REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (8). Páginas 50–70.
Diciembre de 2009. ISSN: 1886-516X. D.L: SE-2927-06.
URL: http://www.upo.es/RevMetCuant/art31.pdf

Análisis bayesiano para la diferencia de dos proporciones usando R

Gutiérrez Rojas, Hugo Andrés
Centro de Investigaciones y Estudios Estadísticos (CIEES)
Universidad Santo Tomás (Bogotá, Colombia)
Correo electrónico: hugogutierrez@usantotomas.edu.co

Zhang, Hanwen
Centro de Investigaciones y Estudios Estadísticos (CIEES)
Universidad Santo Tomás (Bogotá, Colombia)
Correo electrónico: hanwenzhang@usantotomas.edu.co

RESUMEN

Este artículo presenta una colección de funciones computacionales que son utilizadas en la implementación de un análisis bayesiano exhaustivo para la diferencia de dos proporciones. Con este fin, se discute la estimación puntual, la estimación mediante intervalos de credibilidad y la inferencia predictiva desde dos escenarios: el primero basado en las densidades exactas *a priori* y *a posteriori* (construidas mediante la primera función hipergeométrica de Appell) y el segundo basado en densidades simuladas (mediante un algoritmo de cadenas de Markov con métodos de Monte Carlo). La implementación de estas funciones se realiza en el programa estadístico R, porque es un *software* libre, funciona bien en múltiples plataformas y permite enmarcar estas funciones bajo un objeto computacional denominado "paquete".

Palabras clave: estimación; funciones en R; inferencia bayesiana; proporciones.
Clasificación JEL: C11; C12; C63.
2000MSC: 62C10; 62F03; 62P20; 90-08.

Artículo recibido el 13 de octubre de 2009 y aceptado el 30 de noviembre de 2009.

Bayesian Analysis for the Difference of Two Proportions Using R

ABSTRACT

In this paper we present a collection of functions that can be used to implement a comprehensive Bayesian analysis of a difference of two proportions. For instance, point estimation, credibility intervals and predictive inference are discussed in both scenarios, the *priori* and *posteriori* exact densities (based in the first Appell hypergeometric function) and the simulated densities (based in a Markov chain Monte Carlo algorithm). We have chosen to implement the suite of functions using the R statistical software because it is freely available, runs on multiple platforms and allows to compress the functions into a single computational object named "package".

Keywords: Bayesian inference; estimation; proportions; R functions.
JEL classification: C11; C12; C63.
2000MSC: 62C10; 62F03; 62P20; 90-08.

1. Introducción

En las últimas décadas el enfoque bayesiano ha sido uno de los tópicos más desarrollados en la ciencia estadística. Los avances computacionales, como los algoritmos de cadenas de Markov basados en simulaciones de Monte Carlo (MCMC), han hecho que la utilización de los métodos bayesianos sea cada vez mas común por parte del investigador. Sin embargo, como afirman Agresti & Min (2005), las técnicas bayesianas no son muy usadas cuando se trata de la inferencia de tablas 2×2, siendo éste uno de los problemas más comunes en la práctica, específicamente el análisis de la diferencias de proporciones.

Este artículo está enfocado al caso en que las distribuciones *a priori* para cada una de las dos proporciones se rige por una distribución de tipo Beta[1], siguiendo los importantes resultados de Pham-Gia & Turkkan (1993), en donde se encontró la distribución exacta *a posteriori* de la diferencia de proporciones, la cual está en términos de la primera función hipergeométrica de Appell. Sin embargo, en el software estadístico `R`, a nuestro conocimiento actual, no existen funciones, rutinas o paquetes que calculen dicha función; en otras palabras, no es posible realizar inferencias bayesianas exactas para el problema en cuestión, aunque sí es posible hacerlo utilizando procedimientos de simulación mediante métodos de Monte Carlo.

En algunos softwares comerciales de uso frecuente, existen algunas rutinas que implementan procedimientos bayesianos –no exactos– para realizar inferencias acerca de la diferencia de dos proporciones binomiales. Entre ellos está el procedimiento `BGENMOD` de `SAS` (SAS 2006) y la librería `flexBayes` de `S-PLUS` (Jack, Woodard, Hoffman & O'Connell 2007). Por otra parte, existen softwares de uso libre que también permiten implementar un análisis bayesiano, basado en métodos de Monte Carlo, para el problema en cuestión; entre ellos están `BUGS` (Spiegelhalter, Thomas, Best & Lunn 2004) y `EPIDAT` (Magidson 2004). Sin embargo, los autores abogan por la utilización del software `R` puesto que es de libre acceso, ejecutable en una variedad de sistemas operativos incluyendo Windows, Unix y MacOS. Además, provee una plataforma para la programación de nuevos métodos estadísticos de una manera sencilla, contiene rutinas estadísticas avanzadas que aún no están disponibles en otros softwares y genera potentes gráficos actualizados con el estado de la cuestión. Adicionalmente, el código escrito en `R` puede ser ejecutado en softwares comerciales como `SAS`, `SPSS` y `S-PLUS`.

Basado en lo anterior, este artículo presenta un código computacional eficiente que permite realizar inferencia bayesiana tanto exacta como simulada mediante métodos MCMC[2]. Estas funciones son parte de un nuevo paquete del ambiente computacional de `R` desarrollado por los autores del presente artículo.

El artículo está divido de la siguiente forma: en la Sección 2 se presentan los resultados básicos de la inferencia bayesiana que son usados para la estimación y elaboración de los códigos computacionales. En la Sección 3 se presentan los resultados acerca de la distribución exacta *a posteriori* para la diferencia de proporciones y el algoritmo MCMC,

[1] Cubriendo casos importantes como la distribución uniforme o la distribución no informativa de Jeffreys.
[2] Dado que los resultados finales, aunque similares, no son idénticos.

que utiliza el muestreo de Gibbs, para la distribución simulada. En la Sección 4 se introduce la descripción de cada una de las funciones pertenecientes al paquete propuesto. En la Sección 5, se analiza un conjunto de datos aplicados a la economía del mercadeo empresarial mediante el uso de estas funciones computacionales. Finalmente, en la Sección 6 se abordan algunas conclusiones acerca de la inferencia para la diferencia de proporciones.

2. Fundamentos teóricos

En esta sección, se presentan los resultados básicos de la estadística bayesiana en la inferencia, específicamente en la estimación puntual y la construcción de intervalos de credibilidad. Para detalles más específicos, el lector puede referirse a Gelman, Carlin, Stern & Rubin (1995).

