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:

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

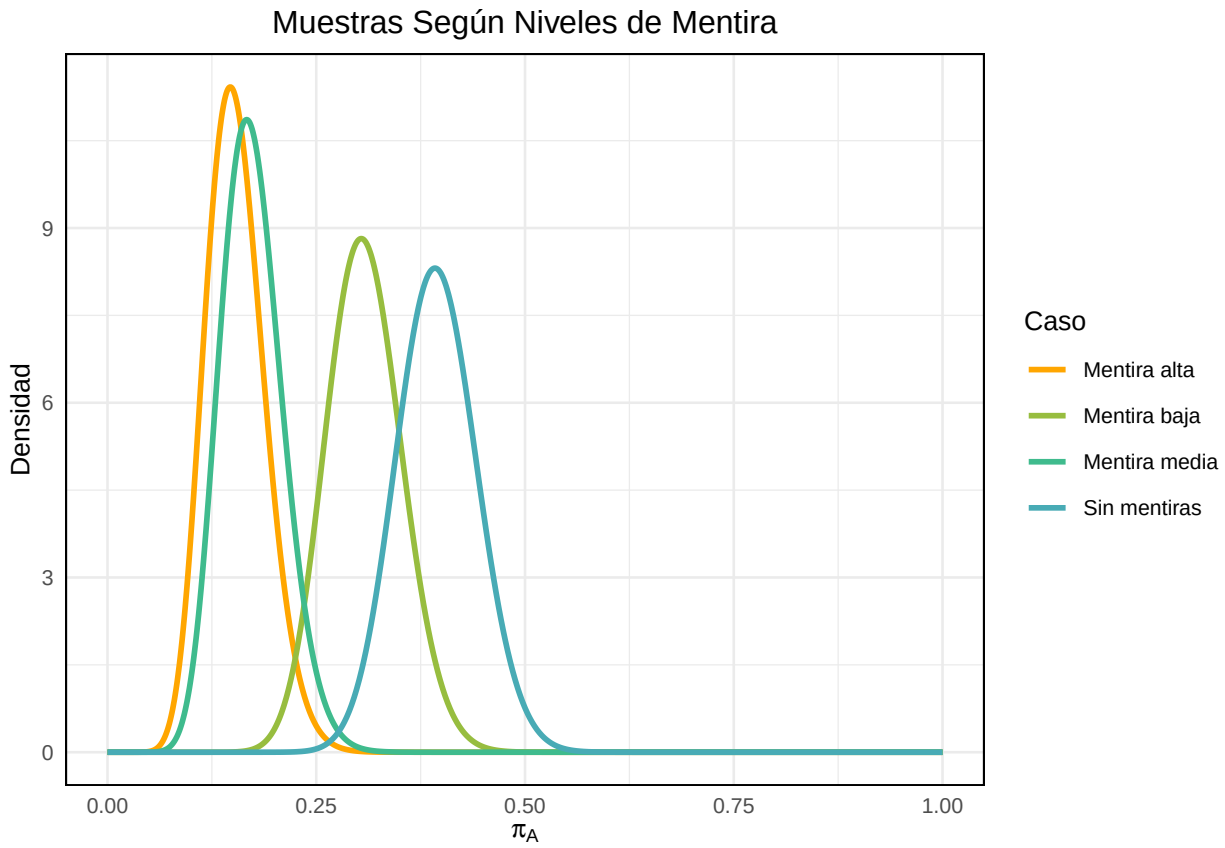
Sustituyendo los parámetros del prior elegido ($\alpha = 2$,$\beta = 2$), se obtiene la distribución posterior para este modelo:

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

Una vez establecido el modelo bayesiano para la estimación de π_a, se procede a evaluar su comportamiento en la práctica mediante simulación. Utilizando el software R para generar conjuntos de datos que representen tanto el escenario ideal donde los estudiantes no mienten, como escenarios alternativos donde se introducen tres disintos niveles de mentira: bajo($\mu = 0.25$), medio($\mu = 0.50$) y alto($\mu = 0.75$). El análisis comparativo de los resultados inferenciales obtenidos en cada caso nos permitirá comprender el impacto de la sinceridad de las respuestas en la estimación del parámetro de interés para una muestra.

```r
#Configuración inicial
set.seed(123)
n <- 100
pi_a <- 0.4
alpha_prior <- 2
beta_prior <- 2

# Simulación de y
```

```r
y <- rbinom(1, n, pi_a)

# Función para calcular posterior por nivel de mentira
posterior <- function(y, n, mu = 0) {
  y_observado <- y - rbinom(1, y, mu)
  alpha_post <- alpha_prior + y_observado
  beta_post <- beta_prior + n - y_observado
  list(y = y, y_obs = y_observado, alpha = alpha_post, beta = beta_post)
}

# Simulaciones para los 4 casos
sin_mentiras <- posterior(y, n, mu = 0)
mentira_baja <- posterior(y, n, mu = 0.25)
mentira_media <- posterior(y, n, mu = 0.5)
mentira_alta <- posterior(y, n, mu = 0.75)
```



