
# Elicitación de una distribución apriori para el modelo logístico

**Jenny Andrea Tangarife Quintero**

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2016

# Elicitación de una distribución apriori para el modelo logístico

**Jenny Andrea Tangarife Quintero**

Trabajo de grado presentado como requisito parcial para optar al título de:
**Magister en Estadística**

Director:
Juan Carlos Correa Morales, Ph.D.

Línea de Investigación:
Estadística Bayesiana

Universidad Nacional de Colombia
Facultad de ciencias, Escuela de Estadística
Medellín, Colombia
2016

**Agradecimientos:**

Gracias a Dios por permitirme finalizar este proyecto, a mis padres por su apoyo incondicional, a mi novio y amigos por su motivación en los momentos en que más lo necesité. A mi director Juan Carlos Correa por compartirme todo su conocimiento y por orientarme siempre con la mejor actitud y dispocisión.

Para empezar un gran proyecto, hace falta valentía. Para culminar un gran proyecto, hace falta perseverancia (Anónimo)

# Resumen

Con el fin de contribuir al desarrollo de los métodos de elicitación, se desarrolló una metodología indirecta para elicitar los parámetros de la regresión logística, haciendo uso de los métodos disponibles para la distribución binomial. En el capítulo 2 se hace una recopilación de las consideraciones importantes al momento de realizar un proceso de elicitación, las heurísticas y sesgos que intervienen en dicho proceso y los desarrollos existentes en la elicitación del parámetro $\pi$ de la distribución binomial, en el capítulo 3 se exploran lo desarrollos que se han dado para la elicitación de la regresión logística, en el capítulo 4 se detalla el método propuesto para este trabajo de investigación y finalmente se da una aplicación de dicho método.

**Palabras clave: Distribución binomial, Distribución Beta, Distribución A priori, Estadística Bayesiana, metodología indirecta**.

# Abstract

With the objective to contribute to the development of elicitation methods, a methodology was designed to elicit the logistic regression parameters, using the available methods for the binomial distribution. On chapter 2 a compilation of the important considerations when doing an elicitation process is made, the heuristic and biases that intervene within the process and the existing developments in the parameter elicitation of the binomial distribution, Chapter 3 explores the developments made for logistic regression elicitation, Chapter 4 details the proposed method for this research and in the end an application of the method is made.

**Keywords: Binomial distribution, Beta distribution, Prior Distribution, Bayesian statistic, Indirect methodology.**

# Contenido

# 1. Introducción

La información extramuestral se puede obtener de diferentes fuentes: datos históricos, información derivada de estudios previos o información proporcionada por uno o varios expertos en el tema de interés mediante un proceso conocido como elicitación, esta última opción es más útil en situaciones en las que los datos históricos que se tienen están incompletos o no son confiables; esta situación suele ser común y es por esto que una parte importante dentro de la estadística Bayesiana es la construcción de métodos de elicitación para hallar dichas distribuciones, adicionalmente muchos trabajos se han encaminado hacia la construcción de métodos de elicitación para los modelos más populares (Hamada et al., 2001), y otros a la comparación de estos diferentes métodos (Umesh, 1988).
En muchas situaciones resulta útil cuantificar información subjetiva que tiene una o varias personas acerca de un tema de interés, para esto se han desarrollado técnicas de elicitación cuyo objetivo es extraer y cuantificar las creencias de un individuo al cual se le denomina experto, un experto es aquella persona que tiene conocimientos sobre un tema de interés y este ha sido obtenido a través de educación, formación y experiencia en el campo (Low et al., 2009), o en términos simples el experto es la persona cuyo conocimiento se quiere elicitar (Correa, 2011). McBride y Burgman (2012) enfatizan que además del conocimiento del experto se debe considerar la dispocision para participar en el proyecto y que no existan conflictos de interés.

En un proceso de elicitación el primer cuestionamiento a abordar es qué significa tener una elicitación exitosa; una elicitación exitosa es aquella que logra representar fielmente la opinión del experto, esto sin importar si el conocimiento de este es cierto o no, es importante diferenciar entre la calidad del conocimiento del experto y la exactitud con la que la distribución de probabilidad construida refleja el conocimiento elicitado. Si el experto es un estadístico o está muy familiarizado con los conceptos estadísticos, entonces no será de gran necesidad direccionar esfuerzos a la construcción de métodos de elicitación para que estos sean de fácil entendimiento para la persona elicitada, pero esto es poco frecuente en la práctica y hace que la obtención de probabilidades subjetivas sea un proceso complejo que requiere de una serie de habilidades (Garthwaite et al., 2005). El uso de un facilitador entrenado es otro punto importante a considerar puesto que este puede ayudar a traducir en probabilidades el conocimiento elicitado, que es finalmente el objetivo de la elicitación.

Garthwaite et al (2005) describen el proceso de elicitación en cuatro etapas separadas:

1. Preparación de la elicitación: selección y entrenamiento de experto(s), identificación de cantidades a elicitar.
2. Elicitación de las cantidades específicas.

3. Ajuste de la distribución de probabilidades.

4. Evaluar la exactitud de la elicitación, con la opción de regresar al punto 2 para elicitar mas cantidades.

Familiarizar a los expertos con el lenguaje de probabilidad en la etapa de preparación de la elicitación ayuda a entender cómo los números pueden ser usados para expresar la incertidumbre y aumenta la conciencia acerca de posibles sesgos en los juicios (O'Hagan, 1998), en la etapa de ajuste de la distribución se debe hacer constante realimentación al experto (Kadane, 1998), en la etapa de evaluación de la exactitud de la elicitación es útil la construcción de escenarios que puedan conducir a valores muy altos o muy bajos de la cantidad incierta, esto ofrece una protección contra el exceso de confianza (Lichtenstein et al., 1980).

