Una propuesta bayesiana para la estimación de la proporción vía *Jackknife* en muestreo probabilístico

Facultad de Ciencias
Departamento de Matemáticas y Estadística
Programa de Estadística

Tania Vanessa Nivia Neira
Estudiante Programa de Estadística

Ph.D(c) Cristian Fernando Tellez Piñerez
Director

Índice

1. Planteamiento del problema

Uno de los parámetros más utilizados en los estudios de mercadeo, marketing y educación, es la proporción. Dicho parámetro, tiene la particularidad que es continuo, pero restringido en el intervalo (0,1) como la distribución uniforme o distribución Beta [Tellez Piñerez et al., 2014].

Hoy en día existen varias investigaciones acompañadas de análisis estadísticos en varias áreas del conocimiento. La razón por la cual se usan este tipo de análisis es porque su propósito es extraer la información más relevante para deducir características de la población. Como ya se mencionó, estas poblaciones surgen de áreas variadas como la ciencia, actuaría, medicina, entre otras. Por ejemplo, para el caso de la medicina es de interés saber el porcentaje de pacientes que sobreviven a una cirugía de corazón abierto, para el caso de actuaría es importante saber el porcentaje de asegurados que fallecen en un periodo determinado y para el caso de la ciencia por ejemplo biológica, es clave saber el porcentaje de bacterias que se dispersan en ciertos lugares en menor tiempo.

Por otra parte, Jackknife es una técnica clásica basada en corregir el sesgo de estimación, siendo esta una aproximación lineal del Bootstrapping. Además, la teoría respalda que la metodología Jackknife es mejor en cuanto a eficiencia comparada con otros métodos de remuestreo.

En la literatura, existen diversas técnicas para estimar este parámetro, sin embargo, ninguno utiliza la técnica Jackknife integrada con la metodología bayesiana. Dada la importancia que tiene la proporción en todo tipo de investigación se requiere ir mejorando cada día más los estimadores de dicho parámetro para generar información más precisa y confiable. A raíz de lo descrito anteriormente, surge la pregunta ¿Es posible estimar una proporción mediante muestreo probabilístico utilizando la metodología Jackknife en el ámbito bayesiano? La anterior pregunta es el fundamento de este trabajo de investigación.

2. Objetivos

2.1. Objetivo general

Construir un estimador para la proporción en muestreo probabilístico utilizando la metodología *Jacknife* y la teoría bayesiana.

2.2. Objetivos específicos

- Incorporar la técnica Jacknife y la teoría bayesiana en el muestreo probabilístico para estimar proporciones.
- Comparar vía simulación en términos de sesgo y varianza el estimador propuesto contra el estimador Bootstrap Bayesiano.
- Realizar una aplicación del estimador propuesto con datos reales.

3. Antecedentes

En la literatura especializada, hay varias formas de estimar la proporción, unas más sofisticadas que otras. Särndal [2003], muestra la forma clásica de estimar una proporción abordado desde el muestreo probabilístico. Por otro lado, uno de los primeros acercamientos a la estimación de parámetros utilizando métodos de remuestreo, fue presentada por Guzman [1978], quienes describen el método de estimación Jackknife, así como sus propiedades y algunas aplicaciones en la inferencia estadística. Uno de los resultados más relevantes de este trabajo, es que esta metodología de estimación permite reducir tanto el sesgo que introduce el remuestreo, como la variación implícita en este. Lo que implica que se obtienen estimadores con poco sesgo y altas eficiencias.

Por otro lado, Gallego [1997] expone mediante los estimadores de razón la utilización de información auxiliar en la estimación de parámetros definidos sobre poblaciones finitas. En este tipo de estimaciones, surge un problema que es el sesgo, esto lleva a la implementación de diseños muestrales tales como πPT y PPT o modificar las expresiones de los estimadores con el fin de tener sesgo reducido y estrategias insesgadas, manteniendo simplicidad del error de muestreo. En el año siguiente Martinez [1998] desarrolla una secuencia finita de estimadores no paramétricos para el número de clústers utilizando el método Jackknife. Adicional a lo anterior, los autores en su trabajo no usan el muestreo y proponen un estimador sesgado, el cual es corregido y ajustado mediante Jackknife generalizado para el número de clúster en una población.

Desde otra perspectiva, Schiaffino [2002], describen y comparan el cálculo de la razón de proporciones en diferentes técnicas como regresión logística, regresión de Breslow-cox, modelo log-binomial y formula de conversión, todo lo anterior sin tener en cuenta el muestreo.

Por su parte, López Gómez [2006], comparó la riqueza de árboles muestreada con la estimada utilizando estimadores no paramétricos como Jackknife, Bootstrap, entre otras. La precisión de los estimadores se realizó mediante la estimación de sesgo y exactitud por medio de la comparación de la riqueza estimada con la riqueza verdadera, como resultado obtuvieron que Jackknife fue el mejor en cuanto a precisión, así que este estimador puede ser el

más adecuado para estimar la riqueza en sistemas manejados como cafetales. Adicional a este año Evwcombe Robert G [2006], presentan métodos para calcular los intervalos de confianza para proporciones y diferencias entre proporciones que superan los procedimientos tradicionales para estos cálculos, estos métodos son basados en el método score.

Un año después, Pacheco [2007] propone un estimador a partir del estimador Jackknife de varianza por Berger y Skinner en el año 2005, este estimador propuesto permite la estimación de la varianza, asumiendo que los parámetros por estimar y sus estimadores se pueden escribir como funciones de medias poblacionales y muestrales, todo esto empleando la metodología Jackknife para muestreo con probabilidades desiguales. En este mismo año Badii Castillo [2007], plantean, como a partir de Jackknife y Bootstrap se puede estimar la precisión de muchos índices en datos muestrales. Estas técnicas fueron evaluadas en un coeficiente muy conocido como es el de Gini de jerarquía en tamaño y en el índice de Jaccard de similitud de comunidades, así, terminan de utilizar Jackknife para calcular el sesgo y el error estándar para una estadística.

