UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2021, Vol. 14, No. 2, pp. 7–25

Una propuesta bayesiana para la estimación de proporciones mediante el *Jackknife* en muestreo probabilístico

A bayesian proposal for the estimation of the proportion via Jackknife in probability sampling

Tania Nivia[a]
tnivian@unbosque.edu.co

Cristian Tellez[b]
ctellezp@unbosque.edu.co

Mario Pacheco[c]
mpachecol@unbosque.edu.co

Resumen

En este artículo se presenta una propuesta de estimación de proporciones poblacionales en muestreo probabilidades de inclusión desiguales que combina el método Jackknife y la inferencia bayesiana. Un estudio de simulación en el que se emplean diferentes distribuciones a priori para la proporción, ρ, es llevado a cabo para examinar como el estimador Jackknife bayesiano propuesto puede tener un mejor comportamiento que otros estimadores clásicos y alternativos en términos de sesgo y errores estándar. Finalmente, un ejemplo de aplicación de la metodología propuesta es presentado usando la Encuesta de Cultura Política Colombiana - 2019.

***Palabras clave*:** Muestreo probabilístico, estimación de proporciones, *Jackknife* bayesiano, *Bootstrap* bayesiano..

Abstract

In this paper, we present a Bayesian proposal for estimating finite population proportions in unequal sampling inclusion probabilities that combines the Jackknife method and Bayesian inference. A simulation study using different a priori distributions for the proportion, ρ, is carried out to examine how the proposed Bayesian Jackknife estimator may perform better than other classical and alternative estimators in terms of bias and standard errors. . Finally, an example of application of the proposed methodology is presented using the Colombian Political Culture

[a]Universidad El Bosque
[b]Universidad El Bosque
[c]Universidad El Bosque

Survey - 2019

***Keywords*:** Probability sampling, proportions estimation, bayesian Jackknife, bayesian Bootstrap..

1. Introducción

Uno de los parámetros más usados en encuestas de opinión pública, marketing, educación e investigaciones gubernamentales es la proporción. Usualmente, las encuestas de corte político o social se enfocan en variables de tipo categórico como lo son el sexo, la raza, el potencial electoral, etc. Existe en la literatura diferentes metodologías alternativas al estimador de Horvitz-Thompson habitual para mejorar las estimaciones de los parámetros de interés. Tal es el caso del estimador de razón, de diferencia o los estimadores de calibración (ver por ejemplo (Singh 2003)). Desde una perspectiva Bayesiana, (Graubardand et al. 2002) proponen realizar inferencias bayesianas para los parámetros superpoblacionales en muestreo por encuestas, (Lazar et al. 2008) proponen una aproximación bayesiana con distribuciones no informativas en muestreo de poblaciones finitas usando variables auxiliares. (Aitkin 2008) propone una aproximación *Bootstrap* Bayesiana para el análisis de poblaciones finitas en diseños de muestra complejos. (Zangeneh & Little 2015) consideran la inferencia bayesiana para el total poblacional de una muestra con probabilidades proporcionales al tamaño heterocedastica. En particular, (Tellez et al. 2014) proponen la estimación de proporciones en muestras complejas usando el método bootstrap bayesiano, este estimador se comporta mejor que otros estimadores clásicos en términos de sesgo y errores estándar.

Por otro lado, el método *Jackknife* es un método introducido por (Quenouille 1949) y llamado así por (Tukey 1958) para la estimación de la varianza de estimadores. El método *Jackknife* es otro tipo de técnica de remuestreo que junto con el *Bootstrap* es ampliamente utilizada para estimar parámetros y el error estándar de las estimaciones en muestreo probabilístico, por ejemplo, (Kovar et al. 1988, Shao & Tu 2012). Algunas actualizaciones incluyen el nuevo estimador de varianza *Jackknife* para muestreo con probabilidades desiguales de (Berger & Skinner 2005). En el proceso de estimación de un parámetro de interés, θ, mediante algún estimador, $\hat{\theta}$, el *Jackknife* consiste en calcular el estimador del parámetro de interés con los valores de la muestra reducida luego de eliminar la observación i-ésima, $\hat{\theta}_{(i)}$. Así, se obtiene una secuencia de estimadores $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, ..., \hat{\theta}_{(n)}$, con los que es posible obtener una estimación puntual Jackknife, una estimación del sesgo del estimador o una estimación de la varianza del estimador original.

En este artículo se presenta una propuesta bayesiana para estimar proporciones de poblaciones finitas mediante el método *Jacknife* en muestreo con probabilidades desiguales. Realizamos un estudio de simulación para mostrar los beneficios de la metodología propuesta analizando el sesgo y el error estándar de las estimaciones. También se presenta un ejemplo de aplicación de la metodología propuesta utilizando la Encuesta de Cultura Política de Colombia - 2019.

2. Inferencia *Jackknife* bayesiana para una proporción

Los métodos de remuestreo (p.ej. *Bootstrap*, *Jackknife*, validación cruzada y Permutaciones) son procedimientos estadísticos que se utilizan con frecuencia para hacer inferencias en muestreo por encuestas. El *Jackknife* es un tipo de técnica de remuestreo que se utiliza para estimar los valores de los parámetros y las desviaciones estándar correspondientes similar al *Bootstrap*. En resumen, podemos definir el procedimento *Jackknife* como sigue: suponga que se dispone de una muestra aleatoria $y_1, y_2, \cdots, y_n$ y un estimador $\hat{\theta}$ del parámetro θ. Sea $\hat{\theta}_k$ el estimador $\hat{\theta}$ evaluado en los $n-1$ elementos que quedan después de omitir el k-ésimo elemento de la muestra, $\hat{\theta}_k = \hat{\theta}(y_1, y_2, \cdots, y_{k-1}, y_{k+1}, \cdots, y_n)$. El primer paso del método consiste en calcular los denominados pseudovalores *Jakknife*

$$\hat{\theta}_{(k)} = n\hat{\theta} - (n-1)\hat{\theta}_k, \tag{1}$$

donde $k = 1, 2, \cdots, n$. El segundo paso consiste en calcular el estimador *Jackknife* $\hat{\theta}_{jk}$ de θ, asociado al estimador $\hat{\theta}$ y a la muestra $y_1, y_2, \cdots, y_n$, definido por (Quenouille 1949) como el promedio de los pseudovalores *Jackknife*, esto es,

$$\hat{\theta}_{jk} = \frac{1}{n}\sum_{k=1}^{n}\hat{\theta}_{(k)} = n\hat{\theta} - \frac{n-1}{n}\sum_{k=1}^{n}\hat{\theta}_k. \tag{2}$$