El uso de la elicitación se ha dado en diferentes campos como solución a muchos problemas. Savage (1971) y De Finetti (1974) proporcionaron fundamentos teóricos para el uso de probabilidades subjetivas en contextos de toma de decisiones donde la incertidumbre necesita ser expresada como una distribución de probabilidad para derivar (y luego maximizar) la utilidad esperada (Garthwaite et al., 2005). El análisis de riesgos y confiabilidad ha hecho aportes importantes a la elicitación de expertos, un caso importante fue el WASH-1400, estudio de seguridad de reactores (United States Nuclear Regulatory Commission, 1975). En este importante análisis de la seguridad de las centrales de generación de energía nuclear, el juicio de expertos se usó ampliamente para desarrollar probabilidades subjetivas de diversos eventos dando paso a una mayor utilización de la opinión de expertos en estudios de políticas públicas (Hora, 2007). La estadística bayesiana ha sido un factor importante para el desarrollo de la teoría de la decisión (Raiffa y Schaifer, 1964) y los investigadores en este campo han hecho importantes contribuciones a las técnicas utilizadas para evaluar probabilidades. Sus motivaciones surgieron de la necesidad práctica para representar incertidumbres a través de probabilidades con el fin de cuantificar los modelos de decisión. Jenkinson (2005) resume una variedad de estudios en los cuales se hizo uso de la elicitación de expertos en diferentes disciplinas, tales como, medicina, análisis de supervivencia, psicología, industria nuclear, veterinaria, agricultura, meteorología, economía, ecología, arqueología y teoría de juegos. La universidad tecnológica de Delft (TUDelft), por muchos años elicitaron mas de 800 expertos que equivale a mas de 80.000 preguntas en una gran variedad de sectores, tales como el sector nuclear, toxicidad de productos químicos, inundaciones, erupciones volcánicas, en el sector aeroespacial, el sector bancario, entre otros (Goossens et al., 2007). Choy et al. (2009) hacen una revisión del uso de la elicitación en ecología y presentan ocho temas claves para tener en cuenta en el diseño de elicitación en contextos de ecología. Koehler (2006) hace una exhaustiva revisión de la literatura de elicitación relevante para la ingeniería de sistemas, además hace ajustes a dicha información considerando las posibilidades y limitaciones reales que plantea el área.

Un método de elicitación es el puente entre las evaluaciones de un experto y la expresión de estas evaluaciones en una forma estadísticamente útil (Garthwaite et al., 2005) y es por esto que se debe prestar especial atención no sólo a las cantidades que se elicitan, tambien el cómo estas cantidades son elicitadas (Kynn, 2008). Cuando se diseña un cuestionario de elicitación es importante tomar en cuenta las consideraciones desde el campo sicológico (estudios sobre las Huerísticas y sesgos).

Los aportes en esta área han sido muy importantes, entre 1960 y 1980 hubo un gran desarrollo en investigación sobre elicitación y esta se caracterizó por participaciones conjunta de estadísticos y sicólogos y la principal conclusión a la que se llegó fue, el hombre tiene una capacidad limitada de procesamiento de la información, esto a su vez implica que su percepción de la información es selectiva y debe recurrir a heurísticas y mecanismos de simplificación congnitivos, todo esto conduce a una serie de problemas en la evaluación de probabilidades subjetivas (Hogarth, 1975).

# 2. Marco Teórico

Una parte importante dentro de la estadística Bayesiana es la construcción de métodos de elicitación para hallar distribuciones de probabilidad, dichas probabilidades subjetivas son una cuantifiación del conocimiento de un experto acerca de un tema de interés (De Finetti, 1974). El proceso de expresar conocimiento en términos de probabilidades no es simple y ha demostrado estar sujeto a algunos tipos de errores repetibles (Hora, 2007). Un protocolo de elicitación estructurado correctamente puede mejorar sustancialmente la calidad de los juzgamientos (Shephard y Kirkwood, 1994). Acerca de cómo se debe desarrollar un proceso de elicitación, en la literatura se pueden encontrar diferentes autores que han dado su protocolo a seguir, Jenkinson (2005) cita a algunos de estos autores:

- Una vez los expertos se han identificado y contratado Phillips (1999) sugiere un proceso de cuatro etapas: Introducción y entrenamiento, acondicionamiento y codificación, la etapa final es la elicitación de la distribución de probabilidad.

- Walls y Quigley (2001) también sugieren un proceso de elicitación con cinco etapas principales y un diagrama de flujo que da mayor detalle. Los principales componentes son similares a los descritos por Clemen y Reilly (2004)

- Clemen y Reilly (2004) sugieren que cualquier protocolo de elicitación debe contener los siguientes siete pasos:

    1. Antecedentes.
    2. Identificación y contratación de expertos.
    3. Motivación de expertos.
    4. Estructuración y descomposición.
    5. Entrenamiento en probabilidad.
    6. Aplicación del método de elicitación y realimentación.

- Garthwaite et al. (2005), sugieren un método de cuatro etapas:

    1. Preparación de la elicitación.
    2. Elicitación de las cantidades especificas.
    3. Ajuste de la distribución de probabilidades.
    4. Evaluación de la exactitud de la elicitación, con la opción de regresar al segundo paso y elicitar mas cantidades, hasta llegar a la distribución de probabilidad adecuada.

De acuerdo a la anterior recopilación se puede ver que diferentes autores coinciden en los pricipales pasos a seguir durante un proceso de elicitación, aunque no en el orden en que estos deben seguirse (Jenkinson, 2005). A continuación se detallarán las diferentes etapas del proceso de elicitación mencionadas anteriormente:

## 2.1. Antecedentes y preparación

En la preparación de un proceso de elicitación, se deben definir las cantidades a elicitar, los objetivos del estudio (qué se quiere medir y por qué), durante esta etapa también se define la estructura de las preguntas que se van a formular al experto, dichas preguntas deben ser acerca de cantidades que sean significativas para el experto, esto sugiere que las preguntas generalmente deberían referirse a cantidades observables en lugar de parámetros no observables, aunque preguntas acerca de parámetros como la proporción o la media pueden ser consideradas como adecuadas (Garthwaite et al., 2005).
Durante esta etapa también es muy importante elegir y entrenar el facilitador, esta persona debe generar confianza entre los expertos, entender el lenguaje, las heurísticas y las dificultades que puede tener el experto al expresar su conocimiento (Correa, 2011), además debe tener habilidades para entrevistar y comprensión en el campo en el que se realiza la elicitación (Clemen y Reilly, 2004).

## 2.2. Identificación y contratación de expertos

Esta etapa es quizás la más importante dentro del proceso de elicitación, el éxito o el fracaso del proceso depende de la personalidad, experiencia y conocimientos técnicos del experto; un experto es reconocido por tener un conocimiento superior acerca de datos, modelos y normas en un área específica o campo (Bonano et al., 1989). A diferencia de un aficionado en cualquier tema de interés, un experto representa un problema en términos de principios formales, resuelve un problema utilizando estrategias conocidas, confía más en el conocimiento procedimental y menos en conocimiento declarativo (Wood y Ford, 1993). Hora (2007) da los siguientes criterios para medir la experiencia de un experto:

- Investigaciones en el área que hallan sido publicadas.
- Trabajos e investigaciones citadas.
- Grados, premios u otro tipo de reconocimiento.
- Recomendación o nominación por parte de grupos de interés público u organizaciones profesionales.
- Cargos desempeñados.
- Membresía de juntas, comisiones, etc.

La nominación por parte de grupos de interés público u organizaciones profesionales, cobra vital importancia cuando el tema de interés es controversial o tiene diversos puntos de vista (Hora y Winterfeldt, 1997), además es importante que el experto esté libre de cualquier sesgo motivacional causado por interés político o económico. En ciertos casos las proximidad física o la disponibilidad serán una consideración importante (Hora, 2007).

