Estadística Bayesiana

Trabajo Práctico: "Modelos Bayesianos para Técnicas de Respuesta Aleatorizada (Warner y Greenberg): Un Estudio Comparativo sobre Apuestas Online en Jóvenes"

Estudiantes: Marcos Leguizamón y Evelin Sánchez Meza

Profesores: Ignacio Evangelista y Tomás Capretto

Fecha: 21/04/2025

Introducción.

La creciente prevalencia de las apuestas en línea entre adolescentes constituye un fenómeno social de interés, con potenciales implicaciones para su bienestar psicológico y desarrollo social. La accesibilidad digital y la exposición constante a publicidad facilitan la adopción de estas prácticas, a menudo en un marco regulatorio aún en desarrollo. La investigación sobre la magnitud de esta participación se enfrenta al desafío inherente del sesgo de respuesta, donde la naturaleza sensible del tema puede inducir a los encuestados a ocultar o distorsionar sus experiencias.

Ante esta limitación metodológica, las técnicas de respuesta aleatorizada emergen como herramientas para la obtención de datos más fiables. Estos métodos introducen un componente aleatorio en el proceso de encuesta, desvinculando la respuesta individual de la pregunta sensible y, por ende, incrementando la probabilidad de obtener respuestas honestas. El presente trabajo práctico se inscribe en el paradigma de la **estadística bayesiana** y se enfoca en la aplicación y comparación de dos técnicas de respuesta aleatorizada específicas: el modelo de Warner y la propuesta de Greenberg.

El marco bayesiano proporciona una estructura natural para la incorporación de incertidumbre y conocimiento previo a través de la especificación de distribuciones *prior*. La información muestral, obtenida mediante la aplicación de las técnicas de respuesta aleatorizada, se modela a través de la función de verosimilitud, permitiendo la actualización de las creencias iniciales en la distribución *posterior*. Este enfoque posibilita la obtención de inferencias probabilísticas sobre el parámetro de interés, en este caso, la proporción (π_A) de estudiantes que participan en apuestas deportivas en línea.

Mediante la implementación de simulaciones computacionales en el entorno R, se evalúa la capacidad de estas técnicas, dentro del marco bayesiano, para proporcionar estimaciones precisas y robustas de la prevalencia de apuestas deportivas online en la población adolescente, incluso en presencia de potenciales sesgos de respuesta que afectarían a las encuestas directas. El objetivo principal consiste en demostrar la utilidad de la integración de las técnicas de respuesta aleatorizada con la inferencia bayesiana como metodología para abordar la investigación de temas sensibles en poblaciones jóvenes.

Modelo Bayesiano

Conforme a lo introducido, el objetivo central es estimar la proporción π_a de estudiantes que participan en apuestas deportivas online. Antes de abordar directamente el desafío del sesgo de respuesta mediante técnicas de respuesta aleatorizada, se establece un modelo bayesiano fundamental. Este modelo inicial se aplicará al escenario de una encuesta realizada mediante pregunta directa, bajo el supuesto de que los estudiantes responden honestamente.

Función de Verosimilitud

Se considera una muestra de n estudiantes seleccionados aleatoriamente. A cada estudiante se le pregunta directamente si ha participado en apuestas deportivas online en un período determinado. Sea Y la variable aleatoria que representa el número de estudiantes que responden afirmativamente ('éxito') en la muestra de n.

Bajo el supuesto de que cada estudiante responde de forma independiente y que la probabilidad de que un estudiante haya participado (y responda afirmativamente, asumiendo honestidad) es π_a , el número de respuestas afirmativas Y sigue una distribución Binomial. La función de verosimilitud, que expresa la probabilidad de observar y respuestas afirmativas dados n y π_a, es:

$$P(Y = y) = \binom{n}{y} \pi_a^y (1 - \pi_a)^{n-y}, \quad \text{para } y = 0, 1, 2, \ldots, n$$

Entonces, la distribución muestral es

$$Y_i \mid \pi_A \sim \text{Binomial}(\pi_A, n)$$

Esta elección se justifica porque:

- El experimento consiste en n ensayos.
- Cada ensayo tiene dos posibles resultados: 'éxito' (responde sí) o 'fracaso' (responde no).
- La probabilidad de 'éxito', π_a, se asume constante para cada estudiante.
- Los ensayos son independientes entre sí.

Distribución a Priori

El parámetro π_a representa una proporción, por lo que su valor debe estar contenido en el intervalo [0,1]. La distribución Beta es una elección usual para modelar creencias a priori sobre una proporción. Se propone la siguiente distribución prior para π_a:

$$\pi_A \sim \text{Beta}(\alpha, \beta)$$

Específicamente, se seleccionan los valores de los parámetros $\alpha = 2$ y $\beta = 2$, es decir,

$$\pi_A \sim \text{Beta}(2, 2)$$

Las razones para esta elección son:

- Soporte Adecuado: La distribución Beta tiene soporte en [0,1], que coincide con el rango posible de π_a.

