UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2017, Vol. 10, No. 2, pp. 207–223

Elicitación de la distribución multinomial a partir de varios expertos[1]

Multinomial distribution elicitation from various experts

Yurledy Montoya[a]
ymontoy@unal.edu.co

Juan Carlos Correa Morales[b]
jccorrea@unal.edu.co

Resumen

El objetivo de este trabajo es implementar una metodología de elicitación mediante el método Delphi, la cual permita estimar el vector de parámetros π de la distribución multinomial a partir de la cuantificación de opinión y creencias de múltiples expertos, buscando así conseguir un resultado diferenciador y de más valor que la suma de aportaciones individuales, con el fin de obtener una única distribución que represente el conocimiento del conjunto de expertos.

***Palabras clave*:** Distribución *a priori*, distribución binomial, distribución multinomial, estadística bayesiana, probabilidad subjetiva.

Abstract

The objective of this work is to implement an elicitation process to estimate the vector of parameters π to multinomial distribution through of Delphi method. Quantification of opinions and beliefs from multiple experts are used to obtain a single distribution that represents the knowledge of all experts that is a different result to get the sum of individual contributions.

***Keywords*:** *a priori* distribution, bayesian statistics, binomial distribution, multinomial distribution, subjective probability.

1. Introducción

El análisis bayesiano se refiere a los métodos para hacer inferencias a partir de datos utilizando modelos de probabilidad para cantidades observables y cantidades

[1]DOI: http://dx.doi.org/10.15332/s2027-3355.2017.0002.02
Montoya, Y., Correa, J. (2017) Elicitación de la distribución multinomial a partir de varios expertos. *Comunicaciones en Estadística*, **10**(2), 207-223.

[a]Magíster en Ciencias-Estadística, Escuela de Estadística, Universidad Nacional de Colombia, sede Medellín.

[b]Profesor Asociado, Escuela de Estadística, Universidad Nacional de Colombia, sede Medellín.

sobre las que se quiere aprender; su característica esencial es el uso explícito de la probabilidad para cuantificar la incertidumbre en inferencias basadas en el análisis de datos estadísticos (Gelman et al. 2003).

A diferencia del punto de vista frecuentista, en la teoría bayesiana no es necesario que un evento sea aleatorio para que se le pueda asignar una probabilidad, el aspecto relevante es que exista incertidumbre sobre el evento de ocurrencia (Mendoza & Regueiro 2011); se puede reducir el nivel de incertidumbre de un fenómeno o experimento aleatorio aumentando el nivel de información con datos históricos observados o también con información subjetiva de expertos en dicho fenómeno. Esta cuantificación de información subjetiva se realiza mediante un proceso conocido como elicitación, el cual se define como el proceso de formular el conocimiento y creencias de un experto acerca de una o más cantidades inciertas en forma de una distribución de probabilidad conjunta, donde el experto es la persona a quien la sociedad sus compañeros atribuyen conocimiento especial sobre el tema que se elicita (Garthwaite et al. 2005). Este enfoque de probabilidad es ampliamente aprovechado por la ,etodología bayesiana, por ello se dice que esta va más allá que la estadística frecuentista al buscar aprovechar toda la información disponible, así se trate de datos observados o de información de otro tipo que ayude a disminuir de manera coherente la incertidumbre (Ruiz 2004).

Bajo el principio estadístico subyacente que entre más información se haya recopilado mejores serán los resultados, parte el deseo de obtener mayor cantidad de información, motivando a elicitar múltiples expertos, pues el resultado de su experiencia combinada puede ser en sí más informativa que cuando se elictan expertos individualmente. Esta distribución que representa la opinión conjunta de los expertos es utilizada como una distribución *a priori* en el análisis bayesiano (Clemen & Winkler 1999). La cuantificación y agregación de creencias de expertos pueden proporcionar información importante para un tomador de decisiones, puede dar lugar a inferencias óptimas justificables de los parámetros de interés (Goossens et al. 2008); por tal motivo, es de mayor ganancia para el análisis considerar un enfoque donde los expertos interactúen como grupo. Un enfoque de elicitación de grupo sencillo y práctico es llevar a los expertos a discutir la cantidad incierta o las cantidades sobre las que sus creencias van a elicitarse y, después de una puesta en común, llegar a una opinión de consenso (Garthwaite et al. 2005).

En la cuantificación de creencias en grupo de expertos se han realizado estudios en los que los investigadores han tratado de elicitar información subjetiva y representarla de una manera rigurosa como parte de una investigación formal, entre ellos se encuentra a Barrera et al. (2011), quienes utilizaron una metodología bayesiana donde implementaron el proceso de elicitación y el método Delphi con el fin de determinar la proporción de estudiantes que desertaron del Instituto Tecnológico Metropolitano de Medellín. Chu & Hwang (2008), proponen un enfoque basado en el método Delphi para elicitar el conocimiento de múltiples expertos en

el diagnóstico del síndrome respiratorio agudo severo. Yau & Chiu (2015) utilizan el método Delphi para identificar y dar prioridad a las opciones a combatir contra la ilegalidad en Hong Kong.

2. Elicitación multivariada

La distribución multinomial juega un papel fundamental en el trabajo aplicado, siendo esta la generalización multivariable de la distribución binomial, surge cuando cada ensayo tiene más de dos posibles resultados. Considérese el caso de n ensayos independientes, que permiten k resultados mutuamente excluyentes $E_1, \ldots, E_k$ cuyas probabilidades respectivas son $\pi_1, \ldots, \pi_k$ (con $\sum_{i=1}^{k} \pi_i = 1$). Denótese $n_1, \ldots, n_k$ la variable aleatoria del número de ocurrencias de los eventos $E_1, \ldots, E_k$ respectivamente en n ensayos con $\sum_{i=1}^{k} n_i = n$. La función de probabilidad de $n_1, \ldots, n_k$ viene dada por Johnson et al. (1997):

$$f(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k} \tag{1}$$

para $n_i = 0, 1, \ldots, n$, pero sujeto a la restricción de $\sum_{i=1}^{k} n_i = n$.