## 2.3. Motivación de expertos

Es importante contar al experto por qué fue seleccionado para el estudio y cómo sus respuestas van a ser usadas, además se le debe motivación e interés por el proyecto (Jenkinson, 2005). Sin motivación es poco probable que los expertos realicen esfuerzos por reconstruir o recordar cosas que el proceso de elicitación requiera (Cannell et al.,1977). El Scoring Rule proporciona un incentivo al experto para registrar bien sus opiniones y para ayudar a capacitarlo en cómo cuantificar sus opiniones con precisión (Garthwaite et al.,2005); se pueden elicitar distribuciones de probalidad para cantidades inciertas cuyo valor sea conocido por el experimentador con el objetivo de comparar las distribuciones de probabilidad estimadas con los datos observados y así proporcionar una medida objetiva de su exactitud.

## 2.4. Estructuración y descomposición

Durante esta etapa la cantidad incierta a elicitar debe ser claramente especificada, la escala de medida definida y el orden de las preguntas establecido para prevenir anclaje en las respuestas. En esta etapa es útil hacer uso de la prueba de clarividencia para determinar si una cantidad incierta está claramente definida, esta consiste en que el experto pueda ser capaz de dar una estimación de la cantidad incierta sin que la escala de medida halla sido especificada (Shephard y Kirkwood, 1994). También se debe explorar la comprensión de los expertos en la causalidad y las relaciones estadísticas entre las variables relevantes (Clemen y Reilly, 2004), esto se puede hacer por medio de mapas mentales o modelos conceptuales, que servirán para representar gráficamente las relaciones causales entre las variables, tambíen puede de ser de gran ayuda un software que construya gráficamente tales modelos y represente las ideas de los expertos (Devilee y Knol, 2011).

## 2.5. Entrenamiento en probabilidad

Pensar acerca de la variabilidad y la incertidumbre como probabilidades puede ser nuevo y bastante incómodo a los expertos (Walker et al., 2001), es por esto que una etapa importante a considerar dentro de un proceso de elicitación es el entrenamiento en probabilidad. Jenkinson (2005) propone que el entrenamiento en probabilidad debe estar compuesto de tres partes: Probabilidad y distribuciones de probabilidad; los conceptos que se aborden en este punto

dependerán del parámetro que se quiera elicitar, información sobre la heurística y sesgos más comunes, además de consejos de como superarlos y elicitaciones de práctica que sirvan como ejemplo, donde el facilitador conoce el verdadero valor del parámetro, pero el experto no. La sesión de entrenamiento debe ser dirigida por alguien con un profundo conocimiento y experiencia en el arte y la ciencia de los procesos de juicio de expertos formales (Bonano et al 1989).

### 2.5.1. Heurística y sesgos

Investigaciones como la presentada por Mosteller y Youtz (1990) ilustran que no es confiable dar evaluaciones de probabilidad precisas en muchos contextos. Una explicación a las deficiencias humanas para dar juicios acerca de probabilidades subjetivas, es que los seres humanos utilizan una serie de heurísticas para dar sus juicios, lo que puede dar lugar a un sesgo grave. Las heurísticas son herramientas para encontrar soluciones a los problemas rápidamente y con estas se pueden o no encontrar la mejor solución (Kynn, 2008). Trabajos acerca de este tema se remontan a la decada de 1970 donde Tversky y Kahneman (1974) propuesieron diferentes sesgos: representatividad, disponibilidad y anclaje y ajuste. Como se mencionó anteriormente la comprensión de la heurística y los sesgos es una herramienta importante en la preparación de una elicitación, por esto se detallará a continuación cada unos de los sesgos propuestos por (Tversky y Kahneman, 1974).

#### 2.5.1.1. Disponibilidad

Se refiere a la tendencia de juzgar la frecuencia de un evento por la facilidad de recordar ejemplos específicos; a sucesos que son más fáciles de recordar se tiende a dar más peso en la formación de juicios de probabilidad (Hora, 2007). Esta heurística se da por factores tales como la familiaridad, la relevancia y lo reciente, además los acontecimientos de interés periodístico también afectan de manera desproporcionada la memoria (Garthwaite et al., 2005). Un ejemplo de esto es cuando se pide estimar la frecuencia relativa de diferentes causas de muerte en los Estados Unidos, la muerte por arma de fuego se podría sobreestimar debido a que las muertes por esta cuasa son mas publicitadas y las muertes por accidente cerebrovascular podrían resultar subestimadas, dado que este tipo de muerte posiblemente se informa sólo si estan relacionadas con personajes reconocidos (Hora, 2007). Algunos de los sesgos provocados por esta heurística son la evaluación de riesgos desproporcionados (debido a la exposición a los resultados negativos, incluso si el evento es en sí poco común) y la correlación ilusoria (Kynn, 2008).

#### 2.5.1.2. Representatividad

Muchas de las preguntas de probabilidad son del tipo: ¿cuál es la probabilidad que un objeto pertenezca a la clase B?, ¿cuál es la probabilidad que el evento A se origine en el proceso B?, ¿cuál es la probabilidad que el proceso B generará el evento A?, para responder a este

tipo de preguntas, se recurriría normalmente a la heurística de la representatividad, en el que las probabilidades son evaluadas por el grado en el que A es representativa de B, es decir, por el grado en que A se asemeja a B. Cuando A es altamente representativa de B, la probabilidad de que A se origina desde B se juzga como alta. Por otra parte, si A no es similar a B, la probabilidad de que A se origina desde B se juzga como baja (Tversky and Kahneman, 1974). Para responder a este tipo de preguntas el experto tendría que recurrir a la evaluación de la probabilidad $P(A|B)$, pero comunmente se presta poca o ninguna atención a la probabilidad incondicional de B (Garthwaite et al., 2005). Un ejemplo de esta heurística es: si se pidiera indicar el hobby de Joe al que se describe como activo, agresivo y entusiasta, sería más probable que la respuesta fuera jugador de fútbol que escritor. Cuando se usa la heurística de la representatividad Joe es juzgado por su estereotipo mas cercano al de un jugador de futbol que el de un escritor (Wilson, 1994).

Tversky (1974) y Tversky y Kahneman (1974) estudiaron los posibles errores ocasionados por la aplicación de esta heurística: la ignorancia de tasa base, insensibilidad al tamaño de la muestra, la percepción errónea del azar y la aleatoriedad, y la confusión de $P(A|B)$ con $P(B|A)$.