Pacheco y Brango [2010], proponen la construcción de intervalos de confianza en muestreo con probabilidades desiguales, empleando un estimador basado en la metodología Jackknife de varianza para muestreo con probabilidades desiguales que funciona mejor que el estimador Jackknife clásico. Todo esto demostrado vía simulación y llegando a la conclusión de que la metodología clásica del Jackknife no es adecuada para entregar intervalos de confianza excesivamente amplios.

Una aproximación a la estimación de la proporción utilizando métodos de remuestreo, fue realizada por Tellez Piñerez et al. [2014], ellos proponen realizar inferencias sobre una proporción en una población finita a partir de muestras con probabilidades desiguales, pero utilizando la metodología Bootstrap. En este mismo año Ríos [2014] propone siete estrategias mediante los cuales modelan la incertidumbre geológica en un yacimiento minero, estas estrategias de modelos se basan en estimar las proporciones de unidades geológicas con estadísticas geográficamente ponderadas y luego aplican simulación plurigaussiana, finalmente para validar los modelos utilizan la metodología Jackknife, todo lo anterior sin hacer uso del muestreo.

Otro trabajo relacionado con las metodologías Bootstrap y Jackknife, fue desarrollado por Ramírez Montoya [2015], quienes encuentran una estrategia para estimar la función de confiabilidad o, también llamada función de supervivencia a través del muestreo y las técnicas antes mencionadas.

Como se pudo observar anteriormente, la metodología Jackknife es ampliamente utilizada, tanto en el muestreo probabilístico como en otras áreas del conocimiento por sus propiedades y fácil utilización. Adicional a lo anterior, en la revisión literaria no se encontró ningún trabajo que integre la estimación de la proporción mediante la técnica Jackknife en muestreo probabilístico con probabilidades desiguales y la teoría bayesiana. Es por esto por lo que el objetivo principal de este trabajo es proponer un estimador para la proporción en muestreo probabilístico utilizando la metodología Jackknife en el ámbito bayesiano.

4. Metodología

Para estimar la proporción utilizando la metodología propuesta en este trabajo, se consideraron tres pasos fundamentales, los cuales se describen a continuación. El primero, consiste en determinar una estrategia general de muestreo con la cual se desee estimar, de manera clásica la proporción y su varianza. En segunda instancia, se deben escoger las distribuciones apriori que nos permitan tener buenas estimaciones de manera bayesiana, por último, integrar las estrategias de muestreo y las metodologías bayesianas que permitan obtener distribuciones posteriores del parámetro de interés, que para nuestro caso es la proporción.

Para el primer aspecto correspondiente a la elección de un diseño muestral, se probará uno de los diseños muestrales más comunes y eficientes como lo es el πPT.

Para la selección de una distribución apriori se examinarán distribuciones previas uniforme y de ser posible aprioris de Jefreys, también se tendrán en cuenta distribuciones conjugadas de la forma $Beta(\alpha, \beta)$, dada la naturaleza del parámetro, es decir, entre cero (0) y uno (1).

Finalmente, se realizará una aplicación con la metodología propuesta *Jackk-*

nife bayesiana en datos reales de tal forma que se ejemplifique el uso de esta a nivel práctico y con ella se evidencie la utilidad de esta a partir de los resultados dados por las medidas de calidad.

5. Marco teórico

En lo siguiente se dará a conocer las definiciones más importantes para el desarrollo de este proyecto junto con las ecuaciones respectivas, todo esto de acuerdo a Särndal [2003] y Gutiérrez [2009].

5.1. Muestreo probabilístico

Según Särndal [2003] se denomina muestreo probabilístico al proceso de selección de una muestra que satisface las siguientes condiciones:

1. Se puede definir el conjunto de muestras, $S = \{s_1, ..., s_M\}$, posibles que se derivan del proceso de muestreo.

2. A cada muestra posible s, le corresponde una probabilidad de selección $p(s)$ conocida.

3. El proceso de selección garantiza que todo elemento de la población tenga una probabilidad de selección mayor a cero.

5.1.1. Marco muestral

El marco muestral es todo material o dispositivo usado para obtener acceso a los elementos de la población de interés. Con este se delimita, identifica y permite tener un acercamiento a los elementos de la población objetivo. En este también se encuentra información auxiliar correspondiente a cada elemento de la población. Por ejemplo, en una encuesta las unidades del marco son los sujetos a los cuales se le aplica la selección por muestreo probabilístico [Särndal, 2003].

5.1.2. Diseño muestral

Un diseño muestral es una función $p(\cdot)$ tal que, $p(s)$ denota la probabilidad de selección de la muestra s. De esta manera se llamará diseño de muestreo al conjunto de probabilidades de selección de todas las muestras posibles. Para

un diseño muestral dado $p(\cdot)$, podemos considerar a cada muestra s como el resultado de una variable aleatoria S, con distribución de probabilidad especificada por $p(\cdot)$. Así, si S es el conjunto de todas las muestras posibles s, entonces,

$$P_r(S = s) = P(s) \tag{1}$$

Para cada $s \in S$, y debido a que $p(s)$ es una distribución de probabilidad en S, se tiene que, $p(s) \geq 0, \forall\ s \in S$.

5.1.3. Probabilidades de inclusión

La inclusión de un elemento k en una muestra es un evento aleatorio indicado por la variable aleatoria I_k, definida por:

$$I_k = \begin{cases} 1 & si, k \in S \\ 0 & si, k \notin S \end{cases} \tag{2}$$

La probabilidad de que un elemento k sea incluido en una muestra, denotado por π_k, es obtenido del diseño muestral dado como:

$$\pi_k = P_r(k \in S) = Pr(I_k = 1) = \sum_{k \in S} P(s) \tag{3}$$

La probabilidad de que un par de elementos k y l sean incluidos en una muestra s será notado por $\pi_{k,l}$ y se obtiene como:

$$\pi = P_r(k, l \in S) = Pr(I_k I_l = 1) = \sum_{k,l \in S} P(s) \tag{4}$$

En particular, $\pi_{k,l} = \pi_{l,k}$ para todo k y l, y $\pi_{k,l} = \pi_k$ si $k = l$.

