instituto de cálculo
UBA - CONICET

Trabajo Práctico

Introducción a la Estadística y Ciencia de Datos
1er Cuatrimestre - 2025

Integrante	LU	Correo electrónico
Carla De Erausquin	126/18	deercarla@gmail.com
Victoria Klimkowski	1390/21	02vicky02@gmail.com
Carlos Fernández	700/16	charlyfn@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

Aclaración general: todos los gráficos que se muestran en el informe cuentan con su versión interactiva en el HTML, así como el código en R para resolver los incisos que así lo requieran. También estará el archivo Rmd (como lo solicita la consigna) pero el HTML también tiene los códigos de los incisos.

1. Método clásico

1.1. (a) Estimador de Máxima Verosimilitud (EMV)

Sea X_i la variable aleatoria que es 1 si el estudiante i responde que sí, y 0 si responde que no. Si todos los encuestados responden con la verdad y son independientes entre sí, entonces:

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

La función de verosimilitud es:

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

Podemos tomar el logaritmo de la verosimilitud ya que es creciente:

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

Como queremos encontrar donde se maximiza, derivamos e igualamos a 0 para encontrar puntos críticos:

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

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

$$\sum x_i - \theta\sum x_i = n\theta - \theta\sum x_i$$

Resolviendo:

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

Para ver que es máximo estudiamos la derivada segunda.

$$\frac{d^2\ell}{d\theta^2} = -\frac{\sum x_i}{\theta^2} - \frac{n - \sum x_i}{(1-\theta)^2}$$

Esta expresión es siempre negativa para $0 < \theta < 1$, ya que tanto los términos $\frac{1}{\theta^2}$ como $\frac{1}{(1-\theta)^2}$ son positivos, y están multiplicados por -1.

$$\frac{d^2\ell}{d\theta^2} < 0 \quad \Rightarrow \quad \text{Máximo local (y global)}$$

Por lo tanto, el valor hallado $\hat{\theta} = \bar{X}$ maximiza la verosimilitud.

Propiedades:

- **Insesgado:**
  Como $X_i \sim \text{Bernoulli}(\theta)$, 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, $\hat{\theta}$ es insesgado.

- **Varianza:**
  Dado que $X_i \sim \text{Bernoulli}(\theta)$, entonces $\text{Var}(X_i) = \theta(1-\theta)$, y como los X_i son i.i.d.:

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

- **Consistencia:**
  Por la ley fuerte de los grandes números, si X_i son i.i.d. con media θ y varianza finita, entonces:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{c.s.} \theta \quad \text{cuando } n\to\infty$$

  Es decir, $\hat{\theta}$ es consistente.

- **Eficiencia:**
  La cota de Cramér-Rao para estimadores insesgados de θ es:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

  Para la distribución Bernoulli, la información de Fisher por observación es:

$$I(\theta) = \mathbb{E}\left[\left(-\frac{\partial}{\partial\theta^2}\log f(X_i;\theta)\right)\right] = \frac{1}{\theta(1-\theta)}$$

  Entonces:

$$\text{Var}(\hat{\theta}) \geq \frac{\theta(1-\theta)}{n}$$

  Como ya vimos que $\text{Var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$, se alcanza la cota. Por lo tanto, $\hat{\theta}$ es eficiente.

1.2. (b) Sesgo y ECM del estimador con respuestas falsas

Supongamos ahora que los estudiantes pueden mentir:

- Si un estudiante se copió, responde no con probabilidad π (y sí con probabilidad $1-\pi$).
- Si un estudiante no se copió, siempre responde "no".

Sea Y_i la respuesta observada ($1 =$ "sí", $0 =$ "no") de la i-ésima persona. La nueva distribución es:

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

Entonces el nuevo estimador $\hat{\theta}$ sigue siendo la proporción muestral de "sí":

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

Esperanza:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[Y_1] = \theta(1-\pi)$$

Sesgo:

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

En promedio, el estimador subestima.

Error Cuadrático Medio (ECM):

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

La varianza de $\hat{\theta}$ bajo este nuevo modelo es:

$$\begin{aligned}
\text{Var}(\hat{\theta}) &= \frac{1}{n}\,\text{Var}(Y_1) \\
&= \frac{1}{n}\Big(E(Y_1^2) - [E(Y_1)]^2\Big) \\
&= \frac{1}{n}\Big(E(Y_1) - [E(Y_1)]^2\Big) \quad \text{(pues } Y_1 \text{ es Bernoulli)} \\
&= \frac{1}{n}\Big(\theta(1-\pi) - [\theta(1-\pi)]^2\Big) \\
&= \frac{1}{n}\theta(1-\pi)[1-\theta(1-\pi)].
\end{aligned}$$

Entonces:

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

No es consistente. Como $\mathbb{E}[\hat{\theta}] = \theta(1-\pi) \neq \theta$ salvo que $\pi = 0$, el estimador está sesgado y convergerá a $\theta(1-\pi)$, no a θ. Por lo tanto, no es consistente. Esto vale por la ley fuerte de los grandes números ya que la varianza es finita.

2. Método de respuesta aleatorizada

A cada entrevistado se le entrega un dado y una moneda, y se le dan las siguientes instrucciones:

- Tirar el dado.
- Si sale 1 o 2, tirar la moneda. Si sale cara, contestar ”sí”. Si sale ceca, contestar ”no”.
- Si sale 3, 4, 5 o 6, contestar con la verdad.

Supongamos que todos los encuestados siguen fielmente estas instrucciones.

2.1. (a) Estimador de momentos, sesgo, ECM y consistencia

Sea θ la proporción verdadera de personas que se copiaron. Sea Yi la respuesta observada (1 si respondió "sí", 0 si respondió "no") para la i-ésima persona. Queremos el estimador de momentos para θ.

- Calculamos la probabilidad de que un encuestado responda "sí":

$$\mathbb{P}(Y_i = 1) = \frac{2}{6} \cdot \frac{1}{2} + \frac{4}{6} \cdot \theta = \frac{1}{6} + \frac{2}{3}\theta$$

- Calculamos la esperanza:

$$\mathbb{E}[Y_1] = \frac{1}{6} + \frac{2}{3}\theta$$

- Igualamos la media muestral a la esperanza y despejamos θ:

$$\bar{Y} = \mathbb{E}[Y_1] = \frac{1}{6} + \frac{2}{3}\theta \quad \Longrightarrow \quad \theta = \frac{3}{2}\Big(\bar{Y} - \frac{1}{6}\Big).$$

- El estimador de momentos es:

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

- **Sesgo:**

  Calculamos la esperanza del estimador:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_{\text{MM}}] &= \mathbb{E}\left[\frac{3}{2}\left(\bar{Y} - \frac{1}{6}\right)\right] \\
&= \frac{3}{2}\left(\mathbb{E}[\bar{Y}] - \frac{1}{6}\right) \\
&= \frac{3}{2}\left(\mathbb{E}[Y_1] - \frac{1}{6}\right) \\
&= \frac{3}{2}\left(\frac{1}{6} + \frac{2}{3}\theta - \frac{1}{6}\right) \\
&= \frac{3}{2}\left(\frac{2}{3}\theta\right) \\
&= \theta.
\end{aligned}$$

  El estimador de momentos es insesgado

- **Varianza:**

  Notemos que cada Y_i es una variable Bernoulli con parámetro

$$p = \mathbb{P}(Y_i = 1) = \frac{1}{6} + \frac{2}{3}\theta.$$

  Por lo tanto:

$$\text{Var}(\hat{\theta}_{\text{MM}}) = \left(\frac{3}{2}\right)^2 \text{Var}(\bar{Y}) = \frac{9}{4} \cdot \frac{\text{Var}(Y_1)}{n} = \frac{9}{4n} \cdot \left(\frac{1}{6} + \frac{2}{3}\theta\right)\left(\frac{5}{6} - \frac{2}{3}\theta\right)$$

- **Error Cuadrático Medio (ECM):**

  Como el estimador es insesgado, el ECM coincide con la varianza:

$$\mathrm{ECM}(\hat{\theta}_{\mathrm{MM}}) = \mathbb{E}\left[(\hat{\theta}_{\mathrm{MM}} - \theta)^2\right] = \mathrm{Var}(\hat{\theta}_{\mathrm{MM}})$$

- **Consistencia:**

  Sabemos que:

  $\bar{Y} \xrightarrow{c.s.} \mathbb{E}[Y] = \frac{1}{6} + \frac{2}{3}\theta$ (por ley fuerte de los grandes números pues la varianza es finita)

  Sea $g(x) = \frac{3}{2}(x - \frac{1}{6})$, una función continua. Entonces, como las funciones continuas preservan la convergencia:

$$\hat{\theta}_{\mathrm{MM}} = g(\bar{Y}) \xrightarrow{c.s.} g\left(\mathbb{E}[Y]\right) = \frac{3}{2}\left(\frac{1}{6} + \frac{2}{3}\theta - \frac{1}{6}\right) = \theta$$

  Luego, es fuertemente consistente.