Método propuesto por Warner

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:

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

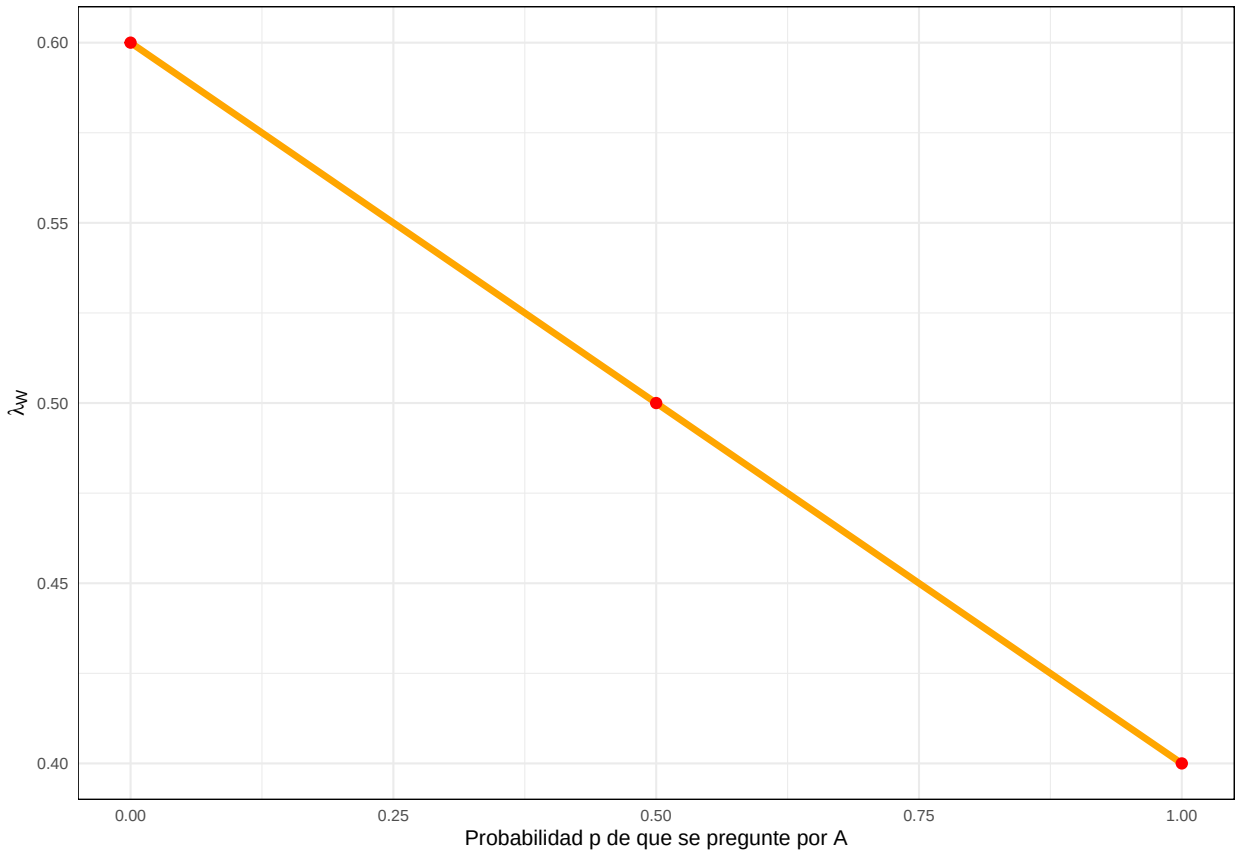
Donde:

- 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 el valor propuesto de $\pi_A = 0.4$, proporción de alumnos que apuestan, se obtuvo lo siguiente:

$$\lambda_W = p \cdot 0.4 + (1 - p) \cdot (1 - 0.4) = p \cdot 0.4 + (1 - p) \cdot 0.6 = 0.4p + 0.6(1 - p) = 0.6 + 0.4p - 0.6p = 0.6 - 0.2p$$

Por lo tanto, se concluyó que la probabilidad de que un estudiante respondiera afirmativamente bajo el método de Warner fue de $\lambda_W = 0.6 - 0.2p$. Es una función linea que depende de la probabilidad de que el encuestado recibiera la pregunta sobre su participación en apuestas.



Como se observo en la figura @fig-lambdaw se puede ver que a medida que la probabilidad de que el encuestado recibiera la pregunta sobre A la probabilidad de obtener respuestas afirmativas disminuye linealmente.

