instituto de cálculo
UBA - CONICET

Trabajo Práctico

10 de Julio de 2025 — Introducción a la Estadística y Ciencia de Datos

Integrante	LU	Correo electrónico
Suarez Ines	890/22	ine.suarez22@gmail.com
Navarro Solana	906/22	solanan3@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

1. Introducción

Se desea saber qué proporción de individuos de una población tienen una cierta característica, pero para ello hay que hacer una pregunta incómoda. Por ejemplo: ¿Qué proporción de estudiantes de la universidad se copió alguna vez en un examen? Llamemos θ a esa proporción desconocida que se desea estimar.

En este trabajo vamos a estimar θ utilizando dos métodos distintos. El primero es el método clásico, que asume que todos los encuestados responden con sinceridad. Luego, analizamos qué ocurre con él y cómo lo afecta que una parte de la población encuestada pueda mentir en su respuesta.

El segundo método es un procedimiento de respuesta aleatorizada, en el que los encuestados responden siguiendo las instrucciones de un experimento con un dado y una moneda. Esto hace que las respuestas no reflejen directamente la verdad individual, sino el resultado de un mecanismo aleatorio.

Finalmente, se compararán ambos métodos para analizar cuál tiene mejores propiedades y cuál se aproxima más al valor real de θ.

2. Análisis de los estimadores

2.1. Método clásico: elegir un conjunto de n estudiantes al azar y preguntarles si se han copiado alguna vez en un examen

(a) **Hallar el EMV de θ asumiendo que todos los encuestados dicen la verdad y estudiar sus propiedades.**

El estimador de máxima verosimilitud es aquel valor de θ que maximiza la función de verosimilitud, es decir, el valor que hace más probable haber observado los datos que efectivamente se observaron.

Dado que cada encuestado responde de forma independiente y con una respuesta binaria (1 si se copió alguna vez, 0 si no), podemos modelar las respuestas como variables aleatorias independientes con distribución de Bernoulli de parámetro θ.

Supongamos que se observa una muestra $X_1, X_2, \ldots, X_n$ de respuestas, donde cada $X_i \in \{0, 1\}$ indica si el estudiante respondió afirmativamente. La función de verosimilitud es:

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

Equivalente a:

$$L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Tomando logaritmos:

$$\log L(\theta) = \Big(\sum x_i\Big) \log\theta + \Big(n - \sum x_i\Big) \log(1-\theta)$$

Para obtener el máximo derivamos respecto de θ e igualamos a cero

$$\frac{d}{d\theta} \log L(\theta) = \Big(\sum x_i\Big) \cdot \frac{1}{\theta} - \Big(n - \sum x_i\Big) \cdot \frac{1}{1-\theta}$$

$$\frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1-\theta} = 0$$

Despejamos y obtenemos que la función se maximiza en:

$$\hat{\theta}_{\text{EMV}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Es decir, el EMV de θ es la proporción de respuestas afirmativas en la muestra.

Ahora vamos a analizar las propiedades de este estimador.

- **Sesgo:** El sesgo de un estimador es la diferencia entre su esperanza y el valor verdadero del parámetro que se desea estimar. Sea $\hat{\theta}_{\text{EMV}} = \frac{1}{n}\sum_{i=1}^{n} X_i$, donde cada $X_i \sim \text{Bernoulli}(\theta)$. Entonces:

$$\mathbb{E}[\hat{\theta}_{\text{EMV}}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\cdot\theta = \theta$$

  Por lo tanto, $\hat{\theta}_{\text{EMV}}$ es un estimador **insesgado**, ya que su esperanza coincide con el valor verdadero del parámetro.

- **ECM:** El error cuadrático medio (ECM) de un estimador se define como:

$$\text{ECM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta}-\theta)^2]$$

  En general, el ECM se puede descomponer como:

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

  Como acabamos de ver que el estimador es insesgado, el término de sesgo es cero. Por lo tanto:

$$\text{ECM}(\hat{\theta}_{\text{EMV}}) = \text{Var}(\hat{\theta}_{\text{EMV}}) = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] = \frac{1}{n^2}\cdot n\cdot\theta(1-\theta) \ = \frac{\theta(1-\theta)}{n}$$

  Es decir, el ECM en este caso coincide con la varianza del estimador. Este resultado implica que el ECM disminuye a medida que aumenta el tamaño muestral n, lo cual es una propiedad deseable.

- **Consistencia:** Un estimador es consistente si converge casi seguramente al verdadero valor del parámetro cuando el tamaño de la muestra tiende a infinito. En este caso, como $\hat{\theta}_{\text{EMV}} = \frac{1}{n}\sum_{i=1}^{n} X_i$, se trata de una media muestral de variables i.i.d. Bernoulli.
  Por la **Ley Fuerte de los Grandes Números** (LFGN), sabemos que:

$$\hat{\theta}_{\text{EMV}} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{c.s.} \mathbb{E}[X_i] = \theta$$

  cuando $n \to \infty$. Es decir, el estimador converge casi seguramente a θ, por lo tanto, es un estimador **fuertemente consistente**.

(b) **Sea π la probabilidad (desconocida) de que un estudiante que se copio en algun examen diga que no en la encuesta y supongamos que la probabilidad de que un estudiante que no se copio en ningun examen diga que si en la encuesta es cero. Hallar el sesgo y el ECM del estimador de θ hallado en el ıtem anterior. ¿Es consistente?**

En los cálculos realizados anteriormente vimos que $\hat{\theta}_{\text{EMV}}$ es un muy buen estimador, siempre y cuando todos los encuestados respondan con sinceridad. Sin embargo, ahora queremos analizar cómo se modifican estas propiedades si tenemos en cuenta que existe una probabilidad π de que un estudiante que se copió en algún examen responda que *no* en la encuesta. Es decir, los encuestados pueden mentir.

Bajo este nuevo supuesto, las respuestas observadas ya no son las verdaderas, sino que pueden estar distorsionadas. Ahora nuestras variables X_i que toman valor 1 si el encuestado responde que sí y 0 si responde que no, ya no tienen distribucion Bernoulli(θ), sino que la probabilidad de observar una respuesta afirmativa se modifica del siguiente modo:

$$\mathbb{P}(X_i = 1) = \mathbb{P}(\text{se copió y dice que sí}) + \mathbb{P}(\text{no se copió y dice que sí})$$

Como solo los estudiantes que se copiaron pueden mentir (con probabilidad π), y quienes no se copiaron siempre dicen la verdad (nunca dicen que sí por error), tenemos:

$$\mathbb{P}(X_i = 1) = (1-\pi)\theta + 0\cdot(1-\theta) = (1-\pi)\theta$$

Por lo tanto, cada X_i sigue una distribución Bernoulli de parámetro $(1-\pi)\theta$.