Las probabilidades π_k son llamadas probabilidades de inclusión de primer orden, y las probabilidades $\pi_{k,l}$ son llamadas probabilidades de inclusión de segundo orden.

5.1.4. El estimador de Horvitz-Tompson para el total

Considere una población U de N elementos. Sea s una muestra de tamaño fijo n sacada sin reemplazo de U de acuerdo a un diseño muestral dado $p(\cdot)$,

el interés es estimar el total poblacional de la variable de estudio y.

$$t = \sum_{k \in U} y_k \tag{5}$$

En función de los valores π_k. Para tal fin se tiene el estimador π de Horvitz-Thompson para el total:

$$\hat{t}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \check{y}_k \tag{6}$$

con $\check{y} = \frac{y_k}{\pi_k}$.

La varianza del estimador Horvitz-Thompson (π) es:

$$V(\hat{t}_\pi) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \check{y}_k \check{y}_l \tag{7}$$

con $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ y $\pi_{kl} > 0$ para todo k y $l \in U$.

Un estimador insesgado para $V(\hat{t}_\pi)$ es:

$$\hat{V}(\hat{t}_\pi) = \sum_{k \in U} \sum_{l \in U} \check{\Delta}_{kl} \check{y}_k \check{y}_l I_k I_l \tag{8}$$

donde $\check{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}}$, con $\pi_{kl} > 0$ para todo k y $l \in U$.

5.1.5. Intervalos de confianza

Un intervalo de confianza es un intervalo aleatorio de la forma:

$$IC(s) = [t_{inf}(s), t_{sup}(s)] \tag{9}$$

donde los límites del intervalo $t_{inf}(s)$ y $t_{sup}(s)$ son dos estadísticas tales que $t_{inf}(s) < t_{sup}(s)$, para todo s.

Para el total poblacional t y $\hat{t}$ un estimador puntual de este, un intervalo de confianza para t de aproximadamente el $100(1-\alpha)\,\%$ es:

$$\hat{t} \pm z_{1-\alpha} \sqrt{\hat{V}(\hat{t}_\pi)} \tag{10}$$

5.1.6. Estimación de parámetros más complejos

Cuando un parámetro puede ser expresado como una función de q totales poblacionales, $t_1, t_2, ..., t_q$, tales que:

$$\theta = f(t_1, t_2, ..., t_q) \tag{11}$$

donde, $t_j = \sum_{k \in U} \check{y}_{kl}$, con $j = 1, 2, ..., q$. Cuya estimación es:

$$\hat{\theta} = f(\hat{t}_{1\pi}, \hat{t}_{2\pi}, ..., \hat{t}_{q\pi}) \tag{12}$$

donde, $\hat{t}_{j\pi} = \sum_{k \in S} \check{y}$, con $j = 1, 2, ..., q$.

Luego, podemos encontrar una aproximación a este estimador a través de la técnica de linealización de primer orden de Taylor de la forma:

$$\hat{\theta} = \theta_0 = \theta + \sum_{j=1}^{q} a_j (\hat{t}_{j\pi} - t_j) \tag{13}$$

donde,

$$a_j = \frac{\partial f(t_1, t_2, ..., t_q)}{\partial t_j} \tag{14}$$

Así, es posible aproximar la varianza de $\hat{\theta}$ con la varianza de la aproximación $\hat{\theta}_0$, como sigue:

$$AV(\hat{\theta}) = AV(\hat{\theta}_0) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \check{u}_k \check{u}_l \tag{15}$$

con, $u_k = \sum_{j=1}^{k} a_j \check{y}_{kj}$. Luego, para encontrar el estimador de $AV(\hat{\theta})$, hacemos:

$$\hat{u}_k = \sum_{j=1}^{k} \hat{a}_j \check{y}_{kj} \tag{16}$$

por tanto, $\hat{V}(\hat{\theta})$ es:

$$\hat{V}(\hat{\theta}) = \sum_{k \in S} \sum_{l \in S} \check{\Delta}_{kl} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l} \tag{17}$$

5.1.7. Estimación de una razón

Uno de los parámetros de mayor interés en la práctica es la razón entre dos totales poblacionales:

$$R = \frac{t_y}{t_z} \tag{18}$$

Con estimador,

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z} \tag{19}$$

de lo anterior tenemos que $f(t_1, t_2) = t_1/t_2$, por tanto,

$$a_1 = \frac{1}{t_2} = \frac{1}{t_z} \tag{20}$$

$$a_2 = \frac{t_1}{t_2^2} = \frac{t_y}{t_z^2} \tag{21}$$

así, con $y_{1k} = y_k, y_{2k} = z_k$

$$u_k = \frac{y_k - Rz_k}{t_z} \tag{22}$$

$$\hat{u}_k = \frac{y_k - \hat{R}z_k}{\hat{t}_z} \tag{23}$$

se puede aproximar la varianza de R con:

$$AV(\hat{R}) = \frac{1}{t_z^2} \sum_{k \in U} \sum_{l \in U} \check{\Delta}_{kl} \frac{y_k - Rz_k}{\pi_k} . \frac{y_l - Rz_l}{\pi_l} \tag{24}$$

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_z^2} \sum_{k \in U} \sum_{l \in U} \check{\Delta}_{kl} \frac{y_k - \hat{R}z_k}{\pi_k} . \frac{y_l - \hat{R}z_l}{\pi_l} \tag{25}$$

5.1.8. Estimación de una proporción

La estimación es un caso particular de la razón, haciendo $y_k = 1, \forall k \in U$.

