UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2014, Vol. 7, No. 1, pp. 31–48

Inferencia *Bootstrap* bayesiana para una proporción en muestreo con probabilidades desiguales

Bootstrap Bayesian inference for a proportion in unequal probabilities sampling

Cristian Fernando Tellez[a]
cftellezp@libertadores.edu.co

Stalyn Yasid Guerrero[b]
syguerrero@correro.unicordoba.edu.co

Mario Pacheco[c]
mariopachecolopez@gmail.com

Resumen

En este artículo se propone el método *bootstrap* bayesiano para realizar inferencias sobre una proporción ρ en una población finita a partir de una muestra con probabilidades desiguales. Vía simulación se determinó que, a partir de una adecuada elección de la distribución *a priori* de ρ, la metodología propuesta obtiene estimaciones con sesgos tan pequeños como los obtenidos mediante el π-estimador clásico. Adicional a esto, se obtuvo menor varianza e intervalos de confianza con niveles de confianza más altos y de menor longitud en comparación con el π-estimador clásico y el estimador BPSP propuesto por Chen et al. (2010). Finalmente se ejemplifica la implementación de la metodología.

***Palabras clave*:** muestreo probabilístico, *Bootstrap* bayesiano, estimación de una proporción, estimador BPSP.

Abstract

This paper describe Bayesian bootstrap method, it is to realize inferences for finite population proportion ρ based on unequal probability sampling. Through Simulation we found that based on an appropriate a priori distribution to ρ with the proposed methodology it is possible to get estimate less-biased like that obtain by the clasic π-estimator. Also, we get less-variance and confidence intervals with highest confidence levels and it has fewer length when we compared it with the

[a]Docente Tiempo completo, Fundación Universitaria los Libertadores, Colombia.
[b]Egresado Universidad de Córdoba, Colombia.
[c]Docente Ocasional Universidad de Córdoba, Colombia.

classic π-estimator and BPSP estimator that was proposed by Chen et al. (2010). Lastly, an example is performed using the development methodology.

***Keywords*:** probability sampling, Bayesian Bootstrap, proportion estimation.

1. Introducción

Un parámetro de interés considerado en muchos estudios estadísticos (investigaciones sociales, económicas, estudios de mercadeo, entre otros) es la proporción. La teoría de muestreo probabilístico clásica asociada a la estimación de dicho parámetro, se basa en funciones no lineales de otros parámetros (como el total poblacional y el total de un dominio), mientras que el enfoque bayesiano lo considera como una variable aleatoria que se puede modelar usando distribuciones de probabilidad de variables aleatorias en el espacio $(0\,,\,1)$, como la distribución uniforme $(0\,,\,1)$ o la distribución beta (α,β), entre otras.

En la literatura especializada es poco lo que se encuentra acerca de la integración entre el muestreo probabilístico y la teoría bayesiana, de igual forma, lo que se halla solo lo hace de manera parcial para el muestreo aleatorio simple o muestreo aleatorio simple estratificado. Por ejemplo, Chen et al. (2010) proponen un estimador *spline* penalizado predictivo bayesiano (BPSP, por sus siglas en inglés) para una proporción en poblaciones finitas bajo muestreo con probabilidades desiguales. De otro lado, Pfeffermann & Royall (1982), en su trabajo centran toda la atención en los supuestos necesarios para la robustez de los procedimientos estadísticos y así poder predecir el total de la característica de interés a la población.

La finalidad de este artículo es mostrar una herramienta para la estimación de proporciones que integre las teorías de estadística bayesiana y el muestreo probabilístico. La herramienta seleccionada es el método *bootstrap* bayesiano, puesto que una característica distintiva de la estadística bayesiana es la forma explícita de tener en cuenta la información previa; sin embargo, uno de sus problemas que se encuentra en la necesidad de asumir la forma paramétrica de la distribución que genera los datos. Mediante la técnica *bootstrap* bayesiano es posible evitar este supuesto.

2. Inferencia *Bootstrap* bayesiana para una proporción

Considere $U = \{u_1, u_2, ..., u_k, ..., u_N\}$, una población finita de tamaño N, en donde cada unidad u_i $(i = 1, 2, ..., N)$ tiene asociada una variable dicótoma y_i, que toma el valor 0 cuando la observación no posee la característica de interés y 1 cuando la posee. Una muestra aleatoria s es seleccionada de U, de acuerdo con un diseño de muestreo probabilístico. En la muestra, la variable de interés y es observada para todos los elementos seleccionados. El interés consiste en estimar la distribución de

probabilidad posterior para el parámetro ρ_y definido como $\rho_y = \sum_{i \in k} \frac{y_i}{N}$, haciendo uso de los valores de la muestra y de las probabilidades de inclusión inducidas por el diseño muestral.

La metodología *bootstrap* bayesiana considera que el parámetro ρ_y está en función de la distribución acumulada de la que proviene la muestra aleatoria s, la cual ha sido seleccionada con un diseño muestral particular y con la que se ha estimado ρ_y, haciendo uso del estimador de Horvitz-Thompson definido como:

$$\hat{\rho}_{y\pi} = \frac{1}{\hat{N}} \sum_{i \in s} y_i / \pi_i$$

con $\hat{N} = \sum_{i \in s} \frac{1}{\pi_i}$ y $\pi_i = Pr\,(i \in s)$

Supongamos entonces que la distribución de probabilidad condicional $\xi(y \mid \rho_y)$ de y existe; esta es, a su vez, la verosimilitud de y en función de ρ_y. Sea $\xi(\rho_y)$ la densidad *a priori* del parámetro ρ_y. Por el teorema de bayes se tiene:

$$\xi(\rho_y \mid y) \propto \xi(y \mid \rho_y)\xi(\rho_y) \tag{1}$$

donde $\xi\,(\rho_y \mid y)$ es la distribución posterior de ρ_y dada la observación de y en la muestra.

Al observar la forma de la distribución posterior de ρ_y se debe pensar en la escogencia de una distribución *a priori* para ρ_y, y en un supuesto distribucional para y condicionado al parámetro ρ_y.