Consecuentemente, la probabilidad de que un estudiante respondiera negativamente (es decir, "no") fue $1 - \lambda_W$. Sustituyendo el resultado encontrado en el punto anterior se obtiene :

$$1 - \lambda_W = 1 - (0.6 - 0.2p) = 1 - 0.6 + 0.2p = 0.4 + 0.2p$$

En contraposición la probabilidad de obtener respuestas negativas esta relacionada linealmente con la probabilidad de que el encuestado recibiera la pregunta sobre A, de manera positiva. Por lo que al aumentar la probabilidad de el encuestado recibiera la pregunta sobre A, la probabiliad de obtener respuestas negativas aumenta.

5. Modelo propuesto para la generación de datos con el Método de Warner

$$Y_i \sim \text{Bernoulli}(\lambda_W), \quad \text{donde} \quad \lambda_W = p \cdot \pi_A + (1 - p) \cdot (1 - \pi_A)$$

Este modelo refleja que cada respuesta Y_i es generada a partir del mecanismo de Warner, que alterna aleatoriamente entre preguntar si el estudiante pertenece a la categoría A (apostadores) o a su complemento A^c, lo cual permite preservar el anonimato y obtener inferencias válidas sobre π_A.

Dado que se supone que $\pi_A = 0.4$, entonces el modelo propuesto queda de esta forma:

$$Y_i \sim \text{Bernoulli}(\lambda_W), \quad \text{donde} \quad \lambda_W = 0.6 - 0.2p, 0 \leq p \leq 1$$

```r
n <- 100              # tamaño muestral
pi_A <- 0.4           # proporción real de apostadores
p <- 0.7              # probabilidad de que se haga la pregunta directa

lambda_W <- p * pi_A + (1 - p) * (1 - pi_A) # Cálculo de lambda_W

respuestas <- rbinom(n, size = 1, prob = lambda_W) # Simulación Bernoulli(lambda_W)
```

6. Posterior exacto bajo prior uniforme

Queremos obtener la distribución posterior de π_A, que representa la proporción real de estudiantes que apuestan, utilizando el método de Warner.

Partimos de un **prior uniforme** sobre π_A, es decir:

$$\pi_A \sim \text{Beta}(1, 1)$$

La densidad de esta distribución es constante en el intervalo $[0, 1]$:

$$p(\pi_A) = \frac{\pi_A^{\alpha-1}(1-\pi_A)^{\beta-1}}{B(\alpha, \beta)}$$

Bajo el mecanismo de Warner, un estudiante recibe con probabilidad p la pregunta directa (si apuesta), y con probabilidad $1-p$, la pregunta contraria (si no apuesta). Entonces, la probabilidad de que un estudiante **responda afirmativamente** es:

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

Supongamos que observamos y respuestas afirmativas en una muestra de tamaño n. Con esto se obtiene la siguiente verosimilitud:

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

$$p(y|\pi_A) = \binom{n}{y} \lambda_W^y (1-\lambda_W)^{n-y} \cdot \frac{\pi_A^{\alpha-1}(1-\pi_A)^{\beta-1}}{B(\alpha, \beta)}$$

$$p(\pi_A|y) \propto \lambda_W^y (1-\lambda_W)^{n-y} \cdot \pi_A^{\alpha-1}(1-\pi_A)^{\beta-1}$$

$$p(\pi_A|y) \propto \lambda_W^y (1-\lambda_W)^{n-y} \cdot \pi_A^{1-1}(1-\pi_A)^{1-1}$$

Como los exponentes de π_A y $(1-\pi_A)$ son ambos 0, la ecuación se simplifica a:

$$p(\pi_A|y) \propto \lambda_W^y (1-\lambda_W)^{n-y}$$

Para normalizar esta distribución, necesitamos una constante Z, que depende de la integral sobre el dominio de π_A. Es decir, tenemos que calcular la constante de normalización Z, que será una integral:

$$p(\pi_A|y) = \frac{N(\pi_A)}{Z}$$

donde Z es la constante de normalización, que puede escribirse como:

$$Z = \int_0^1 \lambda_W^y (1 - \lambda_W)^{n-y} \, d\pi_A$$

La integral Z tiene una forma conocida, y se puede resolver usando la función **beta incompleta**. La integral de Z es la siguiente:

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

donde $B(x; a, b)$ es la **función beta incompleta**, definida como:

$$B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1} \, dt$$

Finalmente, la distribución posterior se expresa como:

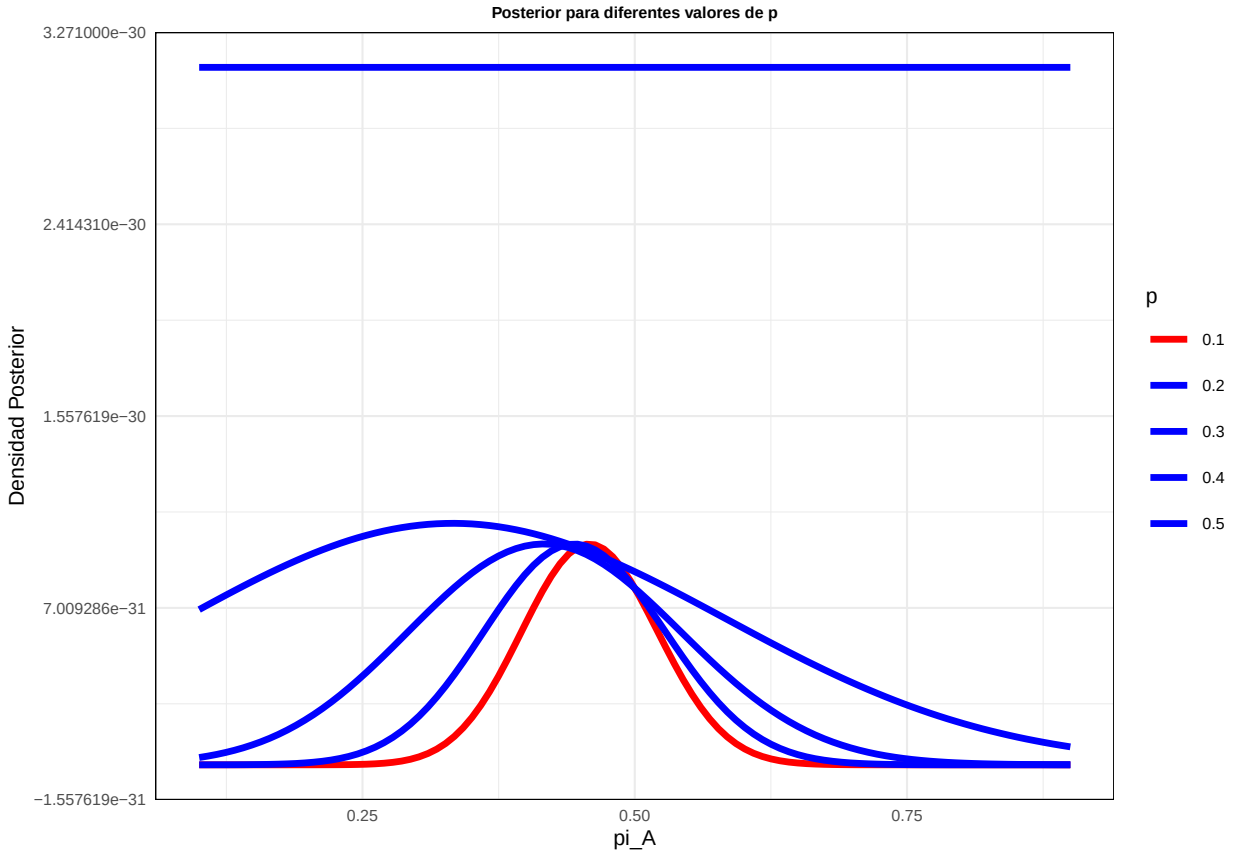
$$p(\pi_A|y) = \frac{\lambda_W^y (1 - \lambda_W)^{n-y}}{Z}$$

El posterior exacto está dado por esta expresión, con la constante de normalización Z calculada mediante la función beta incompleta.