El enfoque bayesiano, además de especificar un modelo para los datos observados $\mathbf{y} = (y_1, \ldots, y_n)$, dado un vector de parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, usualmente en forma de densidad condicional $f(\mathbf{y}|\boldsymbol{\theta})$, supone que $\boldsymbol{\theta}$ es aleatorio y que tiene un densidad *a priori* $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$, donde $\boldsymbol{\eta}$ es un vector de hiper-parámetros. De esta forma, la inferencia concerniente a $\boldsymbol{\theta}$ se basa en una densidad *a posteriori* $p(\boldsymbol{\theta}|\mathbf{y})$, dada por:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta})\ d\mathbf{u}}. \tag{1}$$

La expresión[3] $\int f(\mathbf{y}|\mathbf{u})\pi(\mathbf{u}|\boldsymbol{\eta})\ d\mathbf{u}$ puede ser considerada constante si no depende de $\boldsymbol{\theta}$ y suponiendo que $\mathbf{y}$ es fijo. De esta manera, (1) se convierte en:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta}). \tag{2}$$

Gelman, Carlin, Stern & Rubin (1995) menciona que esta expresión se conoce como la densidad *posterior* no-normalizada y encierra el núcleo técnico de la inferencia bayesiana. Con las anteriores expresiones, es posible calcular la probabilidad *a priori* de que $\boldsymbol{\theta}$ esté en una determinada región G como

$$Pr(\boldsymbol{\theta} \in G) = \int_G \pi(\boldsymbol{\theta}|\boldsymbol{\eta})\ d\boldsymbol{\theta} \tag{3}$$

y también es posible calcular la probabilidad *a posteriori* de que $\boldsymbol{\theta}$ esté en la región G dados los datos observados como

$$Pr(\boldsymbol{\theta} \in G|\mathbf{y}) = \int_G p(\boldsymbol{\theta}|\mathbf{y})\ d\boldsymbol{\theta}. \tag{4}$$

En términos de inferencia predictiva, existen dos etapas que cubren las «actuales» suposiciones acerca del vector de parámetros $\boldsymbol{\theta}$. En una primera etapa –antes de la observación

[3]Las integrales se deben cambiar por sumas en el caso discreto.

de los datos– la suposición «actual» de $\boldsymbol{\theta}$ está dada por la densidad *a priori* $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$. En estos términos, la distribución predictiva *a priori* de $\mathbf{y}$ está dada por:

$$p(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\eta})\ d\boldsymbol{\theta}. \tag{5}$$

En esta primera etapa es posible calcular, con fines confirmatorios (Carlin & Louis 1996), la estimación puntual para el vector $\boldsymbol{\theta}$ dada por alguna medida de tendencia central para la distribución $\pi(\boldsymbol{\theta}|\boldsymbol{\eta})$. En particular, si se escoge la media, entonces:

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) = \int \boldsymbol{\theta}\ \pi(\boldsymbol{\theta}|\boldsymbol{\eta})\ d\boldsymbol{\theta}. \tag{6}$$

También es posible calcular una región C de 100(1-α) % de credibilidad[4] para $\boldsymbol{\theta}$ que, en esta primera etapa, es tal que

$$1 - \alpha \leq Pr(\boldsymbol{\theta} \in C) = \int_C \pi(\boldsymbol{\theta}|\boldsymbol{\eta})\ d\boldsymbol{\theta}. \tag{7}$$

La segunda etapa –después de la recolección de los datos– actualiza las suposiciones acerca de $\boldsymbol{\theta}$, puesto que ahora éste sigue una distribución *a posteriori* dada por (1). Por lo tanto, la distribución predictiva *a posteriori* de $\mathbf{y}$ está dada por

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int f(\tilde{\mathbf{y}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\ d\boldsymbol{\theta}, \tag{8}$$

donde $f(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ es la función de verosimilitud evaluada en nuevos valores $\tilde{\mathbf{y}}$.

De esta forma, es posible calcular la estimación puntual para el vector $\boldsymbol{\theta}$, dados los datos observados. Ésta está dada por alguna medida de tendencia central para la distribución $p(\boldsymbol{\theta}|\mathbf{y})$. En particular, si se escoge la media, entonces:

$$\tilde{\boldsymbol{\theta}}(\mathbf{y}) = \int \boldsymbol{\theta}\ p(\boldsymbol{\theta}|\mathbf{y})\ d\boldsymbol{\theta}. \tag{9}$$

La región C de 100(1-α) % de credibilidad es tal que:

$$1 - \alpha \leq P(\boldsymbol{\theta} \in C|\mathbf{y}) = \int_C p(\boldsymbol{\theta}|\mathbf{y})\ d\boldsymbol{\theta}. \tag{10}$$

Suponiendo que existen dos modelos $M1$ y $M2$ candidatos para $\mathbf{y}$, se define el factor de Bayes en favor del modelo $M1$ como la razón de las densidades marginales de los datos para los dos modelos y es posible demostrar que equivale a la siguiente expresión:

$$FB = \frac{p(\mathbf{y}|M1)}{p(\mathbf{y}|M2)} = \frac{Pr(M1|\mathbf{y})/Pr(M2|\mathbf{y})}{Pr(M1)/Pr(M2)}. \tag{11}$$

Para evaluar esta última expresión es necesario recurrir a las expresiones (3) y (4). El factor de Bayes solo está definido cuando la integral de la densidad marginal de $\mathbf{y}$ bajo cada modelo converge.

[4]La interpretación de las regiones de credibilidad bayesianas difiere de la interpretación de las regiones de confianza frecuentistas. La primera se refiere a la probabilidad de que el verdadero valor de $\boldsymbol{\theta}$ esté en la región. La segunda se refiere a la región de la distribución muestral para $\boldsymbol{\theta}$ tal que, dados los datos observados, se podría esperar que el 100(α) % de las futuras estimaciones de $\boldsymbol{\theta}$ no pertenecieran a dicha región.

El muestreo de Gibbs

Una forma popular de simular valores de una distribución *a posteriori* es mediante los métodos MCMC, los cuales establecen una cadena de Markov irreducible y aperiódica para la cual la distribución estacionaria es igual a la distribución *a posteriori* de interés (Albert 2007). Basado en la anterior, cuando la distribución *a posteriori* del vector de parámetros es difícil de encontrar teóricamente, el esquema del muestreo de Gibbs proporciona una herramienta muy útil cuando esta distribución es muy complicada y resulta difícil o costoso computacionalmente simular las observaciones directamente de la distribución exacta *a posteriori*. Gamerman & Lopes (2006) describen el siguiente algoritmo para llevar a cabo el muestreo de Gibbs:

1. Obtener teóricamente, para $k = 1, \ldots, K$, las distribuciones condicionales *a posteriori*, de cada uno de los parámetros, $p(\theta_k|\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_K, \mathbf{y})$.

2. Fijar el número de iteraciones J, suficientemente grande para asegurar la convergencia de la cadena, y fijar valores iniciales para el vector de parámetros $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_K^{(0)})'$.

3. Obtener un nuevo valor $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \ldots, \theta_K^{(j)})'$, a partir de $\boldsymbol{\theta}^{(j-1)}$, a través de sucesivas iteraciones de valores:

$$\begin{aligned}
\theta_1^{(j)} &\sim p(\theta_1 \mid \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_K^{(j-1)}, \mathbf{y}), \\
\theta_2^{(j)} &\sim p(\theta_2 \mid \theta_1^{(j-1)}, \theta_2^{(j-1)}, \ldots, \theta_K^{(j-1)}, \mathbf{y}), \\
&\vdots \\
\theta_K^{(j)} &\sim p(\theta_K \mid \theta_1^{(j-1)}, \theta_2^{(j-1)}, \ldots, \theta_{K-1}^{(j-1)}, \mathbf{y}).
\end{aligned}$$

4. Repetir el paso anterior hasta completar el número de iteraciones.

Una vez completado el algoritmo, para el k-ésimo parámetro θ_k, se dispone de J valores: $\theta_k^{(1)}, \ldots, \theta_k^{(J)}$. Una forma de asegurar que los valores que entren en el análisis sean valores pertenecientes a la cadena convergida es escoger los valores a partir de un punto m; es decir, $\theta_k^{(m+1)}, \ldots, \theta_k^{(J)}$. Con base en estos valores, se procede al cálculo de la esperanza del parámetro θ_k, definida como $\sum_{j=m+1}^{J} \theta_k^{(j)}/(J-m)$, y este simple promedio puede considerarse como una estimación puntual del parámetro θ_k. También es posible calcular el intervalo de credibilidad de $(1-\alpha)\,\%$ para este parámetro como (l, u), donde l y u son los percentiles $\alpha/2$ y $1-\alpha/2$ del conjunto de valores $\theta_k^{(m+1)}, \ldots, \theta_k^{(J)}$.