Vamos a analizar las propiedades del mismo estimador que teníamos antes, pero considerando que las respuestas observadas pueden estar afectadas por el hecho de que algunos encuestados mienten.

$$\hat{\theta}_{\text{EMV}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Sesgo:**
  En este caso la cuenta de la esperanza del estimador sigue la misma lógica que antes con la diferencia de que $\mathbb{E}[X_i] = (1-\pi)\theta$:

  $$\mathbb{E}[\hat{\theta}_{\text{EMV}}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\cdot(1-\pi)\theta = (1-\pi)\theta$$

  Por lo tanto, el sesgo es:

  $$\text{Sesgo} = \mathbb{E}[\hat{\theta}_{\text{EMV}}] - \theta = (1-\pi)\theta - \theta = -\pi\theta$$

  El estimador está sesgado hacia abajo. El sesgo aumenta cuanto mayor es la proporción de estudiantes que mienten (π) o cuanto mayor es θ.
- **ECM:**
  El error cuadrático medio (ECM) del estimador se calcula como:

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

  Ya que $X_i \sim \text{Bernoulli}((1-\pi)\theta)$, su varianza es:

  $$\text{Var}(X_i) = (1-\pi)\theta\cdot(1-(1-\pi)\theta)$$

  Entonces:

  $$\text{Var}(\hat{\theta}) = \frac{(1-\pi)\theta\cdot(1-(1-\pi)\theta)}{n}$$

  y el ECM queda:

  $$\text{ECM}(\hat{\theta}_{\text{EMV}}) = \frac{(1-\pi)\theta\cdot(1-(1-\pi)\theta)}{n} + (\pi\theta)^2$$

  Esto significa que el ECM está compuesto por dos partes: la varianza del estimador, que disminuye al aumentar el tamaño de la muestra n, y el cuadrado del sesgo, que permanece constante independientemente de n. Por lo tanto, aunque aumentemos el tamaño muestral, el ECM no podrá reducirse a cero debido al sesgo generado por las respuestas mentirosas. Esto implica que el estimador no puede ser perfectamente preciso en presencia de mentiras, incluso con muestras grandes.
- **Consistencia:**
  El estimador sigue siendo una media muestral, por lo tanto por LFGN:

  $$\hat{\theta}_{\text{EMV}} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{c.s.} \mathbb{E}[X_i] = (1-\pi)\theta$$

  Como la convergencia no es hacia θ sino hacia $(1-\pi)\theta$, el estimador ya no es consistente para el parámetro que se desea estimar.

2.2. Método de respuesta aleatorizada: a cada entrevistado se le entrega un dado y una moneda y se le dan las siguientes instrucciones

- Tirá el dado
- Si sale 1 o 2 tirá la moneda. Si sale cara contestá sí y si sale ceca contestá no.
- Si sale 3, 4, 5 o 6 contestá la verdad.

Queremos modelar las variables aleatorias Y_i que representan la respuesta observada de un encuestado i, donde:

$$Y_i = \begin{cases} 1 & \text{si responde "sí"} \\ 0 & \text{si responde "no"} \end{cases}$$

Sea θ la proporción verdadera de personas que se copiaron alguna vez. Entonces, la probabilidad de que un encuestado responda "sí"se puede descomponer en dos escenarios, según el resultado del dado:

- **Caso 1 (responde al azar):** Si al lanzar el dado sale 1 o 2, lo cual ocurre con probabilidad $\frac{2}{6}$, el encuestado lanza una moneda. La probabilidad de que salga cara (y por lo tanto responda "sí") es $\frac{1}{2}$. Entonces, la probabilidad total de responder "sí" en este caso es:

$$\mathbb{P}(\text{dado} = 1 \text{ o } 2) \cdot \mathbb{P}(\text{cara}) = \frac{2}{6} \cdot \frac{1}{2}$$

- **Caso 2 (responde la verdad):** Si al lanzar el dado sale 3, 4, 5 o 6, con probabilidad $\frac{4}{6}$, el encuestado responde la verdad. La probabilidad de que efectivamente se haya copiado (y por lo tanto responda "sí") es θ. Por lo tanto, la contribución de este caso a la probabilidad total de responder "sí" es:

$$\mathbb{P}(\text{dado} = 3,4,5,6) \cdot \mathbb{P}(\text{se copió}) = \frac{4}{6} \cdot \theta = \frac{2}{3} \cdot \theta$$

Sumando ambos casos, obtenemos la probabilidad total de que un encuestado responda "sí":

$$\mathbb{P}(Y_i = 1) = \underbrace{\frac{2}{6} \cdot \frac{1}{2}}_{\text{azar}} + \underbrace{\frac{2}{3} \cdot \theta}_{\text{responde la verdad}} = \frac{1}{6} + \frac{2}{3}\theta$$

Así, la variable aleatoria Y_i que representa la respuesta observada sigue una distribución:

$$Y_i \sim \text{Bernoulli}\left(\frac{1}{6} + \frac{2}{3}\theta\right)$$

(a) **Hallar el estimador de momentos de θ. Calcular su sesgo y su ECM. ¿Es consistente?**

El método de los momentos consiste en igualar la esperanza de Y_i con la media muestral:

$$\mathbb{E}[Y_i] = \frac{1}{6} + \frac{2}{3}\theta \quad \Rightarrow \quad \bar{Y}_n = \frac{1}{6} + \frac{2}{3}\hat{\theta}_{\text{MM}}$$

Despejando, obtenemos el estimador de momentos:

$$\hat{\theta}_{\text{MM}} = \frac{3}{2}\bar{Y}_n - \frac{1}{4}$$

- **Sesgo:**

  Calculamos el sesgo como:

$$\text{Sesgo}(\hat{\theta}_{\text{MM}}) = \mathbb{E}[\hat{\theta}_{\text{MM}}] - \theta$$

  Usando linealidad de la esperanza:

$$\mathbb{E}[\hat{\theta}_{\text{MM}}] = \frac{3}{2}\mathbb{E}[\bar{Y}_n] - \frac{1}{4} = \frac{3}{2}\left(\frac{1}{6} + \frac{2}{3}\theta\right) - \frac{1}{4} = \theta$$

  Entonces, el estimador es **insesgado**.

- **ECM:**
  Dado que es insesgado, el ECM coincide con la varianza:

$$\text{ECM}(\hat{\theta}_{\text{MM}}) = \text{Var}(\hat{\theta}_{\text{MM}}) = \left(\frac{3}{2}\right)^2 \cdot \text{Var}(\bar{Y}_n) = \frac{9}{4} \cdot \frac{1}{n} \cdot \text{Var}(Y_i)$$

  Como $Y_i \sim \text{Bernoulli}(p)$ con $p = \frac{1}{6} + \frac{2}{3}\theta$, tenemos:

$$\text{Var}(Y_i) = p(1-p)$$

  Por lo tanto, el ECM del estimador de momentos es:

$$\text{ECM}(\hat{\theta}_{\text{MM}}) = \frac{9}{4n} \cdot p(1-p) \quad \text{donde} \quad p = \frac{1}{6} + \frac{2}{3}\theta$$

  Expresado directamente en función de θ:

$$\text{ECM}(\hat{\theta}_{\text{MM}}) = \frac{9}{4n} \cdot \left(\frac{1}{6} + \frac{2}{3}\theta\right)\left(\frac{5}{6} - \frac{2}{3}\theta\right)$$

  Nuevamente, el ECM coincide con la varianza del estimador. Este resultado implica que el ECM disminuye a medida que aumenta el tamaño muestral n, lo cual es una propiedad deseable.

- **Consistencia:**
  Queremos demostrar que el estimador de momentos

$$\hat{\theta}_{\text{MM}} = \frac{3}{2}\bar{Y}_n - \frac{1}{4}$$

  es consistente para θ.
  Primero, observamos que este estimador es una función continua de la media muestral $\bar{Y}_n$. Es decir:

$$\hat{\theta}_{\text{MM}} = h(\bar{Y}_n) \quad \text{con} \quad h(x) = \frac{3}{2}x - \frac{1}{4}$$

  Sabemos que $\bar{Y}_n$ es la media de variables i.i.d. con distribución Bernoulli, por lo tanto, como vimos antes, por la Ley fuerte de los Grandes Números:

$$\bar{Y}_n \xrightarrow{c.s.} \mathbb{E}[Y_i]$$

  Ya demostramos que:

$$\mathbb{E}[Y_i] = \frac{1}{6} + \frac{2}{3}\theta$$

  Aplicamos la continuidad de h (la composición de funciones continuas preserva la convergencia) y obtenemos:

$$\hat{\theta}_{\text{MM}} = h(\bar{Y}_n) \xrightarrow{c.s.} h\left(\frac{1}{6} + \frac{2}{3}\theta\right)$$

  Ahora hacemos las cuentas:

$$h\left(\frac{1}{6} + \frac{2}{3}\theta\right) = \frac{3}{2}\left(\frac{1}{6} + \frac{2}{3}\theta\right) - \frac{1}{4} = \theta$$

  Por lo tanto:

$$\hat{\theta}_{\text{MM}} \xrightarrow{c.s.} \theta$$

  Esto demuestra que $\hat{\theta}_{\text{MM}}$ es un estimador **fuertemente consistente** para θ.

