

# Trabajo Práctico– IECD

Grupo 11

2025-12-8

## Parte I: Test perfecto (baseline)

**1.1 Observe**  $T_{per} \sim Bi(n, \theta)$

Consideremos una población en la que cada individuo puede estar o no enfermo. Sea la prevalencia verdadera de la enfermedad

$$\theta = P(Y = 1)$$

donde  $Y$  es la variable aleatoria que indica el estado verdadero de una persona. Esto es,  $Y = 1$  si la persona está enferma y  $Y = 0$  si no lo está.

Se asume, por enunciado, que el test diagnóstico es perfecto, es decir, tanto la sensibilidad como la especificidad valen uno ( $Se = Sp = 1$ ). En este caso, el resultado del test coincide exactamente con el estado verdadero del individuo.

Sea  $Y_1, \dots, Y_n$  una muestra aleatoria de tamaño  $n$  tomada de la población. Suponemos que las variables  $Y_1, \dots, Y_n$  son independientes e idénticamente distribuidas, con

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

Definimos la variable aleatoria que cuenta cuántas personas enfermas hay en la muestra como

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

Dado que  $T_{per}$  es la suma de  $n$  variables independientes con distribución Bernoulli( $\theta$ ), se sigue que  $T_{per}$  tiene una distribución binomial. En efecto, para  $k = 0, 1, \dots, n$  se cumple que

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

Por lo tanto, concluimos que

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

✉

## 1.2 Estimador de Máxima Verosimilitud de $\theta$

Como en el punto anterior, recordemos que bajo un test perfecto se tiene

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

y que podemos representar los datos mediante variables Bernoulli

$$Y_1, \dots, Y_n \sim i.i.d. \text{Bernoulli}(\theta),$$

donde  $Y_i = 1$  si la persona  $i$  está enferma y  $Y_i = 0$  si no lo está.

Sea  $N_1 = \sum_{i=1}^n Y_i$  la cantidad de individuos enfermos en la muestra y  $N_0 = n - N_1$  la cantidad de individuos sanos.

La función de verosimilitud viene dada por

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

La log-verosimilitud resulta entonces

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

Buscamos  $\theta \in (0, 1)$  tal que

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0.$$

Derivando:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{N_0}{1-\theta}(-1) + \frac{N_1}{\theta} = -\frac{N_0}{1-\theta} + \frac{N_1}{\theta}.$$

Igualamos a cero:

$$-\frac{N_0}{1-\theta} + \frac{N_1}{\theta} = 0 \iff \frac{N_1}{\theta} = \frac{N_0}{1-\theta}.$$

Despejando:

$$N_1(1-\theta) = N_0\theta \iff N_1 - N_1\theta = N_0\theta$$

$$\iff N_1 = (N_0 + N_1)\theta = n\theta.$$

Por lo tanto,

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

Finalmente, como  $N_1 = T_{per}$ , obtenemos

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

La segunda derivada confirma que se trata de un máximo, ya que

$$\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\frac{N_0}{(1-\theta)^2} - \frac{N_1}{\theta^2} < 0, \quad \forall \theta \in (0, 1).$$

◻

### 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

```
#fijamos la semilla que se va a utilizar durante todo el tp.
set.seed(43)

tita <- 0.25
ns <- c(10, 20, 50, 100, 1000, 10000) #valores de n para los cuales vamos a evaluar la cobertura
Nrep <- 5000 #empirica
z <- qnorm(0.975)

cubrimientoN <- 0

#para cada n:
for (j in seq_along(ns)){
  n <- ns[j]
  cubrimiento <- 0

  #repite el experimento 5000 veces.
  for(i in 1:Nrep){
    muestra <- rbinom(1,n,tita)

    estTita <- muestra/n
    termino <- z * sqrt((estTita*(1-estTita))/n)
    intervalo <- c(estTita - termino, estTita + termino)

    cubrimiento[i] <- intervalo[1] <= tita && tita <= intervalo[2]
  }

  #promedio de veces que el valor real de tita cayó en el intervalo
  cubrimientoN <- cubrimientoN + mean(cubrimiento)
}
```

```

cubrimientoN[j] <- mean(cubrimiento)
}

cubrimientoEmpirico <- data.frame(n = ns, cubrimiento = cubrimientoN)

```

Y se obtuvieron estos resultados:

```
knitr::kable(cubrimientoEmpirico, digits = 4, caption = "Cubrimiento empírico del intervalo")
```

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 Estimador $\hat{p}$ de $p$ con $T_{per}$

En esta parte definimos

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

donde  $T$  denota el resultado del test diagnóstico aplicado a un individuo:  $T = 1$  si el test da positivo y  $T = 0$  si el test da negativo. Suponemos que

$$T_1, \dots, T_n \sim i.i.d. Bernoulli(p),$$

de modo que  $P(T_i = 1) = p$  y  $P(T_i = 0) = 1 - p$  para todo  $i = 1, \dots, n$ .

Definimos la variable aleatoria

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

que representa la cantidad de tests positivos observados en la muestra de tamaño  $n$ . Por ser suma de variables Bernoulli independientes con parámetro  $p$ , se tiene que

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

Un estimador para  $p$  puede ser la proporción de individuos de la muestra cuyo test dio positivo. Por lo tanto, definimos

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

Este estimador coincide además con el estimador de máxima verosimilitud (EMV) de  $p$  en el modelo binomial, ya que la función de verosimilitud

$$L(p) \approx p^{T_{per}} (1-p)^{n-T_{per}}$$

se maximiza en  $p = T_{per}/n$ .

## 2.0.2 Relación entre $p$ , $\theta$ , $Se$ y $Sp$

Recordemos que  $\theta = P(Y = 1)$  denota la prevalencia de la enfermedad en la población, mientras que  $Y = 1$  indica que el individuo está enfermo y  $Y = 0$  que está sano. La sensibilidad y especificidad del test diagnóstico se definen como

$$Se = P(T = 1 | Y = 1) \quad y \quad Sp = P(T = 0 | Y = 0),$$

donde  $T$  es el resultado del test ( $T = 1$  si el test es positivo,  $T = 0$  si es negativo). Además,  $P(Y = 1) = \theta$  y  $P(Y = 0) = 1 - \theta$ .

La probabilidad de obtener un resultado positivo en el test es

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

Aplicando la ley de probabilidad total, se tiene

$$P(T = 1) = P(T = 1 | Y = 1)P(Y = 1) + P(T = 1 | Y = 0)P(Y = 0).$$

Sustituyendo las definiciones previas,

$$p = Se \cdot \theta + P(T = 1 | Y = 0)(1 - \theta).$$

Como  $Sp = P(T = 0 | Y = 0)$ , se sigue que

$$P(T = 1 | Y = 0) = 1 - Sp.$$

Por lo tanto,

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

Equivalentemente, podemos escribir

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

## 2.0.3 Comportamiento de $p$ en función de $\theta$ , $Se$ y $Sp$ .

Se analiza cómo varía la probabilidad de obtener un resultado positivo en el test diagnóstico,

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

cuando se modifican la prevalencia verdadera  $\theta$ , la sensibilidad  $Se$  y la especificidad  $Sp$  del test. Recordemos que, a partir de la ley de probabilidad total, se obtuvo la relación