Asumiendo ahora la población finita de interés como $U = u_1, u_2, \cdots, u_k, \cdots, u_N$, de tamaño N, en donde cada unidad u_i, $i = 1, 2, ..., N$ tiene asociada una variable dicotómica y_i, que toma el valor 1 cuando el individuo posee una característica determinada y 0 cuando no la posee. Una muestra aleatoria s es seleccionada de U de acuerdo a un diseño de muestreo probabilístico arbitrario. Siguiendo el trabajo de (Tellez et al. 2014) y la teoría de estimación de parámetros diferentes al total en (C.E et al. 2003), considerando que en la muestra la variable de interés y es observada para todos los elementos seleccionados, el interés se centra en estimar la distribución de probabilidad posterior de la proporción ρ_y haciendo uso de los valores de la muestra y de las probabilidades de inclusión dadas por el diseño muestral.

Para implementar la metodología *Jackknife* Bayesiana, consideraremos que el parámetro ρ_y es estimado a partir de la muestra probabilística s, la cual ha sido seleccionada con un diseño muestral arbitrario $p(\cdot)$ haciendo uso del estimador de Horvitz-Thompson como

$$\hat{\rho}_{y\pi} = \frac{1}{\hat{N}}\sum_{i\in s}\frac{y_i}{\pi_i} \tag{3}$$

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

Consideremos que la distribución de probabilidad condicional $\xi(y|\rho_y)$ 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 que

$$\xi(\rho_y|y) \propto \xi(y|\rho_y)\xi(\rho_y) \tag{4}$$

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

Ahora bien, se debe pensar en una distribución a priori para ρ_y y en un supuesto distribucional para la variable y condicionado al parámetro ρ_y. En cuanto a la distribución a priori para ρ_y, existe un gran número de posibilidades entre distribuciones previas informativas y no informativas, como la distribución uniforme, la distribución beta o cualquier distribución que tenga como soporte el intervalo $(0, 1)$. Se debe tener en cuenta que en la teoría de muestreo usualmente no se hacen supuestos distribucionales para y condicionado al parámetro, lo que hace a la metodología *Jackknife* Bayesiana fundamental en la metodología propuesta, la cual consiste en obtener las distribuciones $\xi(y|\rho_y)$ y $\xi(\rho_y|y)$ de forma empírica.

2.1. Distribución posterior de ρ con a priori informativa

Dada la información a priori sobre el parámetro de interés ρ_y, $\xi(\rho_y)$, 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 llegar a un valor cercano de ξ utilizando un estimador de densidades, por ejemplo, $\hat{\xi}(y|\rho_y)$ y hallar un estimador de la distribución posterior de la forma

$$\xi(\rho_y|y) \propto \xi(\rho_y)\hat{L}(y_1, \cdots, y_n|\rho_y), \tag{5}$$

donde $\hat{L}(y_1, \cdots, y_n|\rho_y)$ representa la estimación *Jackknife* 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 de la muestra $y_1, y_2, \cdots, y_n$ se construyen B poblaciones artificiales U^*, puede ser replicando los y_i tantas veces como su factor de expansión $1/\pi_i$, siguiendo el principio de representatividad, a partir de un diseño de muestreo $P(\cdot)$.

2. Seleccionar algunas muestras de U^* denotadas por s^* de cada una de las U^* con el mismo diseño muestral usado para seleccionar la muestra original s de U. A partir de las muestras restantes se calcula el estimador $\rho^*_{y\pi b}$:

$$\hat{\rho}^*_{y\pi b} = \frac{1}{\hat{N}^*} \sum_{i \in s^*} \frac{y^*_{ib}}{\pi^*_{ib}} \tag{6}$$

   Donde $\hat{N}^* = \sum_{i \in s^*} \frac{1}{\pi^*_i}$, π^*_i es la probabilidad de inclusión de los elementos en la muestra *Jackknife* y y^*_{ib} es el j-ésimo elemento de la b-ésima muestra Jackknife.

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

$$f_B(u) = \frac{1}{Bh_B} \sum_{b=1}^{B} K\left(\frac{u - (\hat{\rho}^*_{y\pi b} - \hat{\rho}_{y\pi})}{h_B}\right) \tag{7}$$

Donde la función K es llamada comúnmente como el kernel y en general, es una función de densidad continua, unimodal y simétrica alrededor de 0 (Hollander et al. 2013) muestra las densidades Kernel más usadas. 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

$$\hat{L}_B(\hat{\rho}_{y\pi}|\rho_y) = \frac{1}{Bh_B} \sum_{b=1}^{B} K\left(\frac{2\hat{\rho}_{y\pi} - \rho - \hat{\rho}^*_{y\pi b}}{h_B}\right) \tag{8}$$

4. La distribución posterior resultante $\xi(\hat{\rho}_{y\pi}|\rho_y)$ es entonces proporcional a $\xi(\rho_y)\hat{L}(\hat{\rho}_{y\pi}|\rho_y)$ 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:

$$\xi(\rho_y|y) = c(y)x\xi(\rho_y)x\hat{L}(y_1, \cdots, y_n|\rho_y) \tag{9}$$

donde $c(y)$ se puede obtener como:

$$c(y) = \frac{1}{\int \xi(\rho_y)x\hat{L}(y_1, \cdots, y_n|\rho_y)d\rho_y} \tag{10}$$

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 $\rho^1_y, \rho^2_y, \cdots, \rho^m_y$ a través de la distribución posterior $\xi(\rho_y|y)$ de la siguiente manera:

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

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

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

Las funciones comúnmente utilizadas para minimizar dichos errores son la función de pérdida cuadrática, la función de pérdida error absoluto y la función de pérdida escalonada (Box & Tiao 2011).