En cuanto a la distribución *a priori* para ρ_y existe una gama de posibilidades entre distribuciones previas informativas y no informativas, tales como la distribución uniforme y la distribución *beta* o cualquier distribución que tenga como soporte el intervalo $(0, 1)$. En cuanto al supuesto distribucional para y condicionado al parámetro ρ_y se debe tener en cuenta que en la teoría de muestreo no se hacen dichos supuestos, por lo que se dice que son de libre distribución. Es por esto último que la metodología *bootstrap* bayesiana juega un papel fundamental en la metodología propuesta, la cual consiste en realizar una obtención de $\xi(y \mid \rho_y)$ y $\xi(\rho_y \mid y)$ de forma empírica.

2.1. Distribución posterior de ρ con *a priori* informativa

Según Shao & Tu (1995), el método *bootstrap* bayesiano evita asumir una forma paramétrica de la distribución que genera los datos. Si se está interesado en el parámetro ρ_y y la información *a priori* sobre ρ_y está resumida en $\xi(\rho_y)$ y si $y_1, y_2, ..., y_n$ representan las observaciones de la variable de interés en la muestra con densidad desconocida ξ, entonces es posible aproximar a ξ utilizando un estimador de densidades, por ejemplo, $\hat{\xi}(y \mid \rho_y)$ y hallar un estimador de la distribución posterior como:

$$\xi(\rho_y \mid y)\,\alpha\,\xi(\rho_y)\hat{L}(y_1, ..., y_n \mid \rho_y) \tag{2}$$

donde $\hat{L}(y_1, ..., y_n \mid \rho_y)$ representa la estimación *bootstrap* de la función de verosimilitud, proporcional a $\hat{\xi}$. A continuación se presenta la secuencia de pasos necesarios para determinar $\hat{L}$:

1. Usando los datos muestrales $y_1, y_{2,...,}y_n$, se construye una población artificial U^*. Una forma de construir dicha población consiste en replicar los y_i tantas veces como su factor de expansión ($\frac{1}{\pi_i}$), siguiendo el principio de representatividad.

2. Seleccionar una serie de muestras *bootstrap* de U^* denotadas por s^* con un diseño idéntico al usado para seleccionar la muestra original s de U. Repetir B veces para cada muestra *bootstrap* $s_b^*(b = 1, 2, ..., B)$, calcular el π estimador $\rho_{y\pi b}^*$:

$$\hat{\rho}_{y\pi b}^* \quad = \quad \frac{1}{\hat{N}^*} \sum_{i \in s^*} y_{ib}^* / \pi_{ib}^*.$$

   Donde $\hat{N}^* = \sum_{i \in s^*} \frac{1}{\pi_i^*}$, π_i^* es la probabilidad de inclusión de los elementos en la muestra *bootstrap* y y_{ib}^* es el i-ésimo elemento de la b-ésima muestra *bootstrap.*

3. Con los anteriores estimadores $\rho_{y\pi 1}^*, ..., \rho_{y\pi B}^*$ calcular el estimador de densidad kernel definido como:

$$f_B(u) = \frac{1}{Bh_B} \sum_{b=1}^{B} K\left(\frac{u - (\hat{\rho}_{y\pi b}^* - \hat{\rho}_{y\pi})}{h_B}\right) \tag{3}$$

   Donde la función K es llamada función núcleo *(kernell)*, y en general, es una función de densidad continua, unimodal y simétrica alrededor de 0. El parámetro h_b se conoce como parámetro suavizador.

   Haciendo $u = \hat{\rho} - \rho_y$ en la ecuación anterior, $\hat{f}_B(\hat{\rho} - \rho_y)$ es una estimación de la densidad muestral de $\hat{\rho}_{y\pi}$ dado ρ_y. Evaluándola en $x = \hat{\rho}_{y\pi}$ resulta como función de ρ_y para ser usada como verosimilitud

$$\hat{L}_B(\hat{\rho}_{y\pi} \mid \rho_y) = \frac{1}{Bh_B} \sum_{b=1}^{B} K\left(\frac{2\hat{\rho}_{y\pi} - \rho - \hat{\rho}_{y\pi b}^*}{h_B}\right) \tag{4}$$

4. La distribución posterior resultante $\xi(\hat{\rho}_{y\pi} \mid \rho_y)$ es entonces proporcional a $\xi(\rho_y)\hat{L}((\hat{\rho}_{y\pi} \mid \rho_y)$ y la constante de normalización se puede hallar mediante integración numérica.

De esta forma es posible construir un estimador bayesiano de la distribución posterior de ρ_y como:

$$\xi(\rho_y \mid y) = c(y) \times \xi(\rho_y) \times \hat{L}(y_1, ..., y_n \mid \rho_y)$$

donde $c(y)$ se puede obtener por integración numérica como

$$c(y) = \frac{1}{\int \xi(\rho_y) \times \hat{L}(y_1, ..., y_n \mid \rho_y) d\rho_y}$$

La función K se llama función núcleo (o *kernel*) y, en general, es una función de densidad continua, unimodal y simétrica alrededor de 0. El parámetro h_B se conoce como parámetro de suavizamiento. Hollander & Wolfe (1999) muestra las densidades Kernel más usadas. En este artículo no se consideró la metodología *bootstrap* con *a priori* no informativa dado que sus resultados son muy similares al método *bootstrap* clásico Shao & Tu (1995).

2.2. Inferencia bayesiana sobre la proporción

Para realizar estimaciones de un parámetro mediante inferencia bayesiana, se requiere de una muestra aleatoria obtenida a partir de una distribución posterior dada. En este caso, se genera una muestra aleatoria $\rho_y^1, \rho_y^2, ..., \rho_y^m$ a través de la distribución posterior $\xi(\rho_y \mid y)$ de la siguiente manera [1]:

1. Generar $p_1, p_2, ..., p_m$ valores a partir de una distribución con soporte $(0, 1)$, sin pérdida de generalidad, la distribución uniforme $(0, 1)$.