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

Trabajaremos con los valores de referencia indicados en el enunciado:

$$Se = 0.9, \quad Sp = 0.95, \quad \theta = 0.25,$$

y, en cada caso, variaremos una de las cantidades manteniendo las otras dos fijas.

```
## Valores fijos
Se_fijo <- 0.9
Sp_fijo <- 0.95
theta_fijo <- 0.25

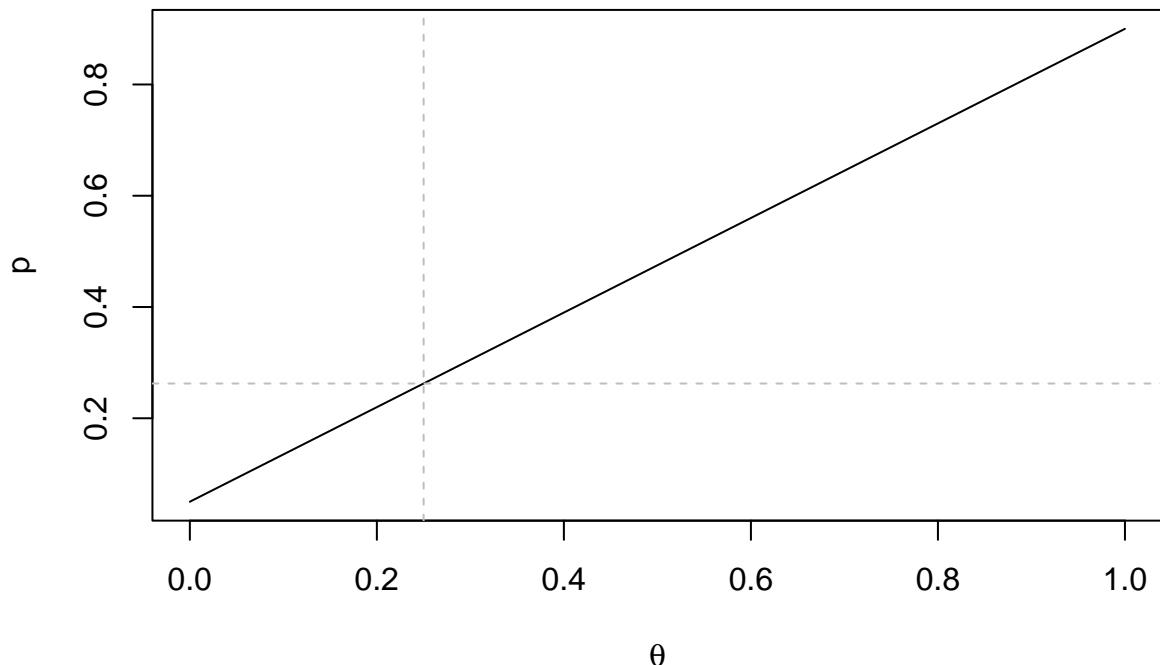
## a
theta_grid <- seq(0, 1, length.out = 1000)
p_theta <- Se_fijo * theta_grid + (1 - Sp_fijo) * (1 - theta_grid)
```

```

plot(theta_grid, p_theta, type = "l",
      xlab = expression(theta),
      ylab = "p",
      main = "Probabilidad de test positivo vs. theta")
abline(v = theta_fijo, col = "gray", lty = 2)
abline(h = Se_fijo * theta_fijo + (1 - Sp_fijo) * (1 - theta_fijo),
      col = "gray", lty = 2)

```

## Probabilidad de test positivo vs. theta



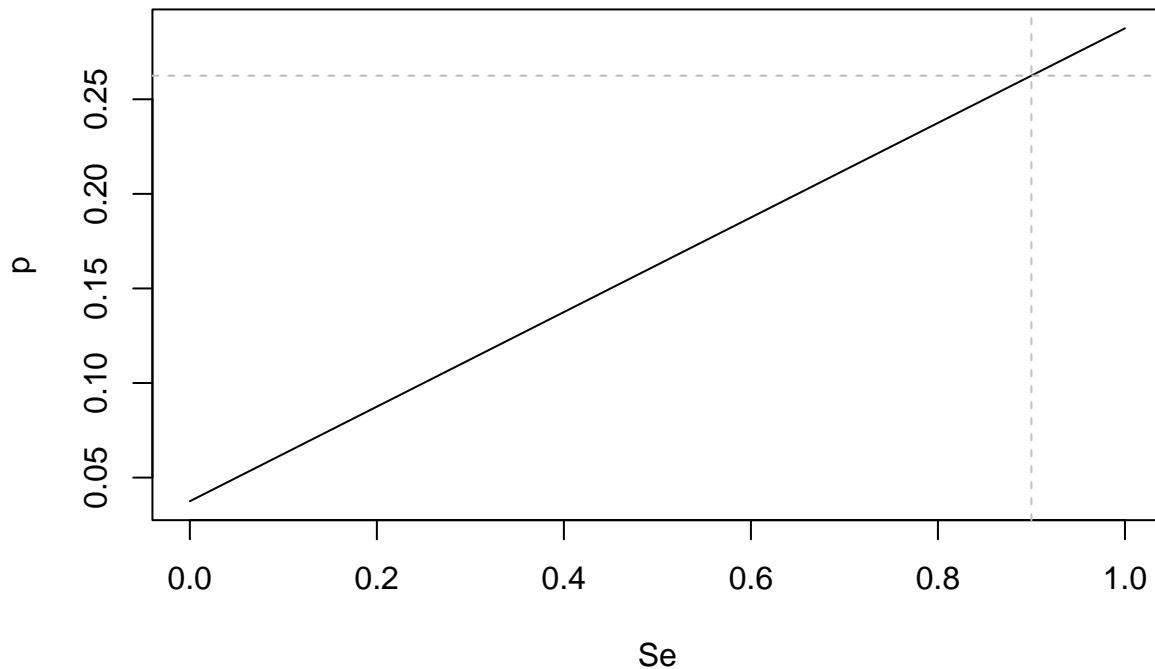
```

## b
Se_grid <- seq(0, 1, length.out = 1000)
p_Se <- Se_grid * theta_fijo + (1 - Sp_fijo) * (1 - theta_fijo)

plot(Se_grid, p_Se, type = "l",
      xlab = "Se",
      ylab = "p",
      main = "Probabilidad de test positivo vs. sensibilidad")
abline(v = Se_fijo, col = "gray", lty = 2)
abline(h = Se_fijo * theta_fijo + (1 - Sp_fijo) * (1 - theta_fijo),
      col = "gray", lty = 2)

```

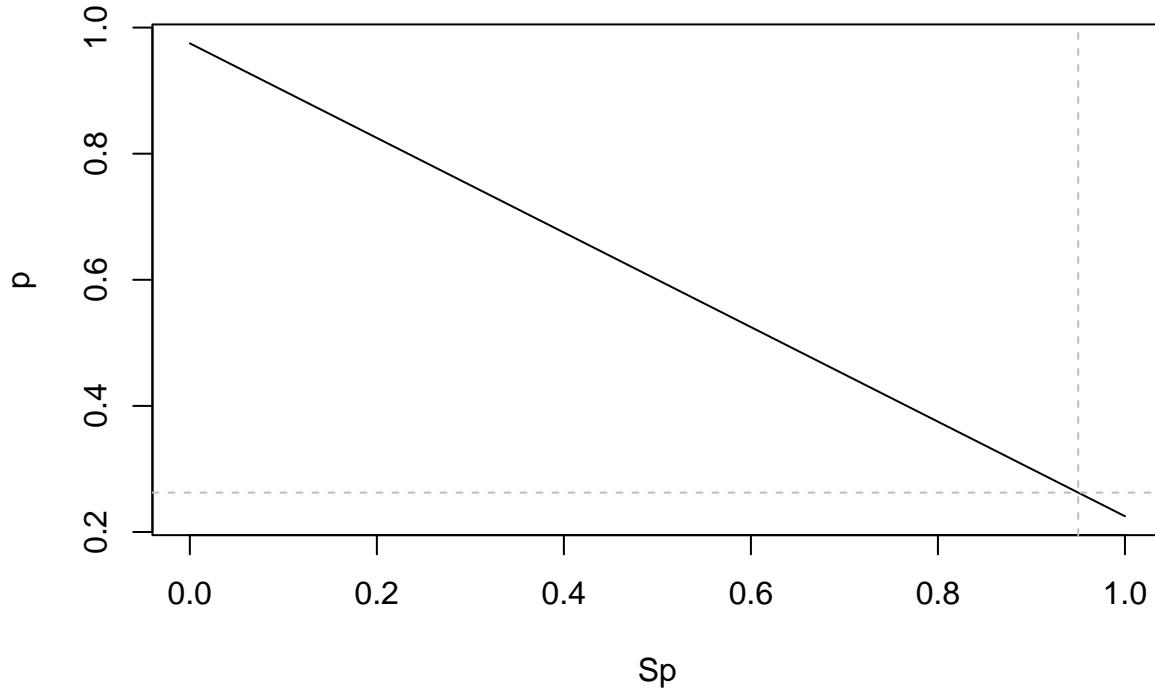
## Probabilidad de test positivo vs. sensibilidad



```
## c
Sp_grid <- seq(0, 1, length.out = 1000)
p_Sp <- Se_fijo * theta_fijo + (1 - Sp_grid) * (1 - theta_fijo)

plot(Sp_grid, p_Sp, type = "l",
      xlab = "Sp",
      ylab = "p",
      main = "Probabilidad de test positivo vs. especificidad")
abline(v = Sp_fijo, col = "gray", lty = 2)
abline(h = Se_fijo * theta_fijo + (1 - Sp_fijo) * (1 - theta_fijo),
       col = "gray", lty = 2)
```

