

# Trabajo Práctico – IECD

Grupo 11

2025-12-08

# Introducción

## Objetivo del trabajo

Análisis de propiedades estadísticas de tests diagnósticos:

1. **Test perfecto** (baseline)
2. **Test imperfecto** con  $S_e$  y  $S_p$  conocidos
3. **Dos muestras** (pre-post intervención)

# **Parte I**

Test perfecto (baseline)

## 1.1 Distribución de $T_{per}$

### Motivación

- ▶ Estudiamos primero un caso ideal: el test diagnóstico **no comete errores**.
- ▶ Esto equivale a suponer:

$$Se = 1, \quad Sp = 1.$$

- ▶ En este escenario, el resultado del test coincide exactamente con el estado verdadero  $Y$ .
- ▶ Objetivo del punto 1.1:

Mostrar que  $T_{per} \sim \text{Binomial}(n, \theta)$ .

## Modelo del estado verdadero

- ▶ Cada individuo  $i$  puede estar enfermo ( $Y_i = 1$ ) o sano ( $Y_i = 0$ ).
- ▶ Modelamos:

$$Y_i \sim \text{Bernoulli}(\theta), \quad i = 1, \dots, n,$$

donde  $\theta = P(Y = 1)$  es la prevalencia verdadera.

- ▶ Como  $Se = Sp = 1$ :
  - ▶ lo que observamos en el test coincide exactamente con  $Y_i$ ;
  - ▶ no hay error de medición.

## Número de individuos enfermos en la muestra

- Definimos:

$$T_{per} = \sum_{i=1}^n Y_i.$$

- Interpretación:

- $T_{per}$  cuenta la cantidad de personas enfermas en la muestra.
- Es la suma de  $n$  variables Bernoulli independientes con el mismo parámetro  $\theta$ .

## Resultado

- ▶ La suma de  $n$  variables Bernoulli independientes con parámetro  $\theta$  sigue una binomial:

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

- ▶ Para  $k = 0, 1, \dots, n$ :

$$P(T_{per} = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}.$$

De esta forma,  $T_{per} \sim \text{Binomial}(n, \theta)$

## 1.2 Estimador de Máxima Verosimilitud

### Planteo

- ▶ Seguimos en el caso de **test perfecto**:

$$Y_i \sim \text{Bernoulli}(\theta), \quad i = 1, \dots, n,$$

independientes e idénticamente distribuidas.

- ▶ Definimos:

$$T_{per} = \sum_{i=1}^n Y_i.$$

- ▶ Nuestro objetivo en este punto es encontrar el **estimador de máxima verosimilitud (EMV)** de  $\theta$ .
- ▶ Notación útil:
  - ▶  $N_1 = \sum_{i=1}^n Y_i = T_{per}$ : número de enfermos en la muestra.
  - ▶  $N_0 = n - N_1$ : número de sanos en la muestra.

## Función de verosimilitud

- Dada la muestra  $Y_1, \dots, Y_n$ , la función de verosimilitud es:

$$L(\theta) = \prod_{i=1}^n P_\theta(Y_i) = \prod_{i=1}^n \theta^{Y_i} (1-\theta)^{1-Y_i}.$$

- Reagrupando términos:

$$L(\theta) = \theta^{\sum_{i=1}^n Y_i} (1-\theta)^{\sum_{i=1}^n (1-Y_i)} = \theta^{N_1} (1-\theta)^{N_0}.$$

- Trabajamos con la **log-verosimilitud**:

$$\ell(\theta) = \log L(\theta) = N_1 \log(\theta) + N_0 \log(1-\theta).$$

- ▶ Derivamos la log-verosimilitud respecto de  $\theta$ :

$$\ell'(\theta) = \frac{d}{d\theta} [N_1 \log(\theta) + N_0 \log(1 - \theta)] = \frac{N_1}{\theta} - \frac{N_0}{1 - \theta}.$$

- ▶ Buscamos  $\hat{\theta}$  tal que  $\ell'(\hat{\theta}) = 0$ :

$$\frac{N_1}{\hat{\theta}} - \frac{N_0}{1 - \hat{\theta}} = 0 \implies \frac{N_1}{\hat{\theta}} = \frac{N_0}{1 - \hat{\theta}}.$$

- ▶ Despejando:

$$N_1(1 - \hat{\theta}) = N_0\hat{\theta} \implies N_1 = (N_0 + N_1)\hat{\theta} = n\hat{\theta}.$$

- ▶ Por lo tanto:

$$\hat{\theta}_{per} = \frac{N_1}{n}.$$

## Verificación de máximo e interpretación

- ▶ Segunda derivada:

$$\ell''(\theta) = -\frac{N_1}{\theta^2} - \frac{N_0}{(1-\theta)^2} < 0 \quad \text{para todo } \theta \in (0, 1).$$

- ▶ Luego, el punto crítico encontrado corresponde a un **máximo global** de la verosimilitud.
- ▶ Como  $N_1 = T_{per}$ , podemos escribir:

$$\hat{\theta}_{per} = \frac{N_1}{n} = \frac{T_{per}}{n}.$$

## Interpretación

- ▶ El EMV de  $\theta$  es simplemente la **proporción muestral de individuos enfermos**.
- ▶ En el caso de test perfecto, el estimador natural de la prevalencia es la cantidad de enfermos observados / tamaño de la muestra.

## 1.3 Sesgo, Varianza, Error cuadrático medio, consistencia y distribución asintótica

### ► Sesgo:

El estimador  $\hat{\theta}_{per} = \frac{T_{per}}{n}$  es insesgado porque

$$\mathbb{E}[\hat{\theta}_{per}] = \mathbb{E}\left[\frac{T_{per}}{n}\right] = \frac{1}{n}\mathbb{E}[T_{per}] = \frac{n\theta}{n} = \theta.$$

Usamos que  $T_{per} \sim \text{Binomial}(n, \theta)$ , por lo que  $\mathbb{E}[T_{per}] = n\theta$ .

► **Varianza:**

$$\text{Var}(\hat{\theta}_{per}) = \text{Var}\left(\frac{T_{per}}{n}\right) = \frac{1}{n^2} \text{Var}(T_{per}) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}.$$

También utilizamos que  $T_{per}$  es binomial.

► **Error cuadrático medio (ECM):**

Como el sesgo es cero,

$$\text{ECM}(\hat{\theta}_{per}) = \text{Var}(\hat{\theta}_{per}) + \mathbb{E}^2(\hat{\theta}_{per}) = \frac{\theta(1-\theta)}{n}.$$

► **Consistencia:**

Observamos que  $\hat{\theta}_{per} = \frac{T_{per}}{n} = \frac{1}{n} \sum_{i=1}^n Y_i$ ,  
donde  $\mathbb{E}[Y_i] = \theta$  y  $\text{Var}(Y_i) = \theta(1 - \theta)$ .

Por la Ley Fuerte de los Grandes Números,

$$\hat{\theta}_{per} \xrightarrow{c.s.} \theta.$$

Por lo tanto,  $\hat{\theta}_{per}$  es **fuertemente consistente**.

► **Distribución asintótica:**

Por el Teorema Central del Límite,

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

## 1.4 Intervalo de confianza para $\theta$

- ▶ Sabemos que por TCL:

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

- ▶ Además, por ley débil de los grandes números:

$$\hat{\theta}_{per} \xrightarrow{P} \theta$$

- ▶ Entonces, combinando los resultados anteriores con Slutsky:

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

Como  $P(Z \geq 1.96) = 0.025$ , vale lo siguiente:

$$P\left(-1.96 \leq \sqrt{n} \frac{\hat{\theta}_{per} - \theta}{\sqrt{\hat{\theta}_{per}(1 - \hat{\theta}_{per})}} \leq 1.96\right) \xrightarrow{n \rightarrow \infty} 0.95.$$

- ▶ Finalmente, se deduce el siguiente intervalo de nivel asintótico 0.95:

$$IC_{0.95}(\theta) = \hat{\theta}_{per} \pm 1.96 \sqrt{\frac{\hat{\theta}_{per}(1 - \hat{\theta}_{per})}{n}}$$

## 1.5 Cubrimiento empírico

- ▶ Para evaluar el cubrimiento empírico del intervalo dado en 1.4, se llevó a cabo una simulación Monte Carlo para los siguientes valores de  $n$ :

10, 20, 50, 100, 1000, 10000

- ▶ Y se obtuvieron estos resultados:

Table 1: Cubrimiento empírico del intervalo

n	cubrimiento
10	0.9232
20	0.8992
50	0.9362
100	0.9462
1000	0.9472
10000	0.9486

- ▶ Los resultados obtenidos son coherentes con la teoría. Al tratarse de un intervalo asintótico, el cubrimiento empírico converge hacia el nivel nominal del 0.95 conforme aumenta el tamaño muestral  $n$ .