- Reflejo de Creencias Previas: Una Beta(2,2) es simétrica en torno a 0.5, lo que representa una postura inicial de que valores alrededor del 50% son los más probables a priori. Sin embargo, al ser α y β mayores que 1, la densidad de probabilidad es menor en los extremos (cerca de 0 y 1) comparado con una Beta(1,1). Esto incorpora la creencia a priori de que es poco probable que ningún estudiante participe (π_a=0) o que todos lo hagan ($\pi_a = 1$), sugiriendo que la prevalencia real se encuentra en la zona intermedia, pero manteniendo incertidumbre.

Por lo tanto, el modelo propuesto es:

$$Y_i \mid \pi_A \sim \text{Binomial}(\pi_A, n)$$

$$\pi_A \sim \text{Beta}(2, 2)$$

Distribución Posterior

Aplicando el Teorema de Bayes, la distribución posterior de π_a, que actualiza las creencias iniciales con la información de los datos, es proporcional al producto del prior y la verosimilitud:

$$p(\pi_A \mid y, n) \propto p(\pi_A) \times p(y \mid n, \pi_A)$$

$$p(\pi_a \mid y, n) \propto \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi_a^{\alpha-1}(1-\pi_a)^{\beta-1}\right] \times \left[\binom{n}{y}\pi_a^{y}(1-\pi_a)^{n-y}\right]$$

$$p(\pi_a \mid y, n) \propto \pi_a^{\alpha+y-1}(1-\pi_a)^{\beta+n-y-1}$$

Dada la elección de distribuciones conjugadas, con un prior Beta y una verosimilitud Binomial, la distribución posterior resultante es también una distribución Beta, con parámetros actualizados:

$$\pi_A \mid y, n \sim Beta(\alpha + y, \beta + n - y)$$

Sustituyendo los parámetros del prior elegido ($\alpha = 2$,$\beta = 2$), se obtiene la distribución posterior para este modelo:

$$\pi_A \mid y, n \sim Beta(2 + y, 2 + n - y)$$

Se decidió utilizar este prior por las siguientes razones. En primer lugar, el parametro π_A es un parámetro que tiene un recorrido continuo dentro del rango 0 al 1. Por lo cual la función beta cumple con este requisito. En segundo lugar, se supone que los valores extremos que puede tomar el parámetro π_A son muy poco probables.En resumen, se cree que no todos los alumnos de la escuela participaron en apuestas deportivas online y en contraposición tambien se supone que al menos una parte significativa de estudiantes han participado en apuestas online. Por estas razones se le da mayor peso a los posibles valores que puede tomar π_A alrededor del 0.5.

Por otro lado, se decidió usar una distribución binomial como función de verosimilitud ya que se cuenta con una situación en la que cada estudiante encuestado solo tiene dos opciones de respuesta; Si el estudiante ha participado en apuestas deportivas online(éxito) o no ha participado en apuestas deportivas(fracaso). Ádemas el tamaño de la muestra es fijo y esta dado por n, la cantidad de alumnos encuestados en la escuela. Y la probabilidad con la que los alumnos encuestados respondan afirmativamenta a la participación en apuestas deportivas online es π_A.

Con estos elementos se sabe que el posterior se obtiene al realizar la productoría entre el prior con la verosimilitud. Y dado los el prior y la verosimilitud propuesta se obtendrá un posterior beta. De la siguiente forma :

$$\pi_A \sim \text{Beta}(2 + y, 2 + n - y)$$

4. Probabilidad de respuesta afirmativa según el método de Warner

Se consideró el método propuesto por Warner, en el cual cada estudiante debía responder una de dos preguntas seleccionadas de manera aleatoria mediante un mecanismo probabilístico:

- Con probabilidad p, se le preguntaba si alguna vez había participado en apuestas deportivas en línea (pertenencia al grupo A).
- Con probabilidad $1 - p$, se le preguntaba si **nunca** había participado en apuestas deportivas en línea (pertenencia al complemento A^c).

Dado que el investigador desconocía cuál de las dos preguntas había sido contestada, no se podía saber directamente si un "sí" correspondía a pertenecer o no al grupo A. No obstante, se dedujo la probabilidad total de obtener una respuesta afirmativa, denotada como λ_W, a partir de las siguientes componentes:

$$\lambda_W = p \cdot \pi_A + (1 - p) \cdot (1 - \pi_A)$$

Donde: - p representaba la probabilidad de que el encuestado recibiera la pregunta sobre A (participación en apuestas), - π_A indicaba la proporción real de estudiantes que habían apostado, - $1 - \pi_A$ correspondía a quienes no lo habían hecho.

Utilizando los valores propuestos en el trabajo: - $p = 0.6$, - $\pi_A = 0.4$,

se obtuvo:

$$\lambda_W = 0.6 \cdot 0.4 + 0.4 \cdot (1 - 0.4) = 0.24 + 0.24 = 0.48$$

Por lo tanto, se concluyó que la probabilidad de que un estudiante respondiera afirmativamente bajo el método de Warner fue de $\lambda_W = 0.48$.