(b) **Observar que el estimador de momentos puede dar fuera del [0, 1] y corregirlo para que esto no ocurra.**

El estimador de momentos que obtuvimos es:

$$\hat{\theta}_{\text{MM}} = \frac{3}{2}\bar{Y}_n - \frac{1}{4}$$

Sin embargo, como $\bar{Y}_n$ toma valores en $[0, 1]$, el estimador $\hat{\theta}_{\text{MM}}$ puede tomar valores fuera del intervalo válido $[0, 1]$ para una proporción. Por ejemplo:

- Si todos los encuestados responden "sí", entonces $\bar{Y}_n = 1$, y:

$$\hat{\theta}_{\text{MM}} = \frac{3}{2} \cdot 1 - \frac{1}{4} = \frac{5}{4} > 1$$

- Si todos responden "no", entonces $\bar{Y}_n = 0$, y:

$$\hat{\theta}_{\text{MM}} = \frac{3}{2} \cdot 0 - \frac{1}{4} = -\frac{1}{4} < 0$$

Esto muestra que $\hat{\theta}_{\text{MM}}$ no garantiza valores dentro del intervalo $[0, 1]$, lo cual no es admisible para una proporción.

Estimador corregido:

Para garantizar que el estimador esté en $[0, 1]$, definimos una versión corregida mediante una función partida:

$$\hat{\theta}_{\text{MOD}} = \begin{cases} 0 & \text{si } \hat{\theta}_{\text{MM}} < 0 \\ \hat{\theta}_{\text{MM}} & \text{si } 0 \leq \hat{\theta}_{\text{MM}} \leq 1 \\ 1 & \text{si } \hat{\theta}_{\text{MM}} > 1 \end{cases}$$

Esta corrección asegura que el valor estimado de θ siempre esté dentro del rango válido para una proporción. Aunque este estimador ya no es insesgado, es más realista en contextos prácticos y puede ser preferible especialmente en muestras pequeñas.

(c) **Hallar el EMV de θ. Verificar que coincide con un estimador de momentos "corregido". Probar que es consistente pero no insesgado. ¿Cuanto vale el sesgo si n = 2 y $\theta = 1/4$?**

Dado un vector de respuestas observadas $(y_1, y_2, \ldots, y_n)$, y sabiendo que cada Y_i sigue una distribución:

$$Y_i \sim \text{Bernoulli}\left(\frac{1}{6} + \frac{2}{3}\theta\right)$$

la función de verosimilitud es:

$$L(\theta) = \prod_{i=1}^{n} \left[\left(\frac{1}{6} + \frac{2}{3}\theta\right)^{y_i} \cdot \left(1 - \frac{1}{6} - \frac{2}{3}\theta\right)^{1-y_i} \right]$$

Usando propiedades de las potencias, podemos agrupar la productoria en una sola expresión:

$$L(\theta) = \left(\frac{1}{6} + \frac{2}{3}\theta\right)^{\sum_{i=1}^{n} y_i} \cdot \left(\frac{5}{6} - \frac{2}{3}\theta\right)^{\sum_{i=1}^{n}(1-y_i)}$$

Aplicamos logaritmo:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} y_i \cdot \log\left(\frac{1}{6} + \frac{2}{3}\theta\right) + \sum_{i=1}^{n}(1 - y_i) \cdot \log\left(\frac{5}{6} - \frac{2}{3}\theta\right)$$

El $\hat{\theta}_{\text{EMV}}$ será el valor de θ que maximice nuestra función. Dado que θ es una probabilidad, sabemos que vamos a buscar dentro del conjunto [0,1], ya que son los valores dónde tiene sentido que viva nuestro estimador. Entonces derivamos e igualamos a 0:

$$\ell'(\theta) = \sum_{i=1}^{n} y_i \cdot \frac{\frac{2}{3}}{\frac{1}{6} + \frac{2}{3}\theta} - \sum_{i=1}^{n} (1 - y_i) \cdot \frac{\frac{2}{3}}{\frac{5}{6} - \frac{2}{3}\theta}$$

Definimos nuevamente $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$, entonces:

$$\ell'(\theta) = \frac{2n}{3} \left[\frac{\bar{Y}_n}{\frac{1}{6} + \frac{2}{3}\theta} - \frac{1 - \bar{Y}_n}{\frac{5}{6} - \frac{2}{3}\theta} \right]$$

Igualamos la derivada a cero:

$$\ell'(\theta) = 0 \quad \Longleftrightarrow \quad \frac{\bar{Y}_n}{\frac{1}{6} + \frac{2}{3}\theta} = \frac{1 - \bar{Y}_n}{\frac{5}{6} - \frac{2}{3}\theta}$$

Esta es la ecuación que se resuelve para obtener el estimador de máxima verosimilitud.

Despejamos θ y obtenemos:

$$\theta = -\frac{3}{2} \left(\frac{1}{6} - \bar{Y}_n \right) = \frac{3}{2} \bar{Y}_n - \frac{1}{4} = \hat{\theta}_{\text{MM}}$$

Pero como maximizamos dentro del intervalo $[0, 1]$, el EMV final se define como:

$$\hat{\theta}_{\text{EMV}} = \hat{\theta}_{\text{MOD}} = \begin{cases} 0 & \text{si } \hat{\theta}_{\text{MM}} < 0 \\ \hat{\theta}_{\text{MM}} & \text{si } 0 \leq \hat{\theta}_{\text{MM}} \leq 1 \\ 1 & \text{si } \hat{\theta}_{\text{MM}} > 1 \end{cases}$$

Entonces, podemos concluir que el $\hat{\theta}_{\text{EMV}}$ coincide con el $\hat{\theta}_{\text{MOD}}$.

Ahora vamos a estudiar sus propiedades.

- **Consistencia:**
  Como cada Y_i es una variable aleatoria acotada y con esperanza finita, por la **ley fuerte de los grandes números** se cumple que:

  $$\bar{Y}_n \xrightarrow{c.s.} \mathbb{E}[Y_i]$$

  En este caso, ya sabemos que:

  $$\mathbb{E}[Y_i] = \frac{1}{6} + \frac{2}{3}\theta$$

  Entonces:

  $$\bar{Y}_n \xrightarrow{c.s.} \frac{1}{6} + \frac{2}{3}\theta$$

  Aplicando la función continua $f(x) = \frac{3}{2}(x - \frac{1}{6})$, que es la transformación usada para obtener el estimador, se mantiene la convergencia casi segura (porque toda función continua preserva los límites bajo convergencia casi segura):

  $$\hat{\theta}_{\text{EMV}} = \frac{3}{2} \left(\bar{Y}_n - \frac{1}{6} \right) \xrightarrow{c.s.} \frac{3}{2} \left(\frac{1}{6} + \frac{2}{3}\theta - \frac{1}{6} \right) = \theta$$

  Por lo tanto, concluimos que:

  $$\hat{\theta}_{\text{EMV}} \xrightarrow{c.s.} \theta$$

  lo cual implica que el estimador es **fuertemente consistente** o sea que para muestras de tamaño grande va a coincidir con θ y va a pertenecer al intervalo [0,1] por si solo.