#### 2.5.1.3. Anclaje y ajuste

Tal vez la heurística más usada para la evaluación de la probabilidad es el juicio mediante el anclaje y ajuste. Este sesgo cognitivo describe la tendencia humana común a confiar demasiado en la primera información que recibe para tomar decisiones. En el caso de la elicitación es común iniciar con una estimación y luego pedir al experto ajustar hacia arriba o hacia abajo, para dar la estimación final, pero en muchos casos los expertos tienden a quedarse muy cerca de esta estimación inicial, dando lugar a resultados que no reflejan lo suficientemente bien su conocimiento. El valor de partida, que generalmente se denomina el anclaje, podría ser sugerido por la naturaleza del problema, pero independientemente de la fuente del valor de partida, el ajuste es generalmente demasiado pequeño (Slovic, 1972). Este tipo de sesgo ocasiona que diferentes puntos de partida produzcan diferentes resultados (Tversky y Kahneman, 1974).
Tversky y Kahneman (1974), realizaron un experimento en el cual se pidió estimar diferentes cantidades expresadas en porcentajes (por ejemplo, el porcentaje de paises africanos en las naciones unidas). Se dieron diferentes valores iniciales elegidos al azar y primero se les pidió decidir si el valor que se les había dado era muy alto o muy bajo, y luego ajustarlo hasta llegar a su estimación deseada. Los sujetos a los que se les dio un punto de partida alto terminaron con estimaciones significativamente más altas que los sujetos a los que se dio un punto de partida bajo. Por ejemplo, las estimaciones de la mediana de la proporción de los países africanos en las Naciones Unidas fue del 25 % para los sujetos que recibieron 10 % como punto de partida y el 45 % para aquellos que recibieron el 65 % como punto de partida (Garthwaite et al., 2005).

## 2.6. Aplicación del método de elicitación

La función de una técnica de elicitación es extraer y cuantificar el juicio individual sobre cantidades inciertas y por lo tanto la selección de la técnica puede ser una decisión crucial, antecedentes de éxito de otros investigadores con diversas técnicas puede proporcionar una guía para la selección. El método de elicitación se debe seleccionar en función de su costo, el experto y la forma de su conocimiento en la materia, si un experto se siente familiarizado y cómodo con una técnica en particular sería una buena razón para elegir dicha técnica (Chesley, 1975). Un registro de la elicitación debería llevarse, idealmente para todas las preguntas que se formulen, junto con las respuestas de los expertos, así como el proceso por el cual una distribución de probabilidad se ajustó a esas respuestas, también es conveniente considerar un facilitador, este ayudará al experto en la formulación de sus conocimientos en forma probabilística y evitar los diferentes sesgos existentes (Garthwaite et al., 2005).

### 2.6.1. Técnicas de elicitación

Existen diferentes métodos de elicitación y se pueden clasificar en enfoques directo e indirecto (Kadane et al., 1980; Kadane, 1980); el enfoque directo, también llamado estructural, elicita información específica sobre los parámetros; se elicita directamente el conocimiento de un experto, es por esto que suele aplicarse en la elicitación de parámetros que se entienden de forma intuitiva como la media o la proporción, con esta técnica y a pesar de su naturaleza simple, se pueden producir los peores resultados, sobre todo cuando el experto no está familiarizado con los conceptos de probabilidad (Denham y Mengersen, 2007). El enfoque indirecto, también llamado predictivo, puede facilitar el proceso de elicitación de parámetros que requieran mayor conocimiento de la teoría de probabilidad (Correa, 2011), en este caso al experto se le pide cuantificar su opinión para los valores de la variable respuesta en varios valores fijos de la variable explicatoria (Denham y Mengersen, 2007). Es preciso resaltar que diferentes técnicas de elicitación pueden producir diferentes resultados porque el método usado puede afectar cómo el problema es visto, la exactitud y la consistencia en las respuestas (Winkler, 1967). Además algunos métodos son poco prácticos para ciertos problemas (Chesley, 1975). Sin embargo, un protocolo de elicitación no tiene que ser exclusivamente directo o indirecto, Kadane y Wolfson (1998) presentan ejemplos en los cuales se usó una combinación de los dos enfoques, para algunas variables se usó el enfoque predictivo y para otras el enfoque estructural.

Los métodos de elicitación directa son llamados así dado que la distribución elicitada es clara para el experto, los princiales métodos de elicitación directas son:

- **Función de distribución acumulada (CDF):** Para este método se hacen preguntas al experto del tipo ¿Qué porcentaje de valores del parámetro de interés están por debajo de un valor dado, digamos $q_i$?, se repite el procedimiento para diferentes valores de $q_i$ y con estos se ajusta un modelo paramétrico o un modelo empírico (Correa, 2011).

- **Función de densidad (PDF):** Con el método PDF se traza la función de densidad en vez de la distribición acumulada (Jenkinson, 2005). Aunque este procedimiento puede resultar un tanto más complejo que el anterior, puesto que determinar valores de densida puede no ser tan fácil y resultaría mejor trabajar con histogramas (Correa, 2011).

Los métodos de elicitación indirecta se introdujeron de forma independiente por Ramsey en 1931 y De Finetti en 1937. La principal característica de estos métodos es que no se tiene una relación definida entre la técnica de elicitación empleada y la distribución elicitada (Wikler, 1967), además son usadas en casos donde es difícil para el experto expresar su conocimiento en respuesta a preguntas directas acerca de un parámetro específico. Las principales técnicas de elicitación indirecta son:

- **Método de cuantiles o intervalo creíble:** Este método es comúnmente llamado así porque a un intervalo se le asigna una probabilidad, lo que se hace es preguntar al experto su estimación del parámetro $\pi$ y varios cuantiles de su distribución subjetiva para $\pi$, con estos se traza una distribución acumulada suave y esta será la representación paramétrica de la opinion del experto (Garthwaite et al, 2005).
- **Muestras hipotéticas futuras (HFS):** Una vez el experto ha hecho su evaluación de la cantidad elicitada, el método HFS examina el efecto que tendría un conocimiento adicional de una muestra aleatoria sobre su probabilidad original (Hampton et al, 1973). Por ejemplo, suponga que se quiere evaluar la proporción $\pi$ de estudiantes varones, se pueden realizar preguntas del tipo: "Suponga que se toma una muestra al azar de 100 estudiantes y 60 fueron hombres. Ahora ¿cuál es la probabilidad que un estudiante elegido al azar sea hombre?". Con este método el experto dará su estimación de la proporción en cuestión, basándose en diferentes muestras hipotéticas, luego se usará alguna forma de promediar dichas respuestas (Winkler, 1967). Con esta técnica se debe hallar el tamaño de muestra equivalente del sujeto elicitado, ya que este tamaño representa realmente el nivel de conocimiento acerca del parámetro de interés (Correa, 2011).
- **Información muestral a priori equivalente (EPS):** Este método también considera el efecto de información muestral. Se solicita al experto dar valores para $n$ y $r$, donde r sería el número de hombres elegidos al azar de una muestra de tamaño n, la razón $r/n$, deberá ser muy cercana al valor de $\pi$. Un mayor tamaño de muestra indica una mayor confianza en la estimación por parte del experto (Hampton et al, 1973).
- **Apuestas y loterías hipotéticas:** Una probabilidad puede asociarse con una apuesta y la cantidad que un sujeto esté dispuesto a arriesgar. Cuando se utilizan loterías en la elicitación de probabilidades, se busca que el experto pueda elegir entre dos loterías con igual rentabilidad. La probabilidades elicitadas se pueden ver afectadas por la función de utilidad del experto, y para esto se recomienda condicionar los intereses de los expertos elicitados (Kadene y Winkler, 1988).
- **Método de elicitación de la ruleta:** Este método consiste en asignar a las casillas de la ruleta intervalos o categorías y el experto distribuirá fichas entre estas casillas (bin), la probabilidad asignada a cada casilla se calculará por la proporción de fichas