3. Detalles computacionales

Para llevar a cabo la inferencia bayesiana sobre la diferencia de proporciones, es posible calcular la función de densidad exacta *a priori* y *a posteriori* de este parámetro, o bien usar el muestreo de Gibbs para obtener la función de densidad *a posteriori* simulada.

Sin embargo, el software estadístico R actualmente no implementa rutinas o paquetes que calculen las funciones hipergeométricas de Appell que son usadas para obtener la densidad *a posteriori* de la diferencia de proporciones y, por consiguiente, tampoco existen rutinas o paquetes que lleven a cabo el respectivo procedimiento bayesiano de forma exacta. Por esta razón, en este artículo se desarrollan funciones que permiten efectuar dicho procedimiento. Antes de introducir estas funciones, se presentan los resultados concernientes a la densidad exacta *a priori* y *a posteriori* de la diferencia de proporciones y el muestreo de Gibbs.

3.1. Distribución exacta

Supongamos que la distribución *a priori* para las proporciones es $Beta(a_i, b_i)$ para el parámetro θ_i, con $i = 1, 2$; es decir, $\theta_i \sim Beta(a_i, b_i)$. Como el parámetro de interés es la diferencia de las proporciones $\theta = \theta_1 - \theta_2$, el siguiente resultado de Pham-Gia & Turkkan (1993) provee la solución exacta para encontrar la distribución *a priori* de θ.

Resultado 1. *Sea $\theta_i \sim Beta(a_i, b_i)$ con $i = 1, 2$ variables aleatorias independientes, entonces $\theta = \theta_1 - \theta_2$ tiene la siguiente función de densidad de probabilidad:*

$$\pi(\theta|a_1, b_1, a_2, b_2) = \begin{cases} \dfrac{1}{A} B(a_2, b_1)\theta^{b_1+b_2-1}(1-\theta)^{a_2+b_1-1} \\ \qquad F_1(b_1, a_1+b_1+a_2+b_2-2, 1-a_1, b_1+a_2, 1-\theta, 1-\theta^2) \\ \qquad\qquad\qquad\qquad\qquad\qquad \textit{para} \quad 0 < \theta \leq 1, \\ \dfrac{1}{A} B(a_1+a_2-1, b_1+b_2-1) \qquad\qquad \textit{para} \quad \theta = 0, \\ \dfrac{1}{A} B(a_1, b_2)(-\theta)^{b_1+b_2-1}(1+\theta)^{a_1+b_2-1} \\ \qquad F_1(b_2, 1-a_2, a_1+b_1+a_2+b_2-2, b_2+a_1, 1-\theta^2, 1+\theta) \\ \qquad\qquad\qquad\qquad\qquad\qquad \textit{para} \quad -1 \leq \theta < 0, \end{cases} \tag{12}$$

donde $A = B(a_1, b_1)B(a_2, b_2)$, con $B(a, b)$ la función beta evaluada en a y b; es decir,

$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt. \tag{13}$$

Por otro lado, $F_1(\varphi, \eta_1, \eta_2, \psi, w_1, w_2)$ corresponde a la primera función hipergeométrica de Appell, dada por

$$\frac{\Gamma(\psi)}{\Gamma(\varphi)\Gamma(\psi-\varphi)} \int_0^1 u^{\varphi-1}(1-u)^{\psi-\varphi-1}(1-uw_1)^{-\eta_1}(1-uw_2)^{-\eta_2} du, \tag{14}$$

cuando las partes reales de φ y $\psi - \varphi$ son positivas, tal como lo muestra Bailey (1934).

Ahora supongamos que se observan los valores que toman las variables $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$, que denotan el éxito o fracaso en cada uno de los n_1 y n_2 ensayos independientes.

Entonces, la distribución *a posteriori* de las proporciones es $p(\theta_1|\mathbf{x}) = Beta(\alpha_1, \beta_1)$ y $p(\theta_2|\mathbf{y}) = Beta(\alpha_2, \beta_2)$, donde $\alpha_i = a_i + x_i$ y $\beta_i = b_i + n_i - x_i$ para $i = 1, 2$.

Para el parámetro de interés $\theta = \theta_1 - \theta_2$, es posible hallar su distribución *a posteriori* usando el Teorema 1. Por tanto, la función de densidad *a posteriori* dada las observaciones de θ es:

$$p(\theta|\mathbf{x},\mathbf{y}) = \begin{cases} \dfrac{1}{A}B(\alpha_2,\beta_1)\theta^{\beta_1+\beta_2-1}(1-\theta)^{\alpha_2+\beta_1-1} \\ \qquad F_1(\beta_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, 1-\alpha_1, \beta_1+\alpha_2, 1-\theta, 1-\theta^2) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{para} \;\; 0 < \theta \leq 1, \\ \dfrac{1}{A}B(\alpha_1+\alpha_2-1, \beta_1+\beta_2-1) \qquad\qquad \text{para} \;\; \theta = 0, \\ \dfrac{1}{A}B(\alpha_1,\beta_2)(-\theta)^{\beta_1+\beta_2-1}(1+\theta)^{\alpha_1+\beta_2-1} \\ \qquad F_1(\beta_2, 1-\alpha_1, \alpha_1+\beta_1+\alpha_2+\beta_2-2, \beta_2+\alpha_1, 1-\theta^2, 1+\theta) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{para} \;\; -1 \leq \theta < 0, \end{cases} \tag{15}$$

donde las definiciones de A y F_1 son enunciadas en el Resultado 1.

Dadas las distribuciones *a priori* y *a posteriori* de θ, es posible calcular la estimación puntual *a priori* y *a posteriori* junto con el intervalo de credibilidad a *priori* y *a posteriori*. Para calcular la estimación puntual *a priori*, tenemos

$$E(\theta) = E(\theta_1) - E(\theta_2) = \int_0^1 \theta_1 \pi(\theta_1|a_1, b_1) d\theta_1 - \int_0^1 \theta_2 \pi(\theta_2|a_2, b_2) d\theta_2. \tag{16}$$

Para calcular el intervalo de credibilidad *a priori*, se debe encontrar dos valores l y u tales que:

$$Pr(l \leq \theta \leq u) = 1 - \frac{\alpha}{2}.$$

En la práctica se escoge l y u de tal manera que $Pr(\theta < l) = Pr(\theta > u) = \alpha/2$. En consecuencia, se buscan valores l y u con

$$\int_{-1}^{l} \pi(\theta|a_1, b_1, a_2, b_2) \; d\theta = \int_u^1 \pi(\theta|a_1, b_1, a_2, b_2) \; d\theta = \frac{\alpha}{2}, \tag{17}$$

donde $\pi(\theta|a_1, b_1, a_2, b_2)$ es la función de densidad *a priori* de θ en (12).