2.2.1. Función de pérdida cuadrática para la proporción

Se considera la función $L(\rho_y\rho_c) = (\rho_c - \rho_y)^2$ que se denominará como función de pérdida cuadrática asociada al parámetro ρ_y, y sea ρ_c la estimación considerada para ρ_y. Sean $\rho_y^1, \rho_y^2, \cdots, \rho_y^m$ una muestra aleatoria de tamaño m generada a través de la distribución posterior $\xi(\rho_y|y)$.

La diferencia entre ρ_c y el valor real de ρ_y se hace mínima si ρ_c se estima empleando la siguiente expresión:

$$\rho_c = E(\rho_y|y) = \int_{-\infty}^{+\infty} \rho_y \xi(\rho_y|y) d\rho_y \tag{11}$$

Esta integral se calcula numéricamente puesto que $\xi(\rho_y|y)$ es una función empírica. Por otro lado, la estimación vía Monte Carlo de la media posterior es:

$$\rho_c = \bar{\rho}_y = \frac{\sum_{j=1}^{m} \rho_y^j}{m} \tag{12}$$

y un error estándar estimado es:

$$se_{\rho_c} = \sqrt{\frac{\sum_{j=1}^{m} (\rho_y^j - \rho_c)^2}{(m-1)m}} \tag{13}$$

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 (Tellez et al. 2014) para así poder comparar los resultados entre las estimaciones que utilizan la teoría bayesiana como lo son en su caso Bootstrap y la propuesta en este trabajo que es *Jackknife*.

3.1. Diseño de la simulación

El estudio de simulación pretende evaluar el comportamiento de la metodología propuesta y compararla con el procedimiento *Bootstrap* Bayesiano en términos del sesgo y el error estándar de estimación. El procedimiento consiste en simular dos poblaciones artificiales de tamaño 2000, generando también 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 $71, 72, 73, \cdots, 2070$. Por otro lado, las probabilidades de inclusión en la población son calculadas proporcionales a la variable tamaño, $\pi_i = n * x_i / \sum x_i$, con $x_i = 71, 72, \cdots, 2070$. 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 simulación se tomó la estructura de media de incremento lineal $f(\pi_i) = 3\pi_i$, esto debido a que en el trabajo antes mencionado no se demostró una diferencia significativa entre utilizar estructura de media lineal y exponencial.

Por otra parte, las variables binarias Y_1, Y_2 y Y_3 son generadas de la siguiente manera: Y_1 es igual a 1 si Z es menor o igual a su percentil 10 y 0 en otro caso. De la misma manera se generan las respuestas para Y_2 y Y_3 usando los percentiles 50 y 90, respectivamente. El objetivo es la proporción poblacional para $Y = 1$. Para cada simulación, se genera una población finita y se calcula la verdadera proporción poblacional para $Y = 1$. Luego, se seleccionan muestras aleatorias, de tamaños $n = 30, 50, 100$ y 200 con probabilidades proporcionales al tamaño (πPT) de cada población y se calcula la proporción estimada $\hat{\rho}_{B.B}$ y $\hat{\rho}_{J.B}$ basada en la función de pérdida cuadrática (media posterior).

El anterior proceso se repite 1000 veces y se calcula el sesgo empírico, el sesgo relativo (S.R(%)), la raíz del error cuadrático medio (RECM), el margen de error al 95 % (M.E(%)), las coberturas de los intervalos de credibilidad y la longitud de los mismos.

Sea $\hat{\rho}_j$ una estimación de ρ_j basada en la muestra $j - ésima$, el sesgo empírico, el S.R(%), la raíz del error cuadático medio, y el M.E(%) son:

$$sesgo = \frac{1}{1000} \sum_{j=1}^{1000} (\hat{\rho}_j - \rho) \tag{14}$$

$$S.R = \frac{sesgo}{\rho} * 100\,\% \tag{15}$$

$$RECM = \sqrt{\frac{1}{1000} \sum_{j=1}^{1000} (\hat{\rho}_j - \rho)^2} \tag{16}$$

$$M.E = RECM * Z_{1-\frac{\alpha}{2}} * 100\,\% \tag{17}$$

Cabe resaltar que, dado que el parámetro es un valor entre cero y uno, utilizar el coeficiente de variación estimado en este caso no sería útil. Lo anterior debido a que esta medida al momento de reemplazar el estimador del parámetro tendría como resultado un valor inflado y muy posiblemente se llegaría a la conclusión equivocada. Es por lo anterior que, como medida de calidad, se utilizará el error estándar o el margen de error, los cuales son mostrados en la tabla de resultados de la simulación.

Se planteó utilizar una distribución a priori $beta(\alpha, \beta)$, donde α toma los valores de $\alpha = 25, 50, 100$ y β toma valores de $\beta = 0.1, 0.5$ y 0.9. Como se mencionó