Consecuentemente, la probabilidad de que un estudiante respondiera negativamente (es decir, "no") fue:

$$1 - \lambda_W = 1 - 0.48 = 0.52$$

5. Modelo propuesto para la generación de datos (Método de Warner)

A partir del esquema planteado por Warner y de la probabilidad λ_W calculada previamente, se propuso un modelo razonable para simular cómo se generaban las respuestas en la muestra, respetando la lógica de la técnica de respuesta aleatorizada.

Cada respuesta observada se consideró como el resultado de dos procesos independientes:

1. **Selección de la pregunta mediante un mecanismo aleatorio**:
   A cada encuestado se le asignó una de las dos preguntas posibles:
   - Con probabilidad p, se le preguntaba si alguna vez había participado en apuestas deportivas en línea (pregunta sobre la categoría A).
   - Con probabilidad $1 - p$, se le preguntaba si **nunca** había participado en apuestas deportivas en línea (pregunta sobre la categoría complementaria A^c).
2. **Respuesta del encuestado según su condición real**:
   Se asumió que los encuestados respondieron de manera sincera (ya que el método protege la privacidad individual), por lo tanto:
   - Si se les preguntaba por la categoría A y efectivamente pertenecían a ella, respondían "sí".

- Si se les preguntaba por la categoría A^c y efectivamente **no** pertenecían a A, también respondían "sí".
- En los demás casos, respondían "no".

De esta manera, el resultado observado podía modelarse como una variable aleatoria de Bernoulli, con probabilidad de éxito (respuesta afirmativa) igual a:

$$\lambda_W = p \cdot \pi_A + (1-p) \cdot (1-\pi_A)$$

Para simular los datos generados bajo este modelo, se siguieron los siguientes pasos:

1. Para cada individuo i en la muestra (de tamaño $n = 100$), se generó una variable indicadora $Z_i \sim \text{Bernoulli}(p)$ que definía cuál de las dos preguntas le fue asignada.
2. Se generó una variable indicadora $Y_i \sim \text{Bernoulli}(\pi_A)$ que definía si el individuo pertenecía o no a la categoría de apostadores.
3. A partir de Z_i y Y_i, se determinó la respuesta observada como:

$$R_i = \begin{cases} 1 & \text{si } (Z_i = 1 \wedge Y_i = 1) \text{ o } (Z_i = 0 \wedge Y_i = 0) \\ 0 & \text{en caso contrario} \end{cases}$$

Este modelo reflejó fielmente el mecanismo de respuesta aleatorizada propuesto por Warner, al tiempo que preservó la confidencialidad de las respuestas individuales y permitió realizar inferencias válidas sobre la proporción π_A en la población.

6. Obtención del posterior exacto bajo un prior uniforme

Se trabajó con el método de respuesta aleatorizada propuesto por Warner, bajo el cual se deseó estimar la proporción real de estudiantes que apostaban en línea, denotada como π_a. Se asumió un prior uniforme sobre π_a, es decir:

$$\pi_a \sim \text{Uniform}(0, 1)$$

La probabilidad de que un estudiante respondiera afirmativamente (es decir, "Sí") bajo el mecanismo de Warner fue:

$$\lambda_W = p \cdot \pi_a + (1-p)(1-\pi_a) = (2p-1)\pi_a + (1-p)$$

Dado que se observó un total de y respuestas afirmativas en una muestra de tamaño n, la verosimilitud de los datos fue modelada como una distribución binomial:

$$Y \sim \text{Binomial}(n, \lambda_W)$$

Por lo tanto, la función de verosimilitud para π_a resultó ser proporcional a:

$$\mathcal{L}(\pi_a) \propto \lambda_W^y (1-\lambda_W)^{n-y}$$

Como se utilizó un prior uniforme, el posterior también resultó proporcional a la verosimilitud:

$$p(\pi_a \mid y) \propto [(2p-1)\pi_a + (1-p)]^y \, [1 - (2p-1)\pi_a - (1-p)]^{n-y}$$

La constante de normalización Z fue definida como:

$$Z = \int_0^1 [(2p-1)t + (1-p)]^y \, [1 - (2p-1)t - (1-p)]^{n-y} \, dt$$

Para resolver esta integral, se aplicó el cambio de variable:

$$\lambda = (2p-1)t + (1-p) \Rightarrow t = \frac{\lambda - (1-p)}{2p-1}$$

El diferencial se transformó como:

$$dt = \frac{1}{2p-1} d\lambda$$

Cuando $t = 0$, se obtuvo $\lambda = 1 - p$; y cuando $t = 1$, se obtuvo $\lambda = p$. Entonces, los límites de integración cambiaron de $t \in [0, 1]$ a $\lambda \in [1 - p, p]$.

La integral de normalización quedó expresada como:

$$Z = \int_{1-p}^{p} \lambda^y (1-\lambda)^{n-y} \cdot \frac{1}{2p-1} d\lambda$$

$$Z = \frac{1}{2p-1} \int_{1-p}^{p} \lambda^y (1-\lambda)^{n-y} d\lambda$$

Reconociendo que la integral corresponde a la función beta incompleta, se llegó a:

$$Z = \frac{B(p; y+1, n-y+1) - B(1-p; y+1, n-y+1)}{1-2p}$$

Donde $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1} dt$ es la función beta incompleta.