Cuando los valores de las variables han sido observadas, la estimación puntual *a posteriori* se define como:

$$E(\theta|\mathbf{x},\mathbf{y}) = E(\theta_1|\mathbf{x}) - E(\theta_2|\mathbf{y}) = \int_0^1 \theta_1 p(\theta_1|\mathbf{x}) \; d\theta_1 - \int_0^1 \theta_2 p(\theta_2|\mathbf{y}) \; d\theta_2. \tag{18}$$

De modo similar, el intervalo de credibilidad está dado por dos valores l y u tales que

$$Pr(l \leq \theta \leq u|\mathbf{x},\mathbf{y}) = 1 - \frac{\alpha}{2}.$$

En consecuencia, se buscan valores l y u con

$$\int_{-1}^{l} p(\theta|\mathbf{x},\mathbf{y})\ d\theta = \int_{u}^{1} p(\theta|\mathbf{x},\mathbf{y})\ d\theta = \frac{\alpha}{2}, \tag{19}$$

donde $p(\theta|\mathbf{x},\mathbf{y})$ es la distribución *a posteriori* de θ dada por (15).

Además de estimar el parámetro θ, es posible obtener predicciones acerca de posibles resultados en nuevas muestras observadas por medio de la distribución predictiva *a posteriori* dada en (8). Para tal fin, es necesario calcular la función de verosimilitud de los datos dada por

$$f(\mathbf{x},\mathbf{y}|\theta_1,\theta_2) = \theta_1^{s_x}(1-\theta_1)^{n_1-s_x}\theta_2^{s_y}(1-\theta_2)^{n_2-s_y}, \tag{20}$$

donde $s_x = \sum_{i=1}^{n_1} x_i$, $s_y = \sum_{i=1}^{n_2} y_i$, n_1 y n_2 es el tamaño de muestras de las dos poblaciones y, x_i, y_i son las realizaciones de variables aleatorias con distribución Bernoulli. A continuación se ilustra el cálculo de la función predictiva con un ejemplo sencillo: supongamos que se vuelven a observar dos muestras, ambas de tamaños 1; es decir, ahora existen dos nuevas variables $\tilde{X}$ y $\tilde{Y}$ que denotan el éxito o fracaso en cada ensayo de la muestra, respectivamente. Entonces, la probabilidad de que ambos ensayos tengan como resultado éxito está dada por:

$$Pr(\tilde{X}=\tilde{Y}=1) = \int_0^1 \int_0^1 \theta_1\theta_2 p(\theta_1|\mathbf{x})p(\theta_2|\mathbf{y})\ d\theta_1 d\theta_2. \tag{21}$$

3.2. Distribución simulada

Desde otro punto de vista, es posible aplicar el muestreo de Gibbs para encontrar la estimación puntual y el respectivo intervalo de credibilidad. En este contexto, las distribuciones tanto *a priori* como *a posteriori* de θ_1 y θ_2 son independientes. Por lo tanto, las distribuciones condicionales *a posteriori* de θ_i, con $i = 1, 2$ en la j-ésima iteración son iguales a las respectivas distribuciones *a posteriori* de θ_i con $i = 1, 2$. Es decir, $p(\theta_1|\theta_2^{(j-1)},\mathbf{x},\mathbf{y}) = p(\theta_1|\mathbf{x})$ y $p(\theta_2|\theta_1^{(j-1)},\mathbf{x},\mathbf{y}) = p(\theta_2|\mathbf{x})$. Nótese que, en este caso, los resultados de la j-ésima iteración y los de la $j-1$-ésima iteración son independientes y, por consiguiente, no hay necesidad de fijar valores iniciales. Esto conlleva que, debido a la independencia, la convergencia del algoritmo se tenga desde su primera iteración. De esta manera, el algoritmo del muestreo de Gibbs se convierte en:

1. Fijar el número de iteraciones J.
2. Simular J observaciones de la distribuciones $p(\theta_1|\mathbf{x})$ y $p(\theta_2|\mathbf{y})$ respectivamente de manera que se dispone de valores $\theta_1^{(1)},\cdots,\theta_1^{(J)}$ y $\theta_2^{(1)},\cdots,\theta_2^{(J)}$.
3. Calcular la diferencia entre los valores simulados: $\theta_1^{(1)}-\theta_2^{(1)},\cdots,\theta_1^{(J)}-\theta_2^{(J)}$. Cada uno de estos valores corresponde al valor de la distribución *a posteriori* de $\theta_1-\theta_2$.

Una vez termine el anterior algoritmo, es posible calcular la estimación puntual de $\theta = \theta_1-\theta_2$ como el promedio $\sum_{j=1}^{N}(\theta_1^{(j)}-\theta_2^{(j)})/J$ y el intervalo de credibilidad, donde los límites inferior y superior corresponden a los percentiles $\alpha/2$ y $1-\alpha/2$ de los valores $\theta_1^{(1)}-\theta_2^{(1)},\cdots,\theta_1^{(J)}-\theta_2^{(J)}$, respectivamente.

4. Código en R

4.1. Recursos en Internet

La página WEB `http://CRAN.R-project.org/` es la página oficial del software estadístico R (R Development Core Team 2008). En ésta se encuentran la descarga y la actualización del software `R`, además de numerosas librerías y paquetes específicos como el que se presenta en el presente artículo.

Por otro lado, está la página `http://predictive.wordpress.com/stats/propbayes/`, donde se encuentra la documentación y ayuda completa del paquete que contiene las funciones resultantes de este artículo. Los lectores interesados pueden descargar el paquete en esta página. Alternativamente, es posible importar las funciones del paquete directamente proveyendo la dirección URL apropiada como un argumento a la función fuente de `R`.

4.2. Descripción del código computacional

A continuación se introducen las funciones principales para llevar a cabo la metodología bayesiana exacta, objeto de este artículo, para la diferencia de dos proporciones cuando el tamaño de muestra es moderado para cada una de las dos distribuciones binomiales.

- `F1` Esta función calcula la primera función de Appell dada por (14). La forma de uso de la función es `F1(A,B1,B2,C,X1,X2)`, donde `A`, `B1`, `B2` y `C` corresponden a los valores ϕ, η_1, η_2 y ψ; `X1` y `X2` corresponden a w_1 y w_2, respectivamente. Esta función utiliza la función `integrate` propia del ambiente `R`, la cual hace eficiente computacionalmente el análisis bayesiano para la diferencia de proporciones.

- `plot.dist` Esta función calcula y grafica la distribución exacta *a priori* y *a posteriori* del parámetro de interés, $\theta = \theta_1 - \theta_2$, usando (12) y (15), respectivamente. La forma de uso de la función es `plot.dist(a1,b1,a2,b2,plot)`. Para calcular la distribución *a priori*, `a1`, `b1`, `a2` y `b2` deben corresponder a los parámetros de la distribución *a priori* de θ_1 y θ_2, respectivamente. Para calcular la distribución *a posteriori*, `a1`, `b1`, `a2` y `b2` deben corresponder a los parámetros de la distribución *a posteriori* de θ_1 y θ_2, respectivamente. El argumento `plot` corresponde a la opción de graficar la función, cuando `plot=TRUE`, elabora la gráfica; y cuando `plot=FALSE`, solo realiza los cálculos, omitiendo la gráfica de la función.

- `p.est` Esta función calcula la estimación puntual bayesiana *a priori* o *a posteriori* de $\theta = \theta_1 - \theta_2$ dada por (16) y (18), respectivamente. Ésta usa los resultados de la función `plot.dist`. La forma de uso de la función es `p.est(a1,b1,a2,b2)`, donde los cuatro argumentos corresponden a los mismos `a1,b1,a2,b2` de la función `plot.dist`. La estimación puntual bayesiana, tanto *a priori* como *a posteriori*, se lleva a cabo usando esta función.