La media y varianza de la distribución multinomial están dadas por:

$$E(n_i) = n\pi_i; \qquad Var(n_i) = n\pi_i(1 - \pi_i) \tag{2}$$

Su distribución *a priori* conjugada es la generalización multivariada de la distribución beta, conocida como la distribución Dirichlet. La distribución Dirichlet es la distribución multivariable más simple y apropiada para representar el conocimiento del experto, también es la más conveniente cuando el conocimiento *a priori* del experto es combinado con una muestra multinomial, la distribución Dirichlet es la conjugada *a priori* para los modelos multinomiales y es ampliamente utilizada por su tratabilidad y simplicidad matemática. Se dice entonces que π tiene una distribución Dirichlet con vector de parámetros $\alpha = (\alpha_1, \ldots, \alpha_k)$, denotado por $\pi \sim Di(\alpha)$, y su función de densidad de probabilidad se puede escribir como **?**:

$$f(\pi|\alpha) = \left[\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \right] \prod_{i=1}^{k} \pi_i^{\alpha_i - 1} \tag{3}$$

donde $\alpha_i > 0; i = 1, \ldots, k$; $\sum_{i=1}^{k} \alpha_i = n$ y $\sum_{i=1}^{k} \pi = 1$. La media y la varianza de la distribución Dirichlet vienen dadas por:

$$E(\pi_i|\alpha_i) = \frac{\alpha_i}{n}; \qquad Var(\pi_i|\alpha_i) = \frac{\alpha_i(n - \alpha_i)}{n^2(n + 1)} \tag{4}$$

Ahora, supóngase que $\mathbf{Y} = (Y_1, \ldots, Y_k)^{'}$ tiene una distribución multinomial con parámetro n fijo y $\pi = (\pi_1, \ldots, \pi_k)^{'}$ vector de probabilidad de ocurrencia desconocido del evento; también que la distribución *a priori* de π es una Dirichlet con vector de parámetros $\alpha = (\alpha_1, \ldots, \alpha_k)^{'}$ con $\alpha_i > 0$; $i = 1, \ldots, k$. Entonces, la distribución posterior de π cuando $Y_i = y_i$, $i = 1, \ldots, k$ es una distribución Dirichlet con vector de parámetros $\alpha^* = (\alpha_1 + y_1, \ldots, \alpha_k + y_k)^{'}$.

La distribución Dirichlet cuenta con dos propiedades útiles para su elicitación, la propiedad de distribución marginal y la condicional:

- **Distribución marginal:** sea $m < k-1$ el primer elemento de π que puede ser denotado por $\pi^m = (\pi_1, \pi_2, \ldots, \pi_m)$, y sea $\pi^*_{m+1} = \sum_{i=m+1}^{k} \pi_i = 1 - \sum_{i=1}^{m} \pi_i$. Entonces $\pi^{m+} = (\pi^m, \pi^*_{m+1}) = (\pi_1, \ldots, \pi_m, \pi^*_{m+1})$ toma valores m-dimensional simples. La propiedad marginal es que la distribución de π^{m+} es una $Di(\alpha_1, \alpha_2, \ldots, \alpha_m, \alpha^*_{m+1})$ donde $\alpha^*_{m+1} = \sum_{i=m+1}^{k} \alpha_i = n - \sum_{i=1}^{m} \alpha_i$. Un caso especial de la propiedad marginal afirma que la distribución marginal de π_i es una beta con parámetros α_i y $n - \alpha_i$, es decir, $\pi_i \sim \text{Beta}(\alpha_i, n - \alpha_i)$.

- **Distribución condicional:** para $i = m+1, \ldots, k$ tenemos:

$$\pi_i^{'} = \frac{\pi_i}{1 - \pi_1 - \pi_2 - \ldots - \pi_m} \tag{5}$$

  Nótese que $\pi_i^{'} = (\pi_{m+1}^{'}, \ldots, \pi_m^{'})$ satisface la condición de caer en $(k - m - 1)$. La propiedad condicional es que la distribución condicional de $\pi^{'}|\pi^m$ es una $Di(\alpha_{m+1}, \ldots, \alpha_k)$, utilizando esta propiedad se puede descomponer la distribución Dirichlet en una secuencia de $k-1$ beta condicional.

Entre los métodos que se han realizado para introducir la elicitación de los parámetros de la distribución Dirichlet se encuentra a Dickey (1983) quienes proponen un método para estimar la distribución *a priori* subjetiva de la Dirichlet en un muestreo multinomial su método es definido como un dispositivos de resultados imaginarios. Elfadaly & Garthwaite (2012) proponen un método que está diseñado para elicitar una distribución generalizada Dirichlet que es también una conjugada *a priori* la cual tiene un mayor número de parámetros y, por lo tanto una estructura de dependencia más flexible; en el método propuesto presentan la cuantificación de la opinión de expertos sobre los hiperparámetros de la *a priori* conjugada Dirichlet, basando su método en la elicitación de los parámetros de las distribuciones beta univariantes como la distribución marginal y condicional de la Dirichlet. Zapata et al. (2012), emplean un dispositivo de sobre ajuste, es decir, elicitan más juicios de los mínimos requeridos con el fin de producir una distribución Dirichlet cuidadosamente considerada y asegurar que es, de hecho, un ajuste razonable para el conocimiento del experto; el método se aplicó en una extensión del *software* de elicitación Sheffield, que es un paquete de documentos, plantillas y *software* que

proporcionan los protocolos de elicitación estructurados ajustables a las buenas prácticas de elicitación moderna, esto con el fin de facilitar el proceso de elicitación multivariante. Flórez & Correa (2015) basan su propuesta para elicitar el vector de parámetros π de la distribución multinomial por medio de tres pasos: el primer paso busca estimar el n-equivalente del experto, en el segundo paso el analista estima el vector de probabilidades de ocurrencia de cada categoría y, finalmente, el tercer paso está basado en un método de elicitación estadística que permite que los parámetros de la distribución conjugada Dirichlet sean estimados por medio del n-equivalente y el vector de probabilidades estimado.