2.2. (b) Corrección del estimador de momentos

Primero, notemos que cada Y_i es una Bernoulli, por lo que:

$$0 \leq Y_i \leq 1. \quad \Longrightarrow \quad 0 \leq \sum_{i=1}^{n} Y_i \leq n, \quad \Longrightarrow \quad 0 \leq \bar{Y} \leq 1.$$

$$-\frac{1}{6} \leq \bar{Y} - \frac{1}{6} \leq \frac{5}{6}$$

$$-\frac{1}{6} \cdot \frac{3}{2} \;\leq\; \left(\bar{Y} - \frac{1}{6}\right) \cdot \frac{3}{2} \;\leq\; \frac{5}{6} \cdot \frac{3}{2}$$

$$-\frac{3}{12} \leq \hat{\theta}_{\mathrm{MM}} \leq \frac{15}{12}$$

Por lo tanto, debemos corregir el estimador para que no tome valores fuera del intervalo $[0, 1]$ ya que θ es una probabilidad:

$$\hat{\theta}_{\mathrm{corr}} = \mathrm{mín}\left(\, \mathrm{máx}\left(\hat{\theta}_{\mathrm{MM}},\, 0\right),\, 1\right).$$

De este modo, siempre se asegura que:

$$0 \leq \hat{\theta}_{\mathrm{corr}} \leq 1.$$

2.3. (c) Estimador de máxima verosimilitud (EMV)

Sea

$$Y_i := \begin{cases} 1 & \text{si la respuesta es sí} \\ 0 & \text{si no} \end{cases}$$

$$Y_i \sim \mathrm{Be}(p)$$

Sección 2.3 (c) Estimador de máxima verosimilitud (EMV)

Sabemos que

$$p = \frac{1}{6} + \frac{2}{3}\theta, \quad \text{con } \theta \text{ una probabilidad.}$$

Dado que $0 \leq \theta \leq 1$, entonces:

$$0 \leq \frac{2}{3}\theta \leq \frac{2}{3} \Rightarrow \frac{1}{6} \leq \frac{2}{3}\theta + \frac{1}{6} = p \leq \frac{5}{6}$$

$$\Rightarrow p \in \left[\frac{1}{6}, \frac{5}{6}\right]$$

Calculemos el EMV de p:

$$L(p) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} = p^{\sum_{i=1}^{n} Y_i}(1-p)^{n-\sum_{i=1}^{n} Y_i}, \quad \frac{1}{6} \leq p \leq \frac{5}{6}$$

Tomamos logaritmo pues es creciente:

$$\ell(p) = \sum_{i=1}^{n} Y_i \ln(p) + \left(n - \sum_{i=1}^{n} Y_i\right) \ln(1-p)$$

Derivamos e igualamos a cero:

$$\frac{\sum_{i=1}^{n} Y_i}{p} - \frac{n - \sum_{i=1}^{n} Y_i}{1-p} = 0$$

$$\sum_{i=1}^{n} Y_i(1-p) = \left(n - \sum_{i=1}^{n} Y_i\right) p$$

$$\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Y_i p = np - \sum_{i=1}^{n} Y_i p \Rightarrow \sum_{i=1}^{n} Y_i = np \Rightarrow \hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}_n$$

Como se muestra en la Sección 1(a) es máximo ya que la derivada segunda es negativa, agregando, para este caso particular, que estará truncado. Esto surge porque la verosimilitud se anula fuera del dominio permitido. Entonces, como el dominio de p está limitado:

$$\hat{p}_{\text{MV}} = \text{mín}(\frac{5}{6}, \text{máx}(\frac{1}{6}, \overline{Y}_n))$$

Esto es así porque $p \in \left[\frac{1}{6},\ \frac{5}{6}\right]$ por lo tanto, la probabilidad de que p sea menor que $\frac{1}{6}$ o mayor que $\frac{5}{6}$ es cero:

$$\mathbb{P}\left(p < \tfrac{1}{6} \text{ o } p > \tfrac{5}{6}\right) = 0.$$

Podemos visualizarlo comparando dónde estaría el estimador si el soporte fuera el estándar, es decir, si p estuviera en el intervalo $[0, 1]$.

Entonces:

- Si $\frac{1}{6} < \overline{Y}_n < \frac{5}{6}$, y estamos dentro del dominio, quedaría igual:

$$\hat{p}_{\text{MV}} = \overline{Y}_n$$

- Si $\overline{Y}_n \leq \frac{1}{6}$, el promedio caería por debajo de la cota inferior del soporte que tenemos.

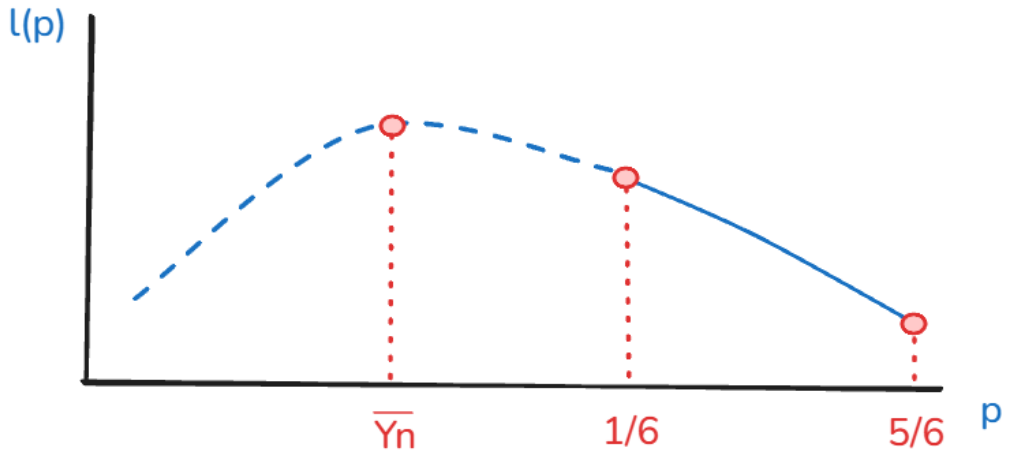


Figura 1: Ilustración l(p) vs p para $\overline{Y}_n \leq \frac{1}{6}$

$$\hat{p}_{\text{MV}} = \frac{1}{6}$$

- Si $\overline{Y}_n \geq \frac{5}{6}$, el promedio caería por arriba de la cota superior del soporte que tenemos.

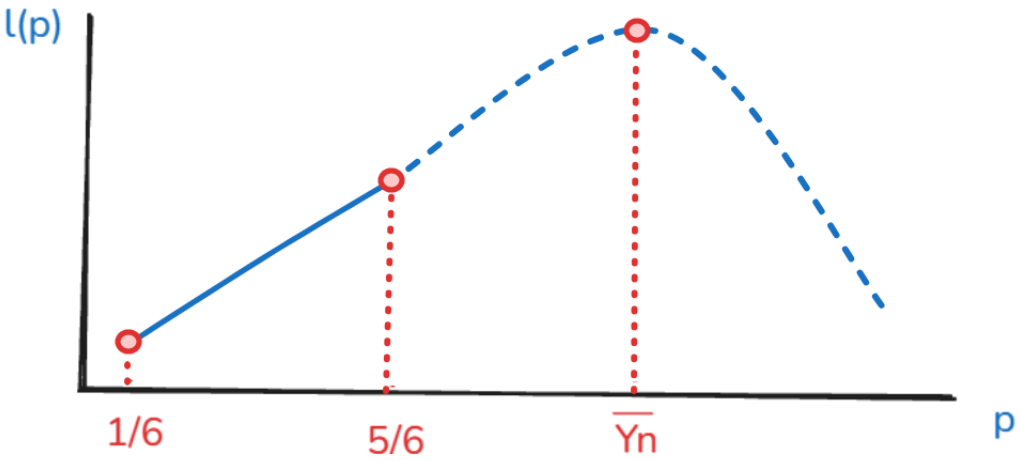


Figura 2: Ilustración l(p) vs p para $\overline{Y}_n \geq \frac{5}{6}$

$$\hat{p}_{\text{MV}} = \frac{5}{6}$$

Recordamos que

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

Sea $g(x) = \frac{3}{2}(x - \frac{1}{6})$ una función biyectiva. Por la invarianza de los EMV, si

$$\hat{p}_{\text{MV}} = \text{mín}(\frac{5}{6}, \text{máx}(\frac{1}{6}, \overline{Y}_n))$$

entonces

$$\hat{\theta}_{\text{MV}} = \text{mín}\left(1, \text{máx}\left(0, \hat{\theta}_{\text{MM}}\right)\right) = \hat{\theta}_{\text{corr}}$$

- **Consistencia**

  Ahora vamos a ver la consistencia de este estimador.