- `percentil` Esta función calcula los percentiles de la distribución *a priori* o *a posteriori* del parámetro $\theta = \theta_1 - \theta_2$, y usa los resultados de la función `plot.dist`. Dada

una probabilidad v, el percentil asociado con v es aquel valor a con $P(\theta < a) = v$. La forma de uso de la función es `percentil(val,a1,b1,a2,b2)`, donde `val` corresponde a la probabilidad v y los restantes cuatro argumentos corresponden a los mismos de la función `plot.dist`. El cálculo del intervalo de credibilidad *a priori* o *a posteriori* se lleva a cabo usando esta función.

- `prob` Esta función calcula la probabilidad *a priori* o *a posteriori* de que $\theta > d$, usando los resultados de la función `plot.dist`. La forma de uso de la función es `prob(val,a1,b1,a2,b2)`, donde `val` corresponde al valor d, y los restantes cuatro `a1,b1,a2,b2` corresponden a los mismos `a1,b1,a2,b2` de la función `plot.dist`. El cálculo del factor de Bayes, dado en (11), se lleva a cabo usando esta función.

- `plot.pred` Esta función calcula y grafica la función de densidad predictiva *a priori* o *a posteriori* de la variable $S_x - S_y$ dada por (5) y (8), respectivamente, cuando se tienen dos nuevas muestras de tamaño n_1 y n_2. La forma de uso de la función es `plot.pred(a1,b1,a2,b2,n1,n2,plot)`. Los primeros cuatro argumentos corresponden a los argumentos de la función `plot.dist`; los dos argumentos siguientes corresponden a los tamaños de muestra n_1 y n_2; y el último argumento `plot` corresponde a la opción de graficación. Cuando `plot=TRUE`, elabora la gráfica; y cuando `plot=FALSE`, solo realiza los cálculos, omitiendo la gráfica de la función predictiva.

- `plot.gibbs` Esta función calcula la estimación puntual, el intervalo de credibilidad *a posteriori* y grafica la función de densidad *a posteriori* para θ, usando el muestreo de Gibbs descrito en la secciones anteriores. La forma de uso de la función es `plot.gibbs(a1,b1,a2,b2,nsim,plot,chain)`. Los argumentos `a1`, `b1`, `a2` y `b2` corresponden a los parámetros de las distribuciones *a posteriori* de θ_1 y θ_2, respectivamente. `nsim` corresponde al número de iteraciones del algoritmo. `plot` corresponde a la opción de graficar la distribución *a posteriori* simulada. Cuando `plot=TRUE`, elabora la gráfica y cuando `plot=FALSE`, omite la gráfica. `chain` corresponde a la opción de graficar los valores simulados de la distribución *a posteriori*; cuando `chain=TRUE`, muestra gráficamente estos valores simulados; cuando `chain=FALSE`, omite la gráfica.

5. Una aplicación al mercadeo empresarial

Pope (1984) afirma que el empaque de un producto juega un papel muy importante en la decisión de compra de los consumidores, pues éste sirve como mecanismo para captar la atención, recordar a los compradores actuales y crea expectativa sobre lo que está adentro, entre otras. Lo anterior implica que un mejor empaque puede ser de significación para un producto, en términos de su posicionamiento en el mercado y/o en el aumento de las ventas del producto. Por esta razón, es indispensable realizar una prueba de empaque antes de lanzar oficialmente un nuevo producto o cambiar la presentación de un producto comercializado en la actualidad. En esta sección, se presenta una aplicación de las funciones creadas en este artículo aplicado a datos reales resultante de una prueba de empaque. La

documentación de estas funciones está descrita en la Sección 4.2 y están disponibles en la página de internet dada en la Sección 4.1.

De esta manera, siguiendo a Magidson (1982), supongamos que una empresa desea cambiar el empaque y la forma de presentación de un producto particular que está regularmente posicionado en el mercado. Para evaluar el impacto de la nueva presentación en la intención de compra del producto, el gerente de *marketing* planea una prueba de empaque por medio de la recolección de información en una sesión de grupo (*focus group*). La prueba fue realizada en 124 consumidores, donde a cada uno de ellos se le pregunta sobre la preferencia entre el empaque nuevo y el actual, en términos de la intención de compra, y los resultados de la prueba de empaque se muestran en la Tabla 1.

Tabla 1: Tabla de conteos resultante de una prueba de empaque.

Empaque	Compra	No compra	Total
Nuevo	32	31	63
Actual	11	24	35

Mediante el análisis estadístico de estos datos se debe responder a la siguiente pregunta: ¿El cambio de empaque afecta la intención de compra de los consumidores en la categoría?

Análisis frecuentista

El análisis estadístico de los datos necesita de un modelo cuyas características generales se dan a continuación. Supongamos que θ_1 es la probabilidad de que se venda un producto con empaque nuevo y que θ_2 es la probabilidad de venta de un producto con empaque actual. Sea $X_i = 1$ si el i-ésimo consumidor encuestado tiene intención de comprar el producto con empaque nuevo y $X_i = 0$ si no tiene intención de comprar el producto con empaque nuevo. De la misma manera, se define $Y_i = 1$ si el i-ésimo consumidor tiene intención de comprar el producto con empaque actual y $Y_i = 0$ en otro caso.

Si asumimos que existe independencia entre y dentro de cada tipo de empaque y que θ_1 y θ_2 son constantes entre los consumidores, entonces su decisión de compra forma una secuencia de ensayos Bernoulli. Definiendo, $S_x = \sum_{i=1}^{63} X_i$ y $S_y = \sum_{i=1}^{35} Y_i$, se concluye que

$$S_x|\theta_1 \sim Binomial(63, \theta_1) \qquad S_y|\theta_2 \sim Binomial(35, \theta_2) \tag{22}$$

Bajo este marco de referencia, casi todos los textos básicos de inferencia estadística (Canavos 1988) proponen que la distribución de la diferencia de dos proporciones muestrales, dada por

$$D = \frac{S_x}{n_1} - \frac{S_y}{n_2}, \tag{23}$$

donde n_1 es el tamaño de muestra de los consumidores de empaque nuevo y n_2 es el tamaño de muestra de los consumidores de empaque actual, es aproximada mediante una

distribución normal de media nula y de varianza $V_\theta(D) = (1/n_1 + 1/n_2)\theta(1-\theta)$. Se supone que los conteos tienen una distribución binomial con el mismo parámetro $\theta = \theta_1 = \theta_2$. Por tanto, para juzgar la hipótesis $\theta_1 = \theta_2$, se construye una nueva variable aleatoria $U = D/\sqrt{V_{\hat{\theta}}(D)}$ que aproximadamente[5] tiene distribución normal estándar. También es posible utilizar la variable U^2 que aproximadamente tiene una distribución chi-cuadrado con un grado de libertad.

Para responder a la pregunta de interés, el investigador estaría tentado a realizar una prueba de diferencia de dos proporciones y a tomar una decisión con respecto al valor p arrojado por dicha prueba. A continuación se muestran los resultados arrojados por la función `prop.test` propia del ambiente computacional del software `R` (R Development Core Team 2008).

```
> n1 <- 63  ;  x1 <- 32
> n2 <- 35  ;  x2 <- 11
> prop.test(c(x1,x2),c(n1,n2))

        2-sample test for equality of proportions with continuity correction

data:  c(x1, x2) out of c(n1, n2)
X-squared = 2.6852, df = 1, p-value =  0.1013
alternative hypothesis: two.sided
```

De esta manera, para un nivel de significación del 5 %, no se rechaza la hipótesis de igualdad de proporciones. En otras palabras, no se encuentra evidencia de que el cambio al empaque nuevo tenga algún efecto sobre la decisión de compra comparado con el empaque actual.