2. Evaluar cada p_i en $\xi(\rho_y \mid y)$, con $i = 1, 2, ..., m$, obteniendo así, la probabilidad de selección de cada valor.

3. Por último, la muestra requerida $\rho_y^1, \rho_y^2, ..., \rho_y^m$ se obtiene tomando una muestra con reemplazo de $p_1, p_2, ..., p_m$ con probabilidad de selección $\xi(p_i \mid y)$ para $i = 1, 2, ..., m$.

Las funciones comúnmente utilizadas para minimizar dichos errores son: la función de pérdida cuadrática, función de pérdida en error absoluto y la función escalonada Box & Tiao (1973).

2.2.1. Función de pérdida cuadrática para la proporción

Se considera una cierta función $L(\rho_y \rho_c) = (\rho_c - \rho_y)^2$ la cual se denotará como función de pérdida cuadrática asociada al parámetro ρ_y, y sea ρ_c la estimación considerada para ρ_y. Sean $\rho_y^1, \rho_y^2, ..., \rho_y^m$ una muestra aleatoria de tamaño m generada a través de la distribución posterior $\xi(\rho_y \mid y)$ mediante el método Metropolis - Hastings. La diferencia entre ρ_c y el valor real de ρ_y se hace mínima si p_c se

[1] Con dicha muestra, lo que se pretende es estimar el parámetro ρ_y que considera un error de estimación el cual debe ser minimizado. Para lograr lo anterior, se debe disponer de una función que relacione la estimación del parámetro ρ_y con el valor real de este.

estima empleando la siguiente expresión:

$$\rho_c = E\left(\rho_y \mid y\right) = \int_{-\infty}^{+\infty} \rho_y \xi\left(\rho_y \mid y\right) d\rho_y \tag{5}$$

Esta integral se calcula númericamente puesto que $\xi\left(\rho_y \mid y\right)$ es una función empírica. Por otro lado, la estimacíón vía Monte Carlo de la media posterior es

$$\rho_c = \overline{\rho_y} = \frac{\sum_{j=1}^{m} \rho_y^j}{m} \tag{6}$$

y un error estándar estimado es:

$$se_{\rho_c} = \sqrt{\frac{\sum_{j=1}^{m} \left(\rho_y^j - \rho_c\right)^2}{(m-1)\, m}} \tag{7}$$

En consecuencia, ρ_c es el estimador puntual de ρ_y cuando tomamos como función de pérdida la función de pérdida cuadrática.

3. Estudio de simulación

Los escenarios de simulación se dispusieron similares a los realizado en el trabajo de Chen et al. (2010) para así poder comparar los resultados entre las estimaciones vía método clásico, el estimador BPSP y las estimaciones hechas por la metodología propuesta en este trabajo.

3.1. Diseño de la simulación

El estudio de simulación pretende evaluar el comportamiento de la metodología propuesta y compararla con el procedimiento clásico y el estimador BPSP en la estimación de una proporción en muestreo probabilístico. El procedimiento consiste en simular dos poblaciones artificiales de tamaño 2000, también se genera una medida de tamaño X para implementar un diseño de muestreo con probabilidad proporcional al tamaño. Los valores que toma esta variable son los enteros consecutivos $71, 72, 73, ..., 2070$.

Por otro lado, las probabilidades de inclusión en la población son calculadas proporcionales a la variable tamaño, $\pi_i = n \times x_i / \sum x_i$, con $x_i = 71, 71, ..., 2070$.. Luego de esto, son generados datos Z de una distribución normal con estructura de media $f(\pi)$ y varianza constante igual a 0.04. Para el proceso de simularon se tomaron dos estructuras de medias: una función de incremento lineal $f(\pi_i) = 3\pi_i$ y una función exponencial $f(\pi_i) = \exp(-4,64 + 26\pi_i)$. En la figura 1 se muestran las distribuciones normales con las dos estructuras de medias.

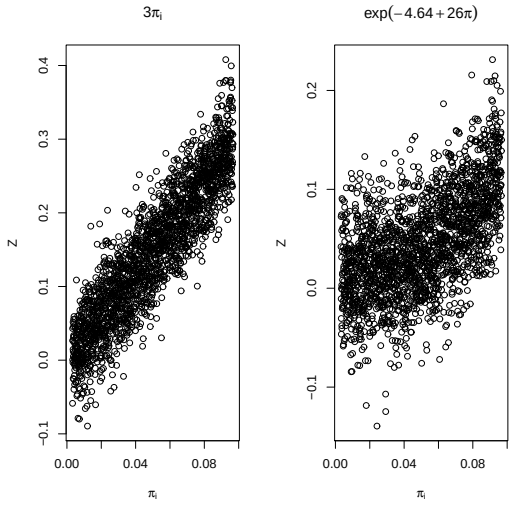

Figura 1: *Distribución normal con estructuras de medias lineal y exponencial. Fuente: elaboración propia.*

De otra parte, las variables respuesta binarias Y_1, Y_2, Y_3 son generadas como sigue: Y_1 es igual a 1 si Z es menor o igual a su percentil 10 y 0 en otro caso. Similarmente, se generán las respuestas Y_2 y Y_3 usando los percentiles 50 y 90. El objetivo inferencial aquí es la proporción poblacional para Y igual a 1.

En cada simulación, se genera una población finita y se calcula la verdadera proporción poblacional, para Y igual a 1. Luego se seleccionan muestras aleatorias, de tamaños $n = 30,\ 50,\ 100\,, 200\,\text{y}\,500$ con probabilidades proporcionales al tamaño (πPT) de cada población y se calcula la proporción estimada $\hat{\rho}$ clásica y *bootstrap* bayesiana basada en la función de pérdida cuadrática (media posterior).

El anterior proceso se repite 1000 veces y se calcula: el sesgo empírico (B), la raíz del error cuadrático medio (RMSE), las longitudes de los intervalos de credibilidad y de confianza y las coberturas de los mismos.