- **Sesgo:**
  El EMV no es insesgado, ya que al truncar valores fuera de $[0,1]$, se introduce una distorsión en la esperanza. Vamos a verlo en el siguiente ejemplo.

  Cálculo del sesgo para $n = 2$ y $\theta = \frac{1}{4}$:
  Sabemos que:
  $$\mathbb{P}(Y_i = 1) = \frac{1}{6} + \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

  Podemos tener 3 situaciones:
  - Ambos encuestados responden que no, el promedio es igual a 0 y el estimador da < 0
  - Un encuestado responde que sí y el otro que no, el promedio es igual a $1/2$ y el estimador da $1/2$
  - Ambos encuestados responden que sí, el promedio es igual a 1 y el estimador da > 1

  Entonces, los posibles valores de $\bar{Y}_2$ (media muestral con $n = 2$) son:
  - 0 con probabilidad $(1 - \frac{1}{3})^2 = \left(\frac{2}{3}\right)^2 = \frac{4}{9}$
  - $\frac{1}{2}$ con probabilidad $2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}$
  - 1 con probabilidad $\left(\frac{1}{3}\right)^2 = \frac{1}{9}$

  Aplicamos la definición del EMV por tramos:

  $$\hat{\theta}_{\text{EMV}} = \begin{cases} 0 & \text{si } \bar{Y}_2 = 0 \\ \frac{3}{2} \cdot \frac{1}{2} - \frac{1}{4} = \frac{3}{4} - \frac{1}{4} = \frac{1}{2} & \text{si } \bar{Y}_2 = \frac{1}{2} \\ 1 & \text{si } \bar{Y}_2 = 1 \end{cases}$$

  Entonces, la esperanza del EMV es:

  $$\mathbb{E}[\hat{\theta}_{\text{EMV}}] = 0 \cdot \frac{4}{9} + \frac{1}{2} \cdot \frac{4}{9} + 1 \cdot \frac{1}{9} = \frac{2}{9} + \frac{1}{9} = \frac{1}{3}$$

  Finalmente, el sesgo es:

  $$\text{Sesgo}(\hat{\theta}_{\text{EMV}}) = \mathbb{E}[\hat{\theta}_{\text{EMV}}] - \theta = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

  Por lo tanto, el EMV tiene sesgo positivo cuando $n = 2$ y $\theta = \frac{1}{4}$.

(d) **Graficar el sesgo del EMV en función del tamaño muestral n suponiendo que $\theta = \frac{1}{4}$.**

Como el sesgo teórico del estimador depende de su esperanza, y esta no siempre puede calcularse de forma exacta (especialmente si el estimador es no lineal o truncado), estimamos el sesgo de forma empírica mediante simulaciones.

Para cada valor de $n \in \{1, 2, \ldots, 100\}$, se generan $B = 1000$ réplicas independientes de muestras de tamaño n, y se calcula el estimador $\hat{\theta}$ en cada una de ellas. Luego, se promedia ese conjunto de estimaciones para aproximar la esperanza de $\hat{\theta}$:

$$\mathbb{E}[\hat{\theta}] \approx \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$$

Con esta aproximación se estima el sesgo como:

$$\widehat{\text{Sesgo}}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)} - \theta$$

Análisis:

- Para tamaños muestrales pequeños (n mas bajo), el sesgo del EMV corregido presenta una mayor variabilidad y puede alejarse mucho de cero. Esto se debe principalmente a dos factores:

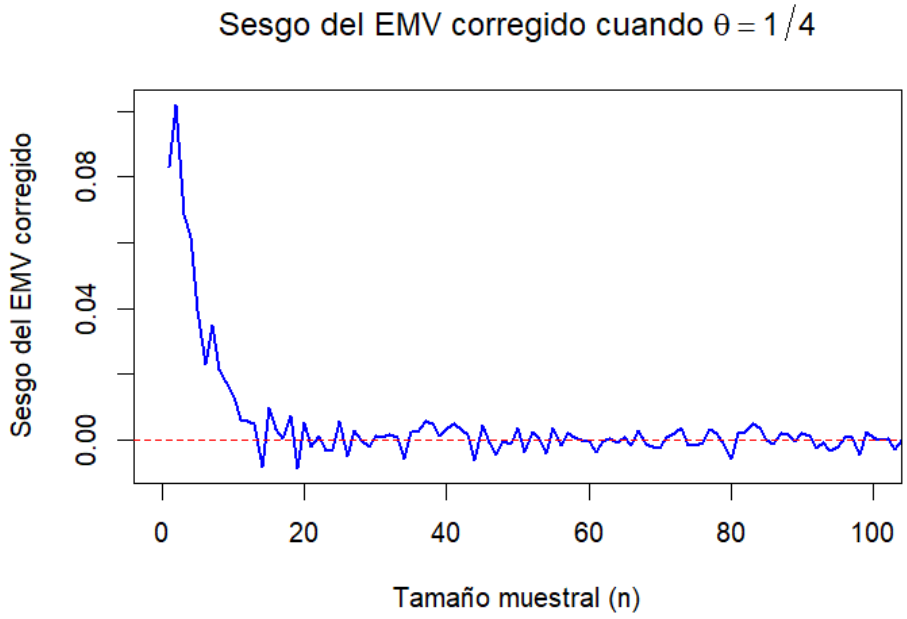

Figura 1: Sesgo EMV

- El **efecto de truncamiento**: al corregir el estimador para que esté en el intervalo $[0, 1]$, se distorsionan los valores extremos, introduciendo sesgo.
- La **alta varianza muestral**: con pocas observaciones, el promedio $\bar{Y}_n$ puede fluctuar considerablemente, lo que afecta el comportamiento del estimador.

- A medida que el tamaño muestral crece ($n \to \infty$), el sesgo del EMV corregido converge rápidamente hacia cero. Esto concuerda con el hecho de que el EMV es un estimador **fuertemente consistente**: aunque sesgado en muestras finitas, es **asintóticamente insesgado**.
- Este comportamiento respalda el uso del EMV corregido en aplicaciones prácticas con muestras moderadas o grandes, ya que garantiza que la estimación se aproxima a θ sin necesidad de conocer π, a diferencia del método clásico.

(e) **Graficar el ECM del EMV en funcion del tamaño muestral n suponiendo que $\theta = 1/4$. Superponer un grafico del ECM del EMV calculado utilizando el metodo clasico suponiendo que la proporcion de estudiantes que se copio alguna vez en un examen y miente cuando se le hace la pregunta directa es $\pi = 1/3$.**

En este item se analiza el comportamiento del error cuadrático medio (ECM) del estimador de máxima verosimilitud corregido (EMV) obtenido mediante el método de respuesta aleatorizada, en comparación con el ECM del estimador clásico basado en una pregunta directa. Se asume que la proporción real de estudiantes que se copiaron alguna vez es $\theta = \frac{1}{4}$, y que la proporción de aquellos que mienten ante una pregunta directa es $\pi = \frac{1}{3}$.