Análisis bayesiano *a priori*

Sin duda alguna, una de las herramientas más poderosas de la inferencia bayesiana es la definición de la distribución *a priori* de los parámetros y como afirma Gelman (2008): de la misma manera que no existe un principio general al definir una verosimilitud para una muestra aleatoria, en el caso frecuentista, tampoco existe un principio general para definir una distribución *a priori*, en el caso bayesiano. Por lo anterior, esta etapa del análisis bayesiano debe ser recorrida con mucho cuidado.

Se supone que la distribución a priori para la proporción de admisión θ_i es $Beta(a_i, b_i)$ con $i = 1, 2$. En la Figura 1 se observan dos candidatos miembros de la familia de las distribuciones Beta; estos son, $Beta(1, 1)$ y $Beta(2, 2)$. Nótese que la distribución $Beta(1, 1)$ se reduce a la distribución uniforme continua sobre el intervalo $(0, 1)$, la cual es una distribución *a priori* no informativa y parece natural pensar que esta distribución pueda adecuarse a este contexto. Por otro lado, la distribución $Beta(2, 2)$ da mayor peso al valor 0.5 para la probabilidad de venta y da menor peso a los valores extremos reflejando así, que

[5] Puesto que $\hat{\theta} = (S_x + S_y)/(n_1 + n_2)$.

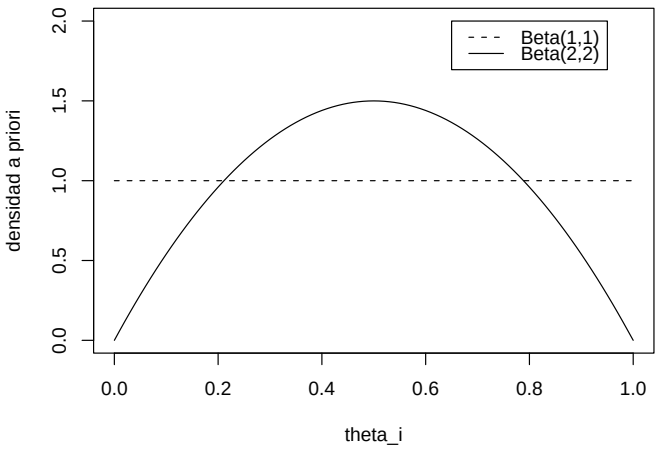

Figura 1: Dos distribuciones *a priori* para la proporción θ_i con $i = 1, 2$.

la probabilidad de vender el producto es la misma sin importar el empaque. Lo anterior conduce a una modesta percepción del investigador hacia el nuevo empaque.

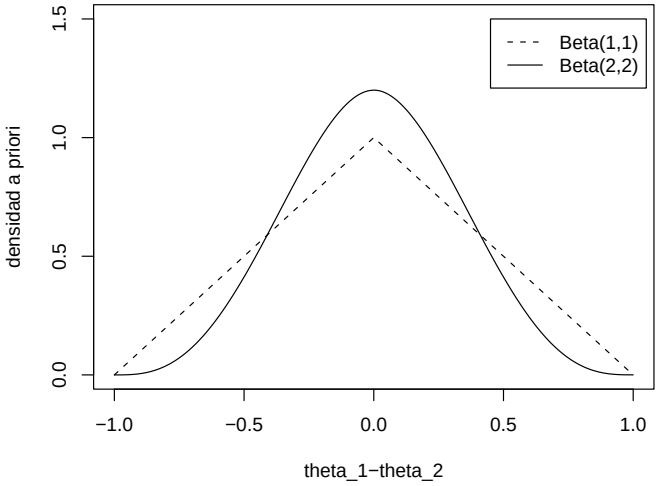

Figura 2: Distribución *a priori* para la diferencia de proporciones $\theta_1 - \theta_2$.

Mediante las distribuciones *a priori* de cada uno de los parámetros, se calcula la distribución *a priori* de la diferencia de proporciones $\theta = \theta_1 - \theta_2$, para cada una de las dos distribuciones mencionadas anteriormente para θ_i con $i = 1, 2$. Este cálculo hace uso del Resultado 1 y del siguiente código. La gráfica de estas distribuciones exactas se muestra en la Figura 2 y se realiza mediante el uso de la función `plot.dist`:

```
> priori1 <- plot.dist(1,1,1,1, plot=FALSE)
> priori2 <- plot.dist(2,2,2,2, plot=FALSE)
```

Nótese que las dos distribuciones *a priori* son simétricas con respecto al valor cero. Sin

embargo, nuestra atención estará centrada en la distribución $Beta(2,2)$ como distribución *a priori* para ambas proporciones; por tanto, si quisiéramos hallar una estimación puntual o por intervalo *a priori* para la diferencia de proporciones θ, simplemente recurriríamos a las funciones `p.est` y `percentil`, respectivamente, tal como lo indica el siguiente código:

```
> a1 <- 2  ;  b1 <- 2
> a2 <- 2  ;  b2 <- 2
> p.est(a1,b1,a2,b2)
[1] -6.589832e-18
> percentil(0.05,a1,b1,a2,b2)
[1] -0.525
> percentil(0.95,a1,b1,a2,b2)
[1] 0.525
```

La estimación puntual *a priori* es muy cercana al valor 0, indicando que la probabilidad de venta con el empaque nuevo debería ser igual a la del empaque actual. El intervalo de credibilidad es simétrico con respecto al valor 0, confirmando la suposición «actual» de que no existe diferencia significativa en los dos tipos de empaques; y, según esa suposición, la probabilidad de que se venden más productos de empaque nuevo que de empaque actual debe ser equivalente a la probabilidad de que se venden más productos de empaque actual que de empaque nuevo. Haciendo uso de la función `prob` se tiene que $Pr(\theta > 0) = Pr(\theta < 0) \approx 0{,}5$.

```
> prob(0,a1,b1,a2,b2)
[1] 0.503
```

Análisis bayesiano *a posteriori*

A continuación se realiza el análisis *a posteriori* para θ, incorporando la información contenida en las muestras observadas (Tabla 1). En primer lugar, se especifican los parámetros de las distribuciones *a posteriori* de θ_1 y θ_2, las cuales son $Beta(34,33)$ y $Beta(13,26)$, respectivamente.

```
> al1 <- a1+x1  ;  be1 <- b1+n1-x1
> al2 <- a2+x2  ;  be2 <- b2+n2-x2
```

A partir de éstas, se calcula la distribución *a posteriori* exacta para $\theta = \theta_1 - \theta_2$ usando (15). La gráfica de esta distribución *a posteriori*, que se realiza usando la función `plot.dist`, se muestra en la Figura 3.

```
> plot.dist(al1,be1,al2,be2, plot=TRUE)
```

Con este simple paso hemos «actualizado» nuestras suposiciones con respecto a θ. La Figura 3 muestra una distribución *a posteriori* que no está centrada en cero y, por consiguiente, pone en tela de juicio la igualdad entre los tipos de empaque. Una estimación

puntual exacta del parámetro θ está dada por la media de la distribución. Haciendo uso de la función `p.est` se encuentra que esta estimación corresponde a un número positivo, sugiriendo que la probabilidad de venta de un producto con el empaque nuevo es superior a la de un producto con el empaque actual.

```
> p.est(al1,be1,al2,be2)
[1] 0.1741294
```

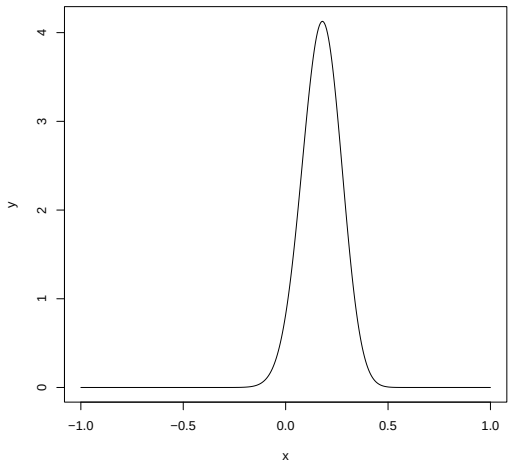

Figura 3: Distribución *a posteriori* para la diferencia de proporciones $\theta_1 - \theta_2$.