$$y_k = \begin{cases} 1 & si, k \in U_d \\ 0 & e.o.c \end{cases} \tag{26}$$

en este caso, $t_z = N$ y $t_y = N_d$, entonces,

$$p = \frac{N_d}{N} \tag{27}$$

con estimador,

$$\hat{p} = \frac{\hat{N}_d}{\hat{N}}. \tag{28}$$

Se puede aproximar la varianza de la proporción y su varianza aproximada por:

$$AV(\hat{p}) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \breve{\Delta}_{kl} \frac{1-p}{\pi_k} . \frac{1-p}{\pi_l} \tag{29}$$

$$\hat{V}(\hat{R}) = \frac{1}{\hat{N}^2} \sum_{k \in U} \sum_{l \in U} \breve{\Delta}_{kl} \frac{1-\hat{p}}{\pi_k} . \frac{1-\hat{p}}{\pi_l} \tag{30}$$

5.2. Jackknife

Los métodos de remuestreo son técnicas basadas en tomar diferentes muestras de una muestra dada y hacer una estimación de un parámetro en cada una de ellas, luego relacionando todas las estimaciones se obtiene un nuevo estimador [Angulo, 2009].

5.2.1. Método Jackknife

El método de Jackknife fue presentado por Quenouille [1949] y es una de las primeras técnicas para obtener estimadores estadísticos fiables.

Supongamos que tenemos una muestra aleatoria $y_1, y_2, \cdots, y_n$ y un estimador t del parámetro θ. Sea t_i el estimador evaluado en los $n-1$ elementos que quedan después de separar el $k - ésimo$ elemento de la muestra $t_k = t(y_1, y_2, \cdots, y_{k-1}, y_{k+1}, \cdots, y_n)$, es decir, esta técnica se centra en ir dejando de lado una observación.

Ahora se construye el estimador la expresión:

$$S_k = nt - (n-1)t_k \tag{31}$$

Donde $k = 1, 2, \cdots, n$ la cual recibe el nombre de pseudovalor. Quenouille define el estimador de Jackknife t_j simple de θ asociado al estimador t y a la muestra $y_1, y_2, \cdots, y_n$ como el promedio de los pseudovalores.

$$t_j = \frac{1}{n}\sum_{k=1}^{n} s_k = nt - \frac{n-1}{n}\sum_{k=1}^{n} t_k \tag{32}$$

Por otra parte, una de las aplicaciones más importantes de la técnica de jackknife es la reducción del sesgo.

5.3. Estadística bayesiana

La estadística bayesiana es un tipo de inferencia estadística en la que las evidencias u observaciones se emplean para actualizar o inferir la probabilidad de que una hipótesis pueda ser cierta. El nombre "bayesiana"proviene de uso frecuente que se hace del Teorema de Bayes durante el proceso de inferencia.

La inferencia bayesiana utiliza aspectos del método científico, que implica recolectar evidencia que se considera consistente o inconsistente con una hipótesis dada.

A medida que la evidencia se acumula, el grado de creencia en una hipótesis se va modificando, con evidencia suficiente, a menudo podrá hacerse muy alto o muy bajo. Así, los que sostienen la inferencia bayesiana dicen que puede ser utilizada para discriminar entre hipótesis en conflicto: las hipótesis con un grado de creencia muy alto deben ser aceptadas como verdaderas y las que tienen un grado de creencia muy bajo deben ser rechazadas como falsas.

Sin embargo, los detractores dicen que este método de inferencia puede estar afectado por un prejuicio debido a las creencias iniciales que se deben sostener antes de comenzar a recolectar cualquier evidencia.

5.3.1. Teorema de bayes

Sean $B_1, B_2, ..., B_k$, eventos mutuamente excluyentes y exhaustivos. Para cualquier evento nuevo A, tenemos:

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}. \tag{33}$$

5.3.1.1. Teorema de bayes para variables aleatorias

Sean X y Y variables aleatorias con funcion de densidad de probabilidad $f(x|y)$ y $g(y)$, de donde se tiene que:

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)d\theta} \tag{34}$$

5.3.1.2. Distribución apriori

En estadística Bayesiana, una distribución de probabilidad apriori de una una cantidad p desconocida, es la distribución de probabilidad que expresa alguna incertidumbre acerca de p antes de tomar en cuenta los "datos".

5.3.1.3. Distribución posteriori

Sean X y θ variables aleatorias con función de densidad de probabilidad $f(x|\theta)$ y $\xi(\theta)$.

$$\xi(\theta|x) = \frac{f(x|\theta)\xi(\theta)}{\int f(x|\theta)\xi(\theta)d\theta} \tag{35}$$

Dentro del marco bayesiano tenemos que:

- X : Datos (escalar o vector o matriz).
- θ : Parámetro desconocido (escalar o vector o matriz).
- $f(x_1 \cdots x_n|\theta)$: Verosimilitud de los datos dado el parámetro (desconocido) θ.
- $\xi(\theta)$: Distribución a-priori de θ.

Así se tiene que:

$$\xi(\theta|x_1 \cdots x_n) = \frac{f(x_1 \cdots x_n|\theta)\xi(\theta)}{\int f(x_1 \cdots x_n|\theta)\xi(\theta)d\theta} \tag{36}$$

Es llamado distribución posteriori.

5.3.2. Inferencia Bayesiana

Cuando se utilizan evidencias y observaciones para establecer que una suposición sea cierta, es lo que se denomina como Inferencia Bayesiana, sumado a lo anterior, esta observa la evidencia y calcula un valor estimado según el grado de creencia planteado en la hipótesis [Páez, 2011].

5.3.2.1. Estimación puntual

Dada una distribución sobre un parámetro particular, digamos θ, requerimos seleccionar un mecanismo para escoger un 'buen' estimador $\hat{\theta}$. Supongamos que θ_0 es el verdadero parámetro, desconocido. Sea d nuestra adivinanza de este valor. Debemos de alguna forma medir el error que cometemos al adivinar a θ_0 mediante d.