## Probabilidad de test positivo vs. especificidad



(a)  $p$  en función de  $\theta$  con  $Se$  y  $Sp$  fijos. En primer lugar, consideramos la función

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

tomando  $Se = 0.9$  y  $Sp = 0.95$  fijos, y dejando variar la prevalencia  $\theta$  en el intervalo  $[0, 1]$ .

El gráfico correspondiente muestra que  $p(\theta)$  es una función lineal y creciente de  $\theta$ . Cuando  $\theta$  es muy pequeña (prevalencias cercanas a cero), la probabilidad de test positivo se aproxima al término de falsos positivos ( $1 - Sp$ ); en este caso,  $p(\theta) \approx 1 - Sp = 0.05$ . A medida que aumenta  $\theta$ , el término asociado a verdaderos positivos  $Se\theta$  domina y  $p(\theta)$  crece. En el extremo  $\theta = 1$ , se tiene  $p(1) = Se$ , es decir, la probabilidad de test positivo coincide con la sensibilidad del test.

El valor específico  $\theta = 0.25$  se destaca sobre la curva para visualizar la probabilidad de test positivo asociada a esa prevalencia, que es

$$p(0.25) = 0.9 \times 0.25 + (1 - 0.95)(1 - 0.25) = 0.225 + 0.05 \times 0.75 = 0.2625.$$

(b)  $p$  en función de  $Se$  con  $\theta$  y  $Sp$  fijos. A continuación, estudiamos  $p$  como función de la sensibilidad  $Se$ , manteniendo fijos  $\theta = 0.25$  y  $Sp = 0.95$ . En este caso,

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

Esta expresión también es lineal y creciente en  $Se$ . El gráfico muestra que, cuando  $Se$  es muy baja, la contribución de verdaderos positivos es pequeña y  $p(Se)$  se mantiene próxima al valor provocado por falsos positivos. Al aumentar la sensibilidad, el test detecta correctamente a una mayor proporción de individuos enfermos y, en consecuencia, la probabilidad total de obtener un test positivo aumenta.

El valor de referencia  $Se = 0.9$  se marca en el eje horizontal, junto con la altura  $p(Se = 0.9)$  en el gráfico, para visualizar el punto correspondiente en la curva.

**(c)  $p$  en función de  $Sp$  con  $\theta$  y  $Se$  fijos.** Finalmente, analizamos el comportamiento de  $p$  en función de la especificidad  $Sp$ , manteniendo fijos  $\theta = 0.25$  y  $Se = 0.9$ . En este caso

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

Aquí  $p(Sp)$  es una función lineal y *decreciente* en  $Sp$ . El gráfico muestra que, cuando la especificidad es baja (es decir, el test comete muchos falsos positivos), el término  $(1 - Sp)(1 - \theta)$  es grande y la probabilidad total de test positivo  $p$  toma valores elevados, incluso si la prevalencia real no es muy grande. A medida que  $Sp$  aumenta, el número de falsos positivos disminuye y, en consecuencia, la probabilidad de observar un test positivo se reduce.

El valor de referencia  $Sp = 0.95$  se indica sobre el eje horizontal, junto con la altura  $p(Sp = 0.95)$ , lo cual permite comparar la situación real del test con otros posibles niveles de especificidad.

#### 2.1.4 Estimador de momentos (MoM).

Por item 2.0.2, sabemos que:

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

Reagrupando:

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

Entonces:

$$\theta = \frac{p + Sp - 1}{Se + Sp - 1}$$

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

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

#### 2.1.5 Sesgo, varianza y ECM.

- **Sesgo:**

Primero se observa que:

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

Luego:

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

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

- **Varianza:**

Primero se observa que:

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

Luego:

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

- **ECM:**

Como el sesgo es 0:

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

- **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 ECM del test perfecto y test impreciso en función de n

- **ECM del test perfecto e imperfecto:**

Recordemos que el ECM del test perfecto es:

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

Y el ECM del test imperfecto es:

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

Podemos ver que el ECM del test perfecto es más chico que el del test imperfecto gracias a un factor:

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

Lo que hace que con los valores anteriormente usados,  $S_e = 0.9$ ,  $S_p = 0.95$ ;

$$S_e + S_p - 1 = 0.85$$

Por lo tanto,

$$\frac{1}{0.85^2} \approx 1.38$$

Osea que el **ECM del test imperfecto es 38% mas grande** que el del test perfecto.

```
# Parámetros
tita <- 0.25
Se <- 0.9
Sp <- 0.95

# p real del test imperfecto
p <- (Se + Sp - 1)*tita + (1 - Sp)

# Rango de n
ns <- 10:1000

# ECM del test perfecto
ECM_perfecto <- tita*(1 - tita) / ns

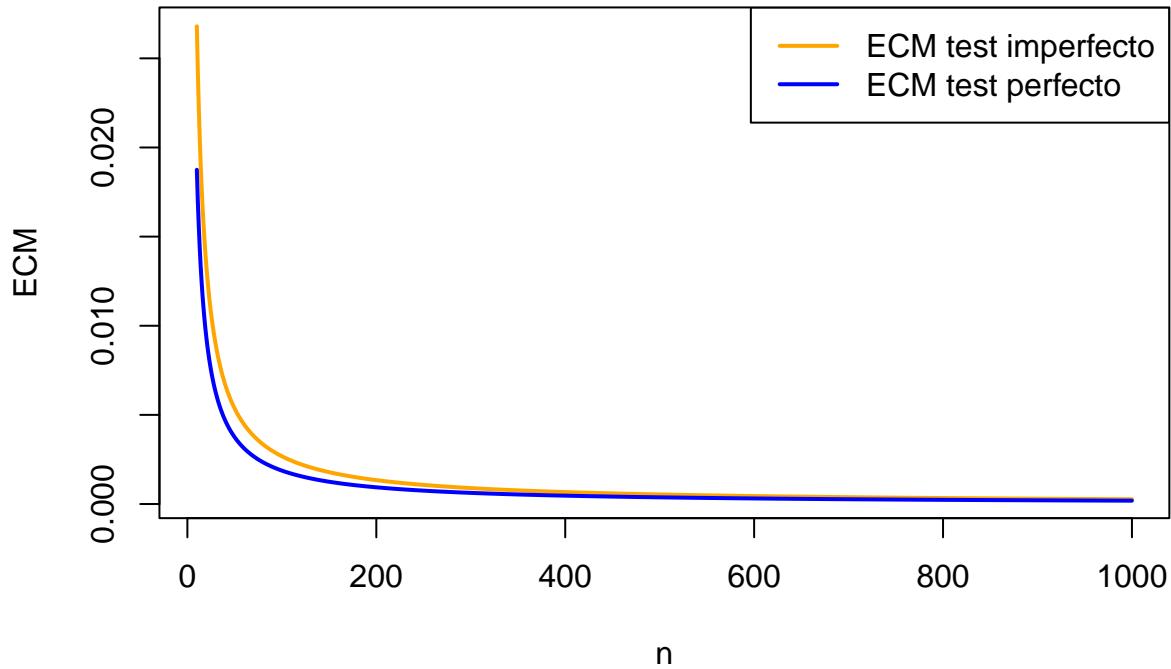
# ECM del test imperfecto
ECM_imperfecto <- p*(1 - p) / (ns * (Se + Sp - 1)^2)

# Gráfico
plot(ns, ECM_imperfecto,
      type="l", col="orange", lwd=2,
      xlab="n", ylab="ECM",
      main="ECM vs n: Test perfecto vs imperfecto")

lines(ns, ECM_perfecto, col="blue", lwd=2)

legend("topright",
       legend=c("ECM test imperfecto", "ECM test perfecto"),
       col=c("orange", "blue"), lwd=2)
```