3. Metodología propuesta

El proceso de elicitación propuesto se realiza mediante seis pasos: en el primer paso, se lleva a cabo el método Delphi por medio de formulación de preguntas y retroalimentación; en el segundo paso, se realiza un análisis descriptivo por medio de clúster; en el tercer paso se integran las opiniones individuales de los expertos con el fin de llegar a un solo vector de parámetros π de la distribución multinomial que represente el conocimiento de todos los expertos; en el cuarto paso, se halla el tamaño de muestra que representa el conocimiento del conjunto de expertos, el cual es empleado en el siguiente paso para realizar la simulación de la distribución multinimomial finalmente, en el sexto paso se halla el vector de parámetros α de la distribución Dirichlet, a partir del cual se obtiene el vector π. El método propuesto es una extensión de la propuesta realizada por Flórez & Correa (2015), donde se pasa de tener la opinión de un solo experto a múltiples expertos.

1. **Método Delphi**

   - **Formulación de las preguntas:** tiene una consecuencia importante sobre el resultado final, ya que si las preguntas reflejan inevitablemente las actitudes culturales, sesgos subjetivos y conocimiento de sus diseñadores, condicionan la compresión correcta por parte del experto de sobre la que se requiere su conocimiento, e influyen, por consiguiente, en la calidad, propiedad y extensión de su respuesta. Por tal motivo, es muy importante confeccionarlas de manera que sean claras y concisas, asegurándose de que son correctamente entendidas y de que no condicionan en absoluto la respuesta (Landeta 1999).

   - **Retroalimentación:** en esta etapa, en la primera ronda se da al experto un tamaño de muestra hipotético para que distribuya esta cantidad entre los niveles de la variable; posteriormente se varía el tamaño de muestra inicial y se le solicita que realice nuevamente el proceso de distribución. Las demás rondas se llevan a cabo por medio de retroalimentación, donde se envía a cada experto por separado la mediana y la desviación estándar obtenida en cada categoría de la ronda anterior por los demás expertos elicitados junto con su respuesta individual; en

este punto se pide al experto reevaluar sus creencias con base en los resultados de los demás expertos.

2. **Análisis descriptivo:** se realiza por medio de clúster conocido también como análisis de conglomerados, es una técnica estadística multivariante que busca agrupar elementos o variables tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. Es un método basado en criterios geométricos y se utiliza fundamentalmente como una técnica exploratoria y descriptiva (Johnson & Wichern 2002). La distancia utilizada en el análisis de conglomerados fue Hellinger, la popularidad de esta distancia se debe a la capacidad de combinar dos propiedades fundamentales en la estimación paramétrica, la eficiencia en la densidad del modelo y las excelentes propiedades de robustez (Toma 2007). La distancia de Hellinger para dos distribuciones multinomiales definidas sobre las mismas k categorías es:

$$Hellinger = HL = \sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2 \tag{6}$$

$$HL = \begin{bmatrix} HL_{1,1} & HL_{1,2} & HL_{1,3} & \dots & HL_{1,n} \\ HL_{2,1} & HL_{2,2} & HL_{2,3} & \dots & HL_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ KL_{n,1} & HL_{n,2} & HL_{n,3} & \dots & HL_{n,n} \end{bmatrix}$$

Donde la matriz de distancia Hellinger (HL) es una matriz simétrica, positiva y su diagonal principal es cero.

3. **Integración de las opiniones individuales:** en esta etapa se obtiene una única distribución para representar el conocimiento del conjunto de expertos; para esto, es fundamental la integración de las respuestas individuales, esta se realiza por medio de una simulación de la distribución multinomial en `R`. Para llevar a cabo esta simulación es necesario realizar dos pasos previos, primero estimar el N-equivalente para cada experto y segundo integrar las opiniones de los expertos por medio de simulación.

   - **N-equivalente para cada experto:** se pide al experto estimar un valor mínimo a y un valor máximo b en el que él considere que se encuentra el verdadero valor de cada nivel de la variable de interés; estos dos valores son utilizados en la fórmula del intervalo de confianza para la distribución multinomial basado en el teorema del límite central, para así despejar el valor de n, obteniendo de esta forma un n para cada nivel de la variable de interés. Finalmente, el N-equivalente asociado al nivel de experticia que tiene el experto elicitado es el n de la categoría que tenga asociado el menor valor.

$$\left(\hat{\pi} - Z_{(\alpha/2k)} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + Z_{(\alpha/2k)} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right) \tag{7}$$

Donde k es el número de niveles que tiene la variable, α es el nivel de significancia, $\hat{\pi}$ es pedido al experto como el valor que él considera más probable que está en el intervalo. Al igualar los valores a y b al límite inferior y superior del intervalo de proporción y despejando n se tiene:

$$n = \frac{4Z^2_{(\alpha/2k)}\hat{\pi}(1-\hat{\pi})}{(b-a)^2} \tag{8}$$

Después de finalizar las rondas de elicitación y hallar el respectivo N-equivalente para cada experto, se define el vector de probabilidades π de cada experto como el vector de probabilidad dado en la última ronda de elicitación, así:

$$\begin{array}{llll}
E_1, & \pi^{(1)} = \left(\pi_1^{(1)}, \pi_2^{(1)}, \ldots, \pi_k^{(1)}\right), & N_1, & w_1 = \frac{N_1}{N_1+N_2+\ldots+N_e} \\
E_2, & \pi^{(2)} = \left(\pi_1^{(2)}, \pi_2^{(2)}, \ldots, \pi_k^{(2)}\right), & N_2, & w_2 = \frac{N_2}{N_1+N_2+\ldots+N_e} \\
\vdots & \vdots & \vdots & \vdots \\
E_e, & \pi^{(e)} = \left(\pi_1^{(e)}, \pi_2^{(e)}, \ldots, \pi_k^{(e)}\right), & N_e, & w_e = \frac{N_e}{N_1+N_2+\ldots+N_e}
\end{array}$$