Sea $\hat{\rho}_j$ una estimación de ρ_j basada en la muestra j-ésima, el sesgo empírico y la raíz del error cuadrático medio son:

$$B = \frac{1}{1000}\sum_{j=1}^{1000}(\hat{\rho}_j - \rho) \tag{8}$$

$$RMSE = \sqrt{\frac{1}{1000}\sum_{j=1}^{1000}(\hat{\rho}_j - \rho)^2} \tag{9}$$

Como distribución *a priori* se tomó una distribución $beta(\alpha, \beta)$ donde α toma los valores de $\alpha = 25, 50, 100$ y para la obtención de los valores del parámetro β, lo

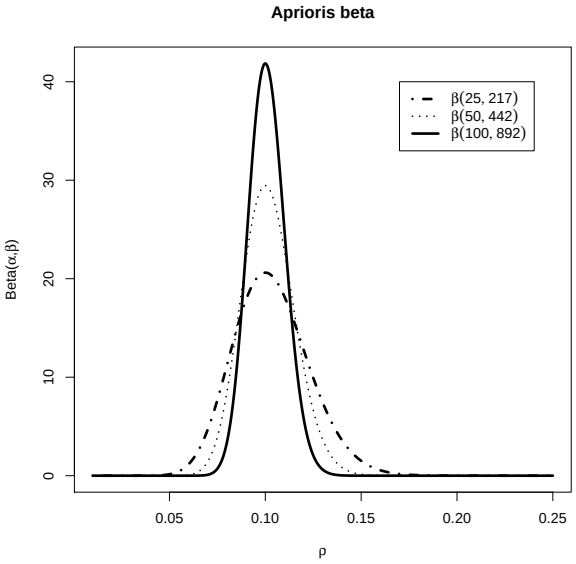

Figura 2: *Distribuciones a priori para $\rho = 0.1$. Fuente: elaboración propia.*

que se realiza es:

1. Fijar α.
2. Igualar la expresión de la media de una distribución *beta* $(\alpha\,,\beta)$ con los valores reales de ρ_y, es decir, $\rho_y = \frac{\alpha-1}{\alpha+\beta-2}$, donde $\rho_y = 0.1\,, 0.5\,, 0.9$.
3. Para cada valor de ρ_y despejar el valor de β.

Los valores de α permiten que la distribución *beta* se concentre en intervalos gradualmente más pequeños, y eso a su vez permita obtener mejores estimaciones de ρ_y.

3.2. Resultado de la simulación

En este apartado se muestran las tablas que contienen los resultados del proceso de simulación antes descrito, con el fin de comparar la metología clásica para la estimación de la proporción, el estimador BPSP y la metodología aquí propuesta. El programa de simulación se desarrolló en el paquete estadístico R versión 2.13.0 (R Core Team 2013).

En la Tabla 1 se compara el sesgo, la RMSE, las longitudes de los intervalos y sus coberturas para las metodología de estimación clásica y la *bootstrap* bayesiana

Tabla 1: *Sesgo, RMSE, cobertura (%) y nivel de confianza con una estructura de media lineal. Fuente: elaboración propia.*

n	ρ	Método	*A priori*	Sesgo	REMC	Cobertura	Amplitud
30	0.1	B.B.	$Beta\,(25, 225)$	-0.00663	0.00025	87.0	0.06541
			$Beta\,(50, 450)$	-0.00540	0.00020	89.2	0.04777
			$Beta\,(100, 900)$	-0.00338	0.00009	91.8	0.03490
		Clásico	-	0.00894	0.00494	83.0	0.20805
	0.5		$Beta\,(25, 25)$	0.00019	0.00048	99.0	0.25094
			$Beta\,(50, 50)$	0.00010	0.00014	99.0	0.18586
			$Beta\,(100, 100)$	0.00015	0.00004	99.0	0.13440
		Clásico	-	0.00260	0.01263	87.0	0.35525
	0.9		$Beta\,(225, 25)$	0.00193	0.00003	99.4	0.06957
			$Beta\,(450, 50)$	0.00152	0.00002	99.8	0.05038
			$Beta\,(900, 100)$	0.00080	0.00001	99.0	0.03621
		Clásico	-	-0.00320	0.00394	81.4	0.18358
50	0.1	B.B.	$Beta\,(25, 225)$	-0.00604	0.00021	92.0	0.06558
			$Beta\,(50, 450)$	-0.00385	0.00012	94.0	0.04859
			$Beta\,(100, 900)$	-0.00230	0.00006	95.6	0.03530
		Clásico	-	0.00625	0.00414	81.2	0.17462
	0.5		$Beta\,(25, 25)$	-0.00134	0.00052	99.0	0.24102
			$Beta\,(50, 50)$	0.00123	0.00018	99.0	0.18132
			$Beta\,(100, 100)$	0.00044	0.00004	99.0	0.13301
		Clásico	-	0.00575	0.00782	87.8	0.28739
	0.9		$Beta\,(225, 25)$	0.00277	0.00005	98.0	0.06791
			$Beta\,(450, 50)$	0.00153	0.00002	99.0	0.05001
			$Beta\,(900, 100)$	0.00094	0.00001	99.2	0.03605
		Clásico	-	0.00011	0.00236	83.0	0.14124
100	0.1	B.B.	$Beta\,(25, 225)$	-0.00381	0.00012	96.4	0.06708
			$Beta\,(50, 450)$	-0.00225	0.00005	97.6	0.04954
			$Beta\,(100, 900)$	-0.00164	0.00003	97.0	0.03567
		Clásico	-	0.00162	0.00231	79.6	0.14307
		BPSP	-	0.00800	0.04720	91.0	-
	0.5		$Beta\,(25, 25)$	0.00146	0.00059	99.0	0.22306
			$Beta\,(50, 50)$	-0.00080	0.00021	99.0	0.17336
			$Beta\,(100, 100)$	0.00044	0.00007	99.0	0.12949
		Clásico	-	0.00302	0.00470	85.6	0.20770
		BPSP	-	-0.00520	0.04770	95.6	-
	0.9		$Beta\,(225, 25)$	0.00226	0.00003	99.4	0.06598
			$Beta\,(450, 50)$	0.00083	0.00001	99.0	0.04939
			$Beta\,(900, 100)$	0.00072	0.00001	99.8	0.03576
		Clásico	-	0.00231	0.00088	84.0	0.09108
		BPSP	-	-0.00290	0.02350	94.5	-

en tamaños de muestra $n = 30$ y 50, cuando el tamaño de muestra aumenta a

100 se incluye la metodología BPSP en la comparación. En forma análoga, en la Tabla 2 se realizan las comparaciones, pero esta vez con tamaños de muestras $n =$ 200 y 500. En ambas tablas se maneja una estructura de media lineal y para cada escenario se varían los parámetros de la distribución *beta* (la cual hace el papel de la distribución *a priori*).