Esto puede ser medido por $(d\theta_0)^2$ o por $|d\theta_0|$ o mediante alguna otra función. Un problema estadístico puede resumirse como (S, Ω, D, L), donde:

S : Es el espacio muestral de un experimento relevante que tiene asociada una variable aleatoria X cuya distribución de probabilidad está parametrizada por un elemento de Ω.

Ω : Espacio parametral (en un sentido amplio).

D : Un espacio de decisiones.

L : Una función de pérdida.

Una vez un problema estadístico ha sido especificado, el problema de inferencia estadística es seleccionar un procedimiento (estadístico), a veces llamado una función de decisión, que nos describe la forma de tomar una decisión una vez un resultado muestral ha sido obtenido.

- Una función de decisión o procedimiento estadístico es una función o estadístico d que mapea de S a D.
- Sea D un espacio arbitrario de decisiones. Una función no negativa L que mapea de ΩD a R es llamada una función de pérdida.
- El valor esperado de $L(\theta, d(X))$ cuando θ es el verdadero valor es llamada la función de riesgo.

$$R(\theta, d) = E_\theta(L(\theta, d(X))) = \int L(\theta, d(x))dP_\theta(x) \tag{37}$$

5.3.2.2. Función de pérdida cuadrática

$$L(\theta, d) = (d - \theta)^2 \tag{38}$$

Miremos el riesgo para esta función de pérdida. Sea:

$$b = E_{\xi(\theta|x)}(\theta) = \int \theta\xi(\theta|x)d\theta \tag{39}$$

el promedio de la distribución posteriori. Entonces,

$$E(L(\theta, d)) \geq \int (b - \theta)^2\xi(\theta|x)d\theta \tag{40}$$

para cualquier valor de d. La desigualdad anterior se convierte en igualdad cuando $d = b$. El estimador bayesiano bajo una función de pérdida cuadrática es la media de la distribución posterior.

5.3.2.3. Función de pérdida error absoluto

Sea $L(d, \theta) = |d\theta|$ la función de perdida asociada a el parametro θ, y d nuestra adivinanza a este valor. El riesgo es minimizado tomando d como la mediana de la distribución posterior, digamos d^*. O sea, la mediana es el estimador bayesiano cuando la función de pérdida es el valor absoluto. Para mostrar esto supongamos otra decisión tal que $d > d^*$. Entonces,

$$|\theta - d| - |\theta - d^*| = \begin{cases} d^* - d & si \quad \theta \geq d \\ d + d^* - 2\theta & si \quad d^* \leq \theta \leq d \\ d - d^* & si \quad \theta \leq d^* \end{cases} \tag{41}$$

Ya que $d + d^* 2\theta > d^* d$ cuando $d^* \leq \theta \leq d$ entonces el siguiente resultado se consigue:

$$E(|\theta - d| - |\theta - d^*|) = (d - d^*)[P(\theta \leq d^*) - P(\theta > d^*)] \tag{42}$$

Esta última desigualdad sigue del hecho que d^* es la mediana de la distribución de θ. La primera desigualdad en este conjunto de ecuaciones será una igualdad si, y solo si, $d^* \leq \theta \leq d = 0$. La desigualdad final será una igualdad si, y solo sí, $P(\theta \leq d^*) = P(\theta > d^*) = \frac{1}{2}$.

Estas condiciones implican que d es también una mediana. Por lo tanto, $E(|\theta d|) \geq E(|\theta d^*|)$ y la igualdad se cumple si, y solo si, d es también mediana.

6. Inferencia Jackknife bayesiana para una proporción

Supongamos $U = u_1, u_2, \cdots, u_k, \cdots, u_N$, una población finita de tamaño N, en donde cada unidad $u_i (i = 1, 2, \cdots, N)$ tiene asociada una variable dicótomica y_i, esta toma el valor de 1 cuando el dato cumple con 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. Varios de los procedimientos son similares a los realizados en el trabajo de Tellez Piñerez et al. [2014].

En la muestra, la variable de interés y es observada para todos los elementos seleccionados, ahora el interés esta en estimar la distribución de probabilidad posterior para el parámetro ρ_y haciendo uso de los valores de la muestra y de las probabilidades de inclusión dadas por el diseño muestral.

La metodología Jackknife bayesiana considera que el parámetro ρ_y está en función de la distribución acumulada, esta proviene la muestra aleatoria s, la cual ha sido seleccionada con un diseño muestral $p(\cdot)$ y con la que se ha estimado ρ_y, haciendo uso del estimador de Horvitz-Thompson para el caso Jackknife definido como:

$$\hat{\rho}_{y\pi} = \frac{1}{\hat{N}} \sum_{i \in s^*}^{n} \frac{y_i}{\pi_i} \tag{43}$$

con $\hat{N} = \sum_{i \in s} \frac{1}{\pi_i}$ 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:

$$\xi(\rho_y|y) \propto \xi(y|\rho_y)\xi(\rho_y) \tag{44}$$

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

Ahora, hay que pensar en una distribución a priori para ρ_y y en un supuesto para la distribución de 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 tiene que tener en cuenta que en la teoría de muestreo no se hacen supuestos distribucionales para y condicionado al parámetro, por lo que se asume cualquier distribución, por lo tanto la metodología Jackknife bayesiana juega un papel fundamental en la metodología propuesta, la cual consiste en realizar una obtención de $\xi(y|\rho_y)$ y $\xi(\rho_y|y)$ de forma empírica.