Para cada tamaño muestral $n \in \{1, 2, \ldots, 300\}$, se estimó empíricamente el ECM del EMV corregido a partir de $B = 1000$ simulaciones, usando la fórmula:

$$\widehat{\text{ECM}} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \theta)^2$$

El ECM del estimador clásico se calculó teóricamente usando:

$$\text{ECM}_{\text{clásico}} = \frac{\theta(1-\pi)[1-\theta(1-\pi)]}{n} + (\theta\pi)^2$$

Análisis:

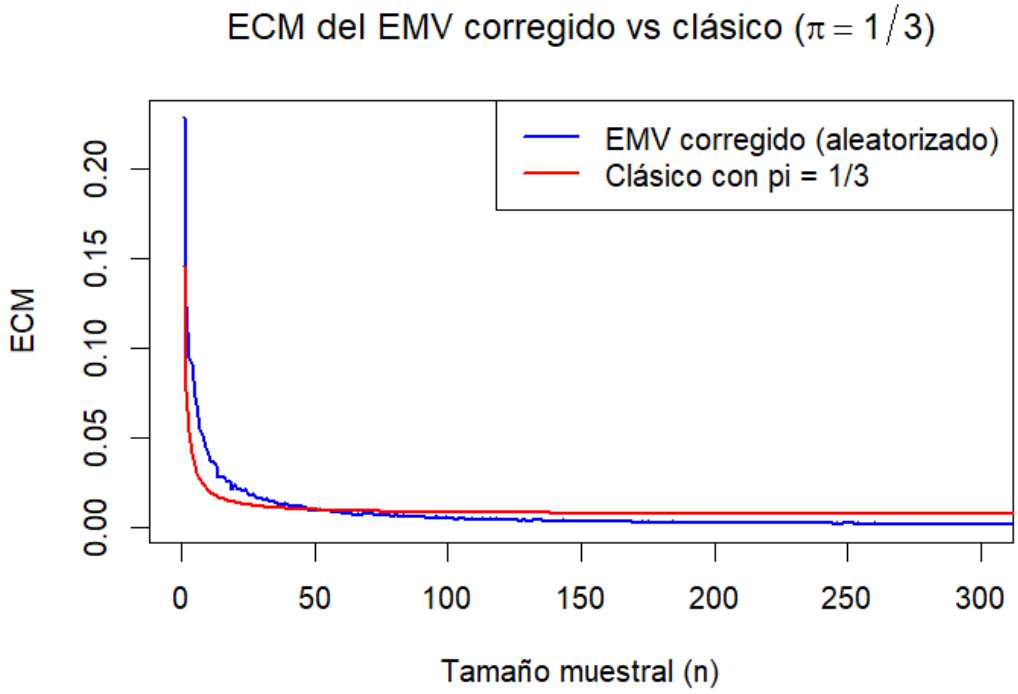

Figura 2: ECM superpuesto

- Para tamaños muestrales pequeños ($n \lesssim 30$), el estimador clásico presenta un ECM menor al del EMV aleatorizado. Esto ocurre porque, aunque el estimador clásico es sesgado (por la presencia de mentiras en las respuestas), su varianza es más baja. En cambio, el método aleatorizado incorpora variabilidad adicional por el uso de mecanismos aleatorios (dado y moneda), lo que incrementa la varianza inicial.
- A medida que n crece, el ECM de ambos métodos disminuye. En particular, el ECM del EMV se reduce rápidamente y se aproxima al del estimador clásico, confirmando así su **consistencia**. Esto es coherente con los resultados teóricos: ambos estimadores tienden a cero a medida que aumenta el tamaño muestral.
- Sin embargo, es importante remarcar que la eficiencia aparente del método clásico en muestras pequeñas depende de que en la simulación se conoce el valor exacto de $\pi = \frac{1}{3}$. En aplicaciones reales, π suele ser desconocido y no se puede corregir el sesgo con precisión.
- Por esta razón, el método de respuesta aleatorizada —aunque menos eficiente para muestras pequeñas— es más robusto en contextos reales. Además, garantiza mayor veracidad en las respuestas.

(f) **Supongamos que se realiza una campaña de concientizacion para reducir la copia. Se quiere saber si luego de ella, la proporcion que se copia es distinta de 1/4. Despues de la encuesta se entrevista a un grupo de 100 estudiantes elegidos al azar de la poblacion de interes mediante el metodo de respuesta aleatorizada y se obtiene que un 20 % contesta que si. Hallar el test del cociente de maxima verosimilitud. Concluir. Deducir un intervalo de confianza para θ. Si conocen otros metodos para hallar intervalos de confianza, aplıquenlos y comparen los resultados.**

Planteamos las siguientes hipótesis:

$$H_0 : \theta = \frac{1}{4} \quad \text{vs.} \quad H_1 : \theta \neq \frac{1}{4}$$

La función de verosimilitud para una muestra $y_1, \ldots, y_n$, con $Y_i \sim \text{Bernoulli}(\frac{1}{6} + \frac{2}{3}\theta)$, es:

$$L(\theta) = \prod_{i=1}^{n} \left(\frac{1}{6} + \frac{2}{3}\theta\right)^{y_i} \cdot \left(\frac{5}{6} - \frac{2}{3}\theta\right)^{1-y_i}$$

El cociente de verosimilitud es:

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta}_{\text{EMV}})}$$

Bajo H_0, se tiene que:

$$-2\log\Lambda \xrightarrow{d} \chi^2_j \quad \text{con } j = \dim(\Theta) - \dim(\Theta_0) = 1 - 0 = 1$$

Aplicamos logaritmo a la razón de verosimilitudes:

$$-2\log\Lambda = -2\left[\log L(\theta_0) - \log L(\hat{\theta}_{\text{EMV}})\right]$$

Utilizando la expresión de la log-verosimilitud:

$$\log L(\theta) = n \cdot \left[\bar{Y}_n \log\left(\frac{1}{6} + \frac{2}{3}\theta\right) + (1 - \bar{Y}_n)\log\left(\frac{5}{6} - \frac{2}{3}\theta\right)\right]$$

Entonces el estadístico queda:

$$-2\log\Lambda = -2n\left[\bar{Y}_n \cdot \log\left(\frac{\frac{1}{6} + \frac{2}{3}\theta_0}{\frac{1}{6} + \frac{2}{3}\hat{\theta}_{\text{EMV}}}\right) + (1 - \bar{Y}_n) \cdot \log\left(\frac{\frac{5}{6} - \frac{2}{3}\theta_0}{\frac{5}{6} - \frac{2}{3}\hat{\theta}_{\text{EMV}}}\right)\right]$$

Dado que:

$$\bar{Y}_n = 0{,}20 \quad \hat{\theta}_{\text{EMV}} = \frac{3}{2} \cdot 0{,}20 - \frac{1}{4} = 0{,}05 \quad \theta_0 = \frac{1}{4}$$

Calculamos los argumentos de los logaritmos con nuestros valores de θ_0 y $\hat{\theta}_{\text{EMV}}$:

$$\frac{1}{6} + \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$
$$\frac{1}{6} + \frac{2}{3} \cdot 0{,}05 = \frac{1}{6} + \frac{1}{30} = \frac{6}{30} + \frac{1}{30} = \frac{7}{30}$$

Entonces:

$$-2\log\Lambda = -2 \cdot 100 \cdot \left[0{,}20 \cdot \log\left(\frac{1/3}{7/30}\right) + 0{,}80 \cdot \log\left(\frac{2/3}{23/30}\right)\right]$$

Calculamos:

$$\frac{1/3}{7/30} = \frac{10}{7}, \quad \log\left(\frac{10}{7}\right) \approx 0{,}3567$$
$$\frac{2/3}{23/30} = \frac{20}{23}, \quad \log\left(\frac{20}{23}\right) \approx -0{,}1398$$

Reemplazando:

$$-2\log\Lambda = -2 \cdot 100 \cdot [0{,}20 \cdot 0{,}3567 + 0{,}80 \cdot (-0{,}1398)]$$
$$= -200 \cdot [0{,}0713 - 0{,}1118] = -200 \cdot (-0{,}0405) = 8{,}10$$

Por la distribución asintótica de nuestro estadístico, la constante de nuestro test para un nivel de significancia $\alpha = 0{,}05$ es:

$$\chi^2_{1,\,0{,}95} \approx 3{,}84$$

Dado que:

$$-2\log\Lambda = 8{,}10 > 3{,}84$$

se rechaza la hipótesis nula al nivel de significancia del 5 %.