asignadas a esta y finalmente la distribución de estas fichas en las casillas entregará una representación gráfica de las creencias del experto (Oakley et al, 2010).

Wikler (1967) reportó diversos grados de éxito con el uso de métodos directos e indirectos, además los sujetos elicitados determinaron la superioridad de los métodos indirectos en cuanto a la comprensión de las técnicas y facilidad de su uso (Hampton et al, 1973).

Mientras que los métodos EPS y HFS están restringidos generalmente a elicitar proporciones, los métodos CDF y PDF pueden usarse para elicitar la distribución de cualquier variable aleatoria continua. Los cuantiles a menudo se elicitan mediante la partición de subintervalos en otros de igual probabilidad. Este método de bisección es muy popular (Correa, 2011).

## 2.7. Realimentación y entrenamiento

La realimentación y el entrenamiento son muy útiles en un proceso de elicitación y hacen parte de las principales bases para mejorar las estimaciones de los expertos. La sobre-confianza, por ejemplo, persistiría por muchas repeticiones de elicitación hasta que el experto reciba realimentación, la práctica y la experiencia por sí solas no remueven los sesgos (Burgman et al., 2007). Jenkinson (2005) recomienda verificar la distribución elicitada con el experto, este es un proceso iterativo, donde se realimenta al experto ya sea con gráficos de distribuciones o resultados sobre las probabilidades que el dio, el experto debe revisar sus estimaciones hasta que sienta que esas declaraciones reflejan verdaderamente su opinión. La inclusión de esta etapa en el protocolo de elicitación ayuda a prevenir la distorsión de las creencias del experto. La realimentación parece ser siempre el enfoque más exitoso para ayudar a los expertos a mejorar la precisión en sus estimaciones. Las principales limitaciones son que generalmente se requiere de un gran número de repeticiones para generar información de calibración útil, y es menos adecuado para eventos que ocurren sola una vez (Burgman et al., 2007).

## 2.8. Elicitación de la distribución Beta

### 2.8.1. Distribución Beta

La distribución Beta es posible para una variable aleatoria continua que toma valores en el intervalo [0,1], lo que la hace muy apropiada para modelar proporciones. Por esta razón es una familia conjugada natural para la distribución Binomial. Esta distribución tiene dos parámetros, $\alpha$ y $\beta$.

Si se define una distribución apriori Beta para el parámetro $\pi$ de la distribución Binomial, se tiene que $\pi|\alpha,\beta \sim beta(\alpha,\beta)$

$$P(\pi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} \tag{2-1}$$

Siempre que se define una familia conjugada como una distribución apriori, la distribución posterior pertenece a la misma familia de distribuciones, por lo tanto, la distribución posterior para el parámetro $\pi$ es :

$$\pi|k,\alpha,\beta \sim Beta(\alpha+k,\beta+n-k) \tag{2-2}$$

Donde $k$ corresponde al número de éxitos en $n$ ensayos bernoulli.

Finalmente la estimación puntual de $\pi$ corresponde a la media de una distribución beta con parámetros $\alpha = \alpha + k$ y $\beta = \beta + n - k$:

$$E(\pi) = \frac{\alpha+k}{\alpha+\beta+n} \tag{2-3}$$

A continuación se describirán los métodos más usados para elicitar los parámetros $\alpha$ y $\beta$ de la distribución Beta. Madden y Hughes (2002) dan una recopilación de los siguientes métodos en un contexto de enfermades en plantas.

#### 2.8.1.1. Métodos discutidos por Pham-Gia, Turkkan y Duong (1962)

- El primero de estos dos métodos requiere una estimación de la media, $\hat{u}$, de la probalidad de éxito, pero la elección de la segunda cantidad es una estimación subjetiva de la desviación absoluta media alrededor de la media, denotada por $\delta_1(\pi)$. Soluciones numéricas de las siguientes ecuaciones dan la estimación de $\alpha$ y $\beta$

$$\hat{u} = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} \tag{2-4}$$

$$_1(\pi) = \frac{2\Gamma(\hat{\alpha}/\hat{\pi})\hat{\pi}^{\hat{\alpha}+1}(1-\hat{\pi})^{((1/\hat{\pi})-1)\hat{\alpha}}}{\Gamma(\hat{\alpha}+1)\Gamma((1/\hat{\pi})-1)\hat{\alpha}} \tag{2-5}$$

Donde $\Gamma(\alpha)$ es la función Gama.

- El segundo método se pide al experto dar una estimación de la mediana, $\hat{M}$, de su distribución de éxitos y una estimación de la desviación absoluta alrededor de la mediana, denotada por $\delta_2(\pi)$. La estimación de los parámetros son halladas resolviendo numéricamente las siguientes ecuaciones.

$$0{,}5 = \int_0^{\hat{m}} \frac{\pi^{(\hat{\alpha}-1)}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha},\hat{\beta})} d\pi \tag{2-6}$$

$$\delta_2(\pi) = \frac{2\hat{M}^{\hat{\alpha}}(1-\hat{M})^{\hat{\beta}}}{(\hat{\alpha}+\hat{\beta})B(\hat{\alpha},\hat{\beta})} \tag{2-7}$$

#### 2.8.1.2. Primer método de Weiler (1965)

Inicialmente se pide al experto dar una estimación subjetiva de la media $\hat{u}$, luego se pide al experto una probabilidad $v$ asociada a su creencia en que el verdadero valor de la media se encuentra en el intervalo $(2\hat{u}, 1)$. En este caso, se pide al experto estimar la probabilidad que el verdadero valor de la media sea más del doble del valor de la media estimada $\hat{u}$.

$$\hat{u} = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} \tag{2-8}$$

$$1 - v = \int_0^{2\hat{u}} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha},\hat{\beta})} d\pi \tag{2-9}$$