# **Parte II**

Test imperfecto con  $S_e$  y  $S_p$  conocidos

## 2.0.1 Estimación de $p$ con $T_{per}$

### Nuevo escenario

- ▶ Ahora el test **no es perfecto**.
- ▶ Características del test:

$$Se = P(T = 1 \mid Y = 1), \quad Sp = P(T = 0 \mid Y = 0).$$

- ▶ El resultado observado del test se llama  $T$  (1 = positivo, 0 = negativo).
- ▶ En este punto definimos:

$$p = P(T = 1)$$

es decir, la **probabilidad de que el test dé positivo**.

## Datos observados

- ▶ Tomamos una muestra de tamaño  $n$ .
- ▶ Para cada individuo observamos el resultado del test:

$$T_1, T_2, \dots, T_n \sim \text{i.i.d. Bernoulli}(p).$$

- ▶ Definimos:

$$T_{per} = \sum_{i=1}^n T_i,$$

que ahora representa **la cantidad de tests positivos**, no la cantidad de enfermos reales.

- ▶ Entonces:

$$T_{per} \sim \text{Binomial}(n, p).$$

## Construcción del estimador

- ▶ La esperanza de la variable binomial es:

$$E[T_{per}] = np.$$

- ▶ El estimador natural es la proporción de tests positivos observados:

$$\hat{p} = \frac{T_{per}}{n}.$$

## Propiedades inmediatas

- ▶ Es el **estimador de máxima verosimilitud** para  $p$ .
- ▶ Es el **estimador de momentos** para  $p$ .
- ▶ Es simplemente:

$\hat{p}$  = proporción muestral de tests positivos.

## Interpretación

- ▶ En el caso imperfecto,  $\hat{p}$  NO es un estimador directo de la prevalencia.
- ▶  $\hat{p}$  mide qué fracción de la muestra obtuvo un resultado positivo en el test
- ▶ Más adelante usaremos la relación entre  $p$ ,  $\theta$ ,  $Se$  y  $Sp$  para estimar  $\theta$ .
- ▶ Esta relación clave será:

$$p = P(T = 1) = Se \cdot \theta + (1 - Sp)(1 - \theta).$$

(Item 2.0.2)

- ▶ Por ahora, lo único que necesitamos es:

$$\hat{p} = \frac{T_{per}}{n}.$$

## 2.0.2 $p$ como función de $\theta$ , $S_e$ y $S_p$ .

### Objetivo

- ▶ Queremos expresar:

$$p = P(T = 1)$$

en función de:

- ▶ la prevalencia verdadera  $\theta = P(Y = 1)$ ,
- ▶ la sensibilidad  $S_e$ ,
- ▶ la especificidad  $S_p$ .
- ▶ Esta relación permitirá, más adelante, **estimar**  $\theta$  usando el test imperfecto.

## Paso 1: Probabilidad total

$$p = P(T = 1)$$

Descomponemos según el estado verdadero  $Y$ :

$$P(T = 1) = P(T = 1 \mid Y = 1)P(Y = 1) + P(T = 1 \mid Y = 0)P(Y = 0).$$

Reemplazamos:

- ▶  $P(T = 1 \mid Y = 1) = Se,$
- ▶  $P(Y = 1) = \theta,$
- ▶  $P(Y = 0) = 1 - \theta,$
- ▶  $P(T = 1 \mid Y = 0) = 1 - Sp.$

Obtenemos:

$$p = Se \cdot \theta + (1 - Sp)(1 - \theta).$$

## Forma final

Podemos escribir:

$$p = (Se + Sp - 1) \theta + (1 - Sp).$$

## Observaciones

- ▶ La expresión es **lineal en  $\theta$** .
- ▶ El coeficiente  $(Se + Sp - 1)$  determina la “calidad” del test.
  - ▶ Si  $Se + Sp - 1 = 0$ , el test **no aporta información** ( $p = 1 - Sp$ ).
  - ▶ Si  $Se + Sp - 1 > 0$ , el test permite recuperar parte de la prevalencia.
- ▶ El término constante  $(1 - Sp)$  es la **probabilidad de falso positivo** cuando  $\theta = 0\$$ .

¿Qué nos dice esta fórmula?

- ▶ El test puede dar positivo por dos razones:
  1. **Verdadero positivo:** la persona está enferma y el test la detecta.
$$Se \cdot \theta$$
  2. **Falso positivo:** la persona está sana pero el test falla.

$$(1 - Sp)(1 - \theta)$$

- ▶ Por eso, la probabilidad total de test positivo es:

$$p = \text{Verdaderos positivos} + \text{Falsos positivos.}$$

## 2.0.3 Comportamiento de $p = P(T = 1)$

### Recordatorio

- ▶ Tenemos:

$$p = P(T = 1) = Se \cdot \theta + (1 - Sp)(1 - \theta).$$

- ▶ También:

$$p(\theta, Se, Sp) = (Se + Sp - 1)\theta + (1 - Sp).$$

- ▶ En este punto queremos ver cómo cambia  $p$  en función de:
  1.  $\theta$  (prevalencia),
  2.  $Se$  (sensibilidad),
  3.  $Sp$  (especificidad),

fijando los otros dos parámetros.

## Caso (a): variamos $\theta$ , fijamos $Se$ y $Sp$

- ▶ Fijamos:

$$Se = 0.9, \quad Sp = 0.95.$$

- ▶ En función de  $\theta$ :

$$p(\theta) = Se \cdot \theta + (1 - Sp)(1 - \theta).$$

- ▶ Esto es una **función lineal y creciente** en  $\theta$ .

- ▶ Valores extremos:

- ▶ Si  $\theta \rightarrow 0$ :

$$p(\theta) \approx 1 - Sp = 0.05 \quad (\text{sólo falsos positivos}).$$

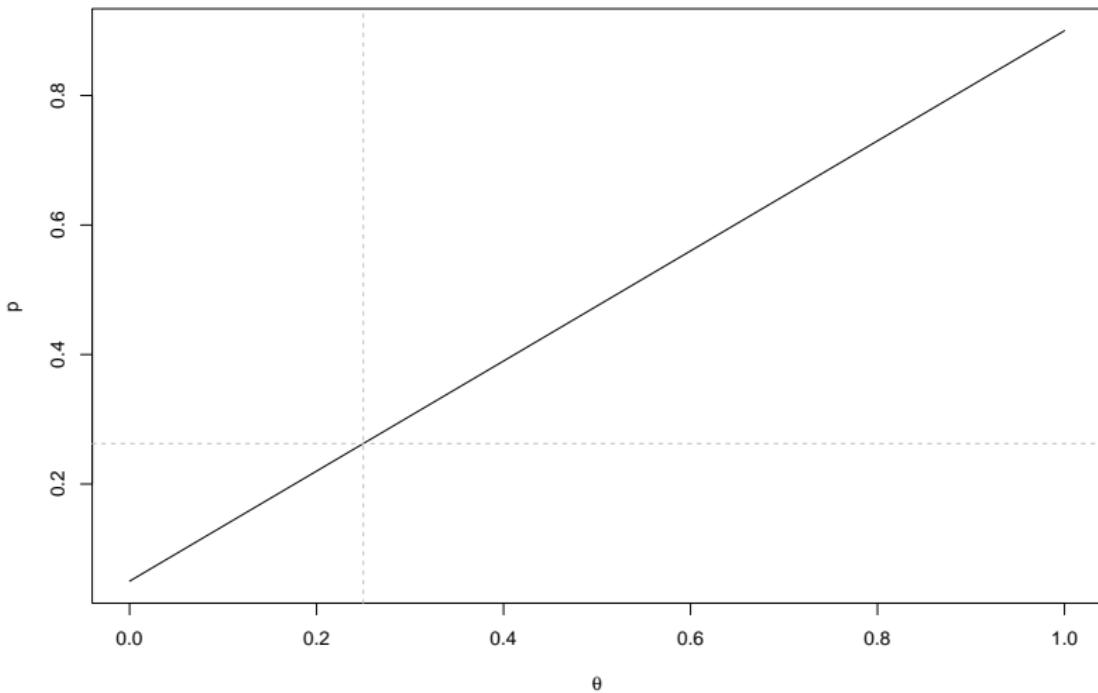
- ▶ Si  $\theta = 1$ :

$$p(1) = Se = 0.9 \quad (\text{todos enfermos, sólo verdaderos positivos}).$$

- ▶ Para  $\theta = 0.25$ :

$$p(0.25) = 0.9 \times 0.25 + 0.05 \times 0.75 = 0.2625.$$

### Probabilidad de test positivo vs. theta



## Caso (b): variamos $Se$ , fijamos $\theta$ y $Sp$

- ▶ Fijamos:

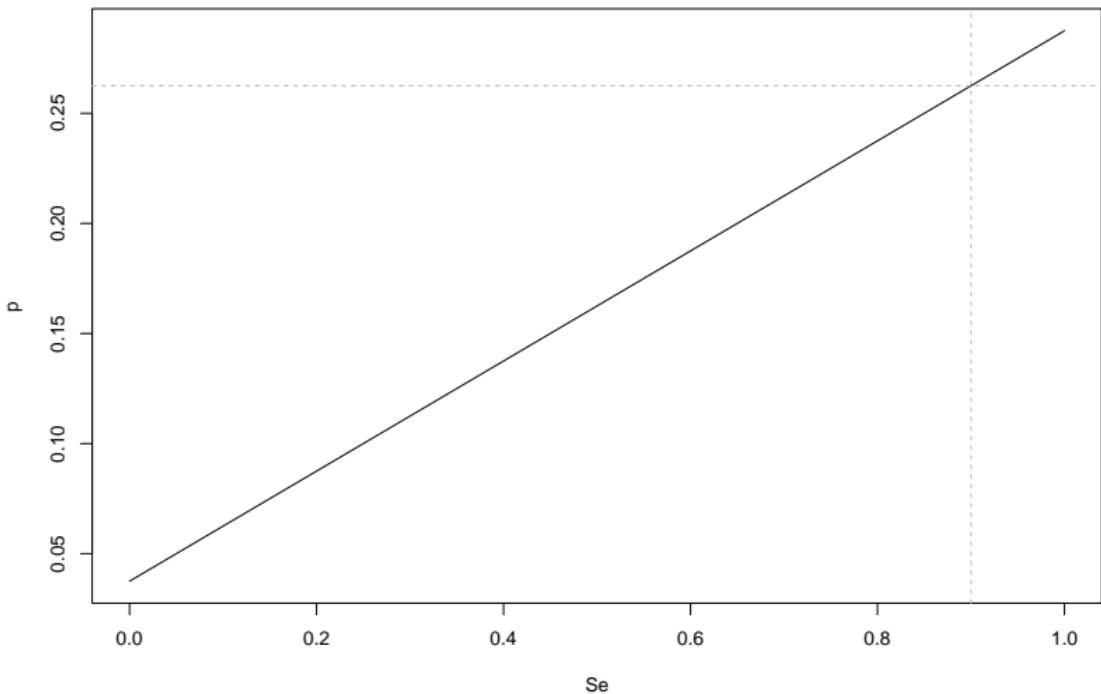
$$\theta = 0.25, \quad Sp = 0.95.$$

- ▶ En función de  $Se$ :

$$p(Se) = Se \cdot \theta + (1 - Sp)(1 - \theta) = 0.25 \cdot Se + 0.05 \cdot 0.75.$$

- ▶ Es una función **lineal y creciente** en  $Se$ .
- ▶ Interpretación:
  - ▶ Si  $Se$  es muy baja, el test casi no detecta enfermos  $\rightarrow p$  se mantiene cerca del nivel de falsos positivos.
  - ▶ A medida que aumenta  $Se$ , sube la proporción de **verdaderos positivos**, y por eso aumenta  $p$ .

### Probabilidad de test positivo vs. sensibilidad



## Caso (c): variamos $Sp$ , fijamos $\theta$ y $Se$

- ▶ Fijamos:

$$\theta = 0.25, \quad Se = 0.9.$$

- ▶ En función de  $Sp$ :

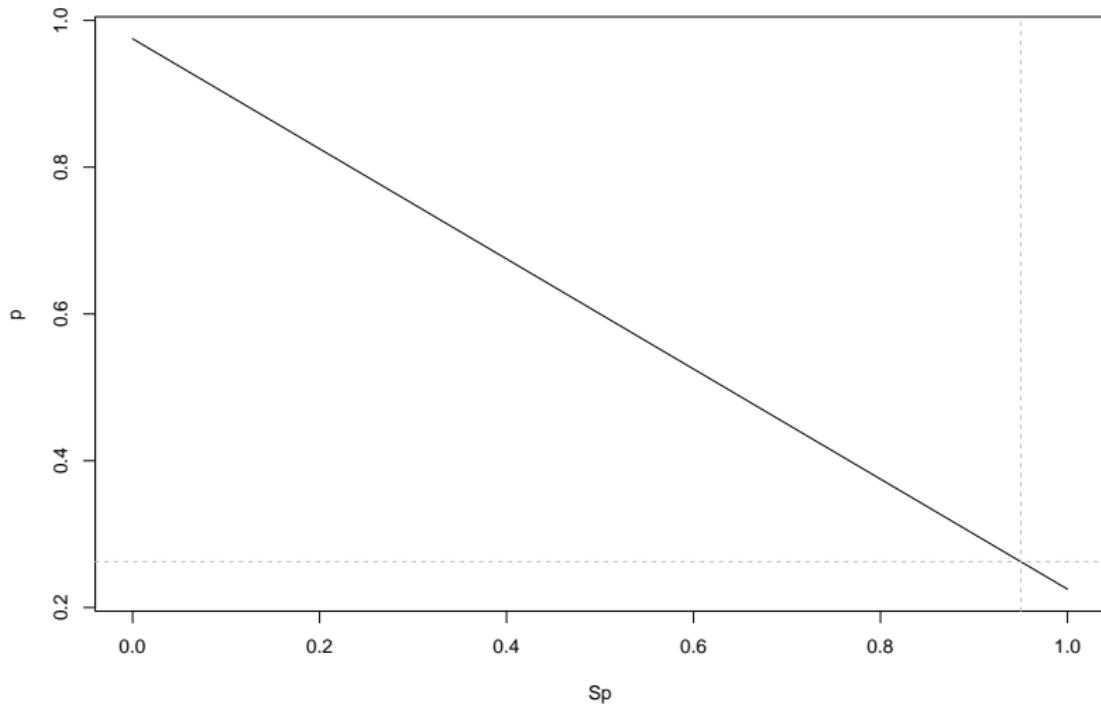
$$p(Sp) = Se \cdot \theta + (1 - Sp)(1 - \theta) = 0.9 \cdot 0.25 + (1 - Sp) \cdot 0.75.$$

- ▶ Es una función **lineal y decreciente** en  $Sp$ .

- ▶ Interpretación:

- ▶ Si  $Sp$  es baja  $\rightarrow$  muchos falsos positivos  $\rightarrow p$  es alto, aunque la prevalencia no sea muy grande.
- ▶ Al aumentar  $Sp$ , se reducen los falsos positivos  $\rightarrow$  disminuye  $p$ .

### Probabilidad de test positivo vs. especificidad



## Comportamiento de $p$

- ▶  $p$  vs  $\theta$  (con  $Se, Sp$  fijos):
  - ▶ Recta **creciente**; va de  $1 - Sp$  hasta  $Se$ .
- ▶  $p$  vs  $Se$  (con  $\theta, Sp$  fijos):
  - ▶ Recta **creciente**; al aumentar  $Se$  suben los verdaderos positivos.
- ▶  $p$  vs  $Sp$  (con  $\theta, Se$  fijos):
  - ▶ Recta **decreciente**; al aumentar  $Sp$  bajan los falsos positivos.