  Por la Sección 2(a):

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

  Sea $g(x) = \text{mín}\Big(1, \text{máx}(0, x)\Big)$. Dado que g es continua, preserva la convergencia casi segura. Por lo tanto,

  $$\hat{\theta}_{\text{MV}} = g\Big(\hat{\theta}_{\text{MM}}\Big) \xrightarrow{\text{c.s.}} g(\theta) = \text{mín}\Big(1, \text{máx}(0, \theta)\Big).$$

  Finalmente, esto es consistente porque θ es una probabilidad, por lo que $\theta \in [0, 1]$ con probabilidad 1, y entonces:

  $$\text{mín}\Big(1, \text{máx}(0, \theta)\Big) = \theta.$$

  Por lo tanto, $\hat{\theta}_{\text{MV}}$ es un estimador consistente de θ.

- **Sesgo cuando n=2**

  Dado que

  $$P(Y_1 = 1) = \frac{1}{6} + \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{3}, \quad P(Y_1 = 0) = \frac{2}{3},$$

  las posibles realizaciones del promedio muestral $\bar{Y} = \frac{Y_1 + Y_2}{2}$ son:

  - $\bar{Y} = 0$ si $Y_1 = Y_2 = 0$, con probabilidad:

    $$P(\bar{Y} = 0) = P(Y_1 = 0)^2 = \left(\frac{2}{3}\right)^2 = \frac{4}{9}.$$

  - $\bar{Y} = 1$ si $Y_1 = Y_2 = 1$, con probabilidad:

    $$P(\bar{Y} = 1) = P(Y_1 = 1)^2 = \left(\frac{1}{3}\right)^2 = \frac{1}{9}.$$

  - $\bar{Y} = \frac{1}{2}$ si $Y_1 \neq Y_2$, con probabilidad:

    $$P(\bar{Y} = \tfrac{1}{2}) = 2 \cdot P(Y_1 = 1) \cdot P(Y_2 = 0) = 2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}.$$

  Evaluamos $\hat{\theta}_{\text{MV}}$ en cada caso:

  - Si $\bar{Y} = 0$, entonces:

    $$\hat{\theta}_{\text{MV}} = \text{mín}\left(1, \text{máx}\left(0, \frac{3}{2}(0 - \tfrac{1}{6})\right)\right) = 0.$$

  - Si $\bar{Y} = \frac{1}{2}$, entonces:

    $$\hat{\theta}_{\text{MV}} = \text{mín}\left(1, \text{máx}\left(0, \frac{3}{2}(\tfrac{1}{2} - \tfrac{1}{6})\right)\right) = \frac{1}{2}.$$

Sección 2.4 (d) Sesgo del EMV en función del tamaño muestral n suponiendo que $\theta = 1/4$.

- Si $\bar{Y} = 1$, entonces:

$$\hat{\theta}_{\text{MV}} = \text{mín}\left(1, \text{máx}\left(0, \frac{3}{2}(1 - \tfrac{1}{6})\right)\right) = 1.$$

Entonces, la esperanza del estimador es:

$$\mathbb{E}(\hat{\theta}_{\text{MV}}) = 0 \cdot \frac{4}{9} + \frac{1}{2} \cdot \frac{4}{9} + 1 \cdot \frac{1}{9} = \frac{2}{9} + \frac{1}{9} = \frac{1}{3}.$$

Como el valor real de θ es $\theta = \frac{1}{4}$, el sesgo es:

$$B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

El estimador $\hat{\theta}$ tiene un sesgo positivo cuando n=2, es decir, en promedio sobreestima a θ.

Si fuera insesgado, lo sería para todo n. Por lo tanto, es un estimador **sesgado**.

2.4. (d) Sesgo del EMV en función del tamaño muestral n suponiendo que $\theta = 1/4$.

Sabemos que el sesgo puede escribirse como:

$$B(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Reescribiendo, nos queda:

$$B(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta = \mathbb{E}\left[\frac{3}{2}\left(\hat{p}_{\text{MV}} - \frac{1}{6}\right)\right] - \theta = \frac{3}{2}\left(\mathbb{E}[\hat{p}_{\text{MV}}] - \frac{1}{6}\right) - \theta$$

Luego, hallemos la esperanza de $\hat{p}_{\text{MV}}$.

Esperanza de $\hat{p}_{\text{MV}}$: El estimador es:

$$\hat{p}_{\text{MV}} = \begin{cases} \frac{5}{6} & \text{si } \overline{Y} \geq \frac{5}{6}, \\ \overline{Y} & \text{si } \frac{1}{6} < \overline{Y} < \frac{5}{6}, \\ \frac{1}{6} & \text{si } \overline{Y} \leq \frac{1}{6}. \end{cases}$$

Entonces, su esperanza se puede escribir como:

$$\mathbb{E}[\hat{p}_{\text{MV}}] = \frac{5}{6} \cdot \mathbb{P}\left(\overline{Y} \geq \frac{5}{6}\right) \; + \; \mathbb{E}\left[\overline{Y} \cdot \mathbb{1}_{\left\{\frac{1}{6} < \overline{Y} < \frac{5}{6}\right\}}\right] \; + \; \frac{1}{6} \cdot \mathbb{P}\left(\overline{Y} \leq \frac{1}{6}\right).$$

Entendamos cada término:

$$\frac{5}{6} \cdot \mathbb{P}\left(\overline{Y} \geq \frac{5}{6}\right) = \frac{5}{6} \cdot \mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \tfrac{5}{6}n\right) = \frac{5}{6} \cdot \sum_{k=\lceil \frac{5}{6}n \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{1}$$

$$\frac{1}{6} \cdot \mathbb{P}\left(\overline{Y} \leq \frac{1}{6}\right) = \frac{1}{6} \cdot \mathbb{P}\left(\sum_{i=1}^{n} Y_i \leq \tfrac{1}{6}n\right) = \frac{1}{6} \cdot \sum_{k=0}^{\lfloor \frac{n}{6} \rfloor} \binom{n}{k} p^k (1-p)^{n-k}. \tag{2}$$

Sección 2.4 (d) Sesgo del EMV en función del tamaño muestral n suponiendo que $\theta = 1/4$.

Para la parte intermedia, para simplificar notación, definimos:

$$A := \left\{ k : \frac{1}{6} < \frac{k}{n} < \frac{5}{6} \right\} = \left\{ k : \lfloor \tfrac{n}{6} \rfloor + 1 \leq k \leq \lfloor \tfrac{5n}{6} \rfloor \right\}.$$

$$\begin{aligned} \mathbb{E}\Big[\overline{Y} \cdot \mathbb{1}_A \Big] &= \sum_{k=0}^{n} \left(\frac{k}{n}\right) \cdot \mathbb{1}_{\{k \in A\}} \cdot \mathbb{P}\Big(\sum_{i=1}^{n} Y_i = k \Big) \\ &= \sum_{k=\lfloor \frac{n}{6} \rfloor + 1}^{\lfloor \frac{5n}{6} \rfloor} \left(\frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{3}$$

Por lo tanto, la expresión final para la esperanza es:

$$\mathbb{E}\Big(\hat{p}_{\text{MV}}\Big) = \underbrace{\frac{1}{6} \sum_{k=0}^{\lfloor \frac{n}{6} \rfloor} \binom{n}{k} p^k (1-p)^{n-k}}_{\text{(segunda ecuación)}} + \underbrace{\sum_{k=\lfloor \frac{n}{6} \rfloor + 1}^{\lceil \frac{5n}{6} \rceil - 1} \frac{k}{n} \binom{n}{k} p^k (1-p)^{n-k}}_{\text{(tercer ecuación)}} + \underbrace{\frac{5}{6} \sum_{k=\lceil \frac{5n}{6} \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}}_{\text{(primer ecuación)}} \tag{4}$$

Ahora sí, queremos graficar el sesgo del EMV usando R:

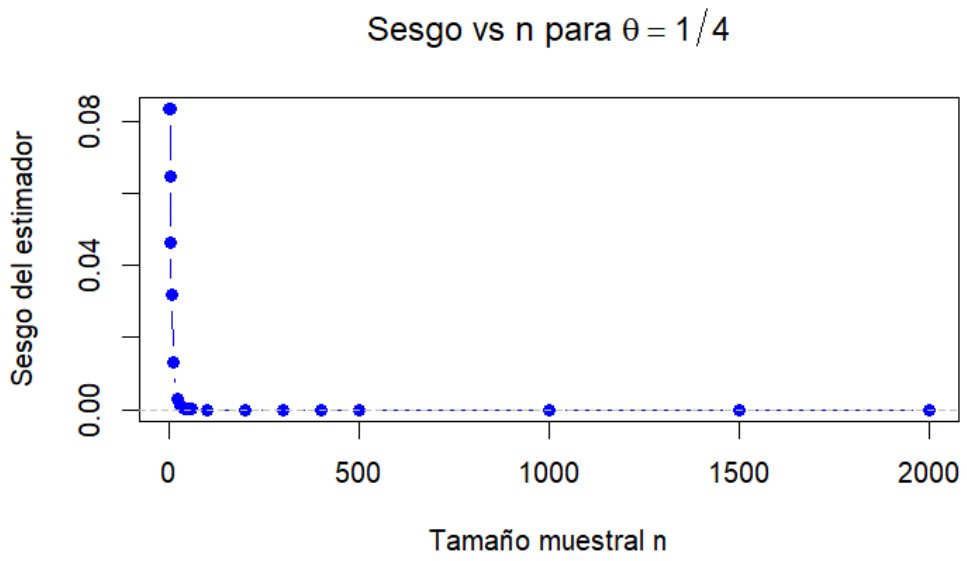


Figura 3: Sesgo del EMV en función del tamaño muestral n suponiendo que $\theta = 1/4$

El gráfico muestra el **sesgo** del estimador de máxima verosimilitud para distintos tamaños muestrales n, con $\theta = \frac{1}{4}$. Es más pronunciado cuando n es pequeño. A medida que n aumenta, el estimador tiende a concentrarse en torno a su valor esperado y el sesgo disminuye, tendiendo a cero.

Del gráfico podríamos deducir que el estimador resulta ser asintóticamente insesgado.

2.5. (e) ECM del EMV en función del tamaño muestral n suponiendo que $\theta = 1/4$.

Ahora queremos hallar el ECM de $\hat{\theta}_{\text{MV}}$. Recordemos que el ECM podemos escribirlo como:

$$\text{ECM}(\hat{\theta}_{\text{MV}}) = \text{Var}(\hat{\theta}_{\text{MV}}) + B(\hat{\theta}_{\text{MV}})^2$$

El sesgo ya lo hallamos, siendo este

$$B(\hat{\theta}_{\text{MV}}) = \mathbb{E}[\hat{\theta}_{\text{MV}}] - \theta = \mathbb{E}\left[\frac{3}{2}\left(\hat{p}_{\text{MV}} - \frac{1}{6}\right)\right] - \theta = \frac{3}{2}\left(\mathbb{E}[\hat{p}_{\text{MV}}] - \frac{1}{6}\right) - \theta$$

Donde la esperanza de $\hat{p}_{\text{MV}}$ es la hallada en la ecuación (4).
Luego, nos queda hallar la varianza de $\hat{\theta}_{\text{MV}}$. Sabemos que:

$$\text{Var}(\hat{\theta}_{\text{MV}}) = \left(\frac{3}{2}\right)^2 \text{Var}(\hat{p}_{\text{MV}}) = \left(\frac{3}{2}\right)^2 \left(\mathbb{E}[\hat{p}^2_{\text{MV}}] - (\mathbb{E}[\hat{p}_{\text{MV}}])^2\right)$$

$$\hat{\theta}_{\text{MV}} = \frac{3}{2}\left(\hat{p}_{\text{MV}} - \frac{1}{6}\right),$$

Luego,

$$\text{Var}(\hat{\theta}_{\text{MV}}) = \frac{9}{4}\left(\mathbb{E}[\hat{p}^2_{\text{MV}}] - (\mathbb{E}[\hat{p}_{\text{MV}}])^2\right)$$

Para ello, expresamos $\mathbb{E}[\hat{p}^2_{\text{MV}}]$ utilizando estadístico inconsciente:

$$\mathbb{E}\left(\hat{p}^2_{\text{MV}}\right) = \frac{1}{36}\sum_{k=0}^{\lfloor \frac{n}{6} \rfloor} \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=\lfloor \frac{n}{6} \rfloor + 1}^{\lceil \frac{5n}{6} \rceil - 1} \binom{n}{k} p^k (1-p)^{n-k} \cdot \left(\frac{k}{n}\right)^2 + \frac{25}{36}\sum_{k=\lceil \frac{5}{6}n \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

Y elevamos al cuadrado la $\mathbb{E}\left(\hat{p}_{\text{MV}}\right)$ calculada en (4).
Por lo tanto, la varianza completa queda:

$$\text{Var}(\hat{\theta}) = \frac{9}{4}\left[\mathbb{E}[\hat{p}^2_{\text{MV}}] - (\mathbb{E}[\hat{p}_{\text{MV}}])^2\right]$$

Podemos graficar el ECM de estimadores de $\theta = 1/4$:

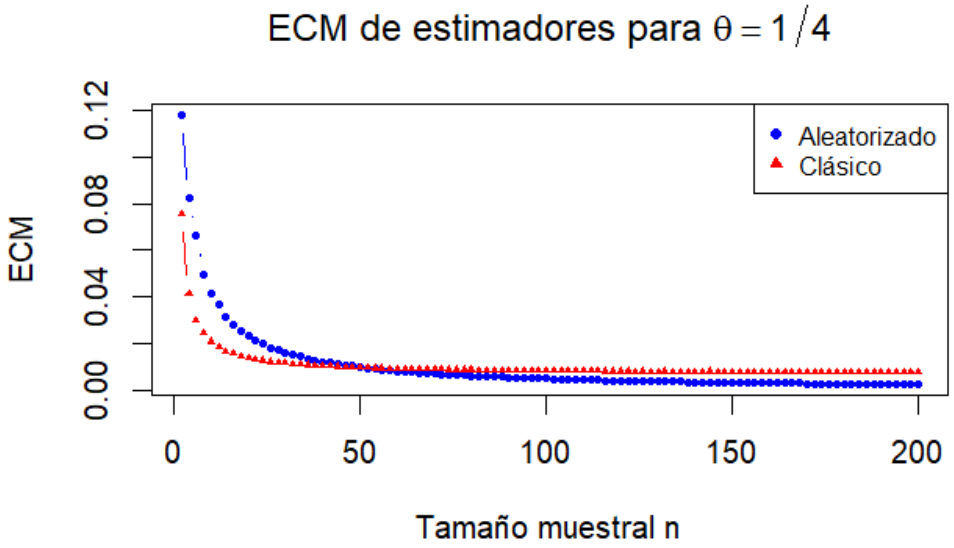


Figura 4: ECM del EMV en función del tamaño muestral n suponiendo que $\theta = 1/4$

El gráfico compara el **Error Cuadrático Medio (ECM)** del estimador de máxima verosimilitud (EMV) obtenido mediante el método de **respuesta aleatorizada** (curva azul), con el ECM del estimador **clásico** (curva roja), suponiendo que la proporción verdadera de individuos con la característica de interés es $\theta = \frac{1}{4}$ y que el porcentaje de estudiantes que mienten en el método clásico es $\pi = \frac{1}{3}$.

Se observa ambos ECM disminuyen a medida que aumenta el tamaño muestral n, reflejando que las varianzas de los estimadores tienden a cero. El ECM del método clásico es más grande que el aleatorizado para muestras chicas pero (ligeramente) más chico a medida que aumenta la muestra. Ambos tienden a estabilizarse a medida que aumenta el tamaño de la muestra, prácticamente constante a partir de $n \approx 50$. El ECM del método aleatorizado tiende a 0 y el ECM del método clásico tiende a su sesgo.

2.6. (f) Test de cociente de máxima verosimilitud e intervalos de confianza.