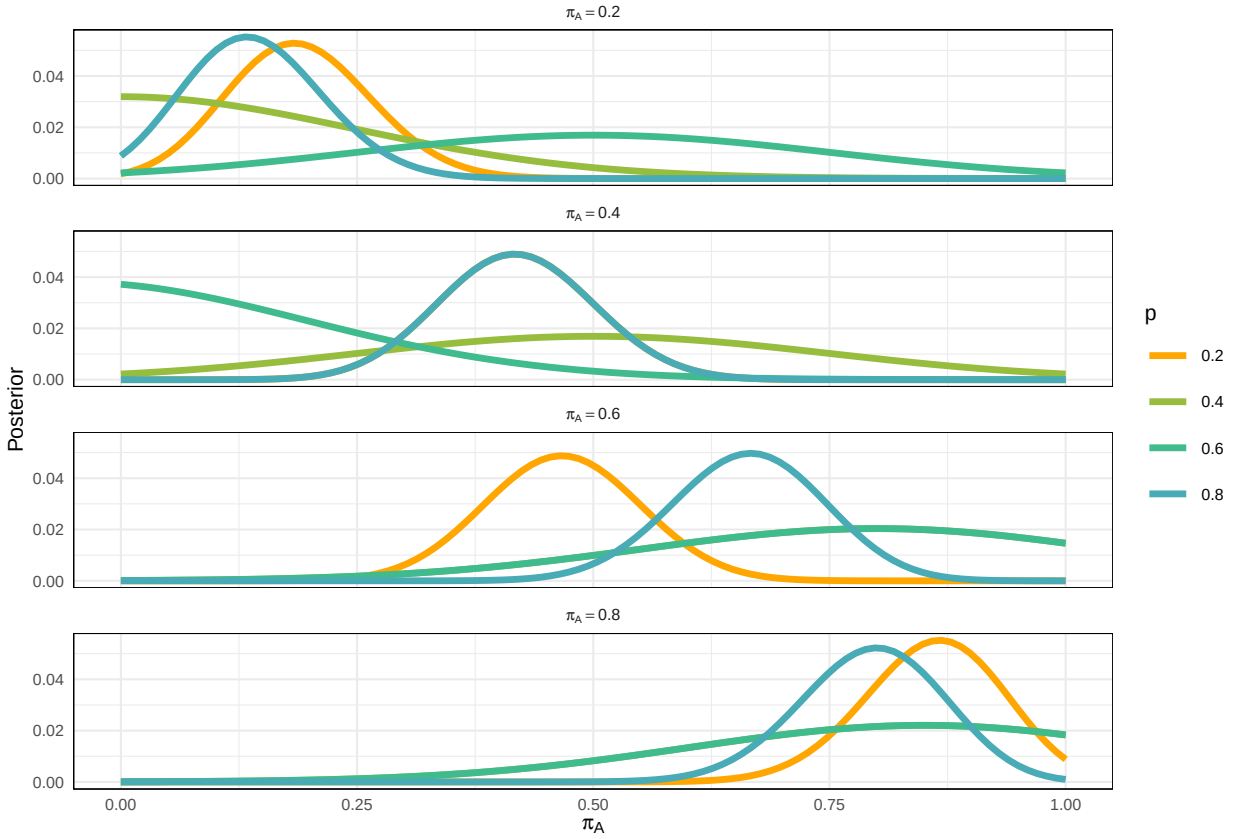
7. Gráfico del posterior para diferentes valores de p

```
pi <- 0.4  # Proporción real de estudiantes que apuestan
n <- 100    # Tamaño de la muestra
p_values <- c(0.1, 0.2, 0.3, 0.4, 0.5)  # Diferentes valores de p

calculate_posterior <- function(p, pi, n, y) {

  lambda <- p * pi + (1 - p) * (1 - pi) # Calcular lambda para el valor dado de p
  pi_grid <- seq(0.1, 0.9, length.out = 100) # Crear la cuadrícula de valores de pi

  lambda_grid <- p * pi_grid + (1 - p) * (1 - pi_grid) # Calcular los valores de lambda para cada pi en

  # Calcular el posterior para cada valor de pi
  posterior <- lambda_grid^y * (1 - lambda_grid)^(n - y) / pbeta(1 - p, y + 1, n - y + 1)

  # Devolver un tibble con los resultados
  return(tibble(p = rep(p, length(pi_grid)), pi = pi_grid, posterior = posterior))
}


y_values <- replicate(100, {
  lambda <- p_values[3] * pi + (1 - p_values[3]) * (1 - pi)  # Usando p = 0.3 como
  rbinom(1, n, lambda)  # Muestra binomial
})

# Calcular el posterior para cada valor de p y para cada muestra
```

```r
posterior_data <- lapply(p_values, function(p_val) {
  # Usar la media de y_values para obtener un posterior promedio
  y_avg <- mean(y_values)
  calculate_posterior(p_val, pi, n, y_avg)
}) %>%
  bind_rows()

# Graficar los posteriores usando ggplot2
ggplot(posterior_data, aes(x = pi, y = posterior, color = as.factor(p))) +
  geom_line(size = 1.2) +
  scale_color_manual(values = c("red", rep("blue", length(p_values)-1))) +
  labs(x = "pi_A", y = "Densidad Posterior", title = "Posterior para diferentes valores de p", color =
  theme_minimal(base_size = 8) +
  theme(panel.border = element_rect(color = "black", fill = NA, size = 0.3), plot.title = element_text(
```