Tabla 2: *Sesgo, RSME, cobertura (%) y nivel de confianza con una estructura de media lineal. Fuente: elaboración propia.*

n	ρ	Método	*A priori*	Sesgo	REMC	Cobertura	Amplitud
200	0.1	B.B.	$Beta\,(25, 225)$	-0.00220	0.00005	98.4	0.06726
			$Beta\,(50, 450)$	-0.00151	0.00003	98.4	0.04966
			$Beta\,(100, 900)$	-0.00080	0.00001	99.4	0.03599
		Clásico	-	0.00111	0.00153	83.6	0.11949
		BPSP	-	0.00510	0.03200	93.8	-
	0.5		$Beta\,(25, 25)$	0.00392	0.00061	99.0	0.19618
			$Beta\,(50, 50)$	0.00302	0.00026	99.0	0.16006
			$Beta\,(100, 100)$	0.00033	0.00009	99.0	0.12351
		Clásico	-	0.00076	0.00217	88.4	0.14852
		BPSP	-	-0.00170	0.03280	94.9	-
	0.9		$Beta\,(225, 25)$	0.00167	0.00003	99.0	0.06088
			$Beta\,(450, 50)$	0.00117	0.00001	99.0	0.04683
			$Beta\,(900, 100)$	0.00063	0.00000	99.0	0.03489
		Clásico	-	0.00279	0.00029	89.0	0.05728
		BPSP	-	-0.00120	0.01550	95.3	-
500	0.1	B.B.	$Beta\,(25, 225)$	0.00041	0.00002	99.6	0.06756
			$Beta\,(50, 450)$	-0.00011	0.00001	99.6	0.04960
			$Beta\,(100, 900)$	0.00005	0.00001	99.0	0.03617
		Clásico	-	0.00873	0.00085	86.8	0.09098
	0.5		$Beta\,(25, 25)$	0.02158	0.00085	99.4	0.14945
			$Beta\,(50, 50)$	0.01690	0.00051	99.8	0.13109
			$Beta\,(100, 100)$	0.01221	0.00026	99.0	0.10813
		Clásico	-	0.03117	0.00171	70.0	0.08885
	0.9		$Beta\,(225, 25)$	0.00627	0.00006	99.8	0.04736
			$Beta\,(450, 50)$	0.00444	0.00003	99.0	0.03947
			$Beta\,(900, 100)$	0.00270	0.00001	99.0	0.03145
		Clásico	-	0.00986	0.00017	66.0	0.02844

En general, las estimaciones de ρ obtenidas mediante la metodología *bootstrap* bayesiana son superiores en las dos tablas en cuanto a un menor RECM, mayor cobertura, una menor amplitud, un sesgo pequeño en comparación con el estimador BPSP y tan pequeño como los obtenidos con el π-estimador clásico. Cabe resaltar que algunos escenarios la amplitud de los intervalos bayesianos fueron ligeramente más grandes que la amplitud de los intervalos clásicos, pero eso es algo menor en comparación con la ganancia en cobertura, sesgos y RMSE.

En las Tablas 3 y 4 se presentan de forma similar las comparaciones realizadas en

Tabla 3: *Sesgo, RSME, cobertura (%) y nivel de confianza con una estructura de media exponencial. Fuente: elaboración propia.*

n	ρ	Método	*A priori*	Sesgo	REMC	Cobertura	Amplitud
30	0.1	B.B.	$Beta\,(25, 225)$	-0.00627	0.00024	90.1	0.06571
			$Beta\,(50, 450)$	-0.00459	0.00015	90.9	0.04830
			$Beta\,(100, 900)$	-0.00475	0.00016	91.0	0.04811
		Clásico	-	0.00336	0.00443	81.2	0.18997
	0.5		$Beta\,(25, 25)$	-0.00014	0.00048	99.8	0.25077
			$Beta\,(50, 50)$	0.00061	0.00016	99.9	0.18540
			$Beta\,(100, 100)$	0.00013	0,00005	99.0	0.13452
		Clásico	-	0.00631	0.01285	85.7	0.35371
	0.9		$Beta\,(225, 25)$	0.00297	0,00005	99.0	0.06862
			$Beta\,(450, 50)$	0.00161	0,00002	99.4	0.05037
			$Beta\,(900, 100)$	0.00101	0,00001	99.7	0.03612
		Clásico	-	-0.00521	0.00453	81.0	0.19195
50	0.1	B.B.	$Beta\,(25, 225)$	-0.00586	0.00021	92.2	0.06538
			$Beta\,(50, 450)$	-0.00337	0.00010	95.4	0.04882
			$Beta\,(100, 900)$	-0.00244	0.00007	96.2	0.03533
		Clásico	-	-0.00309	0.00252	80.8	0.14826
	0.5		$Beta\,(25, 25)$	0.00015	0.00047	99.0	0.24136
			$Beta\,(50, 50)$	0.00066	0.00019	99.0	0.18110
			$Beta\,(100, 100)$	0.00010	0.00004	99.0	0.13299
		Clásico	-	0.00066	0.00712	89.0	0.28565
	0.9		$Beta\,(225, 25)$	0.00312	0.00007	96.6	0.06810
			$Beta\,(450, 50)$	0.00191	0.00003	98.4	0.04986
			$Beta\,(900, 100)$	0.00118	0.00001	99.2	0.03595
		Clásico	-	0.00130	0.00316	77.2	0.15125
100	0.1	B.B.	$Beta\,(25, 225)$	-0.00270	0.00006	98.0	0.06711
			$Beta\,(50, 450)$	-0.00166	0.00003	98.6	0.04948
			$Beta\,(100, 900)$	-0.00087	0.00001	99.2	0.03582
		Clásico	-	0.00172	0.00168	82.0	0.11960
		BPSP	-	0.01700	0.05180	90.8	-
	0.5		$Beta\,(25, 25)$	0.00214	0.00054	99.0	0.22242
			$Beta\,(50, 50)$	0.00094	0.00020	99.0	0.17359
			$Beta\,(100, 100)$	-0.00028	0.00007	99.0	0.12962
		Clásico	-	0.00462	0.00412	89.4	0.20613
		BPSP	-	-0.00140	0.04700	91.1	-
	0.9		$Beta\,(225, 25)$	0.00242	0.00004	99.0	0.06664
			$Beta\,(450, 50)$	0.00175	0.00002	99.6	0.04914
			$Beta\,(900, 100)$	0.00078	0.00001	99.6	0.03585
		Clásico	-	-0.00039	0.00143	83.6	0.10910
		BPSP	-	-0.00100	0.01230	93.0	-