6.1. Distribución posterior de ρ con a priori informativa

Si es de interés el parámetro ρ_y y la información a priori sobre ρ_y dada como $\xi(\rho_y)$ y si $y_1, y_2, \cdots, y_n$ 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 como:

$$\xi(\rho_y|y) \propto \xi(\rho_y)\hat{L}(y_1, \cdots, y_n|\rho_y) \tag{45}$$

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 $\frac{1}{\pi_i}$, 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 usado para seleccionar la muestra original de s de U. A partir de las muestras restantes se calcula el ρ estimador $\rho^*_{y\pi b}$:

$$\hat{\rho}^*_{y\pi b} = \frac{1}{\hat{N}^*} \sum_{i \in s^*} \frac{y^*_{ib}}{\pi^*_{ib}} \tag{46}$$

   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 $\rho^*_{y\pi 1}, \cdots, \rho^*_{y\pi B}$ se calcula el estimador de densidad kernel definido como:

$$f_B(u) = \frac{1}{B h_B} \sum_{b=1}^{B} K\left(\frac{u - (\hat{\rho}^*_{y\pi b} - \hat{\rho}_{y\pi})}{h_B}\right) \tag{47}$$

   Donde la función K es llamada comúnmente como el kernell y en general, es una función de densidad continua, unimodal y simétrica alrededor de 0 Hollander [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

$$\hat{L}_B(\hat{\rho}_{y\pi}|\rho_y) = \frac{1}{B h_B} \sum_{b=1}^{B} K\left(\frac{2\hat{\rho}_{y\pi} - \rho - \hat{\rho}^*_{y\pi b}}{h_B}\right) \tag{48}$$

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:

$$\xi(\rho_y|y) = c(y)x\xi(\rho_y)x\hat{L}(y_1, \cdots, y_n|\rho_y) \tag{49}$$

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

$$c(y) = \frac{1}{\int \xi(\rho_y)x\hat{L}(y_1, \cdots, y_n|\rho_y)d\rho_y} \tag{50}$$

6.2. 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 $\rho_y^1, \rho_y^2, \cdots, \rho_y^m$ 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 $\rho_y^1, \rho_y^2, \cdots, \rho_y^m$ 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, función de pérdida en error absoluto y la función escalonada [Box, 2011].

6.2.1. Función de pérdida cuadrática para la proporción

Se considera la función $L(\rho_y\rho_c) = (\rho_c - \rho_y)^2$ que se denotará como función de pérdida cuadrática asociada al parámetro ρ_y, y sea ρ_c la estimación considerada para ρ_y. Sean $\rho_y^1, \rho_y^2, \cdots, \rho_y^m$ 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:

$$\rho_c = E(\rho_y|y) = \int_{-\infty}^{+\infty} \rho_y \xi(\rho_y|y) d\rho_y \tag{51}$$

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

$$\rho_c = \bar{\rho}_y = \frac{\sum_{j=1}^{m} \rho_y^j}{m} \tag{52}$$

y un error estándar estimado es:

$$se_{\rho_c} = \sqrt{\frac{\sum_{j=1}^{m} (\rho_y^j - \rho_c)^2}{(m-1)m}} \tag{53}$$

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

7. Estudio de simulación

Los escenarios de simulación se dispusieron similares a los realizado en el trabajo de Tellez Piñerez 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.

7.1. Diseño de la simulación

El estudio de simulación pretende evaluar el comportamiento de la metodología propuesta y compararla con el procedimiento Bootstrap Bayesiano.

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, 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:

$$sesgo = \frac{1}{1000}\sum_{j=1}^{1000}(\hat{\rho}_j - \rho) \tag{54}$$

$$S.R = \frac{sesgo}{\rho} * 100\,\% \tag{55}$$

$$RECM = \sqrt{\frac{1}{1000}\sum_{j=1}^{1000}(\hat{\rho}_j - \rho)^2} \tag{56}$$

$$M.E = RECM * Z_{1-\frac{\alpha}{2}} * 100\,\% \tag{57}$$

Cabe resaltar que dado que el parámetro es un valor entre cero y uno, utilizar el CVE en este caso no sería útil, dada la formula de esta medida al momento de reemplazar el estimador del parámetro esto 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 sugiere el error estándar o el margen de error, los cuales son mostrados en la tabla de resultados de la simulación. Un ejemplo para explicar lo anterior es, supongamos que se quiere estimar la proporción de fumadores de una oficina con horario extendido, después de calcular la muestra se tiene que dicha proporción es de 0.11 con un coeficiente de variación del 25 %, un error estándar de 2.8 % y un margen de error aproximadamente de 5.5 %. Se puede ver que si se saca una conclusión con respecto al CVE esta cifra no sería publicable siguiendo con la regla del 15 %, mientras que con el margen de error esta cifra no se debería ignorar, por lo tanto, sería publicado.

Siguiendo con la simulación, se planteo 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 anteriormente para efectos de comparación se utilizaron las mismas distribuciones *a priori* que en el trabajo realizado por Tellez Piñerez et al. [2014][Pg.38], por lo tanto, allí se pueden encontrar los pasos con los cuales se llegaron a estos valores de β.

7.2. Resultado de la simulación

En esta sección se muestran las tablas que contienen los resultados del proceso de simulación antes descrito, con el fin de comparar la metodología Jackknife y la metodología Bootstrap, para la estimación de la proporción.

En los cuadros 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*.

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

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

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

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

Los valores resaltados en color azul 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 *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.

8. Ejemplo de la metodología

Para efectos de aplicación y haciendo uso de la metodología propuesta se consultó la Encuesta de Cultura Política del año 2019 realizada por el DANE que a su vez dicha encuesta se encuentra en la página oficial de este departamento.

La encuesta de Cultura Política indaga sobre la percepción que tienen los

ciudadanos 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".*

Esta encuesta como es de esperar tiene preguntas muy interesantes que llevan al investigador a querer tratar estos datos para dar por ejemplo inferencias respecto a la población, es por esto que en esta sección se escogieron varias de las preguntas de la encuesta para aplicar el estimador Jackknife Bayesiano y posterirmente concluir sobre la población respecto a las preguntas escogidas con un contexto real.

Las preguntas que se escogieron son:
Respecto a la población en general se escogió la pregunta:

- ¿Sabe leer y escribir?