n	ρ	Mét.	A priori	Sesgo	SR(%)	REMC	ME (%)	Cob.	Amp.
30	0,1	J.B	Beta(25,225)	-0,009	-8,850	0,019	3,704	84,4	0,063
			Beta(50,450)	-0,006	-6,060	0,014	2,648	86,5	0,047
			Beta(100,900)	**-0,005**	**-4,550**	0,011	2,170	87,5	0,034
		B.B	Beta(25,225)	-0,009	-8,980	0,019	3,775	83,9	0,063
			Beta(50,450)	-0,007	-6,520	0,015	2,883	85,6	0,047
			Beta(100,900)	-0,005	-4,860	0,012	2,295	86,7	0,034
	0,5	J.B	Beta(25,25)	0,001	0,100	0,025	4,906	99,9	0,247
			Beta(50,50)	-0,001	-0,130	0,015	2,867	99,7	0,184
			Beta(100,100)	0,000	-0,088	0,008	1,635	99,8	0,134
		B.B	Beta(25,25)	**0,000**	**0,064**	0,025	4,924	99,9	0,247
			Beta(50,50)	-0,001	-0,134	0,015	2,891	99,7	0,184
			Beta(100,100)	0,000	-0,092	0,008	1,613	99,8	0,134
	0,9	J.B	Beta(225,25)	0,003	0,313	0,007	1,397	98,1	0,068
			Beta(450,50)	0,002	0,196	0,005	0,904	99,1	0,050
			Beta(900,100)	**0,001**	**0,106**	0,003	0,523	99,8	0,036
		B.B	Beta(225,25)	0,003	0,329	0,008	1,478	97,5	0,069
			Beta(450,50)	0,002	0,211	0,005	0,970	98,7	0,050
			Beta(900,100)	0,001	0,113	0,003	0,566	99,6	0,036
50	0,1	J.B	Beta(25,225)	-0,006	-6,000	0,014	2,826	91,2	0,065
			Beta(50,450)	-0,005	-4,620	0,012	2,411	91,7	0,048
			Beta(100,900)	**-0,003**	**-3,170**	0,010	1,889	94,6	0,035
		B.B	Beta(25,225)	-0,006	-6,030	0,015	2,856	91,3	0,065
			Beta(50,450)	-0,005	-4,760	0,013	2,519	91,7	0,048
			Beta(100,900)	-0,003	-3,090	0,009	1,784	94,6	0,035
	0,5	J.B	Beta(25,25)	-0,001	-0,196	0,025	4,849	99,8	0,237
			Beta(50,50)	0,000	-0,028	0,015	2,934	99,9	0,180
			Beta(100,100)	0,000	-0,010	0,008	1,613	99,9	0,132
		B.B	Beta(25,25)	-0,001	-0,214	0,025	4,827	99,9	0,237
			Beta(50,50)	0,000	-0,040	0,015	2,956	99,9	0,179
			Beta(100,100)	**0,000**	**-0,002**	0,008	1,615	99,9	0,132
	0,9	J.B	Beta(225,25)	0,003	0,292	0,007	1,380	97,9	0,068
			Beta(450,50)	0,002	0,178	0,004	0,864	99,1	0,050
			Beta(900,100)	**0,001**	**0,101**	0,003	0,549	99,1	0,036
		B.B	Beta(225,25)	0,003	0,303	0,007	1,429	97,7	0,068
			Beta(450,50)	0,002	0,188	0,005	0,913	99	0,050
			Beta(900,100)	0,001	0,109	0,003	0,602	98,9	0,036

Tabla 1: sesgo, SR(%), RECM, ME(%), coberturas y longitudes de los intervalos de credibilidad, para simulación con $n = 30, 50$

anteriormente, para efectos de comparación se utilizaron las mismas distribuciones a priori que en el trabajo realizado por (Tellez et al. 2014), por lo tanto, allí se pueden encontrar los pasos con los cuales se llegaron a estos valores de β.

3.2. Resultado de la simulación

En las tablas 1 y 2 se muestran los resultados de la simulación en cuanto a sesgo, SR(%), RECM, ME(%), coberturas y longitudes de los intervalos de credibilidad, todo esto para las estimaciones realizadas por las metodologías mencionadas, con tamaños de muestra de $n = 30, 50, 100, 200$. Utilizando además una estructura de media lineal, un kernel Gausiano y variando los parámetros de las distribuciones *a priori*.

Los valores resaltados en negrita representan el menor sesgo y sesgo relativo(%) en cada posible valor de ρ, en los diferentes tamaños de muestra evaluados. Una buena configuración para la metodología propuesta sería una distribución a priori

n	ρ	Mét.	A priori	Sesgo	SR(%)	REMC	ME (%)	Cob.	Amp.
100	0,1	J.B	Beta(25,225)	-0,005	-4,640	0,012	2,444	94,3	0,066
			Beta(50,450)	-0,003	-2,710	0,008	1,476	95,7	0,049
			Beta(100,900)	-0,002	-2,110	0,006	1,243	96,3	0,035
		B.B	Beta(25,225)	-0,005	-4,650	0,012	2,436	94,3	0,066
			Beta(50,450)	-0,003	-2,750	0,008	1,523	95,7	0,049
			Beta(100,900)	**-0,002**	**-2,090**	0,006	1,201	96,5	0,035
	0,5	J.B	Beta(25,25)	0,000	0,046	0,025	4,802	99,9	0,218
			Beta(50,50)	0,000	0,094	0,017	3,275	100	0,171
			Beta(100,100)	**0,000**	**-0,012**	0,009	1,823	100	0,129
		B.B	Beta(25,25)	0,000	0,038	0,025	4,806	99,9	0,217
			Beta(50,50)	0,000	0,076	0,017	3,267	100	0,171
			Beta(100,100)	0,000	-0,022	0,009	1,840	100	0,128
	0,9	J.B	Beta(225,25)	0,002	0,222	0,006	1,176	99,5	0,065
			Beta(450,50)	0,001	0,131	0,004	0,739	99,5	0,049
			Beta(900,100)	**0,001**	**0,073**	0,002	0,429	100	0,036
		B.B	Beta(225,25)	0,002	0,214	0,006	1,196	99,5	0,066
			Beta(450,50)	0,001	0,130	0,004	0,751	99,6	0,049
			Beta(900,100)	0,001	0,073	0,002	0,437	100	0,036
200	0,1	J.B	Beta(25,225)	-0,003	-2,530	0,008	1,572	97,2	0,067
			Beta(50,450)	-0,002	-1,950	0,007	1,399	97,6	0,049
			Beta(100,900)	**-0,001**	**-0,950**	0,004	0,710	99,2	0,036
		B.B	Beta(25,225)	-0,003	-2,510	0,008	1,564	97,7	0,067
			Beta(50,450)	-0,002	-1,950	0,007	1,401	97,6	0,049
			Beta(100,900)	-0,001	-0,980	0,004	0,733	99,1	0,036
	0,5	J.B	Beta(25,25)	0,004	0,802	0,024	4,714	99,5	0,190
			Beta(50,50)	0,003	0,506	0,017	3,371	100	0,156
			Beta(100,100)	0,002	0,330	0,010	2,052	100	0,122
		B.B	Beta(25,25)	0,004	0,776	0,024	4,692	99,6	0,189
			Beta(50,50)	0,002	0,486	0,017	3,367	100	0,156
			Beta(100,100)	**0,002**	**0,310**	0,010	2,031	100	0,122
	0,9	J.B	Beta(225,25)	0,002	0,234	0,006	1,235	99,9	0,060
			Beta(450,50)	0,001	0,127	0,004	0,778	100	0,046
			Beta(900,100)	0,001	0,078	0,002	0,457	100	0,035
		B.B	Beta(225,25)	0,002	0,222	0,006	1,241	99,9	0,060
			Beta(450,50)	0,001	0,122	0,004	0,782	100	0,046
			Beta(900,100)	**0,001**	**0,076**	0,002	0,463	100	0,035