Finalmente, la densidad posterior exacta se expresó como:

$$p(\pi_a \mid y) = \frac{[(2p-1)\pi_a + (1-p)]^y \, [1 - (2p-1)\pi_a - (1-p)]^{n-y}}{Z}$$

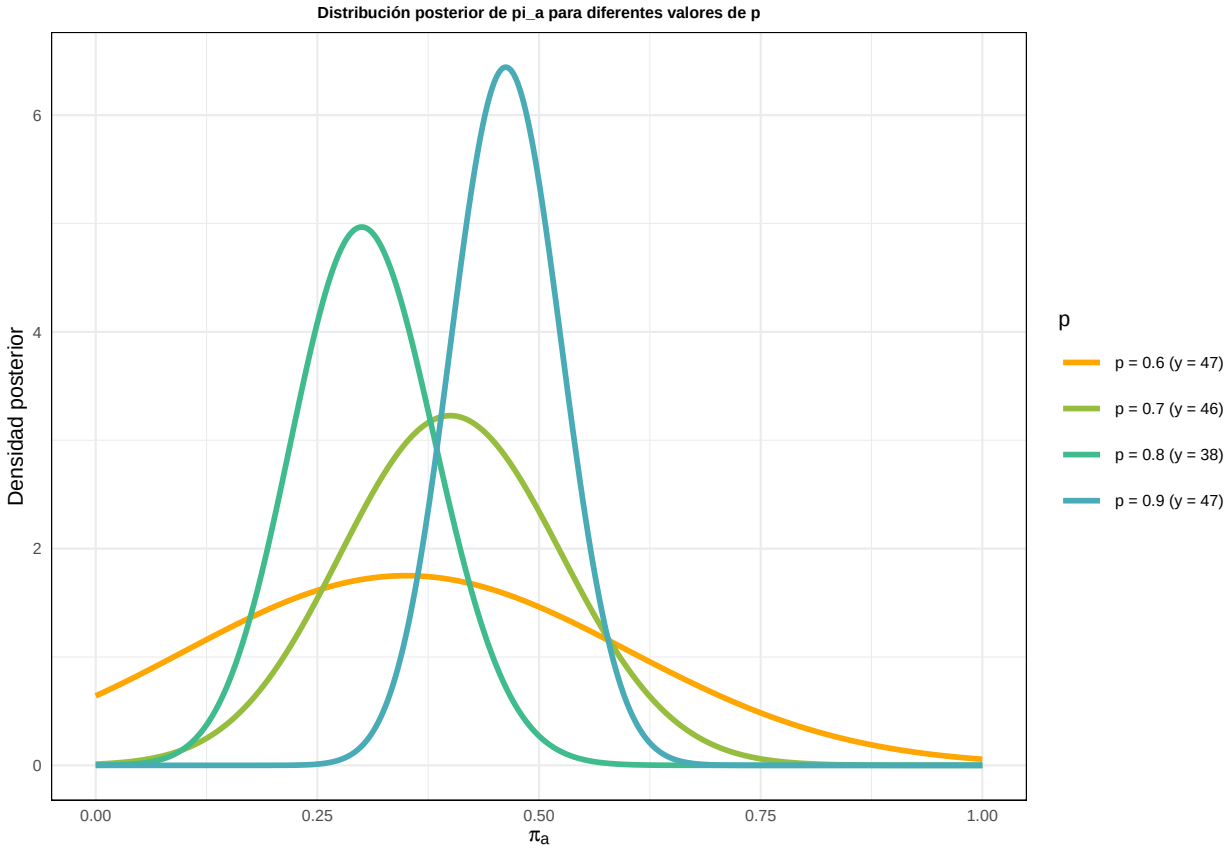
7. Gráfico del posterior para diferentes valores de p

Se analizó el comportamiento de la distribución posterior $p(\pi_a \mid y)$ para distintos valores de p, manteniendo constante la proporción real de apostadores en la población. Se consideró una población de tamaño $N = 1000$, de la cual se simuló una única muestra de tamaño $n = 100$, suponiendo que la proporción real de apostadores era $\pi_a = 0.4$.