las tablas anteriores, solo que, en este caso, la estructura de media es exponencial.

Los resultados obtenidos son muy similares a los anteriores, lo que implica que el cambio de estructura de media no los afecta en gran forma.

Tabla 4: *Sesgo, RSME, cobertura (%) y nivel de confianza con una estructura de media exponencial. Fuente: elaboración propia.*

n	ρ	Método	*A priori*	Sesgo	REMC	Cobertura	Amplitud
200	0.1	B.B.	$Beta\,(25, 225)$	-0.00178	0.00004	99.0	0.06722
			$Beta\,(50, 450)$	-0.00071	0.00001	99.8	0.04995
			$Beta\,(100, 900)$	-0.00059	0.00001	99.8	0.03592
		Clásico	-	0.00246	0.00124	85.6	0.10729
		BPSP	-	0.01340	0.03600	92.5	-
	0.5		$Beta\,(25, 25)$	0.00382	0.00051	99.0	0.19689
			$Beta\,(50, 50)$	0.00377	0.00022	99.0	0.15959
			$Beta\,(100, 100)$	0.00108	0.00010	99.0	0.12383
		Clásico	-	0.00403	0.00232	85.4	0.14884
		BPSP	-	0.00001	0.03210	93.8	-
	0.9		$Beta\,(225, 25)$	0.00247	0.00004	99.0	0.05921
			$Beta\,(450, 50)$	0.00141	0.00002	99.0	0.04610
			$Beta\,(900, 100)$	0.00076	0.00001	99.0	0.03457
		Clásico	-	0.00343	0.00028	86.2	0.05222
		BPSP	-	-0.00007	0.00800	94.5	-
500	0.1	B.B.	$Beta\,(25, 225)$	-0.00007	0.00003	99.8	0.06744
			$Beta\,(50, 450)$	0.00008	0.00001	99.8	0.04970
			$Beta\,(100, 900)$	0.00013	0.00001	99.9	0.03618
		Clásico	-	0.01173	0.00101	87.8	0.09510
	0.5		$Beta\,(25, 25)$	0.02444	0.00093	99.4	0.14864
			$Beta\,(50, 50)$	0.01741	0.00054	99.6	0.13042
			$Beta\,(100, 100)$	0.01304	0.00028	99.0	0.10834
		Clásico	-	0.03306	0.00182	63.8	0.08829
	0.9		$Beta\,(225, 25)$	0.00728	0.00008	99.4	0.04629
			$Beta\,(450, 50)$	0.00534	0.00004	99.8	0.03858
			$Beta\,(900, 100)$	0.00309	0.00001	99.8	0.03114
		Clásico	-	0.01094	0.00018	59.4	0.02700

4. Ejemplo de la metodología

Con el fin de ilustrar la implementación de la metodología aquí propuesta se examinó la base de *calif* que está disponible en la librería *pps* (Gambino 2012) del software estadístico R Core Team (2013), la cual contiene el registro de 1077 observaciones y 6 variables (*condado, población, blanco, amerindio, hispano* y *estrato*). El interés consiste en estimar mediante el π-estimador y la metodología bayesiana la proporción de blanco (Y), que superan el lumbral de 148. El valor real, dada la base de datos, equivale al 5.1067 %.

Se realizó la extracción de una muestra probabilística s con un diseño de muestreo probabilístico proporcional al tamaño de la variable auxiliar (*diseño* $\pi\boldsymbol{PT}$) por estrato (o grupos). Como información auxiliar se utilizó el *logaritmo* de la variable *población* (Log_pob), donde las probabilidades de inclusión de primer y segundo orden fueron calculadas como en Särndal et al. (1992).

Se decide dividir las observaciones en 2 grupos (o estratos) de acuerdo a Log_pob, para lo cual se cálculo la matriz de distancias y se implementó la función `dist` de R con el método de `''euclidean''`. Los resultados obtenidos de la clasificación indican que estos grupos tienen los tamaños de 900 y 177, los cuales denotaremos como $G1$ y $G2$ respectivamente. El tamaño de la muestra considerado es de $n = 30$ observaciones que equivalen a aproximadamente el 2.78 % de la población. Para la obtención de la muestra se realizó una asignación proporcional al tamaño de cada grupo, obteniéndose 25 y 5 observaciones para los $G1$ y $G2$ respectivamente. Finalmente, a fin de realizar la seleción de las muestras se emplea la función `S.piPS` del paquete *TeachingSampling* (Gutiérrez 2012).

Para la muestra seleccionada se estima la proporción mediante el π-estimador; siendo este $\hat{\rho} = 0.0669\,(6.69\,\%)$, con un intervalo de confianza $(0\,,\,0.1725)$.