#### 2.8.1.3. Segundo método de Weiler(1965)

Se da al experto una probabilidad $v$ y se le pide estimar los valores $k_1$ y $k_2$ ($k_1 < k_2$), para los cuales la probabilidad que el verdadero valor de la media $u$ esté contenido en $(0,\ k_1)$ o en $(k_2,\ 1)$ sea igual. Las ecuaciones para los parámetros son:

$$u = \int_0^{k_1} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha},\hat{\beta})} d\pi \tag{2-10}$$

$$1 - v = \int_0^{k_2} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha},\hat{\beta})} d\pi \tag{2-11}$$

#### 2.8.1.4. Método de Fox (1966)

Se requiere que el experto de su estimación del valor modal ($\hat{m}$) de la probabilidad y su probabilidad subjetiva $v$ para el evento que la verdadera moda se encuentre en el intervalo ($\hat{m}$ - $k\hat{m}$, $\hat{m}$ + $k\hat{m}$ ), donde $0 < k < 1$ y es dada al experto por el elicitador. La estimación de los parámetros , $\hat{\alpha}$ y $\hat{\beta}$ se hace mediante dos ecuaciones, con la ecuación (2-13) se calcula la moda de una distribución beta a partir de sus parámetros y la ecuación (2-14) es la integral de la función de densidad sobre el intervalo dado al experto.

$$\hat{m} = \frac{\hat{\alpha} - 1}{\hat{\alpha} + \hat{\beta} - 2} \tag{2-12}$$

$$v = \int_{\hat{m}-k\hat{m}}^{\hat{m}+k\hat{m}} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha},\hat{\beta})} d\pi \tag{2-13}$$

Estas ecuaciones pueden ser resueltas usando la tabla de la función Beta incompleta, o Fox(1966) da una tabla de ($\alpha$,$\beta$) para los valores de ($m$,$v$,$k$).

#### 2.8.1.5. Método de Gross (1971)

Se requiere que el experto de una estimación de la media ($\hat{u}$) de su distribución de probabilidad y su probabilidad subjetiva $v$, para el evento de que la verdadera media se encuentre en el intervalo (0, $ku$), ($k$, es dada al experto), donde ($0 < k < 1$). Las ecuaciones requeridas son la fórmula para la media de la distribución beta y la integral de la densidad sobre el intervalo (0, $ku$).

$$\hat{u} = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} \tag{2-14}$$

$$v = \int_{0}^{k\hat{u}} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha},\hat{\beta})} d\pi \tag{2-15}$$

Igualmente estas ecuaciones pueden ser resueltas usando la tabla de la función Beta incompleta.

#### 2.8.1.6. Método de Waterman (1976)

Se pide al experto estimar, la media $u$, el percentil 5 y 95 de la distribución para $\pi$. Dicha distribución se estima mediante la siguiente ecuación:

$$I_x(x_0, n_0 - x_0) = \int_0^x \frac{\Gamma(n_0)}{\Gamma(n_0 - x_0)} \pi^{x_0-1}(1-\pi)^{n_0-x_0-1} d\pi \tag{2-16}$$

donde $0 < x < 1$

Con los valores elicitados de la media y un percentil podemos consultar el valor de $x_0$ y $n_0$ en las tablas dadas por Waterman et al. (1976) y, así se obtiene la distribución del parámetro $\pi$

#### 2.8.1.7. Método de Elicitación PM (Posterior Mode) (1983)

Método presentado en Charloner y Duncan en el cual se sigue el siguiente procedimiento para estimar los parámetos $\alpha$ y $\beta$:

1. Se especifica el número de ensayos ($n$) a considerar (en muchas aplicaciones n = 20 parece ser un número adecuado).
2. Para cada ensayo se pide el número más probable de éxtios (m) en $n$ ensayos.
3. Se retroalimenta al experto con la gráfica de la distribución Binomial($n$, $m/n$)
4. Se pregunta por los cambios $d_l$ y $d_u$ definidos como sigue:

$$d_l = \frac{p(m-1)}{p(m)}$$
$$d_u = \frac{p(m+1)}{p(m)}$$

   donde $p()$ es la distribución predictiva del experto elicitado.

5. Usando $d_l$ y $d_u$ y condicionando en $m$ resuelva para $\alpha$ y $\beta$

$$d_l = \frac{f(m-1)}{p(m)} = \frac{(n-m)(m+\alpha)}{(m+1)(n-m+\beta-1)}$$
$$d_u = \frac{f(m+1)}{p(m)} = \frac{(n-m+\beta)}{(n-m+1)(m+\alpha-1)}$$

   $f()$ es la función de densidad de la Beta-Binomial. Estas ecuaciones producen dos ecuaciones lineales en $\alpha$ y $\beta$. Llame a la solución de estas ecuaciones $\alpha_1$ y $\beta_1$.

- Se recalcula la moda de la distribución como $\gamma = \frac{\alpha_1 - 1}{\alpha_1 + \beta_1 - 2}$, luego se calcula el intervalo de predicción más pequeño de la distribución Beta-Binomial, con al menos del 50 % de probabilidad. Se debe preguntar al experto por este intervalo , si es muy grande h = -1 , si es adecuado h = 0 y es es muy pequeño h = 1. Y así los nuevos valores de $\alpha$ y $\beta$ son:

$$\alpha_{i+1} = 1 + 2^h(\alpha_i - 1)$$
$$\beta_{i+1} = 1 + 2^h(\beta_i - 1)$$

- Se debe repetir este ultimo paso hasta que $h = 0$
- Repetir el procedimiento para varios valores de $n$ y se combinan dichos resultados.

#### 2.8.1.8. Primer método de Duran y Booker (1988)

Se da al experto una probabilidad $u$ y, le pide estimar el valor de ¸$u$ tal que la probabilidad de que la verdadera media esté en el intervalo $(0, k_u)$ sea igual a $u$.

$$\hat{u} = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} \tag{2-17}$$

$$u = \int_0^{k_u} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha}, \hat{\beta})} d\pi \tag{2-18}$$

#### 2.8.1.9. Segundo método de Duran y Booker (1988)

Es similar al segundo método de Weiler, nuevamente dos valores $k_1$ y $k_2$ son elicitados, pero esta vez estas cantidades representan los extremos superiores de los intervalos tal que la probabilidad de que la verdadera media se encuentre en el intervalo (0,$k_1$) sea $u_1$ y para (0, $k_2$) sea $u_2$

$$u_1 = \int_0^{k_1} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha}, \hat{\beta})} d\pi \tag{2-19}$$

$$u_2 = \int_0^{k_2} \frac{\pi^{\hat{\alpha}-1}(1-\pi)^{\hat{\beta}-1}}{B(\hat{\alpha}, \hat{\beta})} d\pi \tag{2-20}$$