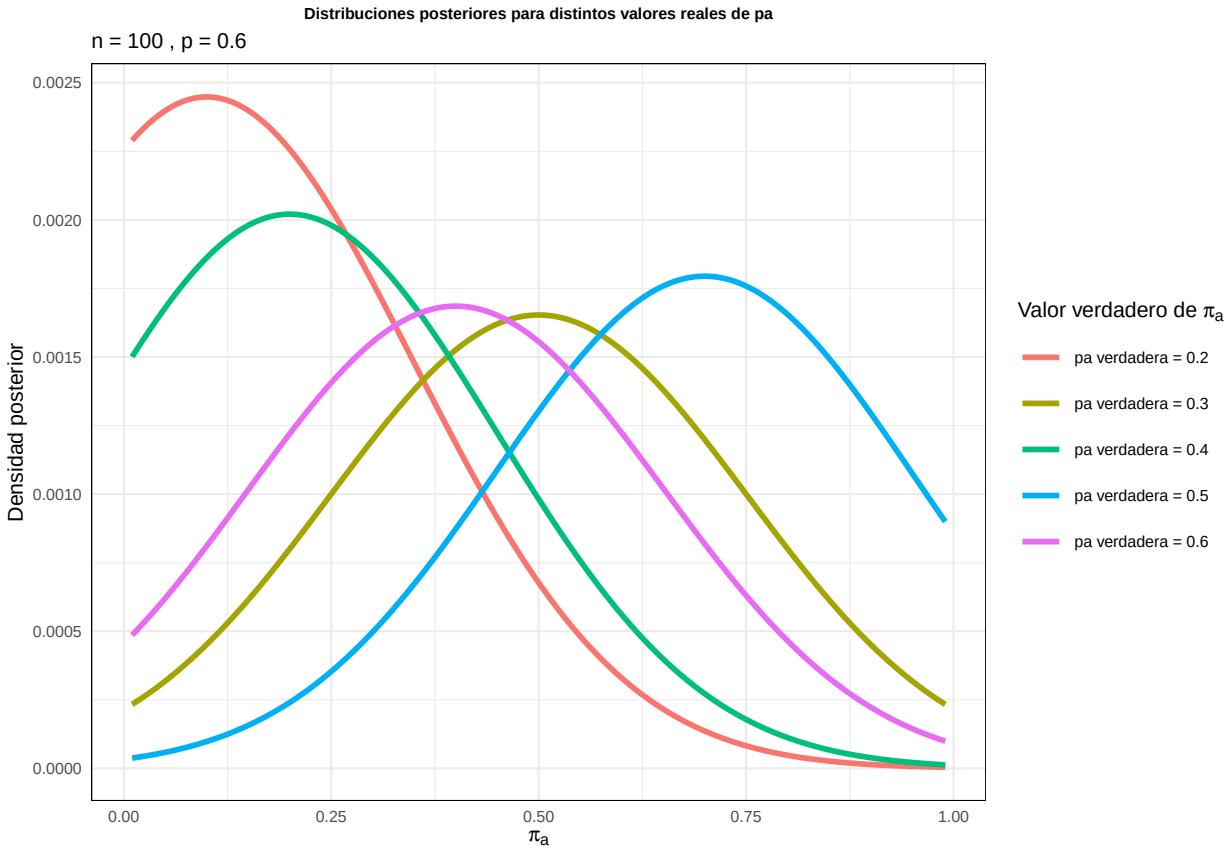
Según lo planteado en el ítem 5, se asumió que los datos se generaban siguiendo una distribución binomial de parámetro $\lambda = (2p-1)\pi_a + (1-p)$, lo cual permitió obtener directamente la variable de interés $Y \sim \text{Binomial}(n = 100, \lambda)$.

Para ilustrar el efecto del parámetro p, se consideraron los siguientes valores: $p = 0.6$, 0.7, 0.8 y 0.9. Para cada uno de ellos se simuló un único valor de y y se calculó el posterior exacto de π_a sobre una grilla de valores en $[0, 1]$.

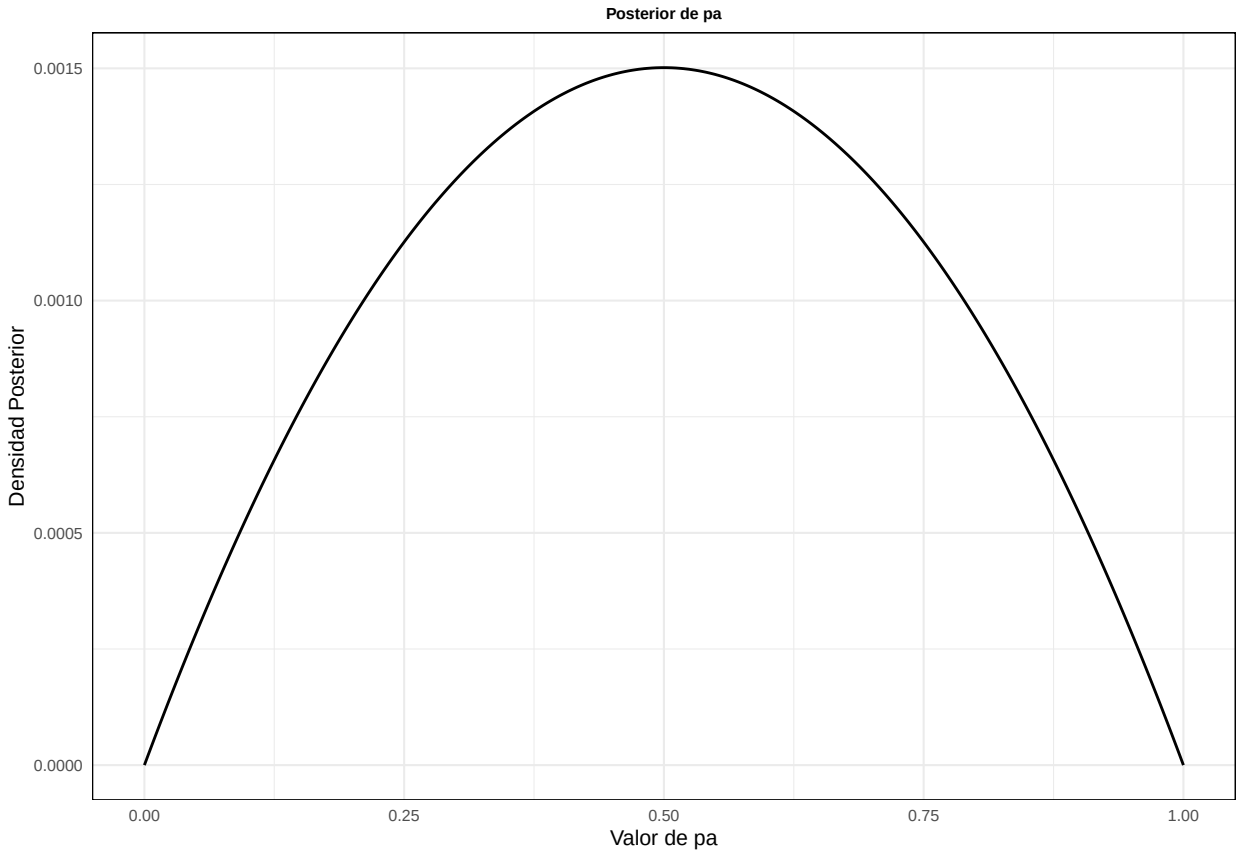
El siguiente código presenta la simulación y los gráficos correspondientes:


Distribución posterior de pi_a para diferentes valores de p
Densidad posterior
0
2
4
6
0.00
0.25
0.50
0.75
1.00
π_a
p
p = 0.6 (y = 47)
p = 0.7 (y = 46)
p = 0.8 (y = 38)
p = 0.9 (y = 47)

punto 8


Distribuciones posteriores para distintos valores reales de pa
n = 100 , p = 0.6
Densidad posterior
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0.00
0.25
0.50
0.75
1.00
π_a
Valor verdadero de π_a
pa verdadera = 0.2
pa verdadera = 0.3
pa verdadera = 0.4
pa verdadera = 0.5
pa verdadera = 0.6

punto 9



Método de Greenberg: Probabilidad de respuesta y función de inferencia

Pregunta 10

Para el método de Greenberg, se estimaron las probabilidades de que un estudiante responda "sí" o "no", en función de la probabilidad de selección de cada pregunta. En este caso, se consideró una probabilidad $p = 0.5$ de que se seleccione la pregunta sensible (la que indaga si el estudiante apuesta en línea).

La probabilidad total de que un estudiante responda "sí" estuvo dada por:

$$\lambda_G = p \cdot \pi_A + (1 - p) \cdot (1 - \pi_B)$$

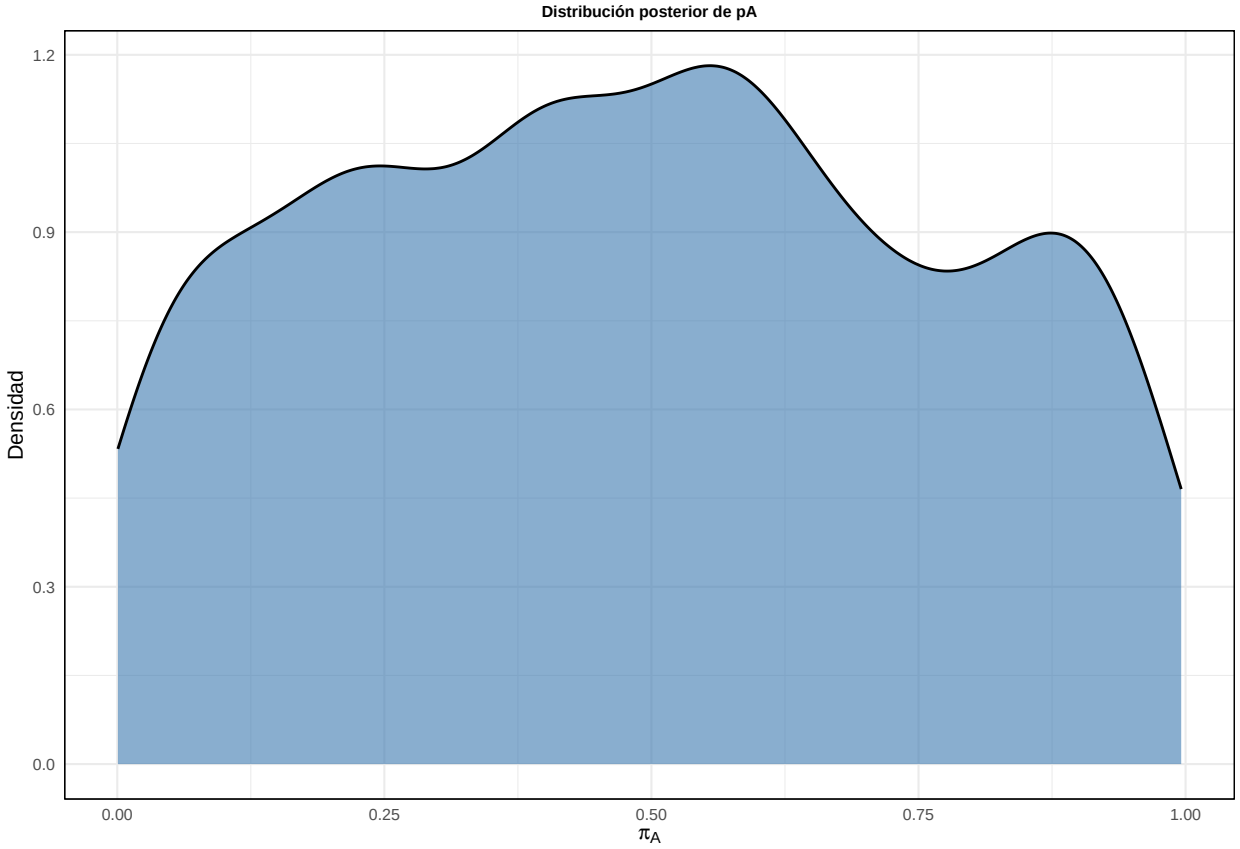
donde: - π_A: proporción real de estudiantes que apuestan. - π_B: proporción de estudiantes que **no apuestan**, usada como control.