## ECM vs n: Test perfecto vs imperfecto



Podemos ver que para todos los tamaños muestrales  $n$ , el ECM del test imperfecto es mayor que el del test perfecto. A medida que  $n$  crece, ambos ECM tienden a 0, pero el imperfecto siempre queda más arriba que el perfecto. Cuanto más pequeño sea  $Se + Sp - 1$  (es decir, cuanto peor sea el test), mayor será la brecha entre las dos curvas.

### 2.1.7 Comparo valores teóricos hallados con los simulados

**Comparación: valores teóricos vs simulados (test imperfecto)** Para el test imperfecto consideramos un modelo donde la variable observada  $T$  proviene de un test con sensibilidad  $Se$  y especificidad  $Sp$ .

La probabilidad real de obtener un test positivo es:

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

de modo que, para un tamaño muestral  $n$ ,

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

El estimador de momentos para  $\theta$  es:

$$\hat{\theta} = \frac{\hat{p}_T - (1 - Sp)}{Se + Sp - 1}, \quad \hat{p}_T = \frac{1}{n} \sum_{i=1}^n T_i.$$

Para cada tamaño muestral

$$n \in \{10, 20, 50, 100, 200\},$$

se simulan  $R = 10000$  réplicas de  $\hat{\theta}$ .

A partir de ellas se estiman:

- el sesgo,
- la varianza simulada,
- el error cuadrático medio (ECM) simulado.

Además, la teoría indica que, como el estimador es insesgado, su varianza es:

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

y el ECM teórico coincide con la varianza teórica.

```
R <- 10000
ns <- c(10, 20, 50, 100, 200)

# Función estimador de momentos
tita_corregida <- function(p_hat, Se, Sp) {
  (p_hat - (1 - Sp)) / (Se + Sp - 1)
}

# resultados
resultados_imp <- data.frame(
  n = ns,
  bias_sim = NA,
  var_sim = NA,
  ecm_sim = NA,
  var_teo = NA,
  ecm_teo = NA
)

for (i in seq_along(ns)) {
  n <- ns[i]
  est_imp <- numeric(R)

  # Probabilidad real del test positivo
  p_Y <- Se*tita + (1 - Sp)*(1 - tita)

  for (r in 1:R) {

    # 1) Genero la muestra de T correctamente
    T <- rbinom(n, 1, p_Y)

    # 2) Estimador de momentos
    p_hat <- mean(T)
    est_imp[r] <- tita_corregida(p_hat, Se, Sp)
  }

  # --- Estadísticos simulados ---
  bias_sim <- mean(est_imp) - tita
```

```

var_sim <- var(est_imp)
ecm_sim <- mean((est_imp - tita)^2)

# --- Valores teóricos ---
p_true <- p_Y
denom <- Se + Sp - 1
var_teo <- (p_true*(1 - p_true)) / (n * denom^2)
ecm_teo <- var_teo # insesgado

# --- Guardar ---
resultados_imp$bias_sim[i] <- bias_sim
resultados_imp$var_sim[i] <- var_sim
resultados_imp$ecm_sim[i] <- ecm_sim
resultados_imp$var_teo[i] <- var_teo
resultados_imp$ecm_teo[i] <- ecm_teo
}

```

- Diferencia entre sesgo teorico y en la simulacion para distintos n:

```

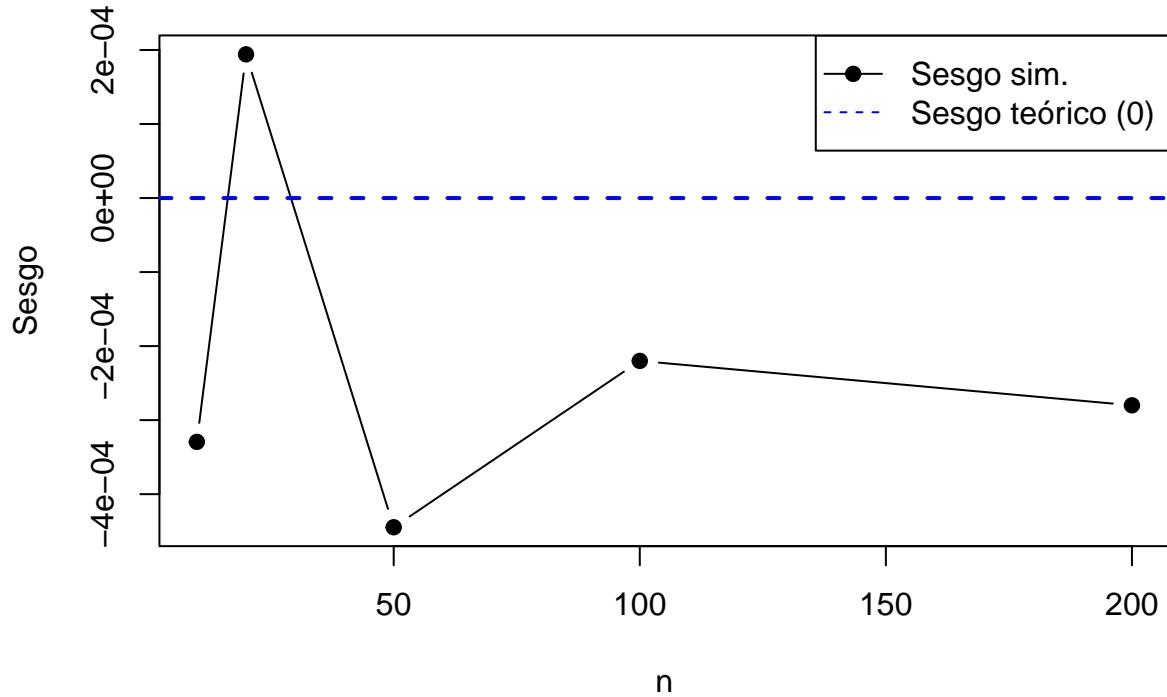
## SESGO (teórico = 0)
plot(resultados_imp$n, resultados_imp$bias_sim,
      type="b", pch=19,
      xlab="n", ylab="Sesgo",
      main="Sesgo: Simulado vs Teórico",
      ylim=c(min(resultados_imp$bias_sim), max(resultados_imp$bias_sim)))

# línea de sesgo teórico = 0
abline(h = 0, col="blue", lwd=2, lty=2)

legend("topright",
       legend=c("Sesgo sim.", "Sesgo teórico (0)"),
       col=c("black","blue"), pch=c(19, NA), lty=c(1,2))