- **Integración de opiniones de los expertos por medio de simulación:** se genera una simulación de la distribución multinomial en el *software* estadístico `R` por medio de la función `rmultinom`, la cual recibe tres argumentos. El primero corresponde al número de simulaciones a realizar, el segundo argumento es el tamaño de muestra que varía para cada experto $N_1, N_2, \ldots, N_e$ y el tercer argumento es el vector $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(e)}$ dado por cada experto. Así, la integración de las opiniones de los expertos se denota como π_i, el cual representa el conocimiento del conjunto de expertos para cada categoría y puede ser hallado de la siguiente forma:

$$\begin{aligned}
\pi_1 &= \frac{\hat{n_1}^{(1)} + \hat{n_1}^{(2)} + \ldots + \hat{n_1}^{(e)}}{N_1 + N_2 + \ldots + N_e} \\
\pi_2 &= \frac{\hat{n_2}^{(1)} + \hat{n_2}^{(2)} + \ldots + \hat{n_2}^{(e)}}{N_1 + N_2 + \ldots + N_e} \\
&\vdots \qquad\qquad\qquad \vdots \\
\pi_k &= \frac{\hat{n_k}^{(1)} + \hat{n_k}^{(2)} + \ldots + \hat{n_k}^{(e)}}{N_1 + N_2 + \ldots + N_e}
\end{aligned}$$

Donde $\hat{n_k}^{(e)}$ representa el promedio del experto e en cada nivel de la varible simulada, por ejemplo $\hat{n_1}^{(2)}$ representa el promedio de las estimaciones simuladas del segundo experto en la primera categoría.Finalmente, el vector de probabilidades que representa el conocimiento del conjunto de expertos esta dado por:

$$\pi = (\pi_1, \pi_2, \ldots, \pi_k) \tag{9}$$

4. **N-global:** en esta etapa se calcula el N-global que represente el tamaño de muestra del conocimiento del total de los expertos. Para esto se realiza un promedio ponderado por los pesos del $N-$equivalente empleando la siguiente expresión:

$$N - \text{global} = \frac{N_1 w_1 + N_2 w_2 + \ldots + N_e w_e}{w_1 + w_2 + \ldots + w_e} \tag{10}$$

5. **Simulación:** en este paso se lleva a cabo la simulación para la distribución multinomial realizada en el *software* estadístico `R` con la función `rmultinom`, la cual recibe tres argumentos. El primero corresponde al número de simulaciones a realizar, el segundo argumento es el tamaño de muestra que representa el conocimiento del conjunto de expertos (N-global) y el tercero es el vector de probabilidades de la distribución del conjunto de expertos representados en la ecuación (9).

```
t(rmultinom(Numsim,Nglobal,prob=c(pi1,pi2,...pik)))/Nglobal
```

6. **Estimación de α_i:** en esta etapa se realiza la estimación del vector de parámetros $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ de la distribución Dirichlet por medio de la propuesta de Flórez y Correa (2015). Dado que cada X_{ij} simulado en el paso anterior sigue una distribución multinomial, entonces $Y_i = X_j/n$ (con $i = 1, 2, \cdots, k$ y $j = 1, 2, \cdots, r$) tiene una distribución Dirichlet con un vector de parámetros $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$, donde n es el tamaño muestral que representa el conocimiento del conjunto de los expertos estimado en el paso 4 (N-Global), k representa el número de categorías y r el número de simulaciones. Si $\alpha_0 = \sum_{i=1}^{k} \alpha_i$, el primer y segundo momento de la distribución Dirichlet vienen dados por:

$$E[Y_i] = \frac{\alpha_i}{\alpha_0}. \tag{11}$$

$$Var[Y_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}. \tag{12}$$

Los valores de (11) y (12) pueden ser estimados a partir de los valores simulados en el paso anterior, reduciéndose el problema en resolver las ecuaciones (11) y (12) en términos de α_0 y α_i. De (11) se tiene que:

$$\alpha_i = \alpha_0 E[Y_i] \tag{13}$$

De (12) se despeja a α_0 en términos de $E[Y_i]$ y $Var[Y_i]$ se tiene que:

$$\alpha_0 = \frac{(E[Y_i] - E[Y_i]^2)}{Var[Y_i]} - 1 \tag{14}$$

Se reemplaza (14) en (13) obteniendo así los valores de cada α_i:

$$\alpha_i = \left(\frac{(E[Y_i] - E[Y_i]^2)}{Var[Y_i]} - 1 \right) E\left[Y_i\right] \tag{15}$$

Normalizando Y_i de manera que se cumpla la restricción $\sum_{i=1}^{k} \bar{Y}_i = 1$ se tiene que:

$$\bar{\bar{y}}_i = \frac{\bar{Y}_i}{\sum_{i=1}^{k} \bar{Y}_i} \tag{16}$$

Así, cada α_i puede ser estimado reemplazando los valores de $\bar{\bar{y}}_i$ y S_i^2 en (15):

$$\alpha_i = \left(\frac{(\bar{\bar{y}}_i - \bar{\bar{y}}_i^2)}{S_{Y_i}^2} - 1 \right) \bar{\bar{y}}_i \tag{17}$$

Finalmente, haciendo uso de la propiedad marginal de la distribución Dirichlet, donde $\pi_i \sim \text{Beta}(\alpha_i, n - \alpha_i)$ con $n = \sum_{i=1}^{k} \alpha_i$, se obtiene así la distribución de cada nivel de la variable de interés.