## 2.1.4 Estimador de momentos (MoM)

- ▶ Por item 2.0.2, sabemos que:

$$p = S_e \theta + (1 - S_p)(1 - \theta)$$

Reagrupoando:

$$p = (S_e + S_p - 1)\theta + 1 - S_p$$

Entonces:

$$\theta = \frac{p + S_p - 1}{S_e + S_p - 1}$$

- ▶ Como  $\mathbb{E}[T_i] = p$ , el estimador de momentos de  $p$  es  
$$\hat{p}_{MoM} = \frac{T_{per}}{n}.$$
- ▶ Finalmente, el estimador plug-in de momentos de  $\theta$  es el siguiente:

$$\hat{\theta}_{MoM} = \frac{\hat{p}_{MoM} + S_p - 1}{S_e + S_p - 1}$$

## 2.1.5 Sesgo, varianza y ECM

### ► Sesgo:

Primero se observa que:

$$\mathbb{E}[\hat{p}_{MoM}] = \mathbb{E}\left[\frac{T_{per}}{n}\right] = p = (S_e + S_p - 1)\theta + 1 - S_p$$

Luego:

$$\mathbb{E}[\hat{\theta}_{MoM}] = \mathbb{E}\left[\frac{\hat{p}_{MoM} + S_p - 1}{S_e + S_p - 1}\right] = \frac{\mathbb{E}[\hat{p}_{MoM}] + S_p - 1}{S_e + S_p - 1} = \theta$$

Por lo tanto  $\hat{\theta}_{MoM}$  es un estimador insesgado de  $\theta$ .

► **Varianza:**

Primero se observa que:

$$\text{Var}[\hat{p}_{MoM}] = \text{Var}\left[\frac{T_{\text{per}}}{n}\right] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Luego:

$$\begin{aligned}\text{Var}[\hat{\theta}_{MoM}] &= \text{Var}\left[\frac{\hat{p}_{MoM} + S_p - 1}{S_e + S_p - 1}\right] = \frac{\text{Var}[\hat{p}_{MoM} + S_p - 1]}{(S_e + S_p - 1)^2} = \\ &\frac{\text{Var}[\hat{p}_{MoM}]}{(S_e + S_p - 1)^2} = \frac{p(1-p)}{n(S_e + S_p - 1)^2}\end{aligned}$$

► **ECM:**

Como el sesgo es 0:

$$\text{ECM}(\hat{\theta}_{MoM}) = \text{Var}[\hat{\theta}_{MoM}] + \mathbb{B}^2[\hat{\theta}_{MoM}] = \frac{p(1-p)}{n(S_e + S_p - 1)^2}$$

## ► Consistencia

Observemos que, por ley fuerte de los grandes números:

$$\hat{p}_{MoM} = \frac{T_{per}}{n} \xrightarrow{cs} p$$

Luego, como  $\hat{\theta}_{MoM}$  es una función continua de  $\hat{p}_{MoM}$ :

$$\hat{\theta}_{MoM} \xrightarrow{cs} \frac{p + S_p - 1}{S_e + S_p - 1} = \theta$$

Por lo tanto,  $\hat{\theta}_{MoM}$  es fuertemente consistente.

## 2.1.6 Comparación ECM: Test perfecto vs imperfecto

### Fórmulas

$$\text{ECM}_{\text{perfecto}} = \frac{\theta(1 - \theta)}{n}$$

$$\text{ECM}_{\text{imperfecto}} = \frac{p(1 - p)}{n(S_e + S_p - 1)^2}$$

Factor de aumento

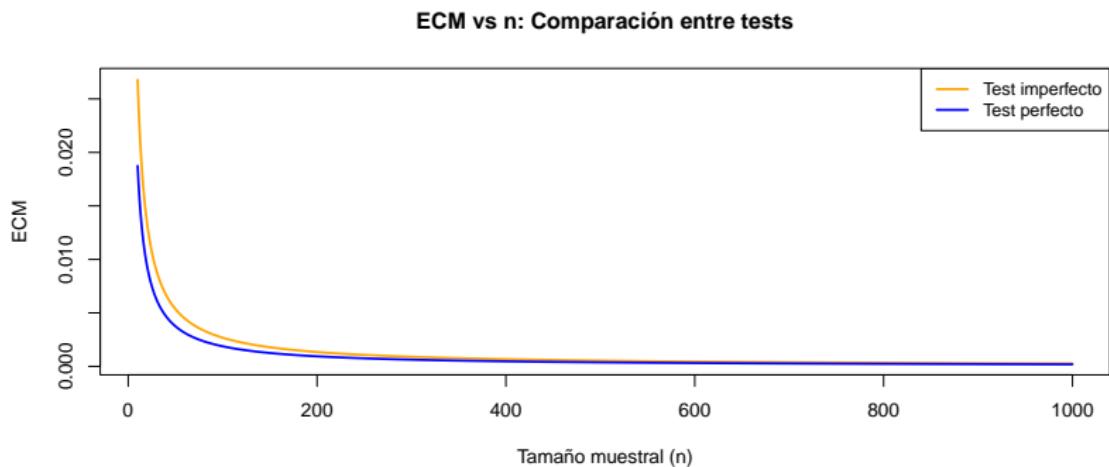
$$\boxed{\frac{1}{(S_e + S_p - 1)^2}}$$

Para  $S_e = 0.9$ ,  $S_p = 0.95$

$$S_e + S_p - 1 = 0.85 \Rightarrow \frac{1}{0.85^2} \approx 1.38$$

**ECM 38% mayor** con test imperfecto

# Comportamiento en función de $n$



## Conclusiones clave

### Observaciones del gráfico

1. **ECM imperfecto > ECM perfecto** para todo  $n$
2. Ambas **decrecen** como  $1/n$

## 2.1.7 Validación: Teórico vs Simulado

### Diseño de simulación

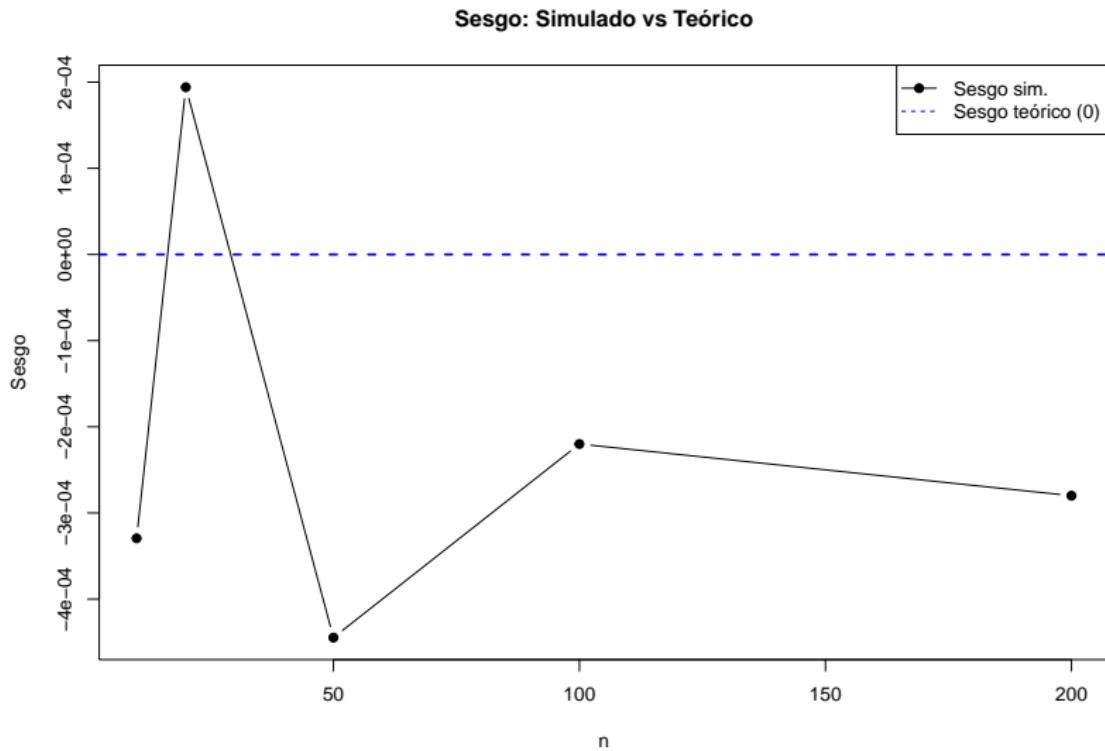
- ▶  $R = 10000$  réplicas
- ▶  $n \in \{10, 20, 50, 100, 200\}$
- ▶  $\theta = 0.25, S_e = 0.9, S_p = 0.95$
- ▶ Estimador:  $\hat{\theta}_{MoM}$