Tabla 2: sesgo, SR(%), RECM, ME(%), coberturas y longitudes de los intervalos de credibilidad, para simulación con $n = 100, 200$

Beta con parámetros $\alpha = 100$ y $\beta = 100$, con esto se obtuvo un sesgo de 0.012 % que a su vez es el menor sesgo en los resultados de la simulación, por otra parte, hay un caso particular en este ejercicio y es que al tener una distribución a priori *Beta* con parámetros $\alpha = 900$ y $\beta = 100$, se presenta los mismos resultados (resaltados en amarillo) en cuanto a sesgo, SR % y cobertura en las dos metodologías, sin embargo, con esta configuración se obtuvo el menor ME % en la simulación y corresponde a la metodología *Jackknife* Bayesiana, esto a la final quiere decir que el estudio es sensible a las distribuciones a priori que se escojan, como en este caso fue al variar los parámetros de la distribución beta.

En general, los resultados muestran que el comportamiento de las metodologías bayesianas comparadas son muy similares en términos de sesgo, siendo ambas buenas para ciertas configuraciones por tener sesgo y margen de error muy cercanas a cero, por otra parte, se puede ver que el estimador propuesto es aproximadamente insesgado, esto a su vez implica que este estimador cumple con las propiedades de buena cobertura y poca amplitud.

Otra manera de ver que los resultados son buenos es de acuerdo a la interpretación que da el (DANE 2008), este departamento menciona que tener una medida de precisión hasta el 7 % es una estimación precisa, entre 8 % y el 14 % es una precisión aceptable, entre 15 % y 20 % una precisión regular y por encima del 20 % se considera una estimación poco precisa, por lo que se puede ver que los resultados obtenidos en su totalidad representan una estimación precisa ya que estos no superan el 7 % en el margen de error.

4. Ejemplo de la metodología propuesta

Para efectos de aplicación y haciendo uso de la metodología propuesta en este artículo, se consultó la Encuesta de Cultura Política del año 2019 realizada por el DANE en Colombia. Dicha encuesta indaga sobre la percepción que tienen los ciudadanos colombianos frente al entorno político del país, como también explora el conocimiento hacia los conceptos de democracia, los mecanismos y espacios de participación ciudadana. También se exploran temas relacionados con el comportamiento electoral, la percepción sobre los partidos políticos y la confianza en las instituciones, destacando además que la población objetivo de este estudio son los ciudadanos mayores de 18 años los cuales son más de 30 millones, por otra parte, un dato a resaltar sobre la historia de esta encuesta es que ha sido aplicada desde hace aproximadamente 13 años. Según el (DANE 2008) *"La encuesta de Cultura Política busca generar información estadística estratégica que permita caracterizar aspectos de la cultura política colombiana, acumulación de capital social, participación en escenarios comunitarios y confianza, basados en las percepciones y prácticas de los ciudadanos sobre su entorno político y social, como insumo para diseñar políticas públicas dirigidas a fortalecer la democracia y la convivencia pacífica colombiana".*

Para efectos de la aplicación de esta metodología, se tomaron algunas preguntas de interés:

- P.1. ¿Sabe leer y escribir?
- P.2. ¿Votó usted en las elecciones presidenciales de 2018?
- P.3. ¿Votaría alguna vez por una mujer?
- P.4. ¿Usted cree que en Colombia existe la libertad de expresar y difundir su pensamiento?
- P.5.¿Usted se informa sobre la actualidad política del país?

La muestra que se utilizó en este ejercicio fue de 18393 entrevistados y 7 variables (las cuales corresponden a las preguntas). El interés de este ejercicio es estimar, mediante el método clásico, *Bootstrap* Bayesiano y *Jackknife* Bayesiano, la proporción de personas que cumplan con la característica de interés en cada una de

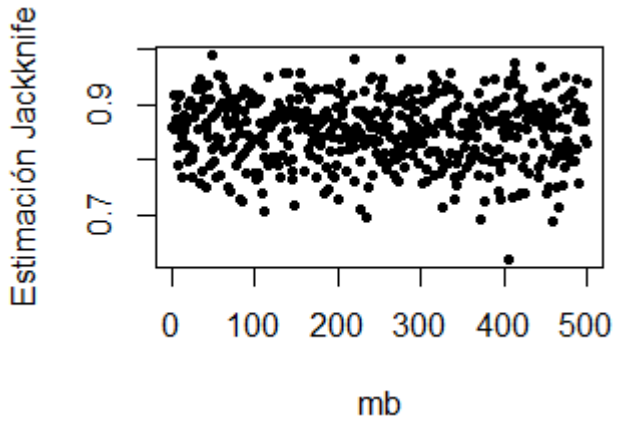

Figura 1: Estimación de la proporción asociada a P.1 por la metodología Jackknife

las preguntas. Estos resultados son comparados en términos de sesgo y margen de error.

El proceso de extracción de las muestras artificiales se realizó con un diseño de muestreo probabilístico tomando las probabilidades de inclusión publicadas. Adicional a lo anterior, se tomó un tamaño de muestra similar al de la base inicial, la cual equivale al 0.05 % de la población.

Teniendo en cuenta lo anterior, el valor estimado para la pregunta ¿Sabe leer y escribir? (que llamaremos P.1) por medio de la metodología clásica es de aproximadamente el 93 %. Por otro lado, utilizando las metodologías bayesianas comparadas en la simulación, se tomaron 500 poblaciones artificiales con reeemplazo de la muestra original y para cada población artificial se obtuvo una muestra artificial de tamaño similar a la muestra inicial, teniendo para cada una de ellas los estimadores respectivos, según la metodología.