La suposición de que sí existe evidencia de que el nuevo empaque afecta de manera positiva a la intención de compra del consumidor se verifica al calcular el intervalo de credibilidad al 95 %, usando la función `percentil`, puesto que este intervalo (0.015,0.33) no contiene al valor cero. Más aún, la probabilidad de que la diferencia de proporciones sea positiva, calculada mediante la función `prob`, resulta ser $Pr(\theta > 0) \approx 0{,}964$.

```
> percentil(0.05,al1,be1,al2,be2)
[1] 0.015
> percentil(0.95,al1,be1,al2,be2)
[1] 0.33
> prob(0,al1,be1,al2,be2)
[1] 0.964069
```

Desde otro punto de vista, el valor crítico que θ debe exceder para que exista diferencia entre los dos tipos de empaque es el valor cero. Dado este valor de corte, es natural comparar las hipótesis $M_1 : \theta > 0$ y $M_2 : \theta \leq 0$. De esta manera, el factor de Bayes en favor de M_1 se calcula fácilmente, usando (11) y la función `prob`, mediante el siguiente código computacional

```
> num <- prob(0,al1,be1,al2,be2)/(1-prob(0,al1,be1,al2,be2))
> den <- prob(0,a1,b1,a2,b2)/(1-prob(0,a1,b1,a2,b2))
```

```
> FB <- num/den
> FB
[1] 26.511
```

Jeffreys (1961) propuso una escala empírica para clasificar la evidencia a favor de M_1 cuando se utilizan los factores de Bayes. Según esta escala, existe una fuerte evidencia de que los efectos de los tipos de empaque sobre la decisión de compra son diferentes a favor del empaque nuevo. La Figura 4 muestra la distribución predictiva *a posteriori*, utilizando la función `plot.pred`, cuando se consideran muestras futuras de tamaño $n_1 = n_2 = 5$, $n_1 = n_2 = 10$ y $n_1 = n_2 = 15$, respectivamente. En estos tres casos la balanza se inclina a favor de la venta de los productos con el empaque nuevo.

```
> plot.pred(al1,be1,al2,be2,1,1,plot=TRUE)
> plot.pred(al1,be1,al2,be2,3,3,plot=TRUE)
```

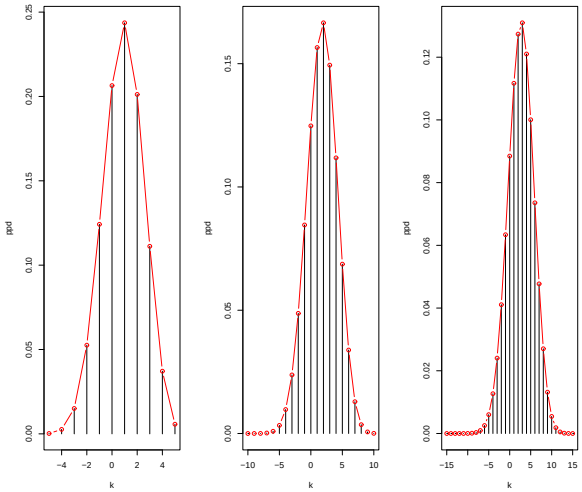

Figura 4: Distribución predictiva *a posteriori* para la diferencia de proporciones en muestras de tamaño 5, 10 y 15.

Los resultados del análisis simulado usando el muestreo de Gibbs son equivalentes a los encontrados con el análisis exacto. Como se observa en la Figura 5, la cadena de Markov converge en la primera iteración y la distribución *a posteriori* es equivalente a la distribución encontrada de manera exacta. Este análisis simulado se realizo mediante la función `plot.gibbs`, la cual también devuelve la estimación puntual para θ y el respectivo intervalo de credibilidad cuyos resultados fueron muy similares a los anteriormente mencionados.

```
> plot.gibbs(al1,be1,al2,be2,10000,plot=FALSE,chain=TRUE)
> plot.gibbs(al1,be1,al2,be2,10000,plot=TRUE,chain=FALSE)
```

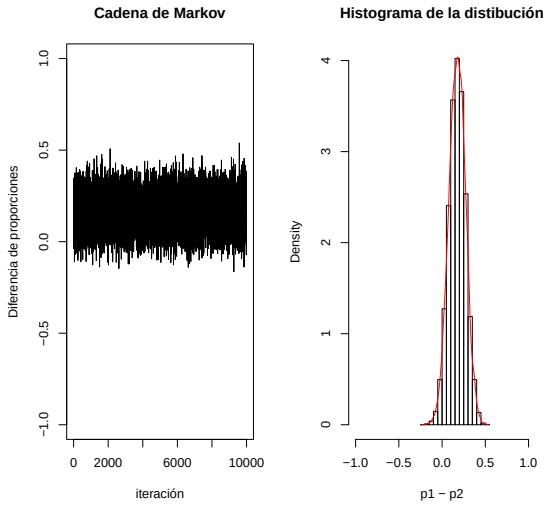

Figura 5: Convergencia de la cadena y respectiva distribución *a posteriori* usando el muestreo de Gibbs.

5.1. Validación del nuevo empaque usando independencia

El problema de evaluar el impacto del cambio de empaque sobre la venta del producto también puede ser resuelto considerando la información recolectada como datos categóricos pertenecientes a una tabla de contingencia 2×2, donde se tienen las categorías en filas y columnas y éstas conducen a la definición de dos variables aleatorias discretas, C y F. Las realizaciones de C se denotan como c_1 y c_2 y las de F como f_1 y f_2 (ver Tabla 2). Bajo esta perspectiva, el investigador desea saber si las filas son independientes de las columnas. Si esto sucede, en el caso de la prueba de producto, es posible concluir que el cambio de empaque no tiene un efecto significativo en la intención de compra del producto.

Tabla 2: Tabla de contingencia 2×2.

	c_1	c_2	Total
f_1	s_x	$n_1 - s_x$	n_1
f_2	s_y	$n_2 - s_y$	n_2

Existen varios métodos estadísticos utilizados para verificar la independencia entre filas y columnas; dos de los más conocidos, en el enfoque frecuentista, son la prueba Ji-cuadrado y la prueba exacta de Fisher. La prueba Ji-cuadrado (*Pearson's Test*) utiliza resultados de teoría asintótica y por tanto solo debe ser utilizada cuando los totales marginales, n_1 y n_2, son grandes. Por otra parte, tampoco es apropiado utilizarlo en tablas de contingencia 2×2 puesto que, en este caso particular, la estadística de prueba «asintótica» tendría un solo grado de libertad. Por otro lado, Fisher propuso una solución a este inconveniente (*Fisher's Exact Test*) la cual guía a la probabilidad «exacta», basada en una distribución hipergeométrica, de obtener un arreglo particular en una tabla 2×2. Sin embargo, esta solución frecuentista tiene problemas de orden práctico (Agresti & Coull 1998).