4. Caso de estudio

La malaria es una enfermedad infecciosa de origen parasitológico febril aguda, se reconoce un espectro de manifestaciones de la enfermedad que va desde procesos asintomáticos, cuadros sintomáticos con escalofrío, fiebre, sudoración y cefalea hasta cuadros severos que pueden llevar a la muerte; es así como se definen dos formas clínicas: malaria no complicada y malaria complicada, esta última se asocian a una mayor mortalidad (Proyecto. 2015). Existen cuatro especies del parásito que infectan a los seres humanos: *plasmodium vivax*, *plasmodium falciparum*, *plasmodium malariae* y *plasmodium ovale*, los más frecuentes son el paludismo por *plasmodium falciparum* y por *plasmodium vivax*, y el más mortal el paludismo por *plasmodium falciparum*. En los últimos años se han presentado algunos casos humanos por *plasmodium knowlesi*, una especie que circula en primates y que aparece en zonas boscosas de Asia Sudoriental, por lo que este parásito ha sido propuesto como el quinto parásito que infecta a los humanos (Instituto 2014).

En Colombia, la malaria continúa siendo un problema grave de salud pública, debido a que cerca del 85 % del territorio rural está situado por debajo de los 1500 metros sobre el nivel del mar y presenta condiciones climáticas, geográficas y epidemiológicas aptas para la transmisión de la enfermedad. Cerca del 60 % de la población colombiana se encuentra en riesgo de enfermar o morir por esta causa. Las especies más frecuentes en zonas endémicas son *plasmodium vivax* y *plasmodium falciparum*. La transmisión de *plasmodium malariae* ocurre en focos dispersos a lo largo de la costa pacífica, principalmente en el departamento del Chocó, y no existe la transmisión de *plasmodium ovale* ni de *plasmodium knowlesi*. También pueden ocurrir casos de infecciones mixtas, definidas como infecciones simultáneas por dos especies, usualmente *plasmodium vivax* y *plasmodium falciparum*. El caso de estudio se realizó con el fin de estimar la distribución del tipo de malaria en pacientes diagnosticados con la enfermedad en la zona costera de Buenaventura.

4.1. Implementación de la metodología propuesta

- **Identificación y selección de expertos:** la identificación de expertos se realizó por medio del Proyecto Malaria Colombia, donde se seleccionaron diferentes áreas de profesionales como expertos, obteniendo así a cuatro expertos para la elicitación, dos biólogos entomólogos, una microscopista y la coordinadora del equipo malaria del Valle del Cauca, como grupo coordinador se contó con la participación de la consultora en sistemas de información monitoreo y evaluación del Proyecto Malaria del Valle del Cauca. Estas personas son consideradas como expertos por su comprensión del tema debido al alto conocimiento y experiencias adquiridas en sus funciones realizadas en el Proyecto Malaria Colombia.

- **Estructuración y descomposición:** el proceso de elicitación comienza dando a cada experto una breve introducción sobre la distribución multinomial y como esta puede ser utilizada en la distribución de los casos de malaria según la especie, proceso que se realiza teniendo en cuenta el conocimiento de cada uno de ellos por medio de elicitación enfocado a una metodología Delphi; por tal motivo, se le explica a cada experto que la variable de interés para el desarrollo de este trabajo es la prevalencia de cada una de las especies de malaria en pacientes diagnosticados con la enfermedad en la zona costera rural de Buenaventura, como también se indica que la escala de medición corresponde al número de pacientes con la enfermedad en cada especie de malaria.

- **Aplicación de la metodología:**

    1. **Formulación de las preguntas:** inicialmente se explica y se valida con cada experto que la pregunta realizada sea comprendida, luego en la primera ronda se dan cinoc muestras hipotéticas, por medio de preguntas como "si se seleccionaran 100 pacientes diagnosticados con malaria de la zona costera rural de Buenaventura, según su conocimiento que proporción de ellos se encuentran en *plasmodium vivax*, *plasmodium falciparum* y malaria mixta". Las rondas posteriores se realizan por medio de retroalimentación, es decir, una vez finalizada cada ronda se envía a cada experto por separado la mediana y la desviación estándar obtenida en cada categoría por los demás expertos junto con su respuesta individual, en este punto el experto puede reconsiderar sus respuesta y realizar algún cambio si él lo considera necesario; si el experto no cambia su respuesta, la información registrada es igual a la ronda anterior. A continuación, se presenta la elicitación obtenida de cada experto en la primera donde se varió el tamaño de muestra hipotético.

       De esta forma, los resultados registrados para cada experto en la primera ronda corresponden a la media de cada categoría de las muestras hipotéticas dadas.

Tabla 1: *elicitación de expertos ronda 1 por tamaño de muestra hipotético. Fuente: elaboración propia.*

Ronda 1	Experto 1						Experto 2					
	Muestra hipotética						Muestra hipotética					
Especie	**100**	**200**	**300**	**700**	**900**	**Media**	**100**	**200**	**300**	**700**	**900**	**Media**
P.vivax	0.090	0.095	0.100	0.057	0.056	0.080	0.200	0.150	0.133	0.143	0.178	0.161
P.falciparum	0.900	0.900	0.897	0.943	0.943	0.917	0.750	0.800	0.833	0.786	0.778	0.789
P.mixta	0.010	0.005	0.003	0.001	0.001	0.004	0.050	0.050	0.033	0.071	0.044	0.050

Ronda 1	Experto 3						Experto 4					
	Muestra hipotética						Muestra hipotética					
Especie	**100**	**200**	**300**	**700**	**900**	**Media**	**100**	**200**	**300**	**700**	**900**	**Media**
P.vivax	0.200	0.130	0.100	0.070	0.110	0.122	0.100	0.100	0.073	0.100	0.111	0.097
P.falciparum	0.750	0.800	0.870	0.900	0.860	0.834	0.820	0.825	0.867	0.829	0.833	0.835
P.mixta	0.050	0.080	0.030	0.030	0.030	0.044	0.080	0.075	0.060	0.071	0.056	0.068

Tabla 2: *Elicitación de expertos por ronda. Fuente: elaboración propia.*

	Ronda 1				Ronda 2			
Especie	**Exp 1**	**Exp 2**	**Exp 3**	**Exp 4**	**Exp 1**	**Exp 2**	**Exp 3**	**Exp 4**
P.vivax	0.080	0.161	0.122	0.097	0.080	0.110	0.122	0.097
P.falciparum	0.917	0.789	0.834	0.835	0.850	0.840	0.834	0.835
P.mixta	0.004	0.050	0.044	0.068	0.070	0.050	0.044	0.068

Posteriormente, se dio retroalimentación a los expertos quienes mantuvieron sus opiniones igual a las de la segunda ronda el total de rondas elicitadas son las mostradas en la tabla anterior.