punto 8



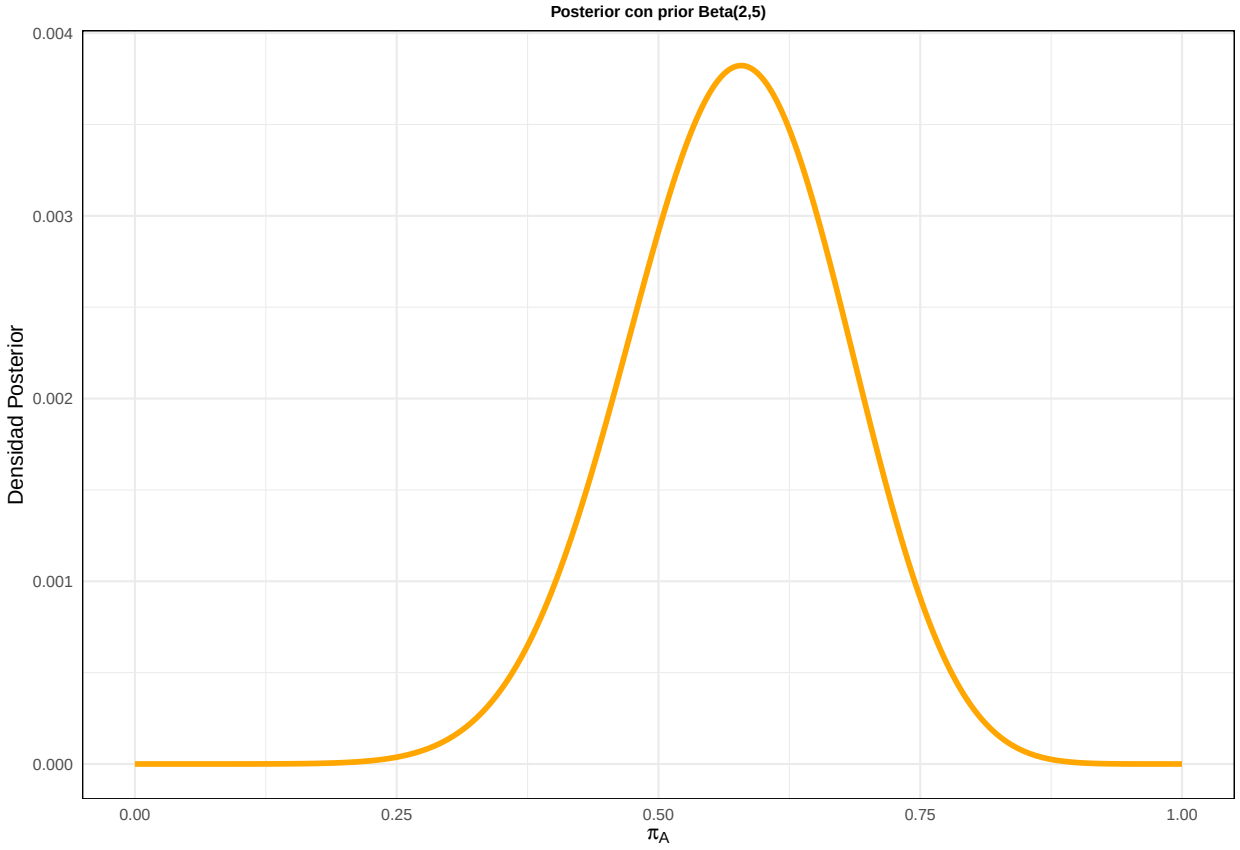
punto 9

```r
inferencia_beta_grid <- function(y, n, p, alpha = 1, beta = 1, grid_length = 1000) {

  pi_grid <- seq(0, 1, length.out = grid_length) # Grilla de valores de pi_A
  lambda <- p * pi_grid + (1 - p) * (1 - pi_grid) # Cálculo de lambda

  likelihood <- dbinom(y, size = n, prob = lambda)# Verosimilitud para cada lambda
  prior <- dbeta(pi_grid, shape1 = alpha, shape2 = beta)# Prior beta

  unnorm_posterior <- likelihood * prior # posterior
  posterior <- unnorm_posterior / sum(unnorm_posterior)# Normalizar

  return(tibble(pi = pi_grid, posterior = posterior))# Retornar la grilla y el posterior
}
```

Se construyó una función en R que aproximó la distribución posterior de π_A utilizando una grilla de valores posibles entre 0 y 1. Para cada valor en la grilla, se calculó la probabilidad $\lambda_W = p \cdot \pi_A + (1-p) \cdot (1-\pi_A)$, que representa la probabilidad de responder "sí" bajo el modelo de Warner. Luego, se evaluó la verosimilitud binomial de los datos observados en función de λ_W, y se multiplicó por la densidad del prior Beta(α, β) evaluado en cada punto de la grilla.

Finalmente, los valores obtenidos se normalizaron dividiendo por la suma total, generando así una aproximación discreta de la densidad posterior. Esta metodología permitió incorporar priors informativos y visualizar cómo se actualizaban las creencias sobre π_A luego de observar los datos.

```
posterior_df <- inferencia_beta_grid(y = 42, n = 100, p = 0.3, alpha = 2, beta = 5)
```



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:

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

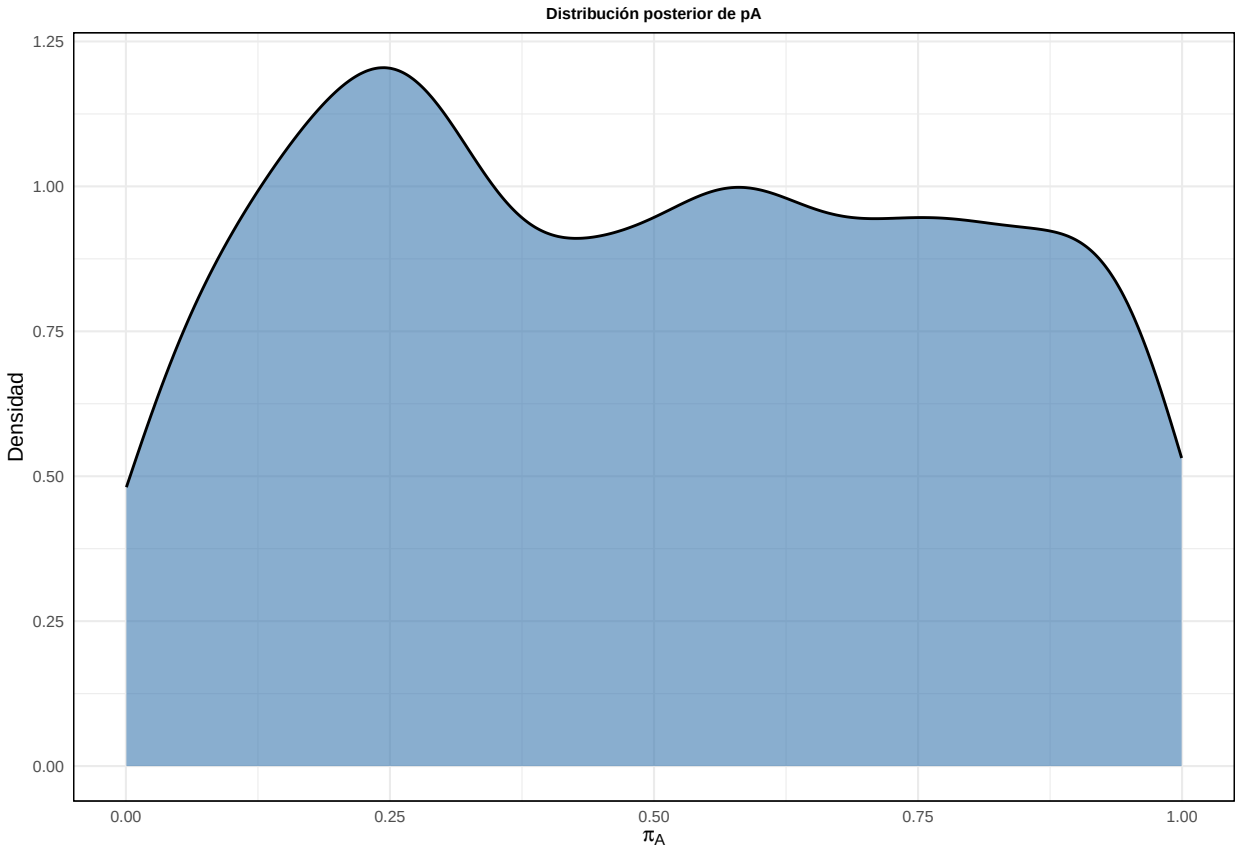
donde: - π_A: proporción real de estudiantes que apuestan. - π_B: proporción de estudiantes que **no apuestan**, usada como control.

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

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

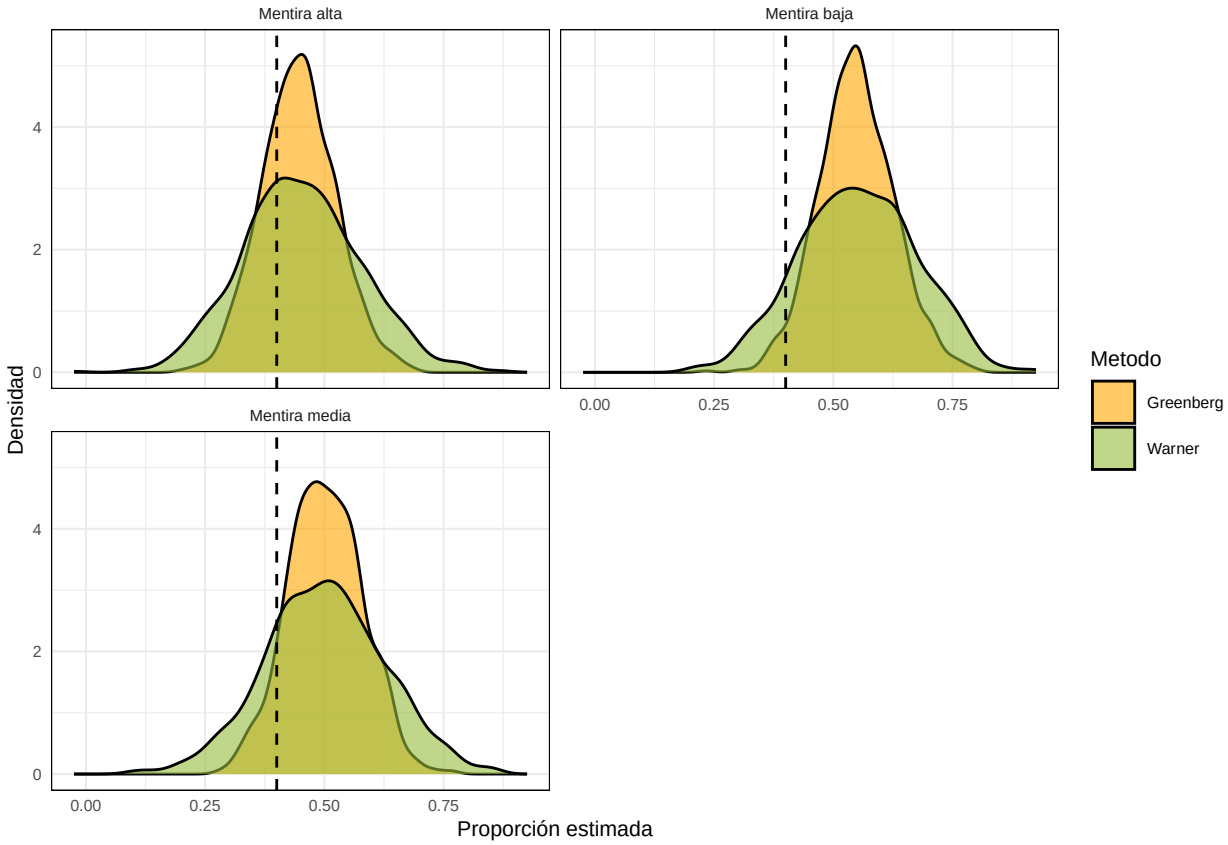
Y se implementó de la siguiente manera:

punto 11


Distribución posterior de pA
Densidad
1.25
1.00
0.75
0.50
0.25
0.00
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.

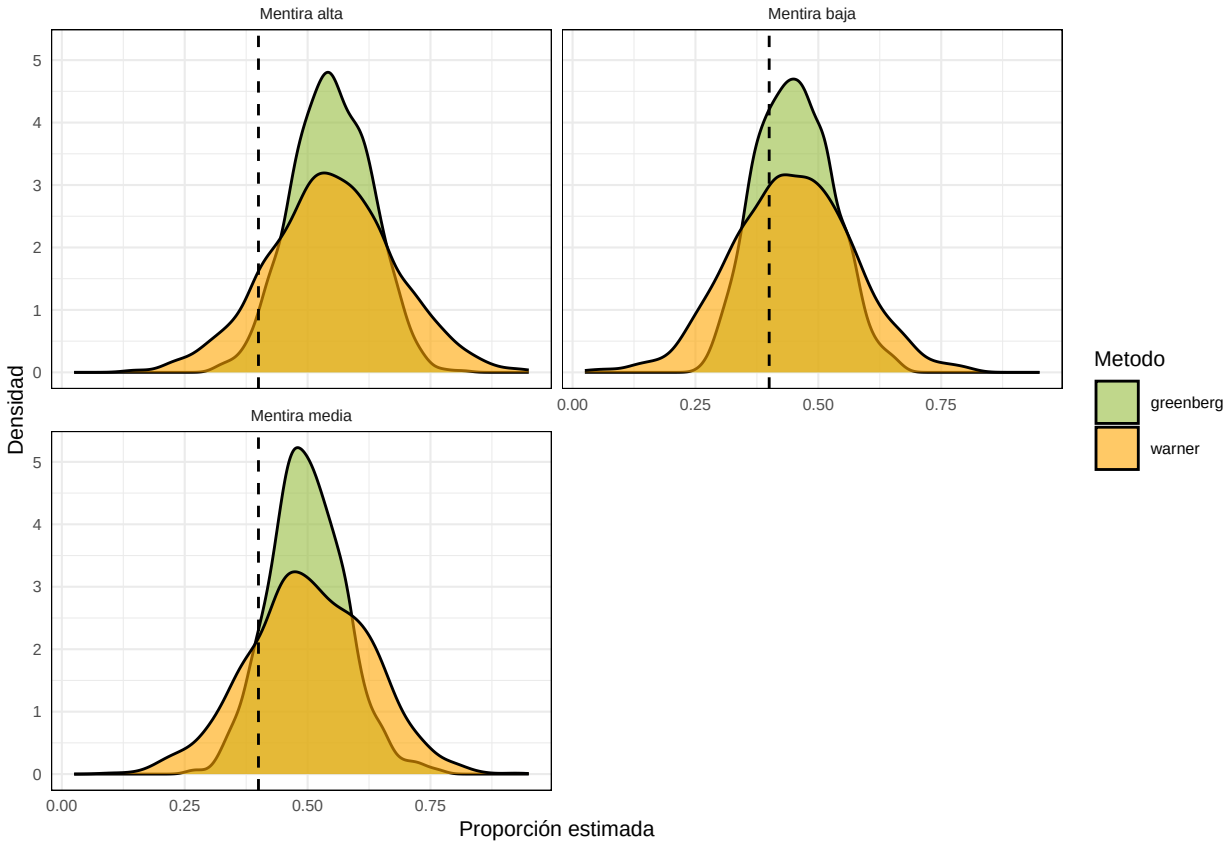


Figura 1: Distribución de las proporciones estimadas según nivel de mentira. Se comparan los métodos de Warner y Greenberg.