp-1)^2}$

Ahora notemos que

$$\bar{X} = \frac{\overline{T_{\text{post}} - T_{\text{prev}}}}{Se + Sp - 1} = \frac{\overline{T_{\text{post}}} - \overline{T_{\text{prev}}}}{Se + Sp - 1}$$

Luego, con todo esto podemos usar TCL

$$\sqrt{n}\frac{\overline{X} - \Delta}{\mathbb{V}ar(X)} \xrightarrow{d} N(0,1)$$

Finalmente, planteamos el siguiente test de nivel asintótico $1 - \alpha$

$$\phi(x) = \begin{cases} 1 & \text{si } \sqrt{n}\frac{|\overline{X}|}{\mathbb{V}ar(X)} \geq z_{\frac{\alpha}{2}}, \\ 0 & \text{si } \sqrt{n}\frac{|\overline{X}|}{\mathbb{V}ar(X)} < z_{\frac{\alpha}{2}}, \end{cases}$$

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

   Aplicamos la prueba basada en $\hat{\theta}_{\text{MoM}}$ con $Se = 0.9$, $Sp = 0.95$, $\theta_{\text{prev}} = 0.2$ y $\theta_{\text{post}} = 0.15$ y $\alpha = 0.05$, y calculamos los valores del estadístico t para los distintos tamaños muestrales, y debido a que su valor varía, a veces una muestra puntual lleva a rechazar H_0 y otras no. Más específicamente, con un $n = 100$ sí se tiene suficiente evidencia estadística para rechazar H_0 con nivel 0.05, mientras que al achicar los tamaños muestrales, ya no se tiene suficiente evidencia como para rechazar a H_0.

   A continuación se pueden observar los estadísticos observados:

n	t_{obs}	Rechaza
10	0.49	No
20	1.13	No
50	0	No
70	0.78	No
100	3.08	Sí

Tabla 3: Estadísticos Observados

   Por otro lado, también es importante recalcar que es posible que el estadístico sea igual a 0, ya que esto ocurre siempre que $\hat{\theta}_{\text{post}} - \hat{\theta}_{\text{prev}} = 0$. En la práctica esto sucede cuando las proporciones observadas de tests positivos en las dos muestras son exactamente iguales (es decir, los conteos positivos en T_{prev} y T_{post} coinciden), de modo que la diferencia corregida por Se y Sp se anula. Además, un estadístico observado igual a cero simplemente indica que no hay evidencia empírica a favor de $\Delta \neq 0$, pero no prueba que $\Delta = 0$ en la población.

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

   Como ya tenemos un test bilateral de nivel asintótico de nivel 0.05 para Δ, su región de aceptación nos permite definir un intervalo de confianza de nivel asintótico 0.95.

$$\mathbb{P}\left(\sqrt{n}\frac{|\bar{X} - \Delta|}{Var(X)} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\mathbb{P}\left(-z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X} - \Delta}{Var(X)} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\mathbb{P}\left(-z_{\frac{\alpha}{2}}\frac{\sqrt{Var(X)}}{\sqrt{n}} \leq \bar{X} - \Delta \leq z_{\frac{\alpha}{2}}\frac{\sqrt{Var(X)}}{\sqrt{n}}\right) = 1 - \alpha$$

$$\mathbb{P}\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sqrt{Var(X)}}{\sqrt{n}} \leq \Delta \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sqrt{Var(X)}}{\sqrt{n}}\right) = 1 - \alpha$$

   Además estudiamos la cobertura de los intervalos con distintos valores de n. Observamos que a partir de las 10 observaciones ya alcanzó niveles de cobertura superiores al 90 % y rapidamente se estabiliza en el nivel deseado, y que un aumento en n representa un aumento en la precisión.

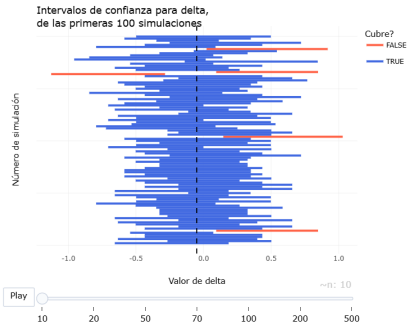

Figura 20: Cobertura del intervalo en las primeras 100 simulaciones con $n = 10$

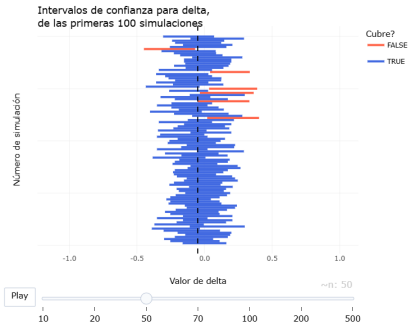

Figura 21: Cobertura del intervalo en las primeras 100 simulaciones con $n = 50$

n	Cobertura empírica
10	93.8 %
20	96.8 %
50	95.6 %
70	94.2 %
100	94 %
200	94.2 %
500	94.8 %

Tabla 4: Coberturas empíricas

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 muesras pequeños y grandes.

   En el gráfico podemos ver como a medida de que aumenta el n, el nivel se acerca rápidamente a 0.05, lo que coincide con lo esperado en un test de nivel asintótico. Aun así, la ganancia en nivel no justifica aumentar el nivel de n a valores tan grandes porque entendemos que en un entorno real esto puede ser muy caro y a veces imposible.

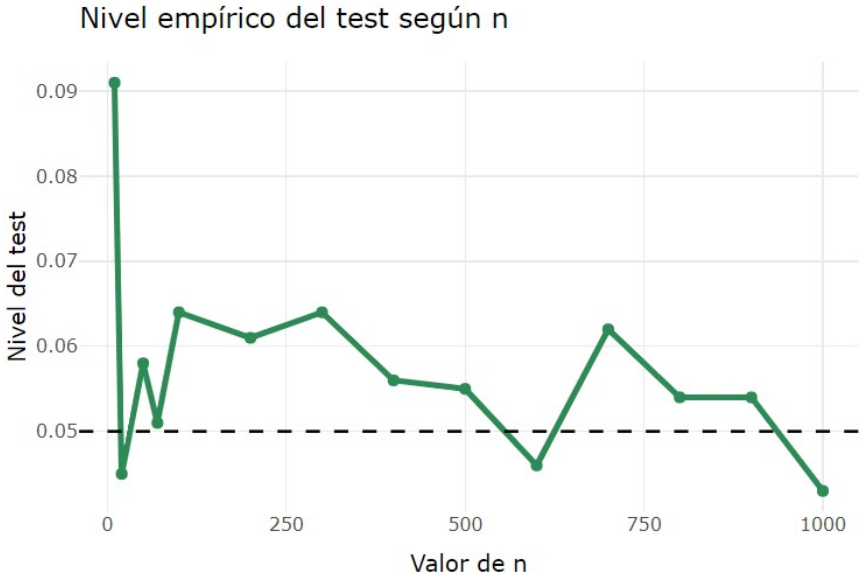

Figura 22: Nivel empírico del test según n