Por otro lado, para estimar la proporción mediante la técnica *bootstrap* bayesiana, se toman 500 muestras con reemplazo de la muestra original s, cada muestra de tamaño 30, esto es, $s_b^* = (y_1^*\,,\,y_2^*, \ldots, y_{30}^*)$, con $b = 1, 2, ..., 500$ (muestra *bootstrap*) y con estas muestras calcular $\hat{\rho}_1^*, \hat{\rho}_2^*, ..., \hat{\rho}_{500}^*$. (véase la figura (3)).

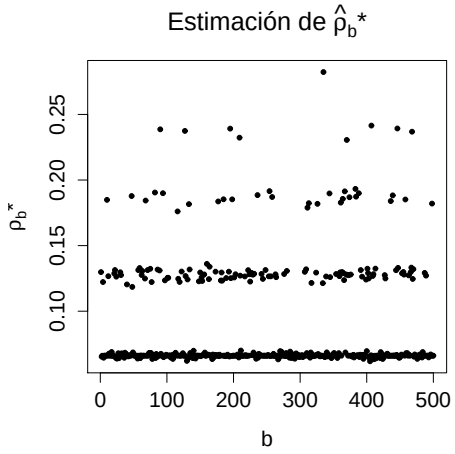

Figura 3: *Proporciones estimadas en las bootstrap. Fuente: elaboración propia.*

Ahora bien, con los 500 valores estimados se calcula la verosimilitud *bootstrap*

$$\hat{L}_B\,(\rho \mid \hat{\rho}) \quad = \quad \frac{1}{500\;(0.0047)} \sum_{b=1}^{500} K\left(\frac{2\;(0.0669) - \rho - \hat{\rho}_b^*}{0.0047}\right)$$

y para el cálculo de la distribución posterior de ρ, sin pérdida de generalidad, fijemos $\alpha = 25$; entonces al resolver $\rho = \frac{\alpha-1}{\alpha+\beta-2}$ se obtiene que $\beta = 457$ por tanto se toma como distribución *a priori* la distribución *beta* $(25\,,\,457)$, la cual es:

$$\xi\,(\rho) \quad \equiv \quad beta(25\,,\,457) \propto \rho^{24}\,(1-\rho)^{456}$$

Utilizando un Kernel Gausiano la distribución posterior de ρ es el producto de la verosimilitud y la distribución *a priori*, siendo esto:

$$\xi\,(\rho \mid y) \propto \hat{L}_B\,(\rho \mid \hat{\rho}) \cdot \rho^{24}\,(1-\rho)^{456}$$

De forma gráfica podemos ver esta distribuciones en la figura (4)

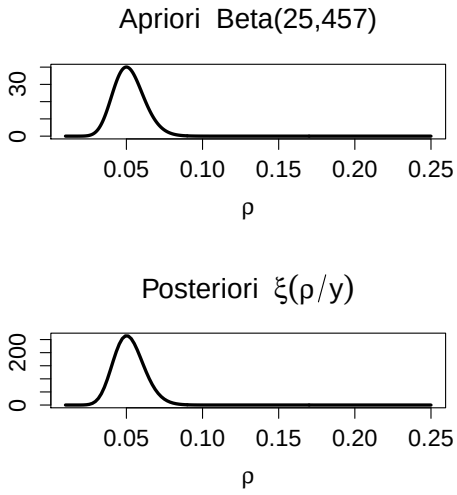

Figura 4: *Distribuciones a priori y a posteriori. Fuente: elaboración propia.*

Como es claro, la distribución porterior no se tiene de manera explícita (dado que la verosimilitud fue aproximada vía Kernel), por tanto, la media posterior, el intervalo de credibilidad y su longitud son calculados de manera empírica, siendo estos respectivamente: 0.0514, $(0.032\,,\,0.071)$ y 0.038.

A manera de conclusión se puede observar que el intervalo de credibilidad tiene una menor longitud en comparación con el intervalo de confianza. Por otro lado, la estimación puntual de ρ utilizando la función de pérdida cuadrática esta mucho más cercana al verdadero valor en comparación con el π-estimador.

Ahora bien, dado que en las simulación se pudo observar que el π-estimador no dio buenos resultados en muestras pequeñas, se decide aumentar el tamaño de

muestra a 200 y poner a prueba las dos metodologías. Los resultados se muestran en la Tabla (5).

	$\hat{\rho}$	*Intervalo*	*Longitud*
$\pi-$estimador	0.026	$(0.0269, 0.0270)$	0.0001
B.B	0.046	$(0.0307, 0.0630)$	0.0322

Tabla 5: *Estimación para una muestra de 200 observaciones. Fuente: elaboración propia.*

Se puede observar que el intervalo de credibilidad tiene una mayor longitud en comparación con el intervalo de confianza, sin embargo este último no contiene al parámetro. Así mismo podemos observar la estimación puntual de ρ utilizando el π-estimador evidenciando que está mucho más alejada del valor real que la estimación realizada mediante la metodología propuesta, lo que implica nuevamente que las estimaciones realizadas por el método *bootstrap* bayesiano son mejores.

5. Conclusiones y recomendaciones

El principal hallazgo consiste en que la estimación de la proporción, usando teoría bootstrap bayesiana, en todos los escenarios probados es mejor en cuanto a: el sesgo, RMSE, longitud del intervalo y cobertura, frente a la estimación hecha mediante teoría clásica y el estimador BPSP. Esto quiere decir, que con una adecuada elección de la distribución *a priori* se pueden encontrar sesgos tan pequeño como los obtenidos mediante el π-estimador, y frente al BPSP es mucho menor. Adicional a esto, se tienen menor RMSE, menor longitud y una mayor cobertura frente a las estimación hecha con la metodología clásica y mediante el estimador BPSP, aunque se cuenten con tamaños de muestras pequeños. Cabe resaltar que esta técnica no es difícil de emplear, puesto que el único supuesto que requiere es tener información previa del parámetro (distribución *a priori*) para su uso, y el conocimiento previo de una proporción a sido bastante estudiado y se han propuesto diferentes metodologías para la elicitación de este.