### Valores teóricos

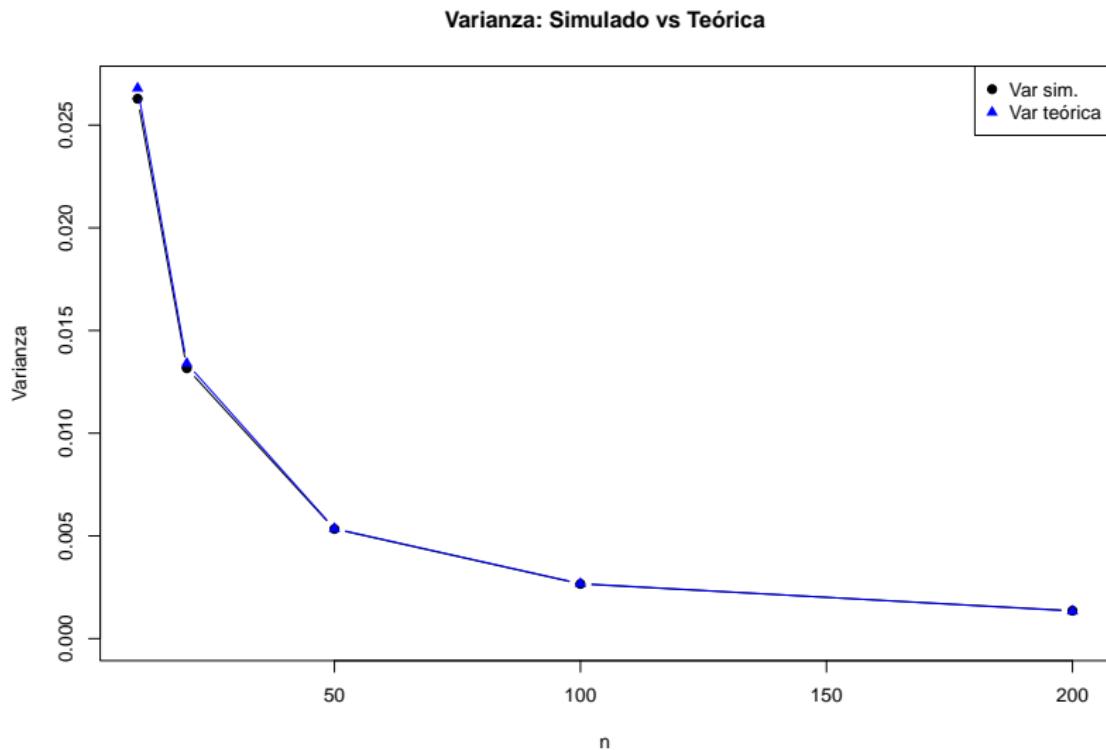
$$\text{Var}_{\text{teórica}} = \frac{p(1-p)}{n(S_e + S_p - 1)^2}$$

$$\text{ECM}_{\text{teórico}} = \text{Var}_{\text{teórica}} \quad (\text{sesgo} = 0)$$

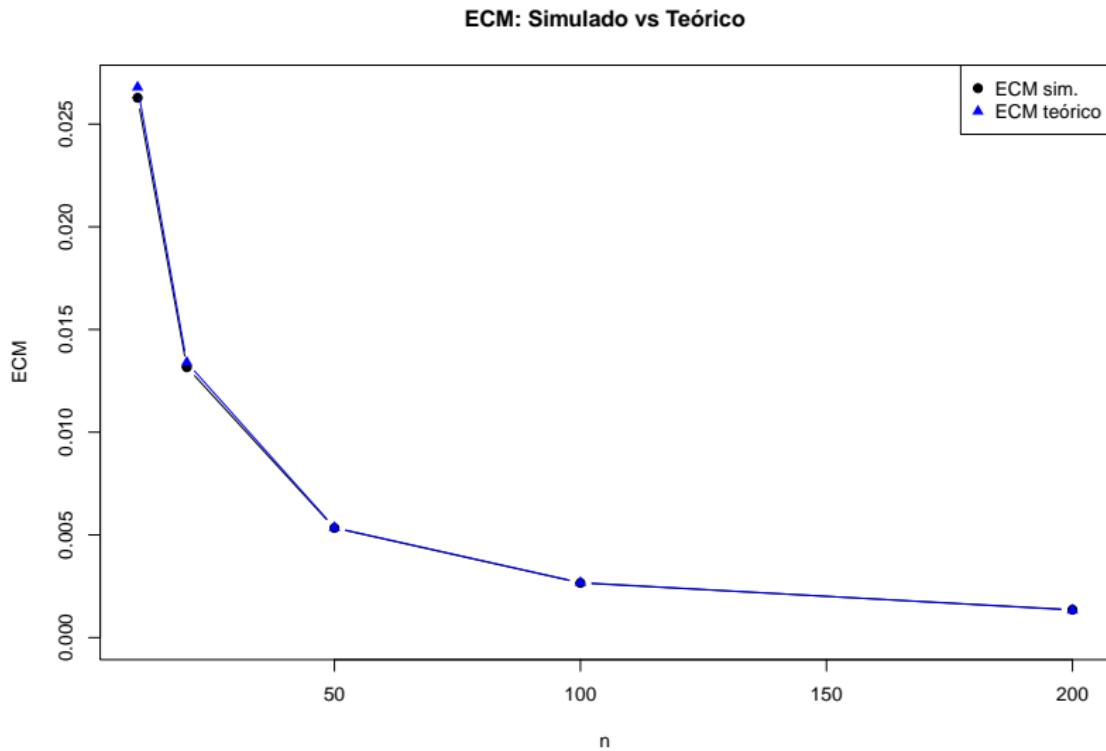
# Sesgo: Simulado vs Teórico ( $=0$ )



# Varianza: Simulado vs Teórica



# ECM: Simulado vs Teórico



# Conclusiones de la validación

## Confirmaciones

1. **Sesgo aprox 0** en simulación → Insesgabilidad verificada
2. **Varianza simulada aprox Varianza teórica** para todo  $n$
3. **ECM simulada aprox ECM teórico** → Fórmulas correctas

## Comportamiento observado

- ▶ **Convergencia:** A mayor  $n$ , mejor ajuste entre lo simulado y lo teorico
- ▶ **Consistencia:** Var/ECM decrecen como  $1/n$
- ▶ **Robustez:** Fórmulas teóricas válidas aún para  $n$  pequeño

## Implicaciones

- ▶ **Expresiones teóricas** son confiables
- ▶ **Estimador MoM** se comporta según la teoría
- ▶ **Simulación** valida aproximaciones asintóticas