Para el caso en que $\pi_A = \pi_B = \pi$, la expresión se redujo a:

$$\lambda_G = p \cdot \pi + (1 - p) \cdot (1 - \pi)$$

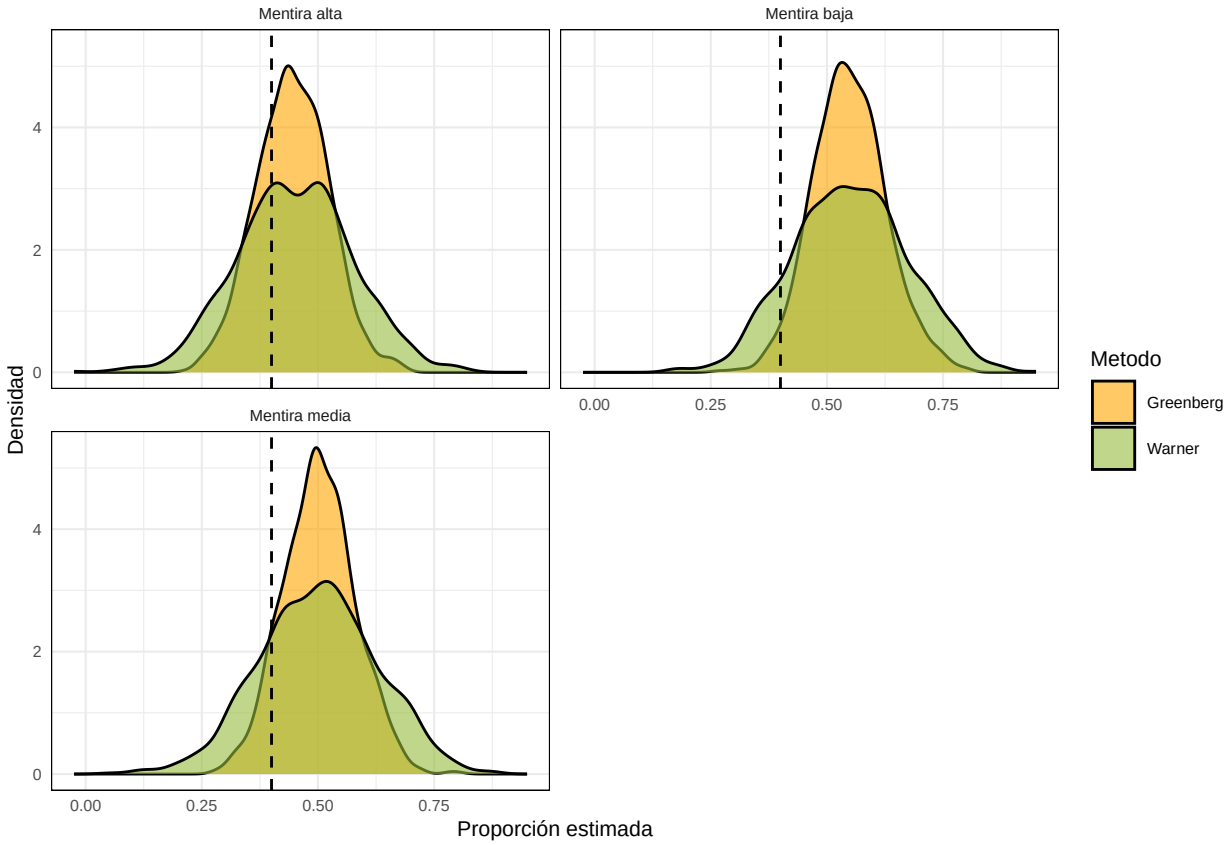
Y se implementó de la siguiente manera:

punto 11


Distribución posterior de pA
Densidad
1.2
0.9
0.6
0.3
0.0
0.00
0.25
0.50
0.75
1.00
π_A