```

## Sesgo: Simulado vs Teórico



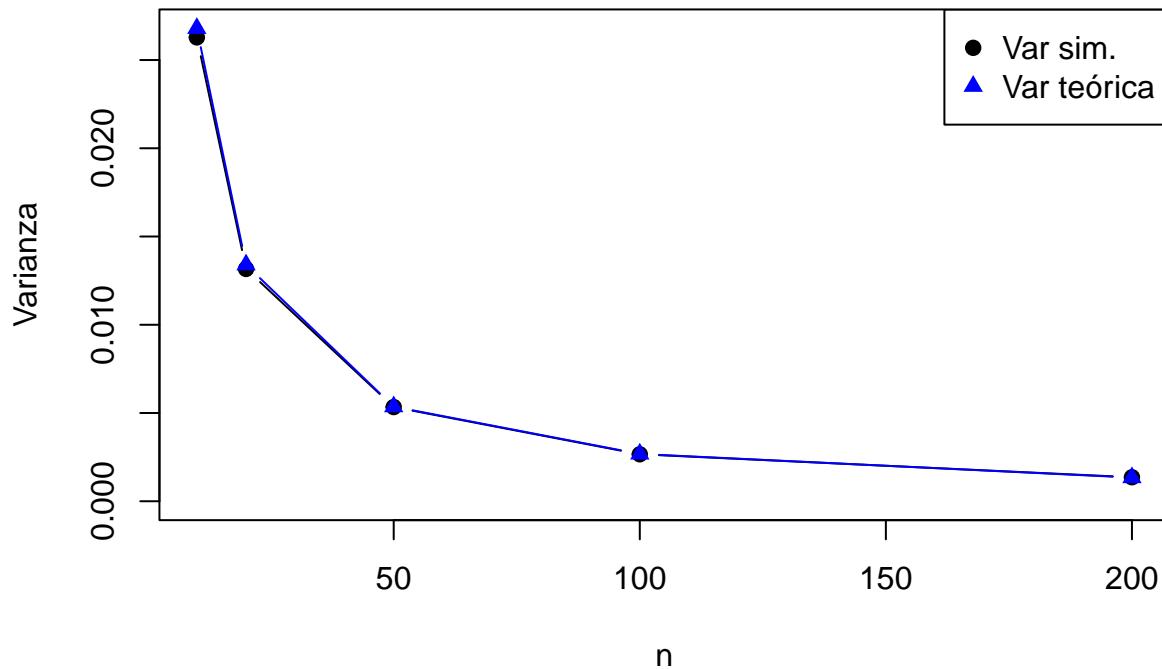
- Diferencia entre varianza teorica y en la simulacion para distintos n:

```
## VARIANZA
plot(resultados_imp$n, resultados_imp$var_sim,
      type="b", pch=19,
      xlab="n", ylab="Varianza",
      main="Varianza: Simulado vs Teórica",
      ylim=c(0, max(resultados_imp$var_sim, resultados_imp$var_teo)))

lines(resultados_imp$n, resultados_imp$var_teo,
      type="b", pch=17, col="blue")

legend("topright",
       legend=c("Var sim.", "Var teórica"),
       col=c("black","blue"), pch=c(19,17))
```

## Varianza: Simulado vs Teórica



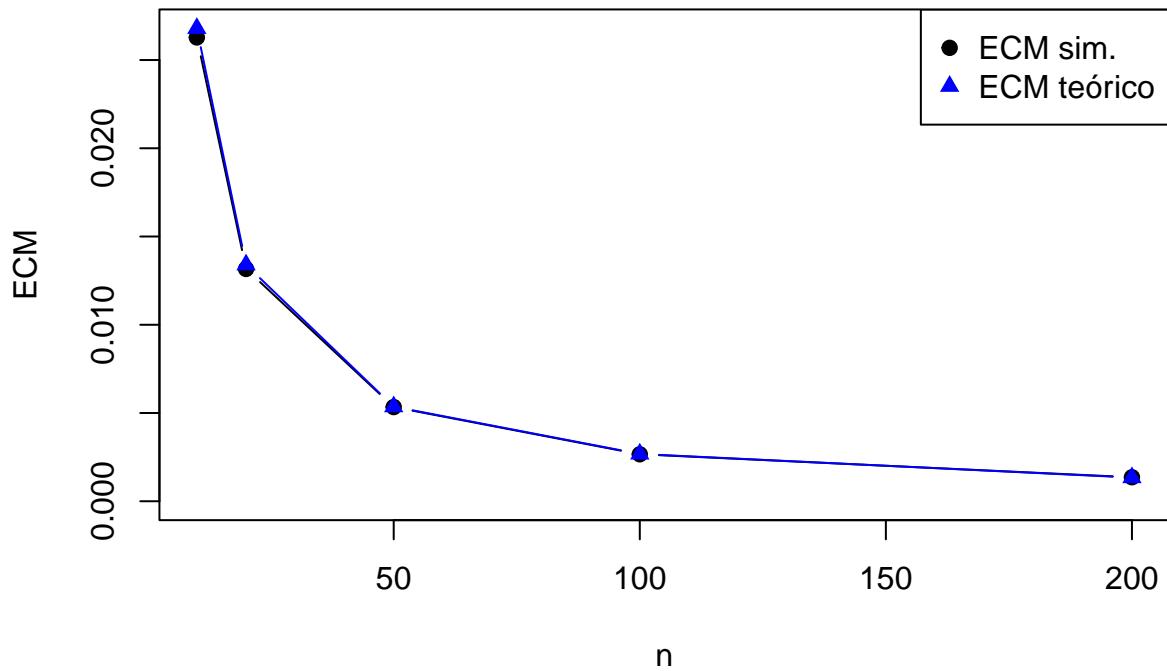
- Diferencia entre ECM teorico y en la simulacion para distintos n:

```
## ECM
plot(resultados_imp$n, resultados_imp$ecm_sim,
      type="b", pch=19,
      xlab="n", ylab="ECM",
      main="ECM: Simulado vs Teórico",
      ylim=c(0, max(resultados_imp$ecm_sim, resultados_imp$ecm_teo)))

lines(resultados_imp$n, resultados_imp$ecm_teo,
      type="b", pch=17, col="blue")

legend("topright",
       legend=c("ECM sim.", "ECM teórico"),
       col=c("black","blue"), pch=c(19,17))
```

## ECM: Simulado vs Teórico



**Conclusiones :**

- Los valores simulados de varianza y ECM están muy cercanos a los valores teóricos para todos los tamaños muestrales considerados, confirmando la validez de las expresiones teóricas.
- A medida que  $n$  aumenta, tanto la varianza simulada como el ECM tienden a disminuir y se aproximan fuertemente al valor teórico (convergencia concordante con la ley de los grandes números y la aproximación asintótica).
- El sesgo empírico del estimador corregido es cercano a cero en todas las réplicas (coincidiendo con la insesgabilidad teórica).