#### 2.8.1.10. Método de Gavasakar (1988)

Este método usa muestras hipotéticas futuras, se le pide al experto imaginar un conjunto de $n_0$ ensayos y dar la moda $m_0$ para el número de éxitos, luego se da al experto una muestra hipotética futura, donde se observaron $s_i$ éxitos en $k_i$ ensayos, y se le pide que imagine otros $n_i$ ensayos (Gavasakar sugiere usar $n_i = 20$) y de su número modal de éxitos, $m_i$. Esto se repite para $I$ muestras hipotéticas futuras. La estimaciones de $\alpha$ y $\beta$ se hallan por el método de mínimos cuadrados, minimizando la siguiente expresión:

$$\sum_{i=0}^{I} \left[ m_i - \left( \frac{(n_i + 1)(\hat{\alpha} + s_i)}{\hat{\alpha} + \hat{\beta} + k_i} - \frac{1}{2} \right) \right]^2 \tag{2-21}$$

donde $k_0 = s_0 = 0$

#### 2.8.1.11. Primer método de León, Vásquez y León (2003)

Este método consiste en tres pasos:

- Se pide al experto realizar una estimación de la media $\hat{u}$ y la moda $\hat{m}$.
- Con las estimaciones del punto anterior resolver para $\alpha$ y $\beta$ las siguientes ecuaciones:

$$\hat{u} = a + (b-a)\frac{\alpha}{\alpha+\beta} \tag{2-22}$$

$$\hat{m} = a + (b-a)\frac{\alpha-1}{\alpha+\beta-2} \tag{2-23}$$

  Donde $a$ y $b$ son los límites inferior y superior que definen el rango de complacencia a pagar según lo determine el experto.
- Mostrar al experto la forma de la distribución resultante de las estimaciones anteriores y validar si se siente cómodo con la distribución o desea realizar algún ajuste.
- Repetir 1-3 hata lograr que el experto este cómodo con la distribución resultante.

#### 2.8.1.12. Segungo método de León, Vásquez y León (2003)

Este método consiste en realizar los siguientes pasos:

- Se pide al experto la estimación de la media $\hat{u}$ y la moda $\hat{m}$ y tres cuantiles.
- Compruebe si el intervalo cerrado, definido por el primer cuartil $q_1$ y el tercer cuartil $q_3$ comprende una región de alta densidad de 50 % para una distribución beta con parámetros $\alpha = \beta = 1$.
- Si la condición en el paso anterior no se cumple, el parámetro $\alpha$ es incrementado en 0.01, y el correspondiente parámetro $\beta$ es generado con las relaciones:

$$\beta = (\alpha-1)\frac{b-a}{d-a} - \alpha + 2 \tag{2-24}$$

  Este paso se repite hasta que el parámetro $\alpha$ y $\beta$ satisfagan las siguientes dos ecuaciones:

$$F(q_2;\alpha,\beta) = 0{,}5 \qquad F(q_3;\alpha,\beta) = 0{,}75 \tag{2-25}$$

  Donde F es una función de distribución acumulada beta. Cuando se logra la convergencia, el intervalo $[q_1, q_3]$ define una región de alta densidad 50 % para los parámetros $(\hat{\alpha}, \hat{\beta})$

- Revisar la consistencia del primer cuantil, es decir que $q_1$ satisfaga $F(q_1, \hat{\alpha}, \hat{\beta}) \approx 0{,}25$ y $p = \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}$
- Si ya que el primer cuantil o la media obtenida a partir de la distribución elicitada se desvía mas de 30 % de lo especificado en el primer paso, se debe pedir al experto volver a evaluar las cantidades elicitadas, hasta conseguir consistencia.

#### 2.8.1.13. Método de aproximación a la distribución Normal (2012)

Método propuesto en Elfadaly y Garthwaite. Inicialmente se pide al experto estimar tres cuartiles, $q_{25}$, $q_{50}$ y $q_{75}$, estos cuartiles se transforman en dos valores iniciales de los parámetros de la distribución Beta:

$$\alpha = cq_50 + 0{,}25 \qquad \beta = c(1 - q_50) + 0{,}25 \tag{2-26}$$

Haciendo uso de la propiedad $Z_{75} - Z_{25} = 1{,}34896$ se tiene que :

$$c \cong \frac{1{,}34896}{4} \left\{ [q_{50}(1-q_{25})]^{1/2} - [q_{25}(1-q_{50})]^{1/2} + [q_{75}(1-q_{50})]^{1/2} [q_{50}(1-q_{75})]^2 \right\}^{-2} \tag{2-27}$$

$Z_{25}$ y $Z_{75}$ son los cuartiles inferior y superior de la distribución normal estándar.

Luego se aplica un método numérico de mínimos cuadrados sobre los valores iniciales de los parámetros $\alpha$ y $\beta$ para optimizarlos. Oakley (2010) y Elfadaly y Garthwaite (2012) reportan el enfoque de mínimos cuadrados como un método para escoger los parámetros que minimicen la función:

$$Q = [F(q_{25}, \alpha, \beta) - 0{,}25]^2 + [F(q_{50}, , \beta - 0{,}5]^2 + [F(q_{75}, \alpha, \beta - 0{,}75)]^2 \tag{2-28}$$

Donde $F(x, \alpha, \beta)$ es la cdf de una distribución Beta con parámetros $\alpha$ y $\beta$ en un punto $x$.

## 2.9. Distribución Normal Truncada

La distribución Beta es la conjugada natural de la distribución Binomial y por esto frecuentemente usada como distribución apriori de la, en algunas situaciones cuando las probabilidades de éxito son muy bajas o muy altas producen parámetros elicitados de la distribución Beta menores que 1 y en este caso esta distribución no es unimodal y

con colas pesadas. Bajo esta situación la distribución Normal truncada podría usarse y garantizaría la unimodalidad.

La distribución Normal truncada es particularmente popular en casos donde se requiere describir patrones no negativos y un límte superior también es necesario (aunque la distribución Beta es muy flexible).

Sea $X \sim N(\mu, \sigma^2)$ y su distribución condicional de $X \in [a, b] \subset \mathbb{R}$. La distribución condicional de $X$ sobre el intervalo $[a, b]$ es la distribución normal truncada. La densidad condicional es:

$$f(x|x \in [a,b]) = \frac{\frac{1}{\sigma}\phi\frac{x-\mu}{\sigma}}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{2-29}$$

Para $a \leq x \leq b$ donde $\phi$ y $\Phi$ representan la densidad y la CDF de una normal estándar respectivamente.