## 2.1.8 Bootstrap para $\hat{\theta}_{MoM}$ ( $n = 10$ )

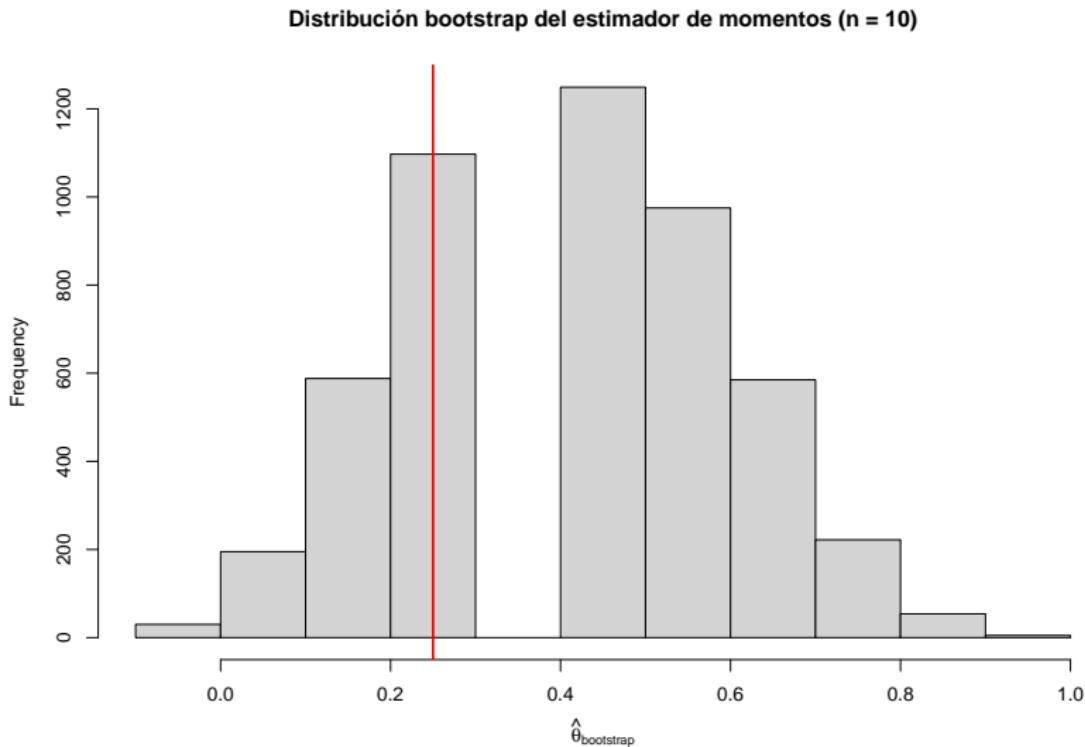
### Parámetros

- ▶  $\theta = 0.25$ ,  $S_e = 0.9$ ,  $S_p = 0.95$
- ▶  $n = 10$ ,  $B = 5000$  réplicas bootstrap
- ▶ Bootstrap no paramétrico

### Método

1. Muestra original:  $T_i \sim \text{Bernoulli}(p)$
2. Remuestreo con reemplazo
3. Recalcular  $\hat{\theta}_{MoM}$  en cada réplica

# Distribución bootstrap ( $n = 10$ )



# Resultados bootstrap

## Estadísticas clave

Media bootstrap = 0.4135

Sesgo bootstrap = 0.1635

Desvío bootstrap = 0.1822

## Características observadas

1. **Alta dispersión** (gran variabilidad)
2. **Sesgo positivo**
3. **Distribución asimétrica** hacia derecha
4. **Valores fuera de [0,1]** posibles
5. Se recomienda aumentar  $n$  (mas centrado al valor real y disminuir var)

## 2.2.9 Intervalos de confianza bootstrap

### Objetivo

Construir intervalos de confianza para  $\theta$  basados en el estimador de momentos utilizando bootstrap no paramétrico.

### Método

- ▶ A partir de los datos  $T_i$ , se calcula  $\hat{\theta}_{MoM}$  en cada remuestreo.
- ▶ Se generan  $B = 1000$  réplicas bootstrap:  
 $\hat{\theta}_{MoM}^{*(b)}$ ,  $b = 1, \dots, B$ .
- ▶ Intervalo bootstrap percentil:

$$IC_{0.95}^{boot} = [Q_{0.025}(\hat{\theta}^*), Q_{0.975}(\hat{\theta}^*)].$$

## Simulación Monte Carlo

- ▶ Replicaciones:  $R = 700$ .
- ▶ Tamaños muestrales:  $n \in \{10, 20, 50, 100, 1000\}$ .
- ▶ Para cada réplica se obtiene:
  - ▶ **Cubrimiento empírico:**  $1\{\theta \in IC^{boot}\}$
  - ▶ **Longitud:**  $IC^{up} - IC^{low}$

## 2.2.10 Intervalo de confianza asintótico

Recordemos que:

- ▶  $\hat{p}_{MoM} = \frac{T_{per}}{n}$
- ▶  $\mathbb{E}[\hat{p}_{MoM}] = p$
- ▶  $\text{Var}[\hat{p}_{MoM}] = \frac{p(1-p)}{n}$
- ▶  $\hat{\theta}_{MoM} = \frac{\hat{p}_{MoM} + S_p - 1}{S_e + S_p - 1}$
- ▶  $\mathbb{E}[\hat{\theta}_{MoM}] = \theta$
- ▶  $\text{Var}[\hat{\theta}_{MoM}] = \frac{p(1-p)}{n(S_e + S_p - 1)^2}$

- ▶ Entonces por TCL:

$$\sqrt{n}(\hat{p}_{MoM} - p) \xrightarrow{D} N(0, (p(1-p)))$$

Defino  $g(x) = \frac{x+S_p-1}{S_e+S_p-1}$  y  $g'(x) = \frac{1}{S_e+S_p-1}$ . Notar que  $g(x)$  es  $C^1$ .

- ▶ Por Método Delta:

$$\sqrt{n}(g(\hat{p}_{MoM}) - g(p)) \xrightarrow{D} N\left(0, \frac{p(1-p)}{(S_e + S_p - 1)^2}\right)$$

Luego,

$$\frac{\hat{\theta}_{MoM} - \theta}{\sqrt{\text{Var}[\hat{\theta}_{MoM}]}} \xrightarrow{D} N(0, 1)$$

- Además  $\hat{p}_{MoM} = \frac{T_{per}}{n} \xrightarrow{cs} p$ , entonces:

$$\sqrt{\frac{\hat{p}_{MoM}(1 - \hat{p}_{MoM})}{p(1 - p)}} \xrightarrow{P} 1$$

debido a que  $h(x) = \sqrt{\frac{x(1-x)}{p(1-p)}}$  es continua en  $(0, 1)$ .

Si  $\widehat{\text{Var}}[\hat{\theta}_{MoM}] = \frac{\hat{p}_{MoM}(1-\hat{p}_{MoM})}{n(S_e+S_p-1)^2}$ , usando teorema de Slutsky:

$$\frac{\hat{\theta}_{MoM} - \theta}{\sqrt{\widehat{\text{Var}}[\hat{\theta}_{MoM}]}} = \frac{\hat{\theta}_{MoM} - \theta}{\sqrt{\text{Var}[\hat{\theta}_{MoM}]}\sqrt{\frac{\hat{p}_{MoM}(1-\hat{p}_{MoM})}{p(1-p)}}} \xrightarrow{D} N(0, 1)$$

- ▶ Por lo tanto, se deduce el siguiente intervalo de nivel asintótico 0.95:

$$IC_{0.95}^\theta = \hat{\theta}_{MoM} \pm 1.96 \sqrt{\widehat{\text{Var}}[\hat{\theta}_{MoM}]}$$

donde  $\widehat{\text{Var}}[\hat{\theta}_{MoM}] = \frac{\hat{p}_{MoM}(1-\hat{p}_{MoM})}{n(S_e+S_p-1)^2}$

## 2.2.11 Comparación entre ambos intervalos

Luego de realizar las simulaciones pertinentes, comparemos los resultados obtenidos.

Table 2: Bootstrap Percentil

n	cubrimiento	longitud_media
10	0.9329	0.5680
20	0.9186	0.4285
50	0.9314	0.2782
100	0.9429	0.2006
1000	0.9514	0.0638

Table 3: Intervalos Asintóticos

n	cubrimiento	longitud_media
10	0.9257	0.5930
20	0.9186	0.4366
50	0.9200	0.2836
100	0.9414	0.2013
1000	0.9457	0.0641

- ▶ El bootstrap parece ser más robusto en muestras pequeñas, lo cual es esperable porque no depende de aproximaciones normales ni del método delta.
- ▶ El método asintótico da intervalos ligeramente más largos.
- ▶ A medida que aumenta  $n$  los resultados convergen a los valores esperados.