### 2.1.8 Muestras bootstrap para el estimador de momentos

Para  $\theta = 0.25$ ,  $S_e = 0.9$  y  $S_p = 0.95$ , construyo muestras bootstrap para observar la distribución del estimador de momentos de  $\theta$  cuando  $n = 10$ .

```

n <- 10      # tamaño de la muestra
B <- 5000    # cantidad de réplicas bootstrap

# Probabilidad real de T = 1 bajo test imperfecto
p_Y <- Se * tita + (1 - Sp) * (1 - tita)

# Genero una muestra original de T
T <- rbinom(n, 1, p_Y)

```

```

# Bootstrap no paramétrico
boot_est <- numeric(B)

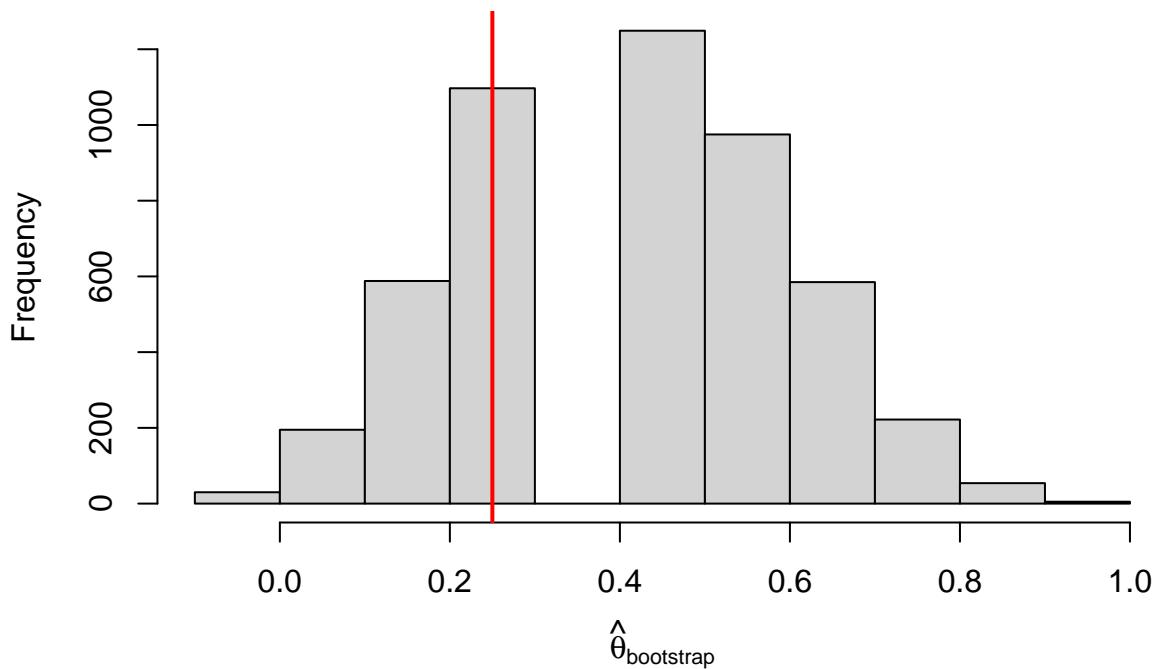
for (b in 1:B) {
  T_boot <- sample(T, size = n, replace = TRUE)
  p_hat_boot <- mean(T_boot)
  boot_est[b] <- tita_corregida(p_hat_boot, Se, Sp)
}

# Histograma
hist(boot_est, breaks = 15, freq = TRUE,
     main = "Distribución bootstrap del estimador de momentos (n = 10)",
     xlab = expression(hat(theta) ["bootstrap"]))

abline(v = tita, col = "red", lwd = 2) # valor verdadero

```

## Distribución bootstrap del estimador de momentos (n = 10)



La distribución bootstrap del estimador de momentos para  $n = 10$  muestra una alta dispersión, lo que muestra una fuerte inestabilidad del estimador cuando el tamaño muestral es chico. Además, el centro de la distribución bootstrap es mayor que el valor real de  $\theta$ , lo que indica la presencia de un sesgo positivo.

Estos resultados muestran que, con  $n = 10$ , el estimador es altamente variable y poco confiable. Por lo tanto, se recomienda aumentar el tamaño muestral para obtener una distribución más concentrada del estimador. Con un  $n$  mayor, el estimador estaría más centrado alrededor del valor real ( $\theta = 0.25$ ) y su varianza disminuiría.

### 2.2.9 Intervalos de confianza bootstrap.

Se quiere calcular intervalos de confianza bootstrap para  $\theta$  basandose en  $\hat{\theta}_{MoM}$ . A partir de la técnica bootstrap no paramétrico obtenemos 1000 estimaciones de  $\theta$  y para construir el intervalo de confianza tomamos como límite inferior al cuantil 0.025 y como límite superior al cuantil 0.975.

También se quieren realizar simulaciones, para evaluar el cubrimiento empírico y la longitud promedio de los intervalos.

En el siguiente código se realiza lo mencionado anteriormente.

```
#Definimos los valores a utilizar.
tita <- 0.25
Se <- 0.9
Sp <- 0.95
p_Y <- Se * tita + (1 - Sp) * (1 - tita)

ns <- c(10, 20, 50, 100, 1000)
R <- 700 #replicas Monte Carlo
B <- 1000 #remuestreos bootstrap

z <- qnorm(0.975)

res_list <- vector("list",length(ns))
names(res_list) <- as.character(ns)

#para cada n:
for (j in seq_along(ns)){
  n <- ns[j]
  cubrimientos <- logical(R)
  longitudes <- numeric(R)

  #repito el experimento R veces
  for (r in 1:R){
    T <- rbinom(n,1,p_Y)
    boot_est <- numeric(B)

    #genero B muestras bootstrap.
    for (b in 1:B){
      T_boot <- sample(T, size = n, replace = TRUE)
      p_hat_boot <- mean(T_boot)
      boot_est[b] <- tita_corregida(p_hat_boot, Se, Sp)
    }

    lower <- quantile(boot_est, probs = 0.025, names = FALSE)
    upper <- quantile(boot_est, probs = 0.975, names = FALSE)

    cubrimientos[r] <- (lower <= tita) && (tita <= upper)
    longitudes[r] <- (upper - lower)
  }

  #resultados para cada n
  res_list[[j]] <- list(
    n = n,
    cubrimiento = mean(cubrimientos),
```

```

    longitud_media = mean(longitudes)
  )

}

res_df <- do.call(rbind, lapply(res_list, function(z){
  data.frame(n = z$n,
             cubrimiento = z$cubrimiento,
             longitud_media = z$longitud_media)
}))

rownames(res_df) <- NULL

```

### 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  $C^1$ .

Entonces, 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.

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

En el siguiente código se calcula el intervalo asintótico mencionado y se realizan simulaciones para evaluar cubrimiento empírico y longitud promedio.

```
res_listAsint <- vector("list", length(ns))
names(res_listAsint) <- as.character(ns)

#para cada n:
for (j in seq_along(ns)){
  n <- ns[j]
  cubrimientos <- logical(R)
  longitudes <- numeric(R)

  #repito el experimento R veces.
  for (r in 1:R){
    T <- rbinom(n, 1, p_Y)

    p_hat <- mean(T)
    tita_hat <- tita_corregida(p_hat, Se, Sp)
    est_Var <- (p_hat*(1-p_hat))/(n*(Se+Sp-1)^2)

    termino <- z*sqrt(est_Var)
    lower <- tita_hat - termino
    upper <- tita_hat + termino

    cubrimientos[r] <- (lower <= tita) && (tita <= upper)
    longitudes[r] <- (upper - lower)
  }

  #resultados para cada n.
  res_listAsint[[j]] <- list(
    n = n,
    cubrimiento = mean(cubrimientos),
    longitud_media = mean(longitudes)
  )
}

resAsint_df <- do.call(rbind, lapply(res_listAsint, function(z){
  data.frame(n = z$n,
             cubrimiento = z$cubrimiento,
             longitud_media = z$longitud_media)
}))

rownames(resAsint_df) <- NULL
```

### 2.2.11 Comparación entre ambos intervalos.

Comparemos los resultados obtenidos.

```
knitr::kable(res_df, digits = 4, caption = "Bootstrap Percentil")
```

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

```
knitr::kable(resAsint_df, digits = 4, caption = "Intervalos Asintóticos")
```

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.