Por lo tanto, concluimos que hay suficiente evidencia estadística para rechazar la hipótesis de que la proporción de estudiantes que se copian es igual a $\theta = 1/4$. Esto sugiere que, tras la campaña de concientización, la proporción puede haber cambiado.

Intervalo de confianza:

Por la dualidad entre test de hipótesis e intervalo de confianza, un intervalo de confianza al nivel $1-\alpha$ para θ está dado por el conjunto de valores de $\theta \in (0,1)$ tales que no rechazamos H_0:

$$-2\left[\log L(\theta) - \log L(\hat{\theta})\right] \leq \chi^2_{1,\,1-\alpha}$$

Despejando θ de esta expresión se consigue el intervalo deseado. En este caso no pudimos hacer las cuentas manualmente por los logaritmos y los exponentes.

Intervalo de confianza para θ usando el TCL y el Teorema de Slutsky

Otro método para hallar un intervalo de confianza para θ es el siguiente: Partimos del Teorema Central del Límite aplicado a la media muestral $\bar{Y}_n$, donde cada $Y_i \sim \text{Bernoulli}(p)$, con

$$p = \frac{1}{6} + \frac{2}{3}\theta.$$

Entonces se cumple que:

$$\frac{\bar{Y}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Definimos una función continua $h(x)$ que representa la raíz de la varianza de $\hat{\theta}_n$:

$$h(x) = \sqrt{\left(\frac{1}{6} + \frac{2}{3}x\right)\left(1 - \frac{1}{6} - \frac{2}{3}x\right)}.$$

Como $\hat{\theta}_n \xrightarrow{c.s.} \theta$ (es un estimador consistente) y $h(x)$ es continua en $[0,1]$, se cumple que:

$$\frac{h(\theta)}{h(\hat{\theta}_n)} \xrightarrow{c.s.} 1.$$

Aplicación del Teorema de Slutsky:

Sea $X_n \xrightarrow{d} X$ y $Y_n \xrightarrow{P} c$, con $c \in \mathbb{R}$. Entonces:

$$X_n Y_n \xrightarrow{d} X \cdot c.$$

Aplicando esto a nuestro caso:

$$\frac{\bar{Y}_n - p}{\frac{1}{\sqrt{n}}h(\hat{\theta}_n)} = \left(\frac{\bar{Y}_n - p}{\frac{1}{\sqrt{n}}h(\theta)}\right) \cdot \frac{h(\theta)}{h(\hat{\theta}_n)} \xrightarrow{d} \mathcal{N}(0,1) \cdot 1 = \mathcal{N}(0,1),$$

Es decir,

$$\boxed{\frac{\bar{Y}_n - p}{\frac{1}{\sqrt{n}} h(\hat{\theta}_n)} \xrightarrow{d} \mathcal{N}(0,1)}.$$

Construcción del intervalo de confianza:
Sea $z_{1-\alpha/2}$ el cuantil correspondiente de la normal estándar. Se tiene:

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\bar{Y}_n - p}{\frac{1}{\sqrt{n}} h(\hat{\theta}_n)} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

Multiplicando por $\frac{1}{\sqrt{n}} h(\hat{\theta}_n)$:

$$\mathbb{P}\left(\bar{Y}_n - z_{1-\alpha/2} \cdot \frac{1}{\sqrt{n}} h(\hat{\theta}_n) \leq p \leq \bar{Y}_n + z_{1-\alpha/2} \cdot \frac{1}{\sqrt{n}} h(\hat{\theta}_n)\right) \approx 1 - \alpha.$$

Como $p = \frac{1}{6} + \frac{2}{3}\theta$, se despeja θ:

$$\theta = \frac{3}{2}\left(p - \frac{1}{6}\right).$$

Aplicando esto a los extremos del intervalo:

$$\left[\frac{3}{2}\left(\bar{Y}_n - z_{1-\alpha/2} \cdot \frac{1}{\sqrt{n}} h(\hat{\theta}_n) - \frac{1}{6}\right),\ \frac{3}{2}\left(\bar{Y}_n + z_{1-\alpha/2} \cdot \frac{1}{\sqrt{n}} h(\hat{\theta}_n) - \frac{1}{6}\right)\right].$$

Pero como:
tenemos finalmente que el intervalo de confianza asintótico para θ es:

$$\boxed{\left[\hat{\theta}_n \pm \frac{3}{2} \cdot z_{1-\alpha/2} \cdot \frac{1}{\sqrt{n}} h(\hat{\theta}_n)\right]}$$

donde:

$$h(\hat{\theta}_n) = \sqrt{\left(\frac{1}{6} + \frac{2}{3}\hat{\theta}_n\right)\left(1 - \frac{1}{6} - \frac{2}{3}\hat{\theta}_n\right)} \quad \text{y} \quad \hat{\theta}_n = \frac{3}{2}\left(\bar{Y}_n - \frac{1}{6}\right)$$

Datos:

- $\bar{Y}_n = 0{,}2$
- $n = 100$
- Nivel de confianza: $1 - \alpha = 0{,}95 \Rightarrow z_{1-\alpha/2} = z_{0{,}975} \approx 1{,}96$

Paso 1: cálculo de $\hat{\theta}_n$:

$$\hat{\theta}_n = \frac{3}{2}\left(\bar{Y}_n - \frac{1}{6}\right) = \frac{3}{2}\left(0{,}2 - \frac{1}{6}\right) = \frac{3}{2} \cdot \frac{1}{30} = \frac{1}{20} = 0{,}05$$

Paso 2: cálculo de $h(\hat{\theta}_n)$:

$$h(\hat{\theta}_n) = \sqrt{\left(\frac{1}{6} + \frac{2}{3} \cdot \hat{\theta}_n\right)\left(1 - \frac{1}{6} - \frac{2}{3} \cdot \hat{\theta}_n\right)}$$

Sustituyendo $\hat{\theta}_n = 0{,}05$:

$$\frac{1}{6} + \frac{2}{3} \cdot 0{,}05 = \frac{1}{6} + \frac{1}{30} = \frac{6}{30} = 0{,}2$$

$$1 - \frac{1}{6} - \frac{2}{3} \cdot 0{,}05 = 1 - \frac{1}{6} - \frac{1}{30} = \frac{30 - 5 - 1}{30} = \frac{24}{30} = 0{,}8$$

Entonces:

$$h(\hat{\theta}_n) = \sqrt{0{,}2 \cdot 0{,}8} = \sqrt{0{,}16} = 0{,}4$$

Paso 3: construcción del intervalo:

$$\hat{\theta}_n \pm \frac{3}{2} \cdot z_{1-\alpha/2} \cdot \frac{1}{\sqrt{n}} \cdot h(\hat{\theta}_n) = 0{,}05 \pm \frac{3}{2} \cdot 1{,}96 \cdot \frac{1}{10} \cdot 0{,}4$$

$$= 0{,}05 \pm \frac{3}{2} \cdot 1{,}96 \cdot 0{,}04 = 0{,}05 \pm 1{,}96 \cdot 0{,}06 = 0{,}05 \pm 0{,}1176$$

$$\Rightarrow \boxed{[-0{,}0676,\ 0{,}1676]}$$

Observación: El límite inferior es negativo, lo cual puede no tener sentido práctico porque $\theta \in [0, 1]$. En ese caso se puede truncar a 0:

$$[0,\ 0{,}1676]$$

Usando el test llegamos a la conclusión de que hay suficiente evidencia estadística para decir que $\theta \neq \frac{1}{4} = 0,25$, por lo tanto, el hecho de que 0,25 no pertenezca al intervalo tiene sentido para este problema.