2. **Análisis descriptivo:** en la figura1 se presentan los resultados del análisis de clúster por medio de la distancia de Hellinger, donde se observa que en la primera ronda la opinión de los expertos se puede clasificar en tres grupos, un grupo donde los expertos dos y tres presentan opiniones similares, el siguiente grupo se compone por el experto cuatro, y, finalmente, en el último grupo se encuentran las opiniones del experto uno las cuales muestran mayor diferencia de los demás expertos. Una vez dada la retroalimentación, se observa en la nueva clasificación que las opiniones del experto uno se asemejan a las del experto cuatro y las opiniones del experto dos y tres siguen permaneciendo en el mismo grupo.
3. **Integración de las opiniones individuales:**
   - N**-equivalente para cada experto:** en esta etapa se realizó la siguiente pregunta a cada experto "si se cuenta con 100 pacientes diagnosticados con malaria en la zona costera rural de Buenaventura, según su conocimiento ¿cuántos como mínimo presentan *plasmodium vivax*?, ¿cuántos como máximo presentan *plasmodium vivax*?, ¿cuál será el valor más frecuente de encontrar pacientes contagiados por *plasmodium vivax*? " Esta pregunta es realizada a cada experto variando la especie de malaria (*plasmodium vivax*, *plasmodium falciparum*, *plasmodium mixta*). La información dada por cada experto se muestra a continuación:
     Reemplazando los valores de la tabla 3 en la ecuación (8) se ob-

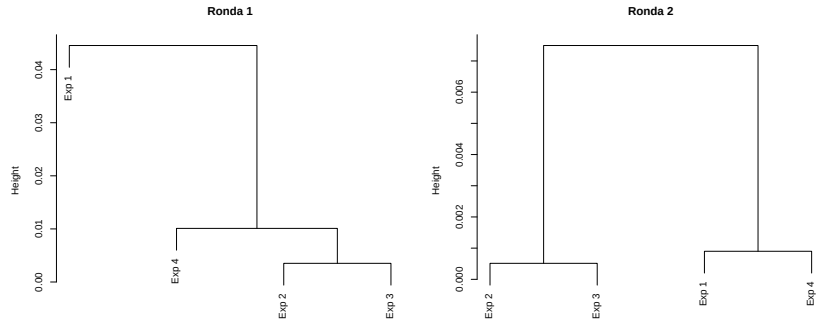

Figura 1: *Clasificación de expertos por ronda. Fuente: elaboración propia.*

Tabla 3: *Resumen de valores elicitacidos para obtener el N-equivalente. Fuente: elaboración propia.*

		Experto 1	**Experto 2**	**Experto 3**	**Experto 4**
Especie	**Muestra**	$\hat{\pi}$ **Intervalo**	$\hat{\pi}$ **Intervalo**	$\hat{\pi}$ **Intervalo**	$\hat{\pi}$ **Intervalo**
P.vivax	100	0.08 (0.03, 0.10)	0.20 (0.05, 0.25)	0.20 (0.12, 0.25)	0.09 (0.08, 0.15)
P.falciparun	100	0.90 (0.80, 0.95)	0.75 (0.50, 0.80)	0.75 (0.60, 0.80)	0.84 (0.80, 0.90)
P.mixta	100	0.01 (0.01, 0.10)	0.05 (0.00, 0.15)	0.04 (0.01, 0.05)	0.07 (0.01, 0.12)

tiene un n para cada experto en cada categoría y, finalmente el N-equivalente para cada experto es el n de la categoría que tenga asociado el menor valor (tabla 3).

Tabla 4: *Estimación N-equivalente. Fuente: elaboración propia.*

	Experto 1	**Experto 2**	**Experto 3**	**Experto 4**
Especie	n	n	n	n
P.vivax	94	92	217	104
P.falciparun	37	48	107	84
P.mixta	57	48	681	34
N-equivalente	$N_1 = 37$	$N_2 = 48$	$N_3 = 107$	$N_4 = 34$

Así, la información obtenida para cada experto es la siguiente:

$$E_1, \quad \pi^{(1)} = (0.080, 0.850, 0.070)\,, \quad N_1 = 37, \quad w_1 = \frac{N_1}{N_1+N_2+N_3+N_4} = 0.163$$
$$E_2, \quad \pi^{(2)} = (0.110, 0.840, 0.050)\,, \quad N_2 = 48, \quad w_2 = \frac{N_2}{N_1+N_2+N_3+N_4} = 0.212$$

E_3, $\pi^{(3)} = (0.122, 0.834, 0.044)$, $N_3 = 107$, $w_3 = \frac{N_3}{N_1+N_2+N_3+N_4} = 0.473$
E_4, $\pi^{(4)} = (0.097, 0.835, 0.068)$, $N_4 = 34$, $w_4 = \frac{N_4}{N_1+N_2+N_3+N_4} = 0.150$