Enfocando la aplicación al tema político se tienen las preguntas:

- ¿Votó usted en las elecciones presidenciales de 2018?
- ¿Votaría alguna vez por una mujer?
- ¿Usted cree que en Colombia existe la libertad de expresar y difundir su pensamiento?
- ¿Usted se informa sobre la actualidad política del país?

La muestra trabajada contiene 18393 datos y 7 variables las cuales corresponden a las preguntas, el factor de expansión de cada uno de los datos y las probabilidades de inclusión. El interés en este trabajo esta en estimar mediante el método clásico, Bootstrap Bayesiano y Jackknife Bayesiano, la proporción de personas que cumplan con la característica dada según la pregunta, por ejemplo es de interés saber la proporción de personas que votaría alguna vez por una mujer, o proporción de personas que se informan sobre la actualidad política del país, estos resultados son comparados en términos de sesgo y M.E %.

El proceso de extracción de las muestras artificiales se hizo con un diseño de muestreo probabilístico teniendo en cuenta las probabilidades de inclusión calculadas como en Särndal [2003], considerando además un tamaño de muestra similar al de la base inicial, la cual equivale a cerca del 0.05 % de la población.

A continuación se muestra la manera en la que se realizaron las estimaciones para cada pregunta y además algunas gráficas que permiten entender los resultados de las estimaciones.

El valor estimado para la pregunta ¿Sabe leer y escribir? (P.1) por medio de la metodología clásica es 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 obtiene una muestra artificial de tamaño similar a la muestra inicial, teniendo para cada una de ellas los estimadores respectivos, según la metodología.

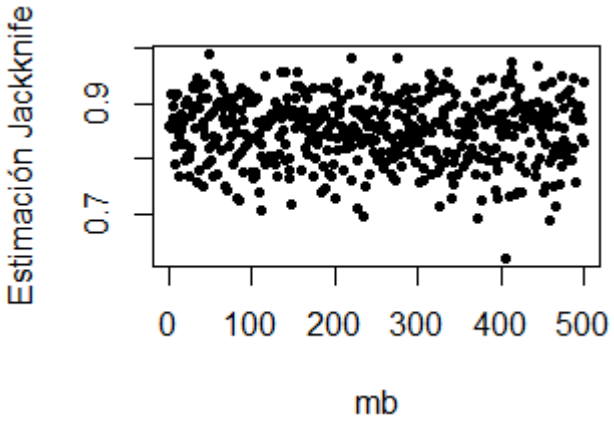

Figura 1: Estimación de la P.1 por la metodología Jackknife

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

$$\hat{L}_J(\rho|\hat{\rho}) = \frac{1}{500(0{,}07173)} \sum_{mb=1}^{500} K\left(\frac{2(0{,}934) - \rho - \hat{\rho}^*_{mb}}{0{,}07173}\right) \tag{58}$$

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 la pregunta 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:

$$\xi(\rho) = beta(90, 5) \propto \rho^{89}(1-\rho)^4 \tag{59}$$

Utilizando un Kernerl Gausiano la distribución posterior de ρ es el producto de la verosimilitud y la distribución *a priori*, quedando como:

$$\xi(\rho|y) \propto \hat{L}_J(\rho|\hat{\rho})\rho^{89}(1-\rho)^4 \tag{60}$$

De forma gráfica se pueden ver estas distribuciones como:

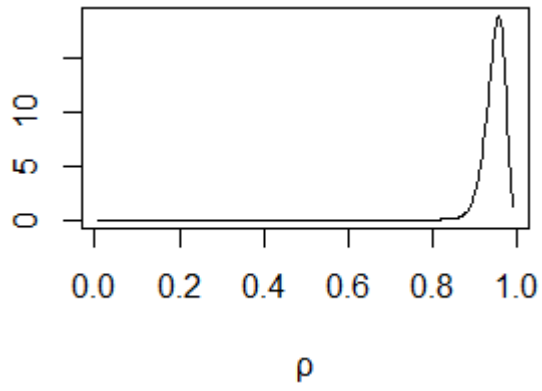

Figura 2: Distribución a priori para la pregunta 1 por medio de Jackknife Bayesiano

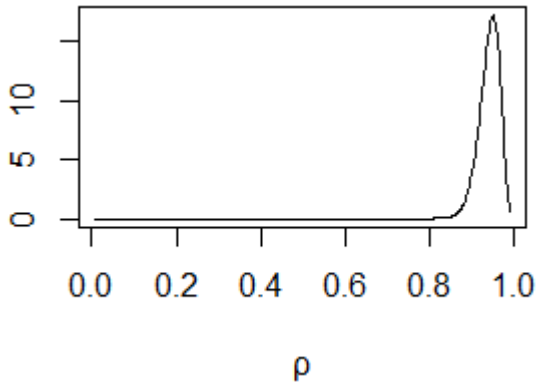

Figura 3: Distribución posteriori para la pregunta 1 por medio de Jackknife Bayesiano

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 estimo que el 95 % de la población sabe leer y escribir,

este 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 %.

Cuadro 3: Resultados de la aplicación

	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

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 agregar el toque bayesiano en estas estimaciones.