(g) **Simular ambos experimentos en R suponiendo $\theta = 1/4$ y p = 1/3, es decir, generar las respuestas de los encuestados por el metodo clasico y por el metodo de respuesta aleatorizada, para n = 10, n = 100 y n = 1000 y calcular el ECM empírico. Presentar los resultados en una tabla y describirlos, verificando que son compatibles con los resultados teoricos hallados en los ıtems anteriores.**

Se simularon ambos métodos de recolección de datos (método clásico y método de respuesta aleatorizada) bajo los parámetros:

- $\theta = \frac{1}{4}$ (proporción real de la característica)
- $\pi = \frac{1}{3}$ (probabilidad de mentir en el método clásico)

Para cada método se generaron múltiples réplicas de muestras de tamaño $n = 10$, 100 y 1000, y se calculó el **Error Cuadrático Medio (ECM)** empírico de los estimadores.

n	ECM_aleatorizado	ECM_clasico
10	0.043128	0.020833
100	0.005058	0.008333
1000	0.000503	0.007083

Figura 3: ECM para distintos valores de n

Análisis:

- Para $n = 10$, el método clásico tiene menor ECM que el aleatorizado. Esto es esperable, ya que aunque el método clásico puede tener sesgo (por el parámetro π), el estimador tiene menor varianza que el aleatorizado en muestras pequeñas.

- Para $n = 100$, el método aleatorizado mejora notablemente su precisión, con un ECM menor que el clásico. Esto refleja su consistencia: a medida que aumenta el tamaño muestral, el estimador aleatorizado se aproxima mejor a θ.
- Para $n = 1000$, el ECM del método aleatorizado sigue disminuyendo, alcanzando un valor muy pequeño. En cambio, el ECM del método clásico se estabiliza debido a que su sesgo no desaparece, ya que $\pi = \frac{1}{3}$ implica que un tercio de quienes se copiaron mienten, introduciendo un sesgo constante que no se reduce con más muestras.

Estos resultados son completamente coherentes con el análisis teórico: el estimador clásico es más preciso para muestras pequeñas, pero está sesgado y no es consistente. El estimador aleatorizado es insesgado y consistente, y su ECM disminuye al aumentar n.

(h) **Hallar un intervalo de confianza para θ por el metodo bootstrap y compararlo con los intervalos hallados en el ıtem (e).**

Para construir un intervalo de confianza para θ utilizamos el **método Bootstrap percentil**, basado en simulaciones a partir de la muestra observada.

Se generaron $B = 1000$ réplicas bootstrap del estimador $\hat{\theta}_{\text{EMV}}$ corregido, que denotamos como $\theta_1^*, \ldots, \theta_B^*$. Estas réplicas se obtienen mediante remuestreo con reemplazo de la muestra original, siguiendo el mismo procedimiento que dio origen al estimador.

Un intervalo de confianza aproximado de nivel $1 - \alpha$ para θ, con $\alpha = 0{,}05$, se construye tomando los percentiles correspondientes de la distribución empírica de las réplicas bootstrap:

$$\left[\theta^*_{(\alpha/2)},\ \theta^*_{(1-\alpha/2)}\right] = \left[\theta^*_{(0{,}025)},\ \theta^*_{(0{,}975)}\right],$$

donde $\theta^*_{(\gamma)}$ representa el percentil γ de la muestra $\theta_1^*, \ldots, \theta_B^*$.

En nuestro caso, utilizamos $n = 100$ y el estimador $\hat{\theta}_{\text{EMV}}$. Este enfoque permite obtener un intervalo que no depende de suposiciones de normalidad y que se ajusta bien al comportamiento empírico del estimador.

```
> print(round(IC, 4))
[1] 0.125 0.395
```

Figura 4: Intervalo de confianza con metodo bootstrap

Comparando con el intervalo obtenido en el item (f), vemos que el intervalo asintótico se construye bajo el supuesto de normalidad del estimador, usando el Teorema Central del Límite y el Teorema de Slutsky. Es simétrico alrededor del estimador puntual y depende de una estimación de la varianza teórica:

$$\hat{\theta}_n \pm z_{1-\alpha/2} \cdot (\text{desvío estimado})$$

Sin embargo, este método puede ser poco confiable si el tamaño muestral no es suficientemente grande, si el estimador está cerca de un borde (por ejemplo, cercano a 0 o 1), o si la distribución del estimador es sesgada o asimétrica.

En este caso, el estimador $\hat{\theta}_n = 0{,}05$ es pequeño, y el intervalo normal incluye valores negativos, lo cual es problemático ya que $\theta \in [0, 1]$. El intervalo fue truncado a partir de 0, perdiendo simetría.

En cambio, el **intervalo bootstrap percentil** se construye empíricamente, tomando muchos remuestreos de la muestra original y calculando el estimador en cada uno.

Este método **no asume simetría ni normalidad** y se adapta mejor a la forma real de la distribución del estimador. En este caso:

- El intervalo bootstrap es más centrado en valores razonables para θ.
- No incluye valores negativos.
- Refleja la asimetría o sesgo presente en la distribución empírica del estimador.

Conclusión: El intervalo bootstrap percentil $[0{,}125; 0{,}395]$ es más amplio, pero también más realista, especialmente en muestras pequeñas o cuando el estimador está cerca de los bordes. En cambio, el intervalo normal $[0; 0{,}167]$, aunque más estrecho, puede subestimar la incertidumbre al basarse en supuestos que no se cumplen del todo en este caso.

(i) **Hallar el nivel de significacion empírico mediante simulaciones y graficarlo en funcion de n.**

Para estimar el **nivel de significación empírico** del test del cociente de máxima verosimilitud bajo el modelo de respuesta aleatorizada, realizamos simulaciones bajo la hipótesis nula $H_0 : \theta = 1/4$, es decir, suponiendo que la proporción real de estudiantes que se copiaron alguna vez en un examen es $\theta = 0{,}25$.

El objetivo es estimar con qué frecuencia el test *rechaza incorrectamente la hipótesis nula* cuando esta es verdadera, es decir, estimar empíricamente el **error de tipo I** o **nivel de significación real** del procedimiento de testeo. En teoría, este nivel debería coincidir con el nivel α (en este caso, 0,05), pero debido a aproximaciones, supuestos o correcciones que se aplican en la práctica, puede diferir. Evaluar esto empíricamente es una forma de validar el test.

Procedimiento

Para cada tamaño muestral n considerado, seguimos los siguientes pasos:

1. **Simulación de datos:** generamos $B = 1000$ muestras independientes de tamaño n bajo el modelo de respuesta aleatorizada, asumiendo $\theta = 1/4$.
2. **Aplicación del test:** sobre cada muestra simulada, aplicamos el test del cociente de máxima verosimilitud para contrastar $H_0 : \theta = 1/4$ contra $H_1 : \theta \neq 1/4$.
3. **Conteo de rechazos:** registramos en cuántas de las B simulaciones el test rechaza la hipótesis nula.
4. **Cálculo del nivel de significación empírico:** el nivel de significación empírico se estima como la proporción de simulaciones en las que se rechazó H_0:

$$\hat{\alpha}_{\text{emp}} = \frac{\text{número de rechazos}}{B}$$

Este procedimiento se repite para distintos valores de n desde 10 hasta 100 incrementando de a 5 con el fin de estudiar cómo varía el nivel empírico en función del tamaño muestral.