Un paso a seguir a este trabajo sería el caso en el cual se tengan encuestas multipropósito y se desee estimar más de una proporción a la vez. Adicional a esto, se puede estudiar el comportamiento de la metodología propuesta cuando se tienen variables auxiliares en el estudio. También se puede implementar esta metodología en parámetros diferentes a la proporción.

Recibido: 21 de enero de 2014
Aceptado: 16 de abril de 2014

Referencias

Box, G. E. P. & Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Massachusetts.

Chen, Q., Elliott, M. R. & Little, R. J. (2010), 'Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling', *Survey Methodology* **36**(1), 23–34.

Gambino, J. G. (2012), *pps: Functions for PPS sampling*. R package version 0.94. *http://cran.r-project.org/package=pps

Gutiérrez, H. A. (2012), *TeachingSampling: Sampling designs and parameter estimation in finite population*. R package version 2.0.1. *http://cran.r-project.org/package=TeachingSampling

Hollander, M. & Wolfe, D. A. (1999), *Nonparametric Statistical Methods*, Cambridge: University Press, Unite State of America.

Pfeffermann, D. & Royall, R. M. (1982), 'Balanced samples and robust Bayesian inference in finite population sampling', *Biometrika* **69**, 401–409.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.r-project.org

Särndal, C. E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer - Verlag, New York.

Shao, J. & Tu, D. (1995), *The jackknife and Bootstrap*, Springer, New York.

A. Códigos del ejemplo en R

```
require(MASS); require(hdrcde) ;require(cubature)
require(pps) ; require(TeachingSampling)

data(calif); head(calif)

Y1=ifelse(calif$white<=148,1,0);table(Y1)/1077
Log_pob=log(calif$population) # Variable auxiliar
d=dist(Log_pob, method="e",) # distance matrix
fit=hclust(d, method="mcquitty")

groups=cutree(fit, k=2)
groups=factor(groups)
levels(groups)=c("G1","G2")
table(groups)
```

```
Y=Y[order(groups)]

# Estimación cuando n=30

n=30
round(table(groups)*n/1077,0) # asignación proporcional al tamañoo del grupo
groups=groups[order(groups)]
pii<-c(Log_pob[groups=="G1"]*25/sum(Log_pob[groups=="G1"]), # Calculo de pi por grupo
Log_pob[groups=="G2"]*5/sum(Log_pob[groups=="G2"]))

Y=cbind(Grupos=groups,pii,Y) # Población ordenada según grupos
head(Y)
MG1=S.piPS(25,pii[groups=="G1"])# Selección de la muestra por grupo
MG2=S.piPS(5,pii[groups=="G2"])

# muestra obtenida

Ym=rbind(Y[MG1[,1],],Y[MG2[,1],])

# estimación clasica

Nest=sum(1/Ym[,2])
num=sum(Ym[,3]/Ym[,2])
pest1=num/Nest
Li=pest1-qnorm(0.95)*sqrt(varp(n,Ym[,2],Ym[,3],pii,pest1))
Ls=pest1+qnorm(0.95)*sqrt(varp(n,Ym[,2],Ym[,3],pii,pest1))

# Construcción de la a priori

# alpha=25
# beta=(24/0.05)-23
# a priori beta(25,457)
# Estimación de rho mediante boot

h=Boot(Ym[,c(3,2)],n,pest1,rho=0.1,alpha1=25,betta1=457)
c(Li=Li,Ls=Ls,lonc=Ls-Li,pest=pest1,Boot=h)

# Estimación cuando n=200
# repetir secuencia anterior con n=200


# Varianza de la proporcion

varp=function(n,pks,ys,pii,pest)\{

# n número de observaciones
```

```
# ys= de submuestreo
# pii= Probabilidades de inclusión
# pest= proporción estimada

pij=((n-1)/n)*(pks%*%t(pks))+((n-1)/n^2)*(pks%*%t(pks^2)+
(pks^2%*%t(pks)))-((n-1)/n^3)*pks%*%t(pks)*sum(pii^2)
pipj = pks%*%t(pks)
Vp = 0
for(i in 1:(n-1)){
for(j in (i+1):n){
Vp = Vp + ((pipj[i,j]-pij[i,j])/pij[i,j])*((ys[i]-pest)/pks[i]-
(ys[j]-pest)/pks[j])^2}}
Vp = (sum(1/pks))^(-2)*Vp
}

# p estimado mediante boot

Boot<-function(y,n,pest,alpha1,betta1,rho=0.1,B=500){
booT<-function(y,n){
pos1=sample(1:n,n,replace=T)
y1bos=y[pos1,]
while(length(which(y1bos[,1]==0))==n){pos1=sample(1:n,n)
y1bos=y[pos1,]}
Nestbos1=sum(1/y1bos[,2])
numbos1=sum(y1bos[,1]/y1bos[,2])
numbos1/Nestbos1  }
pestboot=replicate(B,expr=booT(y,n))
h1=bandwidth.nrd(pestboot)
rejilla=seq(0.01,0.99,length=B)
poste=0
for(i in 1:B){
x<-(2*pest-rejilla[i]-pestboot)/h1
kernelx=dnorm(x)
poste[i]<-1/(h1*B)*sum(kernelx)  }
apriori=dbeta(rejilla,alpha1,betta1)
posteriori=poste*apriori
phi1<-approxfun(rejilla,posteriori)
consta1<-adaptIntegrate(phi1,0.01,0.99)$integral
posteriori1<-(1/consta1)*phi1(rejilla)
muesb=sample(rejilla,1000, prob=posteriori1, replace=T)
p.est=mean(muesb) # estimacion boot proporcion
intcre1=hdr(muesb,95) # intervalo de credibilidad
cont<-ifelse(rho>intcre1$hdr[1] & rho<intcre1$hdr[2],1,0)
Lon.IC=(intcre1$hdr[2]-intcre1$hdr[1])
c(p.est=p.est,Conteo=cont,LonICboot=Lon.IC)  }
```