Test de Hipótesis para θ

Sea $Y_i \sim \text{Be}(p(\theta))$ con $p(\theta) = \frac{1}{6} + \frac{2}{3}\theta$ y con $0 < \theta < 1$
Queremos saber si θ cambió. Planteamos la hipótesis nula y alternativa como:

$$H_0 : \theta = \frac{1}{4}, \qquad H_1 : \theta \neq \frac{1}{4}.$$

Con $p(\theta) = \frac{1}{6} + \frac{2}{3}\theta$ y $0 < \theta < 1$ función de verosimilitud es:

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

El cociente de máxima verosimilitud se define como:

$$\Lambda = \frac{L(\theta_0)}{\sup_\theta L(\theta)},$$

donde $\theta_0 = \frac{1}{4}$.

Calculamos:

$$L\left(\frac{1}{4}\right) = \left(\frac{1}{6} + \frac{2}{3}\cdot\frac{1}{4}\right)^S \left(1 - \left(\frac{1}{6} + \frac{2}{3}\cdot\frac{1}{4}\right)\right)^{n-S} = \left(\frac{1}{3}\right)^S \left(\frac{2}{3}\right)^{n-S},$$

donde $S = \sum_{i=1}^n Y_i$.

Por Sección 2(c) donde aplicamos la Invarianza del Estimador de Máxima Verosimilitud (por tener una función biyectiva)

$$\hat{p}_{\text{MV}} = \text{mín}(\frac{5}{6}, \text{máx}(\frac{1}{6}, \overline{Y}_n))$$

Entonces, el cociente queda:

$$\Lambda = \frac{\left(\frac{1}{3}\right)^S \left(\frac{2}{3}\right)^{n-S}}{\hat{p}_{\text{MV}}^S (1-\hat{p}_{\text{MV}})^{n-S}}.$$

Cuando n es grande, bajo condiciones de regularidad: $-2\log(\Lambda) \sim \chi^2_1$, pues $dim(\Theta) = 1$ y $dim(\Theta_0) = 0$ y 1 - 0 = 1.

$$-2\log(\Lambda) = -2\left[S\log\left(\frac{1/3}{\hat{p}_{MV}}\right) + (n-S)\log\left(\frac{2/3}{1-\hat{p}_{MV}}\right)\right]$$

Luego, si tomamos $\chi^2_{1,\alpha}$ tal que

$$P\left(\chi^2_1 \geq \chi^2_{1,\alpha}\right) = \alpha$$

Nos queda el test asintótico

$$\varphi(x) = \begin{cases} 1 & \text{si } -2\log(\Lambda) > \chi^2_{1,\alpha} \\ 0 & \text{c.c.} \end{cases}$$

Con los datos del problema, donde n=100, S=20 y $\overline{Y}_n = 0.2$

$$T_{obs} = -2\left[20\cdot\log\left(\frac{1}{3\cdot 0{,}2}\right) + 80\cdot\log\left(\frac{2}{3\cdot 0{,}8}\right)\right] = 8{,}738$$

Test de nivel asintótico: si $\alpha = 0{,}05$:

$$\varphi(x) = \begin{cases} 1 & \text{si } -2\log(\Lambda) = T_{obs} > 3{,}841 \\ 0 & \text{c.c.} \end{cases}$$

Conclusión: A nivel $\alpha = 0.05$, rechazamos H_0. Luego, hay evidencia para afirmar $\theta \neq \frac{1}{4}$. Es decir, sirvió la campaña.

Para esta muestra, el p-valor será de aproximadamente 0.00312. Éste se interpreta como la probabilidad más extrema con la que rechazamos H_0. Luego, para niveles mayores al p-valor (incluyendo a 0.05) vamos a rechazar H_0 mientras que para niveles más chicos (por ejemplo, si quisiéramos a nivel 0.001) diremos que no hay evidencia suficiente para rechazar H_0.

Intervalos de confianza para θ

Método 1: TCL plug-in

$X_1, \dots, X_{100} \sim \text{Ber}(p(\theta))$

$\hat{p}_{\text{muestral}} = \bar{Y}, \quad p(\theta) = \frac{1}{6} + \frac{2}{3}\theta$

TCL: $\frac{\bar{Y} - p(\theta)}{\sqrt{\frac{p(\theta)(1-p(\theta))}{n}}} \xrightarrow{D} \mathcal{N}(0,1)$

Como $p(\theta)$ es desconocido, necesitamos estimar su varianza.

Una opción natural es usar el estimador plug-in:

$\hat{\sigma}^2 = \bar{Y}(1-\bar{Y})$

Por lo tanto, definimos el estadístico:

$$Z_n = \frac{\bar{Y} - p(\theta)}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}}$$

Ahora, notamos que:

$$\sqrt{\frac{p(\theta)(1-p(\theta))}{\bar{Y}(1-\bar{Y})}} \xrightarrow{p} 1$$

(pues $\bar{Y} \xrightarrow{p} p(\theta)$ y la función $f(x) = \sqrt{\frac{p(\theta)(1-p(\theta))}{x(1-x)}}$

es continua en el dominio)

Entonces, por el Teorema de Slutsky:

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

$$1 - \alpha = \lim_{n\to\infty} \mathbb{P}\left(-z_{\alpha/2} < \frac{\bar{Y} - p(\theta)}{\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}} < z_{\alpha/2}\right)$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\bar{Y} - z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}} < p(\theta) < \bar{Y} + z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right)$$

Como el parámetro al cual le queremos dar un intervalo de confianza está entre $[\frac{1}{6}, \frac{5}{6}]$ con probabilidad uno, podemos escribirlo como:

$$= \lim_{n\to\infty} \mathbb{P}\left(\text{máx}\left(\frac{1}{6}, \bar{Y} - z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right) < \frac{1}{6} + \frac{2}{3}\theta < \text{mín}\left(\frac{5}{6}, \bar{Y} + z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right)\right)$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\text{máx}\left(0, \frac{3}{2}\cdot\left(-\frac{1}{6} + \left(\bar{Y} - z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right)\right)\right) < \theta < \text{mín}\left(1, \frac{3}{2}\cdot\left(-\frac{1}{6} + \left(\bar{Y} + z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right)\right)\right)\right.$$

Luego, vamos a tener el IC asintótico,

$$IC = \left[\text{máx}\left(0, \frac{3}{2}\cdot\left(-\frac{1}{6} + \left(\bar{Y} - z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right)\right)\right), \text{mín}\left(1, \frac{3}{2}\cdot\left(-\frac{1}{6} + \left(\bar{Y} + z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}\right)\right)\right)\right]$$

Para los datos observados, con n=100 y $\alpha = 0.05$ el intervalo de nivel 0.95 nos queda entonces:

$$\left[\text{máx}\left(0, \left(\frac{76}{625} - \frac{1}{6}\right)\cdot\frac{3}{2}\right), \text{mín}\left(1, \left(\frac{174}{625} - \frac{1}{6}\right)\cdot\frac{3}{2}\right)\right]$$

Luego un IC de nivel 0.95 para θ por el método de TCL plug-in asintótico:

$$\text{IC para } \theta : \boxed{[0,\ 0{,}168]}$$

Método 2: Método Delta

$\sqrt{n}(\bar{Y} - p) \xrightarrow{D} \mathcal{N}(0, p(1-p))$

Usamos método delta: sea $g(x) = \frac{3}{2}\left(x - \frac{1}{6}\right), \quad g'(x) = \frac{3}{2} \neq 0$

$\Rightarrow \sqrt{n}(g(\bar{Y}) - g(p)) \xrightarrow{D} \mathcal{N}(0, \frac{9}{4}p(1-p))$

Vamos a poder usar Slutsky, pues, con un argumento similar al del ítem anterior: $\frac{\sqrt{p(1-p)}}{\sqrt{\bar{Y}(1-\bar{Y})}} \xrightarrow{p} 1$

$\Rightarrow \frac{\sqrt{n}(g(\bar{Y}) - g(p))}{\frac{3}{2}\sqrt{\bar{Y}(1-\bar{Y})}} \xrightarrow{D} \mathcal{N}(0,1)$

$$1 - \alpha = \lim_{n\to\infty} \mathbb{P}\left(-z_{\alpha/2} < \frac{\sqrt{n}(g(\bar{Y}) - g(p))}{\frac{3}{2}\sqrt{\bar{Y}(1-\bar{Y})}} < z_{\alpha/2}\right)$$

$$1 - \alpha = \lim_{n\to\infty} \mathbb{P}\left((\sqrt{n}g(\bar{Y}) - z_{\alpha/2}\cdot\frac{3}{2}\sqrt{\bar{Y}(1-\bar{Y})} < g(p)) < \sqrt{n}g(\bar{Y}) + z_{\alpha/2}\cdot\frac{3}{2}\sqrt{\bar{Y}(1-\bar{Y})}\right)$$