Retomando el caso de la prueba de empaque, usamos los comandos `chisq.test` y `fisher.test`

en R para llevar a cabo estos dos procedimientos de verificación de independencia entre filas y columnas. Es posible observar que, en ambos procedimientos, el valor p es mayor que el nivel de significación usual del 5 %, indicando que hay independencia entre las filas y las columnas. De esta forma, los dos métodos en el enfoque frecuentista coinciden en que la decisión de compra no está influenciada por el tipo de empaque; esto es, el cambio de empaque no tiene efecto sobre la intención de compra de los consumidores.

```
> Datos <- matrix(c(x1,n1-x1,x2,n2-x2),2,2)
> chisq.test(Datos)

        Pearson's Chi-squared test with Yates' continuity correction

data:  Datos
X-squared = 2.6852, df = 1, p-value = 0.1013

> fisher.test(Datos)

        Fisher's Exact Test for Count Data

data:  Datos
p-value = 0.08914
```

Una alternativa bayesiana para analizar la independencia entre filas y columnas en una tabla de contingencia 2×2 es analizar la diferencia de las dos proporciones utilizando el método descrito en las secciones anteriores. En primer lugar, nótese que, por la definición clásica de independencia, C y F son independientes si y solo si:

$$Pr(F = f_i) = Pr(F = f_i \mid C = c_j) \qquad i, j = 1, 2. \tag{24}$$

Por otra parte, supongamos que el análisis bayesiano de diferencia de proporciones arroja como conclusión que θ_1 y θ_2 son estadísticamente iguales[6]. Bajo este supuesto, por el Teorema de Probabilidad Total, se tiene que:

$$\begin{aligned} Pr(F = f_1) &= Pr(F = f_1 \mid C = c_1)Pr(C = c_1) + Pr(F = f_1 \mid C = c_2)Pr(C = c_2) \\ &= Pr(F = f_1 \mid C = c_1)Pr(C = c_1) + Pr(F = f_1 \mid C = c_1)Pr(C = c_2) \\ &= Pr(F = f_1 \mid C = c_1)[Pr(C = c_1) + Pr(C = c_2)] \\ &= Pr(F = f_1 \mid C = c_1), \end{aligned}$$

puesto que $Pr(C = c_1) + Pr(C = c_2) = 1$. Con esto se concluye que $Pr(F = f_1) = Pr(F = f_1 \mid C = c_1)$. Análogamente se tiene también que $Pr(F = f_i) = Pr(F = f_i \mid C = c_j)$ para

[6] En el contexto de la tabla de contingencia, $\theta_1 = Pr(F = f_1 \mid C = c_1)$ y $\theta_2 = Pr(F = f_1 \mid C = c_2)$. También es útil notar que s_x es una realización de la variable S_x, mientras que s_y es una realización de la variable S_y.

$i, j = 1, 2$. Con lo anterior se concluye que si $\theta_1 = \theta_2$, entonces existe independencia entre las filas y columnas.

En el caso de la prueba de empaque, el análisis bayesiano condujo a la conclusión de que $\theta_1 > \theta_2$ puesto que:

1. Siendo $\theta = \theta_1 - \theta_2$, se tiene que $Pr(\theta > 0) \approx 0{,}964$.
2. El factor de Bayes a favor del modelo $M_1 : \theta > 0$ es 26,511, indicando que los datos muestran una fuerte evidencia a favor de que el nuevo empaque afecta de manera positiva la intención de compra.

Por lo anterior, las filas y las columnas de la tabla de contingencia no se consideran independientes; esto es, el cambio de empaque sí influye significativamente en la decisión de compra de los consumidores. En este caso, el empaque nuevo promueve favorablemente la venta del producto comparado con el empaque actual y la recomendación gerencial debería estar enfocada en el lanzamiento del producto con el nuevo empaque.

6. Conclusión

A pesar de los avances teóricos y computacionales en la estadística bayesiana en las últimas décadas, poca atención se ha prestado a uno de los problemas más comunes en la investigación estadística, el análisis de la diferencia de proporciones. Lo anterior implica que, para este problema específico, la distribución exacta *a posteriori*, encontrada por Pham-Gia & Turkkan (1993), sea muy difícil de implementar en la práctica, dada su forma compleja.

En este artículo se plantea la solución computacional a este problema, mediante la creación de una serie de funciones, enmarcadas en el software estadístico R, que permiten realizar un análisis bayesiano exhaustivo: desde la definición de la distribución *a priori* para el parámetro de interés hasta la realización de pruebas de hipótesis bayesianas. Este paquete de funciones no solo se remite a los cálculos numéricos exactos para las distribuciones, sino que también permite representar gráficamente funciones predictivas que reflejan las suposiciones «actuales» acerca de la diferencia de proporciones.

Como una aplicación empírica, se proponen dos soluciones, de tipo bayesiano y frecuentista, a un problema empresarial referente al cambio del empaque de un producto en una categoría de mercado. Como resultado de esta práctica se concluye que las técnicas estadísticas clásicas frecuentistas guían a una conclusión errónea acerca del juzgamiento de la hipótesis de interés. Sin embargo, al utilizar las técnicas bayesianas, aparte de llegar a las conclusiones correctas, es posible obtener información adicional acerca del comportamiento de los parámetros que el enfoque clásico no brinda.

Referencias

Agresti, A. & Coull, B. A. (1998), 'Approximate is better than exact for interval estimation of binomial proportions', *The American Statistician* **52**(2), 119–126.

Agresti, A. & Min, Y. (2005), 'Frequentist performance of bayesian confidence intervals for comparing proportions in 2×2 contingency tables', *Biometrics* **61**, 515–523.

Albert, J. (2007), *Bayesian Computation with R*, Springer.

Bailey, W. N. (1934), 'On the Reducibility of Appell's Function F_4', *Quart. J. Math* **5**, 291–292.

Canavos, G. C. (1988), *Probabilidad y estadística: aplicaciones y métodos*, McGraw-Hill.

Carlin, B. P. & Louis, T. A. (1996), *Bayes and Empirical Bayes for Data Analysis*, 1 edn, Chapman and Hall/CRC.

Fisher, R. A. (1970), *Statistical Methods for Research Workers*, 15 edn, Macmillan Pub. Co.

Gamerman, D. & Lopes, H. F. (2006), *Markov Chain Monte Carlo*, Chapman and Hall/CRC.

Gelman, A. (2008), 'Objections to bayesian statistics', *Bayesian Analysis* **3**(3), 445–450.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995), *Bayesian Data Analysis*, 1 edn, Chapman and Hall/CRC.

Jack, A., Woodard, D., Hoffman, J. & O'Connell, M. (2007), *Bayesian Modeling with S-PLUS and the flexBayes Library*, Insightful Corporation.

Jeffreys, H. (1961), *The Theory of Probability*, Oxford.

Magidson, J. (1982), 'Some common pitfalls in causal analysis of categorical data', *Journal of Marketing Research* **19**, 461–471.

Magidson, J. (2004), 'Epidat 3.0: programa para análisis epidemiológico de datos tabulados', *Revista Española de Salud Pública* **78**(2), 277–280.

Pham-Gia, T. & Turkkan, N. (1993), 'Bayesian analysis of the difference of two proportions', *Communications in Statistics: Theory and Methods* **22**(6), 1755–1771.

Pope, J. L. (1984), *Investigación de mercados. Guía maestra para el profesional*, Grupo Editorial Norma.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

SAS (2006), *Preliminary Capabilities for Bayesian Analysis in SAS/STAT Software*.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2004), *WinBUGS User Manual*.