- **Integración de las opiniones de los expertos:** en esta etapa se realiza una simulación utilizando como vector π la distribución dada por cada experto en la última ronda de elicitación, así la simulación para cada experto puede ser calculada mediante el *software* `R` con el siguiente comando:

```
sim1=t(rmultinom(10000,37,prob=c(0.080,0.850,0.070)))
sim2=t(rmultinom(10000,48,prob=c(0.110,0.840,0.050)))
sim3=t(rmultinom(10000,107,prob=c(0.122,0.834,0.044)))
sim4=t(rmultinom(10000,34,prob=c(0.097,0.835,0.068)))
```

Una vez realizada la simulación se promedian las estimaciones simuladas de cada experto así:

```
promedio1=colMeans(sim1)
promedio2=colMeans(sim2)
promedio3=colMeans(sim3)
promedio4=colMeans(sim4)
```

Se halla el vector de parámetros π_i que representa la opinión conjunta de los expertos:

$$\pi_1 = \frac{\hat{n_1}^{(1)} + \hat{n_1}^{(2)} + \hat{n_1}^{(3)} + \hat{n_1}^{(4)}}{N_1 + N_2 + N_3 + N_4} = \frac{2.97 + 5.31 + 13.10 + 3.32}{37 + 48 + 107 + 34} = 0.109$$

$$\pi_2 = \frac{\hat{n_2}^{(1)} + \hat{n_2}^{(2)} + \hat{n_2}^{(3)} + \hat{n_2}^{(4)}}{N_1 + N_2 + N_3 + N_4} = \frac{31.44 + 40.31 + 89.18 + 28.38}{37 + 48 + 107 + 34} = 0.838$$

$$\pi_3 = \frac{\hat{n_3}^{(1)} + \hat{n_3}^{(2)} + \hat{n_3}^{(3)} + \hat{n_3}^{(4)}}{N_1 + N_2 + N_3 + N_4} = \frac{2.59 + 2.38 + 4.71 + 2.29}{37 + 48 + 107 + 34} = 0.053$$

Por lo tanto, el vector de parámetros $\pi = (0.109, 0.838, 0.053)$ representa el conocimiento del conjunto de los expertos respecto a la especie de malaria dada en la zona costera rural de Buenaventura.

4. N**-global:** el tamaño de muestra que representa el conocimiento del conjunto de los expertos se calcula empleando la ecuación (10):

$$N - \text{global} = \frac{N_1 w_1 + N_2 w_2 + N_3 w_3 + N_4 w_4}{w_1 + w_2 + w_3 + w_4} = 72$$

5. **Simulación y estimación:** empleando el vector de parámetro estimado en el paso tres, se realiza una simulación estadística en `R` por medio de la función `rmultinom` y se estima el vector de parámetros $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ de la distribución Dirichlet, basado en la propuesta de Flórez y Correa en 2015.

```
simt=t(rmultinom(10000,72,prob=c(0.109,0.838,0.053)))/72
```

De esta forma se asegura que $sim \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ por lo que la media y varianza pueden ser calculados de *simt*:

```
medias<-colMeans(simt)
varianza<-sapply(1:3,function(x)var(simt[,x]))
```

Para garantizar que se cumpla la restricción $\sum_{i=1}^{k} \bar{Y}_i = 1$ se normaliza el vector de medias estimado de la siguiente forma:

```
medias<-colMeans(simt)/sum(colMeans(simt))
```

Reemplazando los valores de la media normalizados y la varianza en la ecuación 17 así:

```
alfa<-(((medias-medias^2)/varianza)-1)*medias
medias.alfa<-alfa/sum(alfa)
var.alfa<-(alfa*sum(alfa)-alfa)/((sum(alfa)+1)*sum(alfa)^2)
```

Llegando de esta forma al vector de parámetros $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ de la distribución Dirichlet (tabla 3).

Tabla 5: *Vector de parámetros α distribución Dirichlet. Fuente: elaboración propia.*

	Alfa	Media	Varianza
α_1	7.789	0.109	0.0013
α_2	59.724	0.838	0.0018
α_3	3.806	0.053	0.0006

Finalmente, para encontrar la distribución de la especie de malaria de la zona costera rural de Buenaventura, se hace uso de la propiedad marginal de la distribución Dirichlet, donde $\pi_i \sim Beta(\alpha_i, n - \alpha_i)$ con $n = \sum_{i=1}^{k} \alpha_i$, de esta forma se tiene:

Tabla 6: *distribución de probabilidad marginal por especie de malaria. Fuente: elaboración propia.*

Especie de Malaria	Marginal
P.Vivax	Beta(7.789, 63.52)
P.Falciparun	Beta(59.724, 11.59)
P.Mixta	Beta(3.806, 67.51)

Así la distribución de probabilidad marginal para cada especie de malaria dada en la zona costera rural de Buenaventura se muestra en la figura 2, donde se puede observar que para la especie de malaria mixta la mayor densidad se concentra alrededor de 0.05, para la especie *vivax* la densidad se concentra en 0.10, adicional a esto se ve que las densidades de estas dos especies de malaria se traslapan. Finalmente, se encuentra que la mayor concentración de densidad en malaria se da en la especie

falciparum con 0.85, es decir que alrededor del 85 % de los casos son de esta especie, esto es debido a las condiciones geográficas y la actividad económica de la región, pues esta especie de malaria se desarrolla en zona costera donde las actividades económicas más frecuentes son la pesca, el corte de madera y actividades de agricultura.