Como el parámetro al cual le queremos dar un intervalo de confianza está entre [0,1] con probabilidad uno, podemos escribirlo como:

$$1 - \alpha = \lim_{n\to\infty} \mathbb{P}\left(\text{máx}(0, g(\bar{Y}) - z_{\alpha/2}\cdot\frac{3}{2\sqrt{n}}\sqrt{\bar{Y}(1-\bar{Y})}) < g(p)) < \text{mín}(1, g(\bar{Y}) + z_{\alpha/2}\cdot\frac{3}{2\sqrt{n}}\sqrt{\bar{Y}(1-\bar{Y})})\right)$$

$$1 - \alpha = \lim_{n\to\infty} \mathbb{P}\left(\text{máx}\left(0, g(\bar{Y}) - z_{\alpha/2}\cdot\frac{3}{2\sqrt{n}}\sqrt{\bar{Y}(1-\bar{Y})}\right) < \theta < \text{mín}\left(1, g(\bar{Y}) + z_{\alpha/2}\cdot\frac{3}{2\sqrt{n}}\sqrt{\bar{Y}(1-\bar{Y})}\right)\right)$$

Luego, vamos a tener un IC para asintótico para θ:

$$\left[\text{máx}\left(0, g(\bar{Y}) - z_{\alpha/2}\cdot\frac{3}{2\sqrt{n}}\sqrt{\bar{Y}(1-\bar{Y})}\right), \text{mín}\left(1, g(\bar{Y}) + z_{\alpha/2}\cdot\frac{3}{2\sqrt{n}}\sqrt{\bar{Y}(1-\bar{Y})}\right)\right]$$

Con los datos del problema y $\alpha = 0{,}05$, el IC asintótico queda

$$\left[\text{máx}\left(0, \frac{1}{20} - 1{,}96\cdot\frac{3}{2}\cdot\frac{1}{25}\right), \text{mín}\left(1, \frac{1}{20} + 1{,}96\cdot\frac{3}{2}\cdot\frac{1}{25}\right)\right]$$

$$\text{IC para } \theta : \boxed{[0,\ 0{,}168]}$$

Que es igual al obtenido del método anterior

Método 3: TCL

Sea p $= p(\theta) = \frac{1}{6} + \frac{2}{3}\theta$

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

$$\Rightarrow \frac{\sqrt{n}(\bar{Y} - p)}{\sqrt{\bar{Y}(1-\bar{Y})}} \xrightarrow{D} \mathcal{N}(0,1) \quad \text{por Slutsky, usando que } \frac{\sqrt{p(1-p)}}{\sqrt{\bar{Y}(1-\bar{Y})}} \xrightarrow{p} 1$$

$$\Rightarrow \frac{\sqrt{n}(\bar{Y} - p)}{S_n} \xrightarrow{D} \mathcal{N}(0,1)$$

Donde $S_n^2 = \hat{\sigma}^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2$ que calcularemos con ayuda de R

Tomamos $\alpha = 0{,}05$, sabiendo que n es grande

$$1 - \alpha = \lim_{n\to\infty} \mathbb{P}\left(-z_{\alpha/2} < \frac{\sqrt{n}(\bar{Y} - p)}{S_n} < z_{\alpha/2}\right)$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\bar{Y} - \frac{z_{\alpha/2}S_n}{\sqrt{n}} < p < \bar{Y} + \frac{z_{\alpha/2}S_n}{\sqrt{n}}\right)$$

Podemos, además, restringir al dominio de p (estará por fuera de su dominio con probabilidad 0)

$$= \lim_{n\to\infty} \mathbb{P}\left(\text{máx}(\frac{1}{6}, \bar{Y} - \frac{z_{\alpha/2}S_n}{\sqrt{n}}) < p < \text{mín}(\frac{5}{6}, \bar{Y} + \frac{z_{\alpha/2}S_n}{\sqrt{n}})\right)$$

Aplicamos la función creciente para la transformación de p a θ

$$= \lim_{n\to\infty} \mathbb{P}\left(\text{máx}(\frac{1}{6}, \bar{Y} - \frac{z_{\alpha/2}S_n}{\sqrt{n}}) < \frac{1}{6} + \frac{2}{3}\cdot\theta < \text{mín}(\frac{5}{6}, \bar{Y} + \frac{z_{\alpha/2}S_n}{\sqrt{n}})\right)$$

$$= \lim_{n\to\infty} \mathbb{P}\left(\text{máx}\left(0, ((\bar{Y} - \frac{z_{\alpha/2}S_n}{\sqrt{n}}) - \frac{1}{6})\cdot\frac{3}{2}\right) < \theta < \text{mín}\left(1, ((\bar{Y} + \frac{z_{\alpha/2}S_n}{\sqrt{n}}) - \frac{1}{6})\cdot\frac{3}{2}\right)\right)$$

Entonces, el IC asintótico para θ

$$\left[\text{máx}\left(0, ((\bar{Y} - \frac{z_{\alpha/2}S_n}{\sqrt{n}}) - \frac{1}{6})\cdot\frac{3}{2}\right), \text{mín}\left(1, ((\bar{Y} + \frac{z_{\alpha/2}S_n}{\sqrt{n}}) - \frac{1}{6})\cdot\frac{3}{2}\right)\right]$$

Para los datos observados calculamos $S_n = 0{,}402$ con $\alpha = 0{,}05$ nos queda:

$$\left[\text{máx}\left(0, ((\frac{20}{100} - \frac{1{,}96\cdot 0{,}402}{\sqrt{100}}) - \frac{1}{6})\cdot\frac{3}{2}\right), \text{mín}\left(1, ((\frac{20}{100} + \frac{1{,}96\cdot 0{,}402}{\sqrt{100}}) - \frac{1}{6})\cdot\frac{3}{2}\right)\right]$$

$$\text{IC para } \theta : \boxed{[0,\ 0{,}168]}$$

Método 4: Numérico Otra opción, sería encontrarlo numéricamente (por la dificultad del despeje) teniendo en cuenta el test de máxima verosimilitud de esta misma Sección 2(f).

$$\varphi(x) = \begin{cases} 1 & \text{si } -2\log(\Lambda) > \chi^2_{1,\alpha} \\ 0 & \text{c.c.} \end{cases}$$

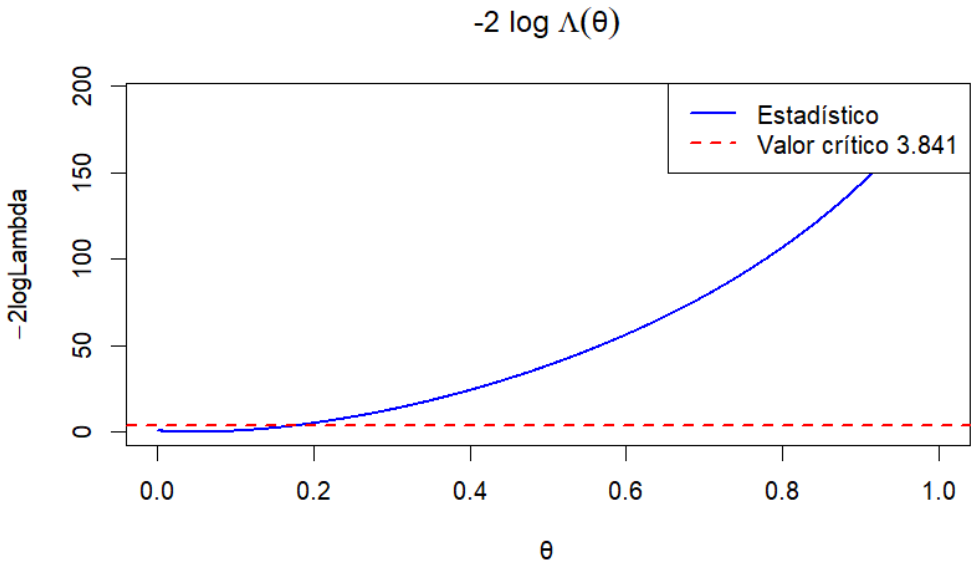
Es decir, con los datos del problema, los θ con los que se pueda establecer una relación con los p (restringidos a su soporte en $[\frac{1}{6}, \frac{5}{6}]$) que cumplan que:

$$-2 \cdot 20 \cdot \log(p) - 2 \cdot 80 \cdot \log(1-p) + 2 \cdot 20 \cdot \log\left(\frac{20}{100}\right) + 2 \cdot 80 \cdot \log\left(\frac{80}{100}\right) \leq \chi^2_{1,\alpha}$$

con $\alpha = 0.05$

Por su complicado despeje, calculamos los p numéricamente. Graficamos la función de recién en función de g(p). Con $g(p) = \frac{3}{2} \cdot (p - \frac{1}{6})$ y $p \in [\frac{1}{6}, \frac{5}{6}]$

Obtuvimos el siguiente gráfico



La intersección en el eje x (es decir el θ que será la cota superior del IC), se puede ver muy claramente en la versión interactiva del HTML. Ésta corresponde, redondeando a 3 cifras significativas, a θ =0.178.