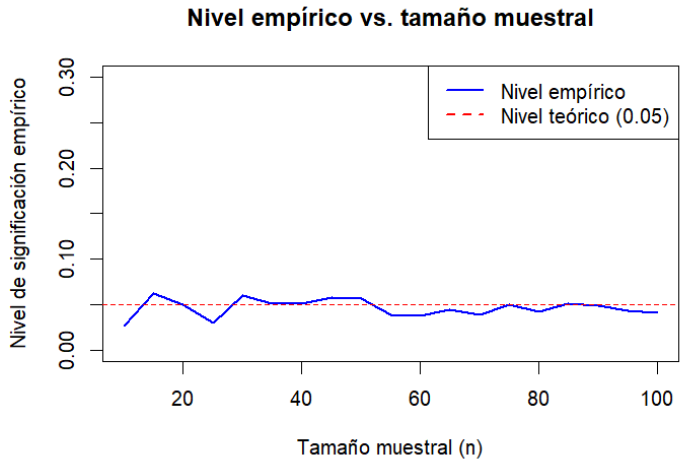

Figura 5: Nivel de significacion empirico

A partir del gráfico se puede observar que:

- El nivel empírico estimado se mantiene relativamente cercano al nivel teórico nominal de $\alpha = 0{,}05$, especialmente a medida que el tamaño muestral crece.

- Para valores pequeños de n, el nivel empírico presenta una mayor variabilidad. Esta fluctuación es esperable debido a la mayor inestabilidad de los tests en muestras pequeñas.
- A partir de $n \approx 40$, las estimaciones del nivel de significación empírico se estabilizan en torno a 0.05, lo que indica que el test se comporta adecuadamente para tamaños muestrales moderados en adelante.

Este análisis valida el uso del test del cociente de máxima verosimilitud bajo el modelo de respuesta aleatorizada, al menos en cuanto a su control del error de tipo I, siempre que se cuente con un tamaño muestral razonablemente grande.

(j) **Hallar la funcion de potencia mediante simulaciones y graficarla para n = 100.**

El objetivo de esta sección es estimar la **función de potencia empírica** del test del cociente de máxima verosimilitud bajo el modelo de respuesta aleatorizada. La función de potencia de un test mide la probabilidad de rechazar la hipótesis nula H_0 cuando la verdadera proporción θ toma distintos valores alternativos.

En este caso, se desea analizar cómo varía la potencia del test en función del valor real de θ, manteniendo fijo el tamaño muestral n, y así evaluar su capacidad para detectar desviaciones respecto de la hipótesis nula $H_0 : \theta = 1/4$.

Procedimiento

Para estimar empíricamente la función de potencia, se siguieron los siguientes pasos:

1. Se eligió un tamaño muestral fijo $n = 100$.
2. Se consideraron distintos valores de θ en el intervalo $[0, 1]$ (por ejemplo, de 0 a 1 en pasos de 0,02).
3. Para cada valor de θ, se generaron $B = 1000$ muestras independientes bajo el modelo de respuesta aleatorizada.
4. A cada muestra se le aplicó el test del cociente de máxima verosimilitud para contrastar $H_0 : \theta = 1/4$ contra $H_1 : \theta \neq 1/4$.
5. Se contó cuántas veces se rechazó H_0 en las B simulaciones, y se estimó la potencia empírica como:

$$\hat{\pi}(\theta) = \frac{\text{número de rechazos de } H_0 \text{ cuando } \theta \text{ es el valor real}}{B}$$

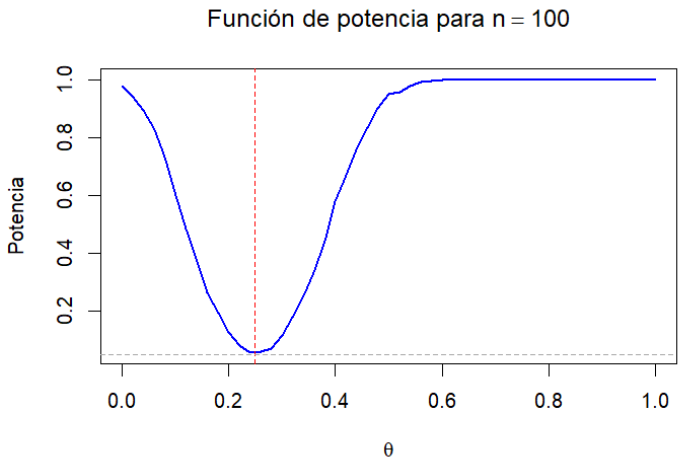

Figura 6: Funcion de potencia

Análisis

- La potencia empírica alcanza su valor mínimo (cercano a 0.05) cuando $\theta = 0{,}25$, que es el valor supuesto bajo la hipótesis nula H_0. Esto es coherente con el nivel de significación del test, ya que refleja la probabilidad de rechazar incorrectamente H_0 cuando esta es verdadera (error de tipo I).
- A medida que θ se aleja de 0,25, la potencia crece rápidamente, lo cual indica que el test tiene buena capacidad para detectar desviaciones de la hipótesis nula. En particular:
  - Para valores bajos de θ (por ejemplo, menores a 0.1) o altos (por encima de 0.4), la potencia se aproxima a 1, lo que implica que el test casi siempre rechaza correctamente H_0 cuando esta es falsa.
  - En el entorno de $\theta = 0{,}25$, hay una región donde la potencia es más baja (pero mayor que 0.05), lo cual es esperable ya que los valores cercanos al nulo son más difíciles de distinguir estadísticamente.
- En general, el gráfico muestra que el test es **consistente**, ya que la potencia tiende a 1 cuando θ se aleja del valor nulo.

Este análisis confirma que el test del cociente de máxima verosimilitud tiene un buen desempeño: controla adecuadamente el error de tipo I y presenta alta potencia ante alternativas razonablemente alejadas del valor nulo.

Conclusión

En este trabajo se analizaron y compararon dos métodos para estimar la proporción θ de estudiantes que se copiaron alguna vez en un examen: el método clásico basado en preguntas directas y el método de respuesta aleatorizada.

El método clásico es simple y eficiente en muestras grandes si se asume sinceridad total en las respuestas. Sin embargo, cuando existe una probabilidad π de que los estudiantes mientan, el estimador clásico se vuelve sesgado y no consistente. Su Error Cuadrático Medio (ECM) deja de disminuir con el tamaño muestral debido a un sesgo que no puede eliminarse sin conocer π.

En contraste, el método de respuesta aleatorizada, aunque introduce variabilidad adicional por el uso de mecanismos aleatorios (dado y moneda), genera un estimador insesgado y consistente. Este método protege la privacidad del encuestado y reduce el incentivo a mentir, lo cual es especialmente útil para estimar proporciones asociadas a conductas sensibles.

Se derivaron los estimadores por momentos y de máxima verosimilitud (EMV) para el modelo aleatorizado, y se analizó su comportamiento teórico y empírico mediante simulaciones. Ambos estimadores resultaron equivalentes tras una corrección por truncamiento en el caso del estimador de MM, garantizando que sus valores pertenezcan al intervalo $[0, 1]$.

A través de simulaciones se verificó que:

- Para tamaños muestrales pequeños, el método clásico puede parecer más eficiente por tener menor varianza, pero su sesgo constante lo perjudica.
- A medida que el tamaño muestral crece, el estimador aleatorizado muestra mejor comportamiento asintótico, con ECM decreciente y sesgo despreciable.
- El test del cociente de verosimilitudes bajo el modelo aleatorizado mostró buen control del error tipo I y alta potencia, validando su uso práctico.
- Los intervalos de confianza construidos con métodos asintóticos y bootstrap ofrecieron alternativas complementarias, siendo este último más robusto frente a sesgos o estimaciones cercanas a los bordes del parámetro.

En conclusión, aunque el método clásico puede resultar más preciso en situaciones ideales, el método de respuesta aleatorizada es más realista, robusto y éticamente recomendable cuando se trata de obtener información veraz sobre conductas sensibles. Este estudio refleja la importancia de diseñar estrategias estadísticas que tengan en cuenta el comportamiento humano en el proceso de recolección de datos.