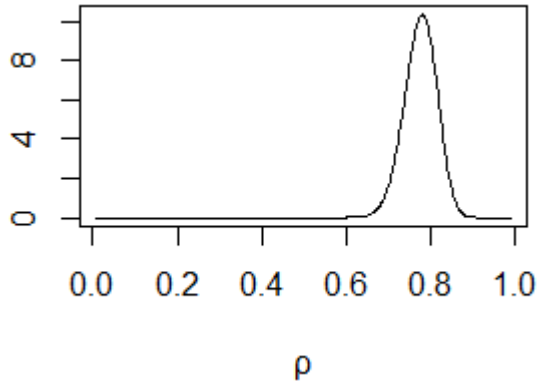

Figura 4: Distribución a priori para la pregunta 2 por medio de Jackknife Bayesiano

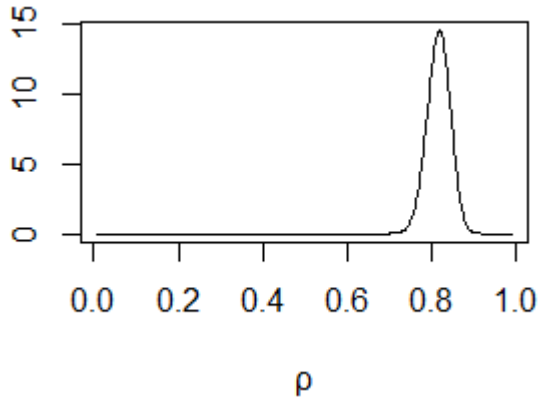

Figura 5: Distribución a posteriori para la pregunta 2 por medio de Jackknife Bayesiano

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 J.B 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.

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.

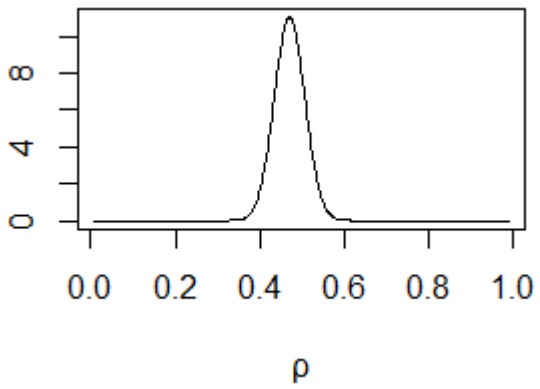

Figura 6: Distribución a priori para la pregunta 4 por medio de Jackknife Bayesiano

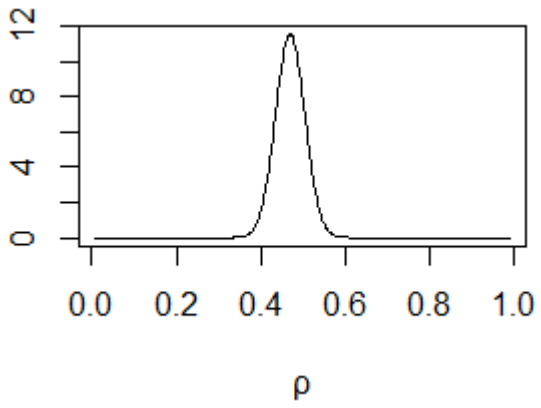

Figura 7: Distribución posteriori para la pregunta 4 por medio de Jackknife Bayesiano

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 metodología propuesta es excelente en términos de sesgo (por ser cercana a cero) comparado con la estimación clásica o B.B, 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 %.

9. Conclusiones y recomendaciones

Recapitulando lo encontrado en las secciones anteriores para la estimación de la proporción, la metodología Jackknife bayesiana construida en este trabajo es mejor a la hora de estimar en términos de sesgo ya que estos son valores muy cercanos a cero, lo que quiere decir que este estimador es aproximadamente insesgado, además en su mayoría los intervalos de credibilidad tienen una amplitud baja con una cobertura por encima del 80 %, comprobado en la simulación.

Como se mencionó en la sección de simulación hay una clasificación establecida por el Departamento Administrativo Nacional de Estadística (DANE), con esta se puede evidenciar que las estimaciones calculadas con la metodología propuesta son precisas de acuerdo al M.E %, a pesar que J.B se somete a varias configuraciones como por ejemplo los parámetros de la distribución *apriori*, este sigue arrojando resultados muy buenos para la estimación de la proporción, adicional a lo anterior, estimar por medio de Jackknife Bayesiano no es complejo, pues solo se requiere tener información previa del parámetro.

Como resultado adicional en este trabajo y como se mencionó en la sección de simulación, los sesgos y los niveles de cobertura del estimador Jackknife son similares 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 Bootstrap.

Referencias

Angulo, E. M. y. F. (2009). Técnica de jackknife y estimadores en un modelo lineal. *Scientia et technica*, 1(41).

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

Box, George EP y Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons.

DANE, C. (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.

Evwcombe Robert G, M. S. C. (2006). Intervalos de confianza para las estimaciones de proporciones y las diferencias entre ellas. *Interdisciplinaria*, 23(2).

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

Gutiérrez, H. A. (2009). *Estrategias de muestreo diseño de encuestas y estimacion de parametros.* Universidad Santo Tomas, Bogota (Colombia).

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

Hollander, Myles Wolfe, D. A. y. C. E. (2013). *Nonparametric statistical methods*, volume 751. John Wiley & Sons.

López Gómez, Ana María, W. L. G. (2006). Evaluación de métodos no paramétricos para la estimación de riqueza de especies de plantas leñosas en cafetales. *Botanical Sciences*, (78).

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

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

Páez, Lesley Ofelia Mesa Lozano y Miller Rivera Davila, J. A. R. (2011). Descripción general de la inferencia bayesiana y sus aplicaciones en los procesos de gestión. *La simulación al Servicio de la Academia*.

Ramírez Montoya, Javier Osuna Vergara, I. R. M. J. y. G. G. S. (2015). Remuestreo Bootstrap y Jackknife en confiabilidad: Caso Exponencial y Weibull. *REVISTA FACULTAD DE INGENIERÍA*, 25(41).

Ríos, T. F. (2014). Kriging y simulación secuencial de indicadores con proporciones localmente variables.

Särndal, Carl-Erik Swensson, B. W. J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schiaffino, A Rodríguez, M. P. M. I. R. E. B. y. C. F. (2002). ¿Odds ratio o razón de proporciones? Su utilización en estudios transversales. *Gac Sanit*.

Tellez Piñerez, C. F., Guerrero, S. Y., and Pacheco, M. (2014). Inferencia Bootstrap bayesiana para una proporción en muestreo con probabilidades desiguales. *Comunicaciones en Estadística*, 7(1):31.