IC para θ : $\boxed{[0,\ 0{,}178]}$

Comparaciones

Podemos ver que se obtiene el mismo IC para θ mediante los primeros 3 métodos y uno levemente más grande mediante la aproximación numérica.

En los casos del primer y tercer método, se parte del mismo resultado asintótico (basado en el Teorema Central del Límite), pero se emplean distintos estimadores de la varianza: uno plug-in (usando $\bar{Y}(1-\bar{Y})$) y el otro el muestral insesgado (usando S_n^2). Sin embargo, dado que ambos son estimadores consistentes de la varianza poblacional, sus diferencias se vuelven asintóticamente despreciables (con n = 100 en este caso).

En el segundo método, se usa el Método Delta para transformar la media muestral $\bar{Y}$ en una estimación para θ. Aunque la construcción del intervalo parece distinta, en el fondo también se basa en el Teorema Central del Límite, ya que parte de la aproximación asintótica normal de $\bar{Y}$. Al aplicar una función continua y derivable a $\bar{Y}$, se obtiene otra variable que también sigue aproximadamente una distribución normal. Por eso, este método termina generando un intervalo con diferencias despreciables a los otros. Notemos que la función g elegida era la transformación de p a θ.

El cuarto método utiliza la distribución asintótica de $-2 \cdot log(\Lambda)$ que se distribuye como una χ^2_1. Con esa premisa, nos conseguimos los valores de p restringidos a su soporte con los que no se rechaza el test propuesto a nivel α. El intervalo contendrá a los valores de p que produzcan una verosimilitud dentro del rango de aceptación. Luego, como existe una biyección entre θ y p, es casi trivial encontrar los θ que produzcan los p necesarios utilizando la misma g(p) del método anterior. Es decir, g(p) = $\frac{3}{2} \cdot (p - \frac{1}{6})$. Al ser un método de aproximación numérica no es sorprendente que sea levemente más grande que los obtenidos por métodos clásicos.

El hecho de que ninguno de los IC encontrados de nivel 0.95 contenga al 0.25 es consistente con haber rechazado que $\theta = 0.25$ con el test de nivel 0.05.

2.7. (g) Comparación de ECM para métodos clásico y aleatorizado.

Se realizaron simulaciones para estimar el Error Cuadrático Medio (ECM) empírico de los estimadores obtenidos por el método clásico y por el método de respuesta aleatorizada, considerando distintos tamaños muestrales $n \in \{10, 100, 1000\}$, bajo el supuesto de que la verdadera proporción de individuos con la característica sensible es $\theta = 0{,}25$, y que en el método clásico, la probabilidad de que un individuo que se copió mienta es $\pi = 1/3$.

Los resultados obtenidos fueron los siguientes, redondeados a tres cifras significativas:

n	ECM clásico	ECM aleatorizado
10	0,020782	0,041558
100	0,008347	0,005012
1000	0,007087	0,000503
2000	0,006995	0,000256
3000	0,007003	0,000165

Podemos observar que, a tres cifras significativas, el ECM aleatorizado es despreciable para n lo suficientemente grandes.

Recordemos que la expresión del ECM del método clásico asumiendo que existe una proporción π de gente que miente es la siguiente:

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

Para pequeñas muestras, el ECM, se observó, es menor para el estimador clásico lo cual se condice con los resultados teóricos obtenidos en incisos anteriores.

Cuando n tiende a infinito, el ECM tiende a $(\theta \cdot \pi)^2$, con $-\theta \cdot \pi$ el sesgo como se muestra en la Sección 1(b). Para este ejemplo, $\theta = \frac{1}{4}$ y $\pi = \frac{1}{3}$. Luego, era esperable que el ECM tendiera al sesgo al cuadrado porque la varianza se vuelve asintóticamente despreciable, es decir,

$$(-\theta \cdot \pi)^2 = \left(\frac{1}{4} \cdot \frac{1}{3}\right)^2 = 0{,}006,$$

lo cual es muy similar a lo observado.

Ahora, para el caso del ECM aleatorizado por la Sección 2(e) vimos también que a mayor tamaño de muestras, el ECM simulado tiende a cero, al igual que el la tabla obtenida.

2.8. (h) Intervalo de confianza para θ por el método bootstrap

Método Bootstrap Percentil

Una alternativa no paramétrica para construir intervalos de confianza consiste en el **método bootstrap percentil**. Este se basa en generar muchas réplicas del estimador $\hat{\theta}$ a partir

de remuestreos con reemplazo de la muestra observada. A partir de la distribución empírica obtenida, se construye un intervalo de confianza tomando los percentiles adecuados.

En este caso, se considera una muestra $y = (y_1, \ldots, y_n)$ de tamaño 100, con 20 respuestas afirmativas (1) y 80 negativas (0). El estimador de interés es:

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

Este estimador se trunca para asegurar que respete el soporte de $\theta \in [0, 1]$. Luego, se procede de la siguiente manera:

- Se generan $B = 100{,}000$ muestras bootstrap con reemplazo a partir de y.
- En cada muestra bootstrap se calcula el estimador $\hat{\theta}^*$.
- A partir de la distribución empírica de los $\hat{\theta}^*_b$, se obtienen los percentiles 2.5 y 97.5.

De esta forma, el intervalo de confianza bootstrap al nivel 95 % es:

$$IC_{\text{boot}} = \left[\hat{\theta}^*_{(0{,}025)},\ \hat{\theta}^*_{(0{,}975)}\right]$$

En este caso:

$$IC_{\text{boot}} = [0,\ 0{,}17]$$

$\alpha = 0{,}05$
$y \leftarrow$ `c(rep(1, 20), rep(0, 80))`
$\hat{\theta} = \frac{3}{2}\left(\hat{p} - \frac{1}{6}\right)$
$\hat{\theta} = \text{mín}(1, \text{máx}(0, \hat{\theta}))$
$B \leftarrow 100000$
$\theta_b \leftarrow$ `replicate(B, {`

```
  sample_y  ← sample(y, size = 100, replace = TRUE)
  mean(sample_y)
})
```

$\text{IC}_{\text{boot}} \leftarrow$ `quantile(theta_b, probs = c(0.025, 0.975))`
IC: $\boxed{[0,\ 0{,}170]}$

Este intervalo es más amplio por el extremo superior que los construidos por métodos clásicos en la Sección 1(f) pero más estrecho que el conseguido por la aproximación numérica. Es más flexible, ya que no requiere suponer normalidad asintótica ni simetría en la distribución del estimador. Además, la restricción al soporte es intrínsica al método en lugar de tenerlo como un agregado. Sin embargo, el IC obtenido solo es 1 % más grande que los métodos clásicos y, razonablemente, 0.25 sigue sin pertenecer al intervalo.

Método Bootstrap Normal

Otra alternativa para construir un intervalo de confianza es el **método bootstrap normal**. Este se basa en suponer que la distribución del estimador $\hat{\theta}$ es aproximadamente normal, con media θ y desvío estándar estimado empíricamente a través del bootstrap.

Sección 2.9 (i) Nivel de significación empírico mediante simulaciones.

En este caso, se considera nuevamente la muestra $y = (y_1, \ldots, y_n)$ de tamaño 100, con 20 respuestas afirmativas (1) y 80 negativas (0), y el mismo estimador:

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

truncado para respetar el soporte $[0, 1]$. Se procede de la siguiente forma:

- Se generan $B = 100{,}000$ muestras bootstrap con reemplazo a partir de y.
- En cada muestra bootstrap se calcula el estimador $\hat{\theta}^*$.
- Se calcula el desvío estándar de la muestra $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.
- Se utiliza la normal estándar para construir un intervalo del tipo:

$$IC_{\text{boot}} = \left[\hat{\theta} - z_{\alpha/2} \cdot \text{sd}(\hat{\theta}^*),\ \hat{\theta} + z_{\alpha/2} \cdot \text{sd}(\hat{\theta}^*)\right]$$

Donde $\hat{\theta}$ es el estimador calculado a partir de la muestra original, y $\text{sd}(\hat{\theta}^*)$ es el desvío estándar de los valores bootstrap.