## 2.3.12 Comportamiento del estimador de momentos

El estimador puede quedar fuera de  $[0, 1]$

Recordemos el estimador de momentos:

$$\hat{\theta}_{MoM} = \frac{\hat{p} + Sp - 1}{Se + Sp - 1}.$$

- ▶ La prevalencia verdadera satisface:

$$0 \leq \theta \leq 1.$$

- ▶ Sin embargo, el estimador  $\hat{\theta}_{MoM}$  **NO necesariamente cumple**

$$0 \leq \hat{\theta}_{MoM} \leq 1.$$

## ¿Por qué ocurre esto?

- ▶ Porque  $\hat{p}$  es una proporción muestral sujeta a variabilidad.
- ▶ Cuando la muestra es pequeña o el test es poco informativo ( $Se + Sp - 1$  cercano a 0), la fórmula puede devolver:
  - ▶ valores **negativos**, o
  - ▶ valores **mayores a 1**.

## Consecuencia

- ▶ Aparece la necesidad de un **estimador truncado** para garantizar valores válidos.

## Casos típicos

1.  **$\hat{p}$  muy bajo** pero el test tiene muchos falsos positivos  
→ el numerador puede volverse negativo.
2.  **$\hat{p}$  muy alto** pero el test tiene baja sensibilidad  
→ el estimador puede superar 1.
3. **Muestras pequeñas ( $n$  chico)**  
→ mayor variabilidad en  $\hat{p}$  → más frecuencias fuera de  $[0, 1]$ .
4. **Test débil** ( $Se + Sp - 1$  pequeño):  
→ divide por un número muy chico → el estimador “explota”.

## Ejemplo concreto

Supongamos un test con:

- ▶  $Se = 0.90$
- ▶  $Sp = 0.95$

Entonces:

$$Se + Sp - 1 = 0.90 + 0.95 - 1 = 0.85.$$

Tomemos una muestra donde la proporción de tests positivos es muy baja:

$$\hat{p} = 0.02.$$

Calculemos el estimador de momentos:

$$\hat{\theta}_{MoM} = \frac{\hat{p} + Sp - 1}{Se + Sp - 1} = \frac{0.02 + 0.95 - 1}{0.85} = \frac{-0.03}{0.85} \approx -0.035.$$

## Interpretación

- ▶ La prevalencia **no puede ser negativa**, pero el estimador de momentos sí puede producir valores  $< 0$ .
- ▶ Esto sucede porque  $\hat{p}$  es muy bajo y el test genera falsos positivos.
- ▶ **Motiva el uso del estimador truncado**, que en este caso devolvería:

$$\hat{\theta}_{trunc} = 0.$$

## 2.3.13 Estudio del estimador truncado

Definición del estimador truncado

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

Este estimador **siempre produce valores válidos.**

Queremos estudiar, para  $n = 10, 100, 1000$ :

1. **Sesgo** de  $\hat{\theta}_{trunc}$
2. **Varianza**
3. **ECM**
4. **Distribución aproximada**

Parámetros usados

- ▶  $\theta = 0.25$
- ▶  $Se = 0.90$
- ▶  $Sp = 0.95$

Simulamos  $N_{rep}$  réplicas (típicamente 10,000).

## Resultados

Se obtiene lo siguiente:

Table 4: Resultados

n	media	sesgo	var	ECM
10	0.2485059	-0.001494118	0.024626528	0.024626298
100	0.2500482	0.000048235	0.002685104	0.002684837
1000	0.2500671	0.000067059	0.000262917	0.000262895

De esta forma, se observa lo siguiente:

- ▶ Sesgo: Para  $n = 10$  el estimador presenta un sesgo notable. Esto se debe a que el estimador de momentos frecuentemente cae por debajo de 0 y es truncado a cero, lo cual desplaza el valor medio del estimador hacia la región central del intervalo. A medida que el tamaño muestral aumenta, la frecuencia de truncamientos disminuye y el sesgo se aproxima a cero. En consecuencia, el estimador truncado es asintóticamente insesgado.
- ▶ Varianza: La varianza disminuye al crecer  $n$ , como es esperable para un estimador consistente. Para  $n = 10$  la varianza es elevada debido a la alta variabilidad de  $\hat{p}$ , mientras que para  $n = 1000$  el estimador muestra muy poca dispersión.

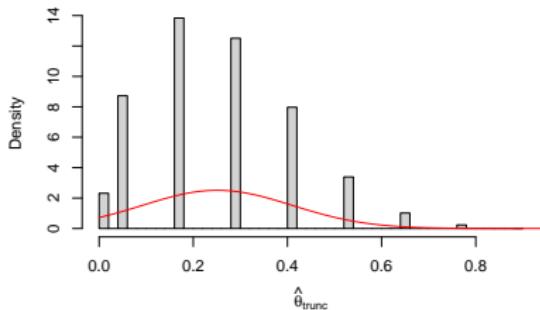
- ▶ Error cuadrático medio: El ECM refleja el compromiso entre sesgo y varianza:

$$ECM = \text{Sesgo}^2 + \text{Var.}$$

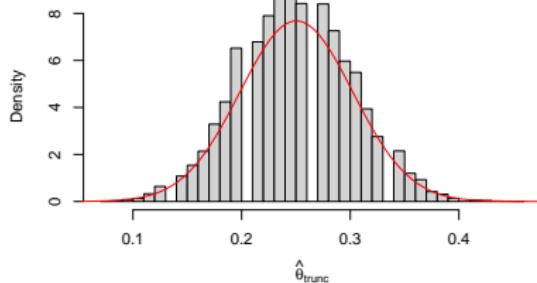
Para tamaños muestrales pequeños, el ECM se ve afectado tanto por el sesgo inducido por el truncamiento como por la alta varianza. Para tamaños muestrales grandes, el ECM resulta muy pequeño, consistente con la convergencia del estimador.

- Distribución asintótica: observemos los siguientes gráficos para la distribución asintótica:

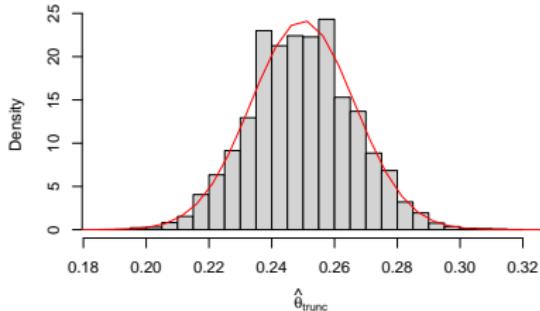
Histograma de  $\hat{\theta}_{\text{trunc}}$ ,  $n = 10$



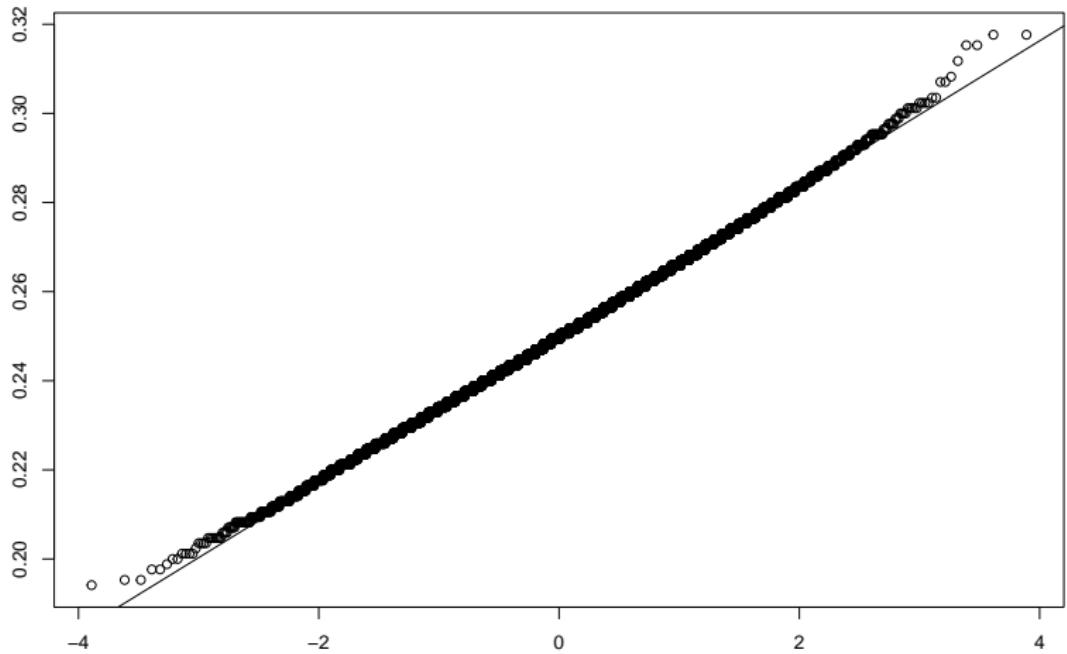
Histograma de  $\hat{\theta}_{\text{trunc}}$ ,  $n = 100$



Histograma de  $\hat{\theta}_{\text{trunc}}$ ,  $n = 1000$



QQ-plot de  $\hat{\theta}_{\text{trunc}}$  para n=1000



Para  $n$  pequeños, la distribución empírica de  $\hat{\theta}_{\text{trunc}}$  muestra acumulación de masa en los puntos 0 y 1, debido al truncamiento. Por ello, la distribución no es aproximadamente normal. Sin embargo, para  $n$  grandes el truncamiento ocurre con probabilidad prácticamente nula, y la distribución empírica se aproxima bien a una normal, coincidiendo en este caso con la distribución asintótica del estimador de momentos no truncado.