Recordemos que el estimador de momentos de la prevalencia se obtiene despejando  $\theta$  en la expresión

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

y reemplazando  $p$  por su estimador  $\hat{p} = T_{\text{per}}/n$ . Esto produce

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

Si bien  $\hat{p} \in [0, 1]$ , la transformación anterior es lineal y puede llevar a que el estimador resultante tome valores fuera del intervalo válido para una probabilidad. En particular,

$$\hat{\theta}_{MoM} < 0 \iff \hat{p} < 1 - Sp,$$

lo cual ocurre en muestras donde se observan menos resultados positivos de los esperados, mientras que

$$\hat{\theta}_{MoM} > 1 \iff \hat{p} > Se,$$

lo cual puede suceder cuando la muestra presenta un número inusualmente alto de resultados positivos.

Ejemplos numéricos Consideremos los valores utilizados en esta parte del trabajo:

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

Entonces,

$$1 - Sp = 0.05, \quad Se + Sp - 1 = 0.85.$$

Ejemplo 1:  $\hat{\theta}_{MoM} < 0$ .

Si en una muestra particular no se observan resultados positivos, entonces

$$\hat{p} = 0.$$

Reemplazando en la fórmula del estimador:

$$\hat{\theta}_{MoM} = \frac{0 + 0.95 - 1}{0.85} = \frac{-0.05}{0.85} \approx -0.0588.$$

En este caso, el estimador toma un valor negativo, aun cuando la prevalencia real sea positiva.

Ejemplo 2:  $\hat{\theta}_{MoM} > 1$ .

Si por azar la muestra produce un valor muy alto de  $\hat{p}$ , por ejemplo

$$\hat{p} = 0.95,$$

entonces

$$\hat{\theta}_{MoM} = \frac{0.95 + 0.95 - 1}{0.85} = \frac{0.90}{0.85} \approx 1.0588.$$

En este caso, el estimador supera el valor máximo admisible para una prevalencia.

Estos ejemplos ilustran que el estimador  $\hat{\theta}_{MoM}$  puede ubicarse fuera de  $[0, 1]$  para ciertas muestras.

### 2.3.13 Estudio del estimador truncado

Recordemos que el estimador de momentos puede tomar valores fuera del intervalo  $[0, 1]$ . Para evitar esta situación, definimos el estimador truncado como

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

Este estimador coincide con  $\hat{\theta}_{MoM}$  en la región central y recorta los valores imposibles asignándolos al borde del intervalo.

Para estudiar sus propiedades, realizamos simulaciones Monte Carlo para  $\theta = 0.25$ ,  $Se = 0.9$ ,  $Sp = 0.95$  y tamaños muestrales  $n = 10, 100$  y  $1000$ . En cada caso generamos  $N_{\text{rep}}$  réplicas independientes, calculamos  $\hat{\theta}_{\text{trunc}}$  y estimamos el sesgo, la varianza, el error cuadrático medio (ECM) y la forma empírica de la distribución.

```
# Parámetros
theta_verdadera <- 0.25
Se_fijo         <- 0.9
Sp_fijo         <- 0.95

# Probabilidad verdadera de test positivo bajo el modelo
```

```

p_verdadera <- Se_fijo * theta_verdadera + (1 - Sp_fijo) * (1 - theta_verdadera)

# Número de réplicas para Monte Carlo
N_rep <- 10000

# Funciones para los estimadores

# Estimador de momentos de theta
theta_mom <- function(T_obs, n, Se = Se_fijo, Sp = Sp_fijo) {
  p_muestral <- T_obs / n
  theta_mom_est <- (p_muestral + Sp - 1) / (Se + Sp - 1)
  return(theta_mom_est)
}

# Estimador truncado de theta
theta_trunc <- function(T_obs, n, Se = Se_fijo, Sp = Sp_fijo) {
  theta_mom_est <- theta_mom(T_obs, n, Se, Sp)
  theta_trunc_est <- pmin(1, pmax(0, theta_mom_est))
  return(theta_trunc_est)
}

## Tamaños muestrales a estudiar
n_valores <- c(10, 100, 1000)

## Data frame para almacenar resultados
resultados <- data.frame(
  n      = n_valores,
  media = NA_real_,
  sesgo = NA_real_,
  var   = NA_real_,
  ECM   = NA_real_
)

for (i in seq_along(n_valores)) {
  n <- n_valores[i]

  # Simular  $T_{obs} \sim Binomial(n, p_{verdadera})$   $N_{rep}$  veces
  T_sim <- rbinom(N_rep, size = n, prob = p_verdadera)

  # Estimaciones truncadas
  theta_trunc_sim <- theta_trunc(T_sim, n)

  # media
  media_hat <- mean(theta_trunc_sim)

  # sesgo
  sesgo_hat <- media_hat - theta_verdadera

  # varianza
  var_hat <- var(theta_trunc_sim)
}

```

```

# ECM
ECM_hat <- mean((theta_trunc_sim - theta_verdadera)^2)

# resultados
resultados$media[i] <- media_hat
resultados$sesgo[i] <- sesgo_hat
resultados$var[i] <- var_hat
resultados$ECM[i] <- ECM_hat
}

knitr::kable(resultados, digits = 9, caption = "Resultados")

```

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

**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:** Hacemos los siguientes gráficos para la distribución asintótica:

```

par(mfrow = c(2, 2))  ## grid para los gráficos

for (n in n_valores) {
  # nueva data para cada n
  T_sim <- rbinom(N_rep, size = n, prob = p_verdadera)
  theta_trunc_sim <- theta_trunc(T_sim, n)

  # Histograma
  hist(theta_trunc_sim,
        breaks = 40,
        main = paste("Histograma de theta_trunc, n =", n),

```

```

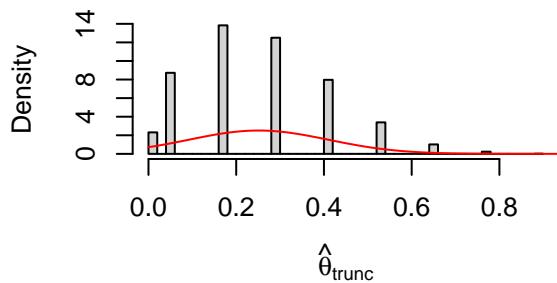
xlab = expression(hat(theta)[trunc]),
probability = TRUE)

media_hat <- mean(theta_trunc_sim)
sd_hat     <- sd(theta_trunc_sim)
x_grid     <- seq(0, 1, length.out = 200)
lines(x_grid, dnorm(x_grid, mean = media_hat, sd = sd_hat), col = "red")
}

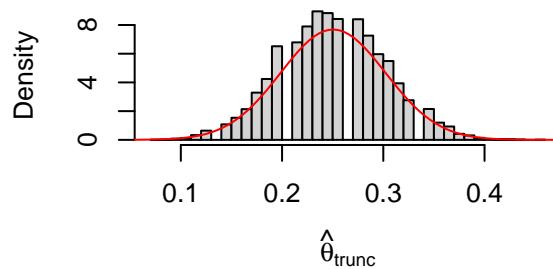
par(mfrow = c(1, 1))