En este caso particular, con $\alpha = 0{,}05$, se tiene:

$$\begin{aligned}
&\hat{\theta} = 0 \\
&\text{sd}(\hat{\theta}^*) \approx 0{,}061 \\
&z_{0{,}025} = 1{,}96 \\
&IC_{\text{boot}} = [0{,}149]
\end{aligned}$$

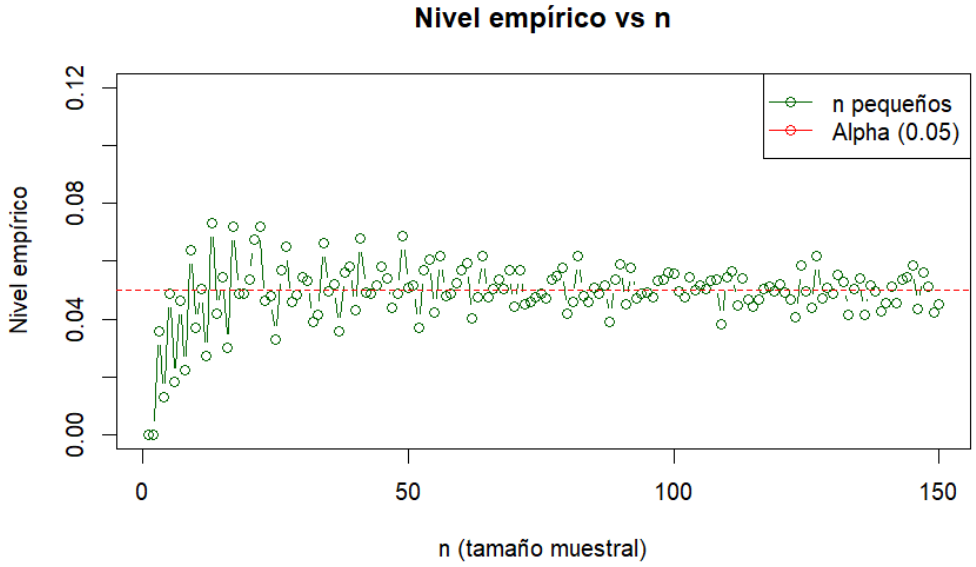
Este intervalo es más estrecho que el obtenido por el método percentil, pero asume normalidad en la distribución del estimador, lo cual puede no cumplirse en todos los casos. Además, el truncamiento para respetar el soporte de θ puede distorsionar la simetría del intervalo. Observamos que 0.25 sigue sin pertencer al intervalo.

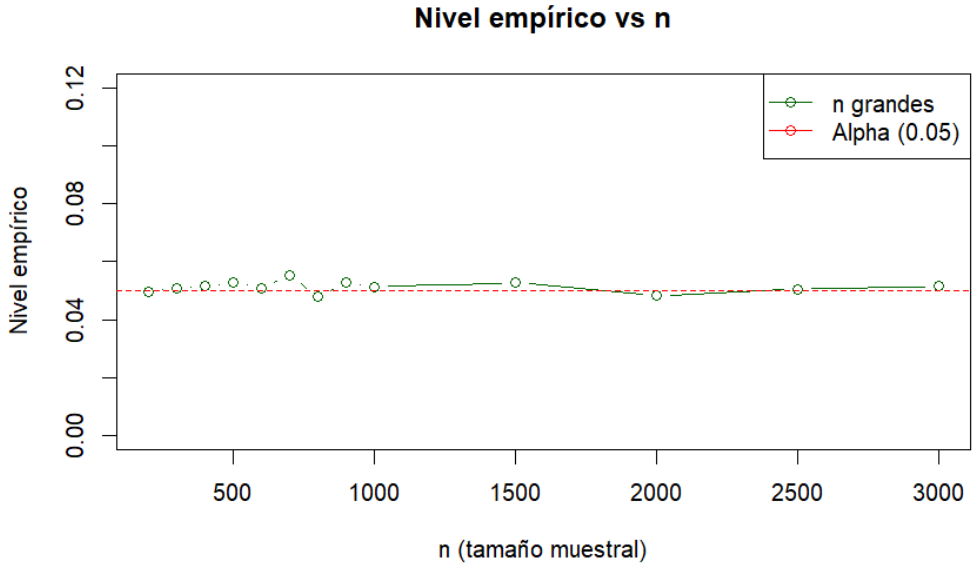
Ambos intervalos fueron generados con *bootstrap*, uno utilizando los percentiles y el otro utilizando la distribución bootstrap normal asintótica. El intervalo de menor longitud fue el por el método *bootstrap* normal y el de mayor longitud por el *bootstrap* percentil. Éste resultó muy similar a los obtenidos por los métodos clásicos y al método de aproximación numérica.

2.9. (i) Nivel de significación empírico mediante simulaciones.

Queremos estimar cuántas veces rechazamos incorrectamente la hipótesis nula cuando ésta es cierta (es decir, el nivel de significación empírico) con $H_0 : \theta = 1/4$ (valor real). Simulamos muchas veces una muestra de tamaño n bajo esta hipótesis.

En cada simulación, aplicamos el test de la Sección 2(f) y vemos la proporción de rechazos. Recordemos que el nivel empírico se define como la proporción de veces que se rechaza la hipótesis nula cuando ésta es verdadera.





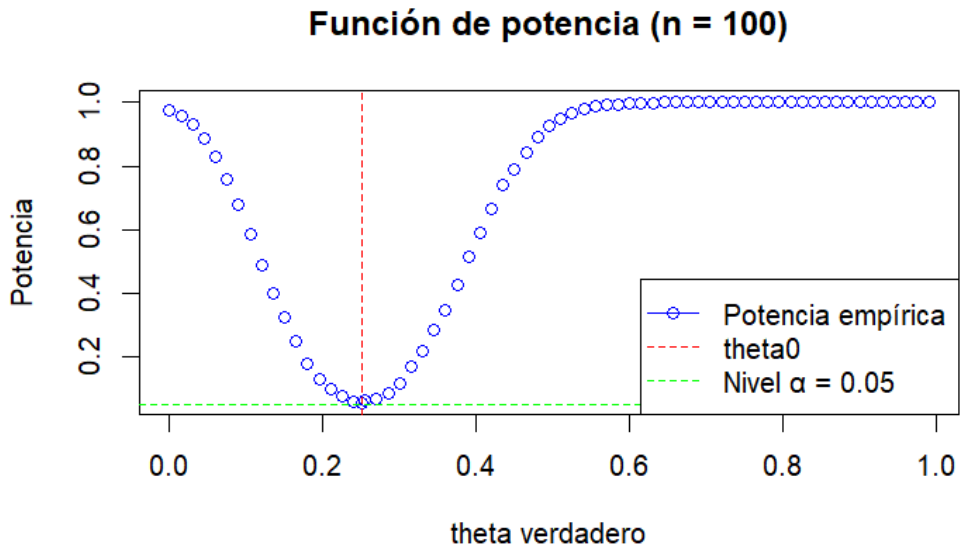
En estos gráfico se observa que:

- Para valores muy pequeños de n, el nivel empírico no es tan estable y, bajo H0, a veces hay más rechazos que el 5 % y a veces hay menos.
- A medida que el tamaño muestral crece, el nivel empírico converge hacia el nivel teórico deseado $\alpha = 0{,}05$ (línea roja punteada) y se mantiene ahí para n muy grandes.

Este resultado muestra que, si bien el test tiene un comportamiento deficiente para muestras pequeñas, su nivel de significación se ajusta al valor teórico a medida que aumenta el tamaño muestral. Por lo tanto, es recomendable utilizar este test con n suficientemente grande.

2.10. (j) Función de potencia mediante simulaciones.

El siguiente gráfico representa la función de potencia del test del cociente de máxima verosimilitud para tamaño muestral $n = 100$.



Se destaca lo siguiente:

- La potencia empírica alcanza valores cercanos a 1 cuando el valor verdadero de θ se aleja mucho de la hipótesis nula $\theta_0 = 0{,}25$, tanto hacia valores bajos como altos.
- El mínimo de la curva ocurre justamente en $\theta_0 = 0{,}25$, como era esperable. En este punto, la potencia del test coincide aproximadamente con el nivel de significación empírico, indicado por la línea verde punteada ($\alpha = 0{,}05$).
- La simetría de la curva alrededor de θ_0 sugiere que el test tiene igual sensibilidad para detectar desviaciones tanto positivas como negativas respecto del valor nulo.