Para finalizar, estos resultados, en conjunto, muestran que el truncamiento introduce sesgo en muestras pequeñas, pero el estimador es asintóticamente equivalente al estimador de momentos original, preservando sus propiedades cuando  $n$  es grande.

# **Parte III**

Dos muestras (pre-post intervención)

- ▶ En esta sección analizamos si la prevalencia verdadera de la enfermedad cambió después de implementar una campaña de vacunación. Para ello se toman dos muestras independientes:
- ▶ Una **antes** de la campaña, de tamaño  $n_{\text{pre}}$ , con número de test positivos  $X_{\text{pre}}$ .
- ▶ Otra **después**, de tamaño  $n_{\text{post}}$ , con número de test positivos  $X_{\text{post}}$ .

En ambos casos se utiliza el mismo test diagnóstico imperfecto, caracterizado por su **sensibilidad  $Se$**  y **especificidad  $Sp$** .

- ▶ Los conteos observados siguen:

$$X_{\text{pre}} \sim \text{Binomial}(n_{\text{pre}}, p_{\text{pre}}), \quad X_{\text{post}} \sim \text{Binomial}(n_{\text{post}}, p_{\text{post}})$$

donde

$$p_A = (Se + Sp - 1)\theta_A + (1 - Sp), \quad A \in \{\text{pre}, \text{post}\}$$

- ▶ Nuestro objetivo es inferir la **diferencia de prevalencias verdaderas**:

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

### 3.1.1 Test de Hipótesis

Se quiere plantear un test de nivel aproximado 0.05 para las siguientes hipótesis, basándose en el estimador  $\hat{\theta}_{MoM}$ :

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

Tenemos:

$$\hat{\theta}_{post} = \frac{\hat{p}_{post} + S_p - 1}{S_e + S_p - 1} \quad y \quad \hat{\theta}_{pre} = \frac{\hat{p}_{pre} + S_p - 1}{S_e + S_p - 1}$$

Y por el item 2.1.10, se deduce que:

$$\hat{\theta}_{post} \approx N(\theta_{post}, \widehat{\text{Var}}(\hat{\theta}_{post})) \quad y \quad \hat{\theta}_{pre} \approx N(\theta_{pre}, \widehat{\text{Var}}(\hat{\theta}_{pre}))$$

$$\Rightarrow \hat{\theta}_{post} - \hat{\theta}_{pre} \approx N\left(\theta_{post} - \theta_{pre}, \widehat{\text{Var}}(\hat{\theta}_{post}) + \widehat{\text{Var}}(\hat{\theta}_{pre})\right)$$

$$U_\Delta(X_{pre}, X_{post}) = \frac{\hat{\theta}_{post} - \hat{\theta}_{pre} - \Delta}{\sqrt{\widehat{\text{Var}}(\hat{\theta}_{post}) + \widehat{\text{Var}}(\hat{\theta}_{pre})}} \approx N(0, 1)$$

- ▶ Tenemos un pivote decreciente en  $\Delta$  con distribución aproximada conocida independiente de  $\Delta$ . Entonces, utilizando método del pivote, resulta el siguiente test de nivel aproximado 0.05.

$$\Phi(X_{pre}, X_{post}) = \begin{cases} 1 & \text{si } |U_0(X_{pre}, X_{post})| > z_{1-\alpha/2} = z_{0.975} \approx 1.96, \\ 0 & \text{c.c.} \end{cases}$$

### 3.1.2 Aplicación en caso ficticio

Vamos a aplicar el test dado en el item anterior a un caso ficticio con  $n_{pre} = n_{post} = 100$ ,  $S_e = 0.9$ ,  $S_p = 0.95$ ,  $\theta_{pre} = 0.2$ ,  $\theta_{post} = 0.15$  y  $\alpha = 0.05$ .

- ▶ El estadístico observado tiene el siguiente valor:

```
## U = -0.3415
```

Por lo tanto, no hay evidencia suficiente como para rechazar  $H_0$  a nivel aproximado 0.05. Esto quiere decir que no se afirma que la vacuna es realmente efectiva. Es razonable porque la diferencia real entre prevalencias es pequeña (0.05).

- ▶ ¿Qué pasa si achicamos la muestra?
- ▶ Para  $n = 50$ , obtenemos el siguiente resultado:

```
## U (n=50) = 1.1625
```

El valor del estadístico observado sigue estando en la región de aceptación. Igualmente, al haber menos muestras la varianza aumenta y el resultado es distinto de 0.

- ▶ Para  $n = 10$ , obtenemos el siguiente resultado:

```
## U (n=10) = 0.5199
```

Para una muestra tan pequeña el test se vuelve inestable y es prácticamente inútil, ya que nuestro test es de nivel asintótico.

### 3.1.3 Intervalo de confianza de nivel asintótico 0.95 para $\Delta$

Usamos el estimador de momentos (no truncado):

$$\hat{\theta}_A = \frac{\hat{p}_A + (Sp - 1)}{Se + Sp - 1}, \quad A \in \{\text{pre, post}\},$$

donde  $\hat{p}_A = X_A/n_A$  y  $X_A$  es el número de tests positivos observados.

El pivote utilizado es:

$$Z = \frac{\hat{\Delta} - \Delta}{\sqrt{\widehat{Var}(\hat{\theta}_{\text{post}}) + \widehat{Var}(\hat{\theta}_{\text{pre}})}},$$

el cual tiene distribución aproximada  $N(0, 1)$ .

El intervalo de confianza asintótico al 95% para la diferencia

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

es

$$\hat{\Delta} \pm 1.96 \sqrt{\widehat{\text{Var}}(\hat{\theta}_{\text{post}}) + \widehat{\text{Var}}(\hat{\theta}_{\text{pre}})}.$$

donde  $\hat{\Delta} = \hat{\theta}_{\text{post}} - \hat{\theta}_{\text{pre}}$ .

```
## DeltaHat = 0.1176
```

```
## IC 95% para Delta: [ -0.3259 , 0.5612 ]
```

El intervalo calculado fue:

$$IC_{95\%}(\Delta) = [-0.3259, 0.5612].$$

## Observaciones

- ▶ **Incluye al 0**, no hay evidencia suficiente para decir que hay un cambio real entre pre y post.
- ▶ **Intervalo bastante ancho**, hay alta incertidumbre.
- ▶ **Límite superior es bastante grande** puede haber aumento grande entre pre y post.
- ▶ **Límite inferior negativo**, tambien puede haber una disminucion moderada.

### 3.1.4 Nivel empírico del test.

Se quiere calcular el nivel empírico del test dado en 3.1.2. Para ello, se define una grilla de valores de  $n$  para calcular el nivel para cada valor. Se sabe que el nivel de un test está definido por el supremo de los errores de tipo 1. Entonces, también se define una grilla de valores de  $\theta$  y se realizan las simulaciones pertinentes.

- ▶ Se obtuvieron los siguientes resultados:

Table 5: Nivel empírico del test

n	nivel
10	0.1098
20	0.0895
50	0.0650
100	0.0585
1000	0.0570
10000	0.0595

Se observa claramente el comportamiento asintótico del test. A medida que aumenta  $n$ , el nivel del test se acerca a 0.05.