```

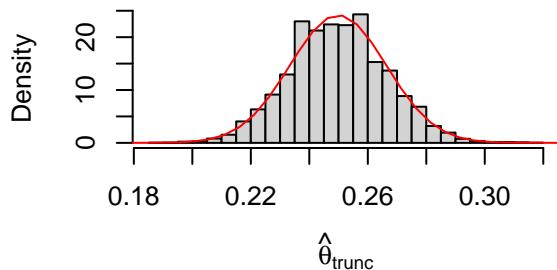
**Histograma de theta\_trunc, n = 10**



**Histograma de theta\_trunc, n = 100**



**Histograma de theta\_trunc, n = 1000**

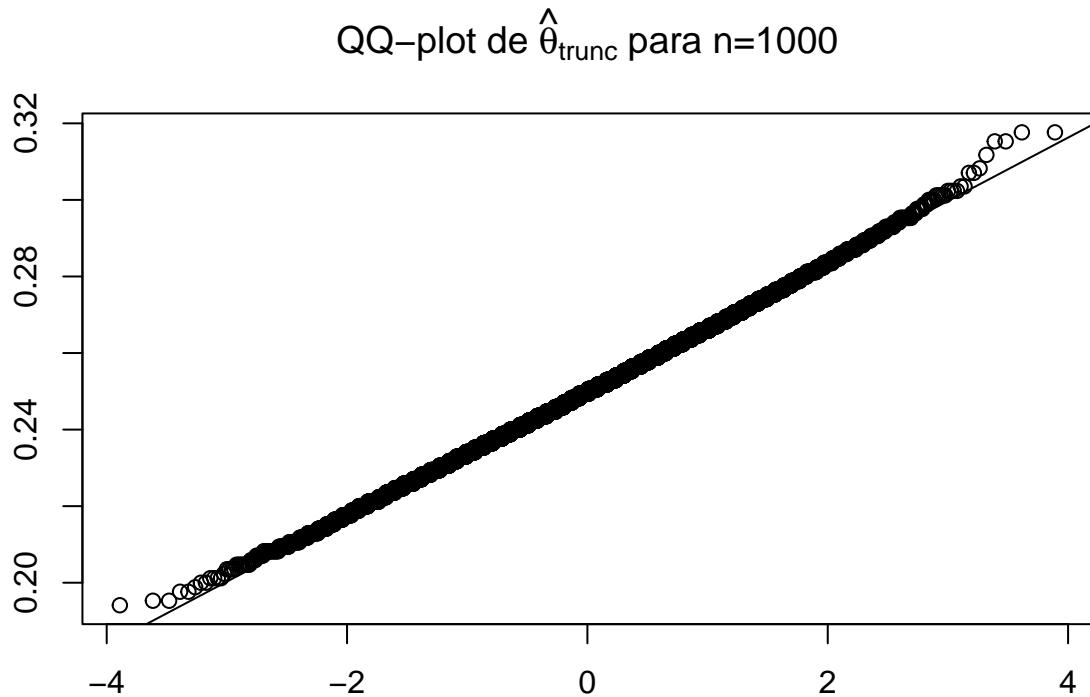


```

n <- 1000
T_sim <- rbinom(N_rep, size = n, prob = p_verdadera)
theta_trunc_sim <- theta_trunc(T_sim, n)

qqnorm(theta_trunc_sim,
       main = expression(paste("QQ-plot de ", hat(theta)[trunc], " para n=1000")),
       xlab= "", 
       ylab= "")
qqline(theta_trunc_sim)

```



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.

En conjunto, estos resultados 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.

### 3 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, basandose en el estimador  $\hat{\theta}_{MoM}$ :

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

Tenemos:

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

Y por el item 2.1.10, se deduce que:

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

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

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

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.

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

### 3.1.2 Aplicación en caso ficticio.

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

```
#Definimos los valores.
n_pre <- 100
n_post <- 100
Se <- 0.9
Sp <- 0.95
tita_pre <- 0.2
tita_post <- 0.15

#Calculamos p_pre y p_post:
p_pre <- (Se + Sp - 1)* tita_pre + (1 - Sp)
p_post <- (Se + Sp - 1)* tita_post + (1 - Sp)
```

```

#Generamos ambas muestras (pre y post):

Xpre <- rbinom(n_pre, 1, p_pre)
Xpost <- rbinom(n_post, 1, p_post)

#Ahora calculamos U(Xpre, Xpost):
p_preHat <- mean(Xpre)
p_postHat <- mean(Xpost)

tita_preHat <- (p_preHat + Sp - 1)/(Se + Sp - 1)
tita_postHat <- (p_postHat + Sp - 1)/(Se + Sp - 1)

Var_preHat <- (p_preHat*(1-p_preHat))/(n_pre*(Se+Sp-1)^2)
Var_postHat <- (p_postHat*(1-p_postHat))/(n_post*(Se+Sp-1)^2)

U <- (tita_postHat - tita_preHat)/(sqrt(Var_postHat + Var_preHat))

```

El estadístico observado tiene el siguiente valor:

U

```
## [1] -0.3414939
```

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 se achica la muestra? A continuación se realizan las mismas simulaciones variando n.

Observemos que pasa para  $n = 50$

U50

```
## [1] 1.162476
```

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.

Y ahora con  $n = 10$ :

U10

```
## [1] 0.5198752
```

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 asintotico 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:

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

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

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

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

```
# Estimador de Delta y su desvío estándar
DeltaHat <- tita_postHat - tita_preHat
se_Delta <- sqrt(Var_preHat + Var_postHat)

# Intervalo asintótico 95%
z <- qnorm(0.975)
IC_lower <- DeltaHat - z * se_Delta
IC_upper <- DeltaHat + z * se_Delta

cat("DeltaHat =", round(DeltaHat, 4), "\n")

## DeltaHat = 0.1176

cat("IC 95% para Delta: [", round(IC_lower, 4), ", ", round(IC_upper, 4), "] \n")

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

Con los datos simulados, el intervalo calculado fue:

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

Podemos ver que el intervalo incluye al 0, por lo tanto no hay evidencia suficiente para decir que hay un cambio real entre pre y post. El intervalo es bastante ancho, se puede decir que hay alta incertidumbre. Pero podemos ver tambien que el limite superior es bastante grande ( $\approx 0.50$ ) lo cual muestra que con nuestros datos todavia es posible un aumento grande entre pre y post. Y como el limite inferior es 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. Este proceso se encuentra en el código siguiente.

```
#Definimos los valores a utilizar.
Nrep <- 2000
ns <- c(10, 20, 50, 100, 1000, 10000)
alpha <- 0.05
Se <- 0.9
Sp <- 0.95
z <- qnorm(0.975)

# Como queremos nivel, definimos prevalencias iguales para pre y post ( $H_0 \delta = 0$ ).
#Además queremos el supremo de los errores de tipo 1, entonces hay que variar tita.
grid_tita <- seq(0.01, 0.99, length.out = 20)

#Vector que va a tener el nivel empírico para cada n.
nivelN <- numeric(length(ns))

#para cada n:
for (r in seq_along(ns)){
  n <- ns[r]

  #vector que va a tener error de tipo 1 para cada tita.
  errorTipoi <- numeric(length(grid_tita))

  #para cada tita:
  for (j in seq_along(grid_tita)){
    tita <- grid_tita[j]
    p <- (Se + Sp - 1)* tita + (1 - Sp)

    #vector que va a tener si rechazo o no el test en cada pos.
    rechazos <- numeric(Nrep)

    #repito el experimento Nrep veces:
    for(i in 1:Nrep){
      Xpre <- rbinom(n,1,p)
      Xpost <- rbinom(n,1,p)

      #Ahora calculamos  $U(X_{pre}, X_{post})$ :
      p_preHat <- mean(Xpre)
      p_postHat <- mean(Xpost)

      tita_preHat <- (p_preHat + Sp - 1)/(Se + Sp - 1)
      tita_postHat <- (p_postHat + Sp - 1)/(Se + Sp - 1)

      Var_preHat <- (p_preHat*(1-p_preHat))/(n*(Se+Sp-1)^2)
      Var_postHat <- (p_postHat*(1-p_postHat))/(n*(Se+Sp-1)^2)
```

```

U <- (tita_postHat - tita_preHat)/(sqrt(Var_postHat + Var_preHat))

#rechazo o no H0.
rechazos[i] <- abs(U) > z
}

#proporción de rechazos. (ignoro nan en caso de que el p estimado sea 0 o 1
#pues sucede un div por 0)
errorTipo1[j] <- mean(rechazos, na.rm = TRUE)

}

nivelN[r] <- max(errorTipo1)
}

nivelEmpirico <- data.frame(n = ns, nivel = nivelN)

knitr::kable(nivelEmpirico, digits = 4, caption = "Nivel empírico del test")

```

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.