En la Figura 1 se puede ver los posibles valores de la estimación de ρ (la proporción poblacional) por medio de la metodología *Jackknife*, siguiendo la idea, ahora se calcula la verosimilitud con los valores anteriormente encontrados, quedando como:

$$\hat{L}_J(\rho|\hat{\rho}) = \frac{1}{500(0.07173)} \sum_{mb=1}^{500} K\left(\frac{2(0.934) - \rho - \hat{\rho}^*_{mb}}{0.07173}\right) \tag{18}$$

En la anterior ecuación, se puede ver la forma en la que se calcula la verosimilitud para realizar la estimación por medio de *Jackknife* Bayesiano para P.1, sin pérdida de generalidad, se fija $\alpha = 90$ y al resolver $\rho = \frac{\alpha-1}{\alpha+\beta-2}$ se obtiene que $\beta = 5$, por lo tanto, queda una distribución a priori como: $Beta(\alpha = 25, \beta = 5)$, visto de otra forma es:

$$\xi(\rho) = beta(90, 5) \propto \rho^{89}(1-\rho)^4 \tag{19}$$

Utilizando un Kernel Gausiano la distribución posterior de ρ es el producto de la

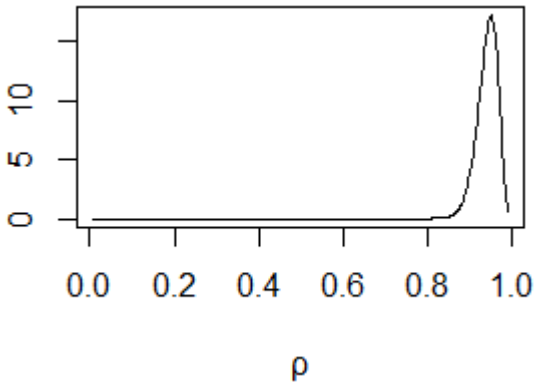

Figura 3: Distribución a posteriori para la proporción poblacional, ρ, de la P.1 por medio de *Jackknife* Bayesiano

verosimilitud y la distribución *a priori*, quedando como:

$$\xi(\rho|y) \propto \hat{L}_J(\rho|\hat{\rho})\rho^{89}(1-\rho)^4 \tag{20}$$

De forma gráfica se pueden ver estas distribuciones como:

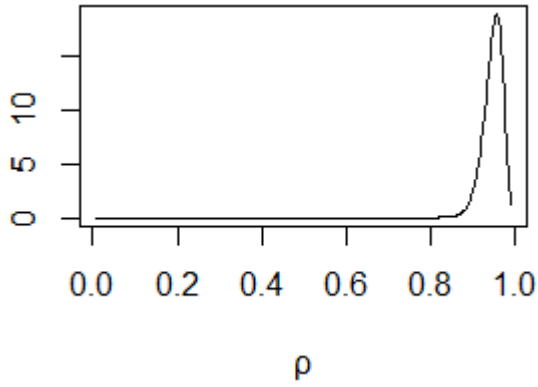

Figura 2: Distribución a priori para la proporción poblacional, ρ, de la P.1 por medio de *Jackknife* Bayesiano

Como es de esperarse, la distribución posterior no se tiene de manera explícita (por la aproximación que se dio vía kernel), así que, la media posterior, el intervalo de credibilidad y su cobertura son calculadas de manera empírica. Continuando con los resultados, se obtuvo una estimación *Jackknife* Bayesiana de $\hat{\rho}_{J.B} = 0.954(95\,\%)$ y en la estimación *Bootstrap* Bayesiana un $\hat{\rho}_{B.B} = 0.954(95\,\%)$.

Para darle un contexto a estos datos, por medio de la metodología propuesta se puede ver que se estimó que el 95 % de la población sabe leer y escribir, este

	Mét. Clásico		Mét. Boot.B		Mét. Jack.B	
	$\hat{\rho}_{Cla}$	M.E %	$\hat{\rho}_{B.B}$	M.E %	$\hat{\rho}_{J.B}$	M.E %
¿Sabe leer y escribir?	0,934	11,224	0,954	2,649	0,954	2,517
¿Votó en las elecciones presidenciales del 2018?	0,787	13,203	0,784	3,470	0,783	3,538
¿Votaría alguna vez por una mujer?	0,948	7,184	0,948	3,025	0,947	2,773
¿Cree que en Colombia existe la libertad de expresión y difundir su pensamiento?	0,466	14,024	0,471	1,511	0,47	1,629
¿Se informa sobre la actualidad política del país?	0,697	10,467	0,671	1,626	0,672	1,565

Tabla 3: Resultados de la aplicación

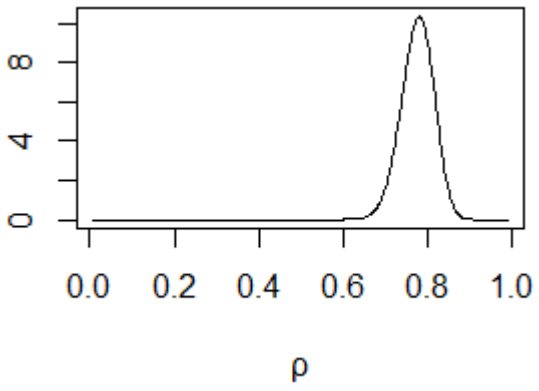

Figura 4: Distribución a priori para la proporción poblacional, ρ, para la P.2 por medio de Jackknife Bayesiano

estimador está acompañado de un margen de error muy pequeño como se puede ver en el Cuadro 3, siendo este valor del 2.5 % y según lo relacionado con la publicación del (DANE 2008) esto se considera una buena estimación, por la regla del 7 %.

Las demás estimaciones se realizaron de manera similar a la explicación anterior, es decir, partiendo de una estimación clásica, seguido de la estimación ya sea *Jackknife* o *Bootstrap* y finalmente se agregó el componente bayesiano en estas estimaciones.