12. Comparación de escenarios: Sin mentira, Mentira (bajo, medio, alto), Warner y Greenberg

En este punto, se compararon los métodos de respuesta aleatorizada de Warner y Greenberg bajo cuatro escenarios distintos, definidos por el nivel de veracidad de los encuestados: no mienten, mentira baja, mentira media y mentira alta. Para cada combinación de método y nivel de mentira, se simuló una única muestra de tamaño fijo con el fin de obtener una estimación puntual de la proporción real de estudiantes que participan en apuestas en línea. Esta comparación permitió observar cómo varía la estimación según el método utilizado y el grado de sinceridad de las respuestas.



13. Simulaciones repetidas (1000 repeticiones)

Para evaluar la estabilidad y precisión de los métodos utilizados, se repitió la simulación un total de 1000 veces. Este enfoque permitió analizar el sesgo y la varianza de cada método bajo diferentes escenarios, proporcionando una mejor comprensión de su desempeño en estimaciones repetidas. Los resultados obtenidos permitieron comparar de manera más robusta las inferencias generadas por cada uno de los métodos aplicados.

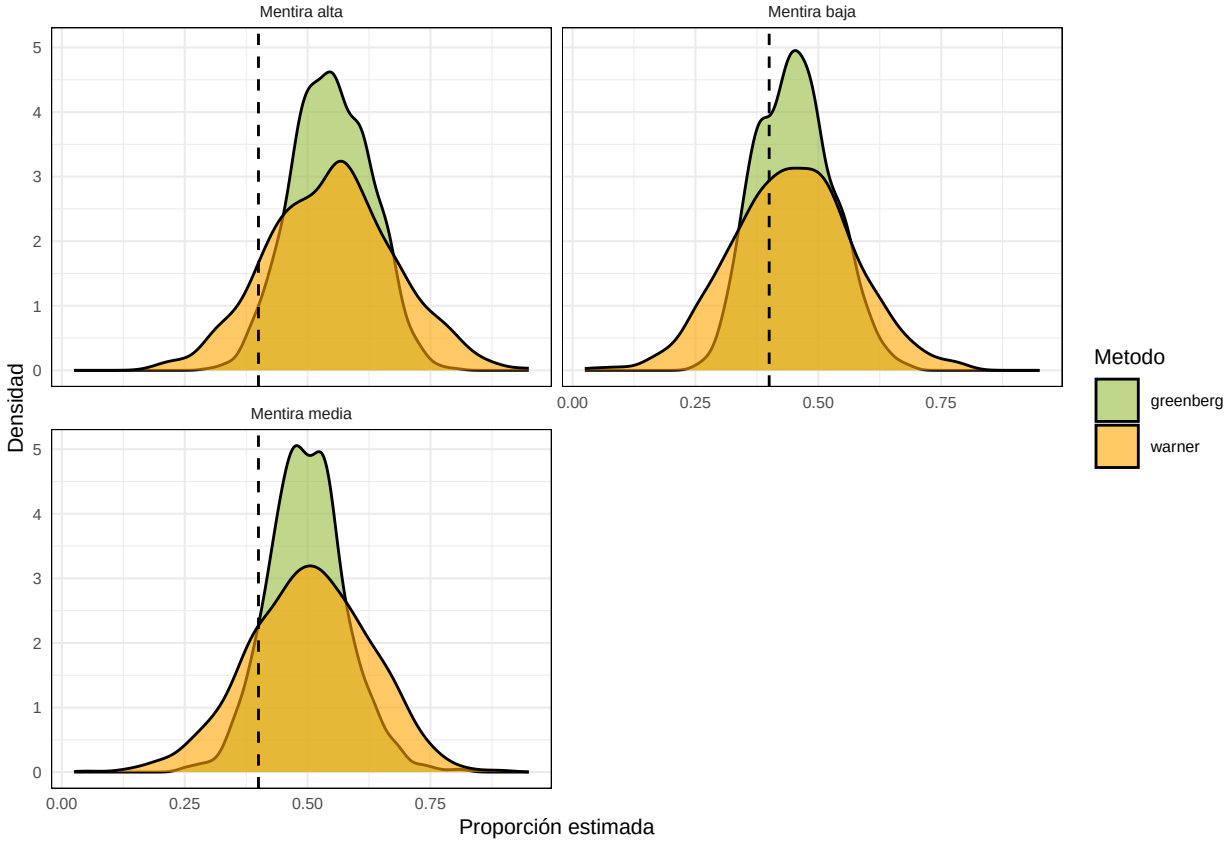


Figura 1: Distribución de las proporciones estimadas según nivel de mentira. Se comparan los métodos de Warner y Greenberg.