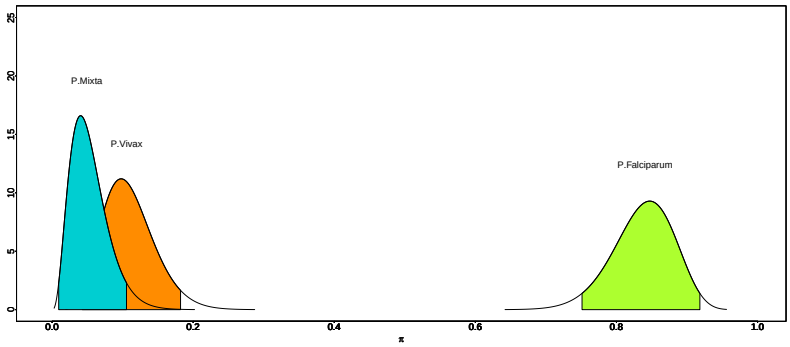

Figura 2: *Densidad por especie de malaria. Fuente: elaboración propia.*

5. Conclusiones

El objetivo principal de este trabajo fue desarrollar una metodología bayesiana que permitiera elicitar el vector de parámetros de la distribución multinomial a partir de varios expertos, lo cual es una ventaja sobre otros métodos de elicitación (donde solo se cuenta con la opinión y conocimiento de un experto), ya que el consenso de un grupo de individuos puede ser superior bajo determinadas condiciones a la suma de los resultados individuales de los miembros que la componen (Landeta 1999). La implementación del proceso de elicitación para trabajar con múltiples expertos fue basada en la metodología Delphi, donde inicialmente se construye una distribución *a priori* para cada experto por separado y luego se integran con el fin de obtener una única distribución que represente el conocimiento de todos los expertos. Es importante resaltar el análisis descriptivo llevado a cabo por medio de clúster en la metodología propuesta, ya que es una herramienta útil que permite al grupo coordinador observar similitudes entre las opiniones *a prioris* de los expertos y el comportamiento que estas van tomando al realizar el *feedback* por cada ronda de elicitación. La cuantificación del conocimiento del experto en términos del tamaño muestral (N-equivalente) se realiza de una forma menos subjetiva hacia el facilitador a la de la propuesta utilizada por Bromaghin (1993) y por Flórez & Correa (2015), ya que no es el facilitador quien califica directamente al experto; en lugar de esto, la estimación se realiza teniendo en cuenta el conocimiento de cada experto por medio de un intervalo de confianza, de esta forma se espera por parte del grupo coordinador que el experto que ellos

consideran con un conocimiento más alto no tenga intervalos de confianza no sean tan amplios en las categorías de la variable de interés.

Recibido: 24 de Junio de 2016
Aceptado: 30 de Agosto de 2017

Referencias

Barrera, C., Sandoval, J. & Sepúlveda, F. (2011), 'Estimación por intervalos de probabilidad a posteriori para la proporción de estudiantes desertores.', *Revista Tecno Lógicas* **27**, 75–87.

Chu, H. & Hwang, G. (2008), 'A Delphi-based approach to developing expert systems with the cooperation of multiple experts.', *Expert Systems with Applications* **34**, 2826–2840.

Clemen, R. & Winkler, R. (1999), 'Combining probability distributions from experts in risk analysis.', *Risk Analysis* **19**, 187–203.

Dickey, J. (1983), 'Multiple Hypergeometric functions: probabilistic interpretations of statistical uses.', *Journal of the American Statistical Association* **78**, 628–637.

Elfadaly, F. & Garthwaite, P. (2012), 'On eliciting some prior distributions for Multinomial Models.', *Department of Mathematics and Statistics, The Open University, UK* .

Flórez, A. & Correa, J. (2015), 'Una propuesta metodológica para elicitar el vector de parámetros π de la distribución Multinomial.', *Comunicaciones en Estadística* **8**, 81–97.

Garthwaite, P., Kadane, J. & O'Hagan, A. (2005), 'Statistical methods for eliciting probability distributions.', *Journal of the American Statistical Association* **100**, 680–700.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2003), *Bayesian data analysis*, 2 edn, Chapman & Hall/CRC, New York.

Goossens, L., Cooke, R., Hale, A. & Rodic-Wiersma, L. (2008), 'Fifteen years of expert judgement at TUDelft.', *Safety Science* **46**, 234–244.

Instituto, N. S. (2014), '(Instituto Nacional de Salud) Protocolo de vigilancia en salud pública.', *http://www.ins.gov.co/lineas-de-accion/Subdireccion-Vigilancia/sivigila/Protocolos %20SIVIGILA/PRO %20Malaria.pdf, tomado el 08/01/2016* .

Johnson, N., Kotz, S. & Balakrishnan, N. (1997), 'Discrete multivariate distributions.', *John Wiley & Sons* **1**, 31–83.

Johnson, R. & Wichern, D. (2002), *Applied Multivariate statistical analysis*, 5 edn, Prentice Hall, Upper Saddle River, New Jersey.

Landeta, J. (1999), *El método Delphi: una técnica de previsión para la incertidumbre*, 1 edn, Editorial Ariel, Barcelona.

Mendoza, M. & Regueiro, P. (2011), *Estadística bayesiana*, Instituto Teconológico de México, New York.

Proyecto., M. C. (2015), '(Proyecto Malaria Colombia) Vigilancia de susceptibilidad a insecticidas de Anopheles (Nyssorhynchus) darlingi, An. (N.) Nuneztovari y An. (N.) Albimanus en localidades centinelas de los departamentos de Antioquia, Cauca, Choco, Córdoba y Valle del Cuaca', *http://www.ins.gov.co/temas-de-interes/Memorias %20Malaria/10.Resistencia %20a %20insecticidas.pdf, tomado el 08/01/2016.* .

Ruiz, A. (2004), 'Apuntes de estadística bayesiana.', *http://www.angelfire.com/ex/proba/Notas/Bayes2.pdf, tomado el 11/03/2016* .

Toma, A. (2007), 'Minimum Hellinger distance estimators for some Multivariate models: influence functions and breakdown point results.', *C. R. Acad. Sci. Paris, Ser I* **345**, 353–358.

Yau, Y. & Chiu, S. (2015), 'Combating building illegality in Hong Kong: a policy Delphi study.', *Habitat International* **48**, 349–356.

Zapata, R., O'Hagan, A. & L, S. (2012), 'Eliciting expert judgements about a set of proportions.', *Journal of Applied Statistics* .