Por medio del estimador propuesto se obtuvo que el 78 % de las personas votaron para las elecciones presidenciales del año 2018. Además se puede observar en las Figuras 4 y 5 las estimaciones por el método *Jackknife* bayesiano en las cuales se ve un mayor volumen hacía valores cercanos al 0.8(80 %), o por otro lado, se puede concluir que el porcentaje de personas que votaron en el año 2018 para las elecciones presidenciales esta entre 65 % y 85 % aproximadamente.

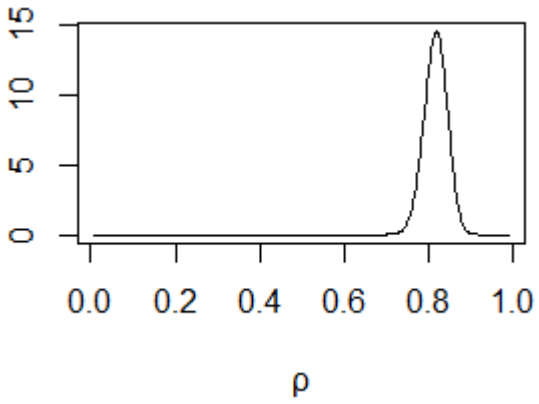

Figura 5: Distribución a posteriori para la proporción poblacional, ρ, para la P.2 por medio de Jackknife Bayesiano

La siguiente pregunta es de interés personal, debido a que en el año 2019 fueron las elecciones de alcaldes en todo Colombia y pasó algo particular en la ciudad de Bogotá, por primera vez ganó una mujer para ser la mandataria en la alcaldía de la ciudad, es por esto por lo que se quiso evaluar la pregunta ¿Votaría alguna vez por una mujer?, esto arrojo una estimación del 0.948(95 %) por medio de la metodología clásica, por Bootstrap Bayesiano fue un 0.948 (95 %) y por el *Jackknife* bayesiano fue un 0.947(95 %), este último con una amplitud de 0.069 y una cobertura del 86 %, se puede ver además que aunque son sesgos despreciables no deja de ser el menor de ellos. Concluyendo que más del 90 % de las personas votarían alguna vez por una mujer.

Las Figuras 6 y 7 hacen referencia a las estimaciones vía *Jackknife* bayesiano a la pregunta ¿Cree que en Colombia existe la libertad de expresión y difundir su pensamiento?, esto tristemente tiene una estimación del 47 % aproximadamente por medio de las tres metodologías, es decir que ni siquiera la mitad de las personas creen que se puede expresar libremente, las estimaciones bayesianas tienen un M.E % inferior el 2 %, lo que implica que son estimaciones precisas según la regla anteriormente mencionada, además de una cobertura del 100 %; dando un contexto a esto, en el país usualmente pasa que las personas que expresan su pensamiento libremente y no va de acuerdo a los pensamientos del gobierno, estas personas terminan borradas del mapa, es por esto por lo que posiblemente es baja la proporción de personas que no creen en la libre expresión en Colombia.

Para terminar, la pregunta ¿Se informa sobre la actualidad política del país? tiene una estimación *Jackknife* bayesiana del 67 %, con una clasificación de estimación precisa según el M.E que es inferior al 7 % y una cobertura del 100 %, esta estimación implica que un poco más de la mitad de las personas se informan de la actualidad política de Colombia.

A manera de conclusión, se puede observar que la estimación de ρ utilizando la

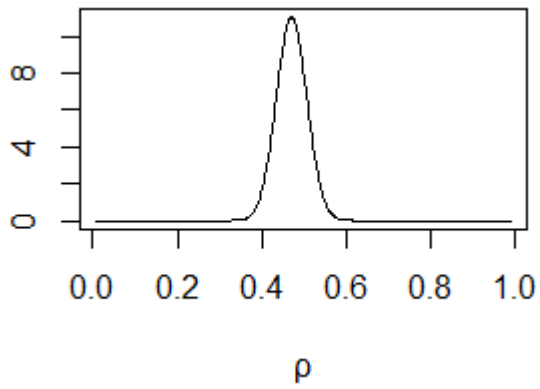

Figura 6: Distribución a priori para la proporción poblacional, ρ, para la P.4 por medio de *Jackknife* Bayesiano

metodología propuesta es excelente en términos de sesgo (por ser cercana a cero) comparado con la estimación clásica o Bootstrap Bayesiana, sin embargo, estos métodos bayesianos no presentan gran diferencia entre ellos, siendo ambos buenos para llegar a estimaciones certeras ya que presentan un margen de error inferior al 7 %.

5. Conclusiones

Fue posible construir un método de estimación puntual e intervalar mediante la metodología propuesta del métos *Jackknife* Bayesiano. Se mostró mediante un pequeño estudio de simulación que esta tiene un mejor desempeño en términos de sesgo y amplitud de los intervalos de estimación (credibilidad) que la estimación clásica. Esto se puede observar en los resultados del estudio de simulación, en los que se aprecia que los sesgos son cercanos a cero, lo que sugiere que el estimador es aproximadamente insesgado, y además, en la mayoría de los escenarios los intervalos de credibilidad tienen una amplitud baja con una cobertura por encima del 80 %. Adicional a lo anterior, los márgenes de error son bajos lo cual es aceptable para tomarlos como dato oficial. Así, el estimador propuesto supera al estimador clásico.

Adicionalmente, los sesgos y los niveles de cobertura del estimador *Jackknife* son cercanos a los del estimador *Bootstrap*, lo que implica que ambos estimadores son buenos al momento de estimar una proporción, resaltando que el estimador *Jackknife* teóricamente tiene mejor comportamiento que el Bootstrap (Shao & Tu 2012), y adicionalmete siendo el *Jackknife* un método menos intensivo computacionalmente.

También fue posible implementar en la práctica la metodología propuesta. Se

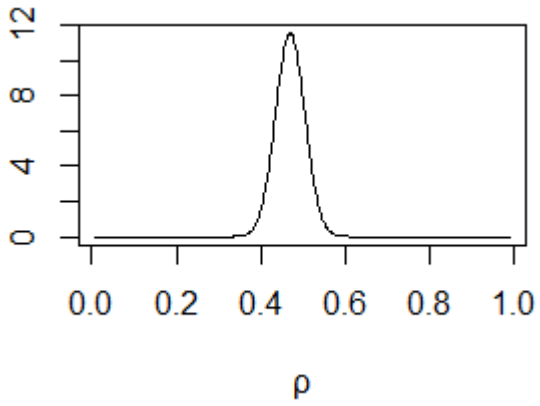

Figura 7: Distribución posteriori para la proporción poblacional, ρ, para la P.4 por medio de *Jackknife* Bayesiano

realizó una aplicación con datos reales en las que se logró hacer inferencias sobre proporciones asociadas a preguntas de la Encuesta de Cultura Política del año 2019 realizada por el DANE en Colombia.

Recibido:
Aceptado:

Referencias

Aitkin, M. (2008), 'Applications of the bayesian bootstrap in finite population inference', *Journal of Official Statistics* **24**(1), 21.

Badii, M., Castillo, J., & Landeros, J. (2007), 'Precisión de los índices estadísticas: Técnicas de jacknife & bootstrap (Precision of statistical indices: Jacknife & bootstrap techniques)'.

Berger, Y. G. & Skinner, C. (2005), 'A jackknife variance estimator for unequal probability sampling', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 79–89.

Box, G. & Tiao, G. (2011), *Bayesian inference in statistical analysis*, Vol. 40, John Wiley & Sons.

C.E, S., B, S. & J, W. (2003), *Model assisted survey sampling*, Springer Science & Business Media.

DANE (2008), *Estimación e interpretación del coeficiente de variación de la encuesta cocensal, https://www.dane.gov.co/files/investigaciones/boletines/censo/ est_interp_ coefvariacion.pdf*, https://www.dane.gov.co/files/investigaciones/boletines/ censo/est_interp_.

Gallego, J. (1997), 'Estimadores de razón: una revisión', *Qüestiió: quaderns d'estadística i investigació operativa* **21**(1).

Graubardand, B., Korn, E. et al. (2002), 'Inference for superpopulation parameters using sample surveys', *Statistical Science* **17**(1), 73–96.

Gutiérrez, H. (2009), *Estrategias de muestreo diseño de encuestas y estimación de parámetros.*, Universidad Santo Tomas, Bogota (Colombia).

Guzmán, A. (1978), 'Algunas consideraciones sobre la naturaleza de la técnica jackknife de estimación y las ventajas', *Estadística* **78**(79).

Hollander, M., D.A, W. & E, C. (2013), *Nonparametric statistical methods*, Vol. 751, John Wiley & Sons.

J, S. & D, T. (2012), *The jackknife and bootstrap*, Springer Science & Business Media.

Kovar, J., Rao, J. & Wu, C. (1988), 'Bootstrap and other methods to measure errors in survey estimates', *Canadian Journal of Statistics* **16**(S1), 25–45.

Lazar, R., Meeden, G. & Nelson, D. (2008), 'A noninformative bayesian approach to finite population sampling using auxiliary variables', *Survey Methodology* **34**(1), 51.

López, A. & Williams, L. (2006), 'Evaluación de métodos no paramétricos para la estimación de riqueza de especies de plantas leñosas en cafetales', *Botanical Sciences* (78).

Martínez, J. (1998), 'Estimación del número de clusters en una población aplicando el jackknife generalizado'.

Meeden, G. (1999), 'A noninformative bayesian approach for two-stage cluster sampling', *Sankhyā: The Indian Journal of Statistics, Series B* .

Mesa, L., Rivera, M. & Romero, J. (2011), 'Descripción general de la inferencia bayesiana y sus aplicaciones en los procesos de gestión', *La simulación al Servicio de la Academia* .

Miller, R. (1974), 'The jackknife-a review', *Biometrika* **61**(1), 1–15.

Pacheco-López, M. & Brango, H. (2010), 'Intervalos de confianza jackknife para cuantiles en muestreo con probabilidades desiguales', *TecnoLógicas* (2).

Pacheco-López M, M. G. (2007), 'Un estimador jackknife de varianza en muestreo en dos fases con probabilidades desiguales', *Revista Colombiana de Estadística* **30**(2).

Quenouille, M. (1949), 'Approximate tests of correlation in time-series', *Journal of the Royal Statistical Society: Series B (Methodological)* **11**(1), 68–84.

Ramírez, J., Osuna, I., Rojas, J. & Guerrero, S. (2015), 'Remuestreo Bootstrap y Jackknife en confiabilidad: Caso Exponencial y Weibull', *Revista Facultad de Ingeniería* **25**(41).

R.G, E. & Merino, C. (2006), 'Intervalos de confianza para las estimaciones de proporciones y las diferencias entre ellas', *Interdisciplinaria* **23**(2).

Ríos, T. (2014), 'Kriging y simulación secuencial de indicadores con proporciones localmente variables'.

Royall, R. & Pfeffermann, D. (1982), 'Balanced samples and robust bayesian inference in finite population sampling', *Biometrika* **69**(2).

Schiaffino, A., Rodríguez, M., Pasarín, M., Regidor, E., Borrell, C. & Fernández, E. (2002), 'Odds ratio o razón de proporciones su utilización en estudios transversales', *Gac Sanit* .

Shao, J. & Tu, D. (2012), *The jackknife and bootstrap*, Springer Science & Business Media.

Singh, S. (2003), *Advanced Sampling Theory With Applications: How Michael Selected Amy*, Vol. 2, Springer Science & Business Media.

Tellez, C., Guerrero, S. & Pacheco-López, M. (2014), 'Inferencia bootstrap bayesiana para una proporción en muestreo con probabilidades desiguales', *Comunicaciones en Estadística* **7**(1), 31.

Tukey, J. (1958), 'Bias and confidence in not quite large samples', *Ann. Math. Statist.* **29**, 614.

Valencia, E. & Mesa, F. (2009), 'Técnica de jackknife y estimadores en un modelo lineal', *Scientia et technica* **1**(41).

Zangeneh, S. & Little, R. (2015), 'Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample', *Journal of Survey Statistics and Methodology* **3**(2), 162–192.