## 2.10. Regresión Logística

Modelo que describe la relación entre una variable respuesta binaria o dicotómica $Y$ y un conjunto de varibles independientes $X$ llamadas covariables. La forma específica del modelo de regresion logística es:

$$\pi(x) = \frac{e^(\beta_0 + \beta_1 x)}{1 + e^(\beta_0 + \beta_1 x)} \tag{2-30}$$

La transformación logit de $\pi(x)$ esta dada por:

$$\begin{aligned} g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ = \beta_0 + \beta_1 x \end{aligned} \tag{2-31}$$

La transformación logit, es lineal en sus parámetros y dependiendo del valor de $x$ puede ser continua y variar entre $-\infty$ y $\infty$. Una observación de la variable respuesta puede expresarse como $y = \pi(x) + \epsilon$. La cantidad $\epsilon$, puede tomar uno de dos posibles valores, si $y = 1$ luego $\epsilon = 1 - \pi(x)$ con probabilidad $\pi(x)$, y si $y = 0$ luego $\epsilon = -\pi(x)$ con probabilidad $1 - \pi(x)$. Esto es, la distribución condicional de la variable respuesta $y$, sigue una distribución binomial con probabilidad dada por la media condicional, $\pi(x)$

(Hosmer et., al, 2013).

y así:

$$p(y_i|X_i,\beta) = \frac{exp(X_i^T\beta)}{1+exp(X_i^T\beta)} \tag{2-32}$$

En el paradigma Bayesiano, el conocimiento de un experto es introducido especificando una distribución previa para los parámetros de regresión $\beta$ (James et al., 2010).

## 2.11. Elicitación en la regresión logística

Los primeros trabajos en elicitación para Modelos lineales generalizados (GLM) fueron propuestos por Bedrick et al. (1996, 1997), ellos propusieron un método de elicitación en el que la distribución predictiva es elicitada en diferentes puntos de diseño y luego combinada para formar una distribución apriori. Algunas formas especifícas de GLM, entre ellas la regresión logística han recibido especial atención e importantes desarrollos se han dado en el área de la ecología, ejemplo de esto, son los métodos presentados a continuación:

- Kynn (2005): Herramienta gráfica interactiva de metodología indirecta llamada elicitor (complemento del software WinBUGS). Se pregunta al experto por dos puntos cualquiera y la mediana, luego se grafica la relación univariante entre la variable respuesta y una covarible (manteniendo todas las otras constantes). Inicialmente se usó para elicitar distribuciones normales a priori para un modelo de regresión logística con el fin de estimar la presencia de especies en un ecosistema. Este método fué inspirado por Bedrick et al. (1996) y Garthwaite y Dickey (1988).
- Martin et al., (2005): Método directo para elicitar opinión de expertos, usando cuestionarios, a partir de multiples expertos, y en dicha ocasión sólo se consideró una covariable para la regresión Poisson. O'Leary et al. (2008) adaptó este enfoque para uno o múltiples expertos y múltiples covariables en el contexto de la regresión logística. Para cada covarible se preguntó a los expertos si el efecto sobre la variable respuesta incrementaba, disminuía o no existía. Este método no requiere conocimiento acerca de probabilidad o distribuciones.
- Garthwaite y Al-Awadhi (2006): Desarrollaron un método en el área de la ecología que modela la distribución muestral mediante un modelo de regresión logística continuo lineal por partes, que es más flexible que el modelo de regresión logística estándar, además se usaron gráficos interactivos para realimentar al experto.
- Denham y Mengersen (2007): Método indirecto para el modelamiento ambiental, este procedimiento hace uso de la naturaleza geográfica de estos problemas e incorpora un sistema de información geográfico (SIG) para suministrar información

acerca de la vegetación, tipos de rocas, precipitaciones, temperatura, etc. La elicitación de expertos en este caso se usó para relacionar todas estas variables con la probabilidad de presencia/ausencia de una especie en peligro de extinción. Durante este ejercicio en lugar de especificar puntos de diseño como números, cada punto de diseño fue una ubicación real en Queensland.

- James et al., (2010): Diseñaron el software elicitator e hicieron una aplicación a través de un estudio que tiene como objetivo desarrollar un modelo de regresión logística para predecir la distribución geográfica de una especie en un contexto ecológico. Esta herramienta extiende el trabajo hecho por Denham y Mengersen (2007), puesto que soporta una variedad de aplicaciones y usos, e igualmente se usó un método indirecto. Se pide al experto para cada caso $k$ con covariables $X_{1k}, X_{2k}, ..., X_{jk}$ conocidas, estimar la probabilidad de éxito $Z_k$, el rango de valores con probabilidad variable (percentiles) y su mejor estimación (moda). Esta información se utiliza para estimar numéricamente $\mu_k$ y $\gamma_k$ en $P(Z_k|x_k)$, posteriormente se proporciona realimentación al experto y se le da la oportunidad de modificar sus creencias. Esto se repite para $k = 1, 2, ..., K$. La información suministrada por el experto para todas las covaribles puede ser combinada para formar el modelo del experto y se utiliza una regresión beta para relacionar los datos del experto $Z_k$ a las covariables (Choy et al., 2009; James et al., 2010).

$$Z_K \sim Beta(\mu_k, \gamma_k), \quad logit(\mu_k) = x_k\beta \tag{2-33}$$

Por medio de la función "Link" el parámetro de forma $a_k$ y el parámetro escala $b_k$ para la probabilidad esperada de éxito es $\mu_k = a_k/\gamma_k$ y el tamaño de muestra efectivo del experto es $\gamma_k = a_k + b_k$.

La gran mayoría de los métodos mencionados en la recopilación anterior, coinciden en el uso de una metodología indirecta de elicitación, a excepción del método presentado por Martin et al., (2005). En el caso de un modelo de regresión, el uso de un método directo requeriría que el experto cuantificara el impacto de un cambio en el valor de la covariable sobre la varible respuesta, siendo aún más complicado el caso de la regresión logística puesto que esta relación no es lineal. En la práctica es poco probable que el experto sea capaz de hacer una estimación directa sobre los parámetros del modelo , incluso si estan bien informados de la relación que se está modelando (Huson y Kinnersley, 2008). Desde el punto de vista del modelista estadístico, un enfoque directo puede resultar más fácil, pero este podría producir resultados menos precisos en comparación con un enfoque indirecto, especialmente para expertos un poco ajenos a los conceptos de probabilidad. Un enfoque indirecto resulta más fácil para el experto; estos se sienten más cómodos estimando cantidades observables, que en un modelo de regresión equivale a una estimación de la variable respuesta, para diferentes valores de las covariables (Choy et al., 2009). Pero a menudo requiere más esfuerzo del modelista en el diseño del método de elicitación y la codificación para transformar respuestas de los expertos en la forma

requerida (James et al, 2010). Kadane y Wolfson (1998) recalcan que el objetivo de la elicitación es que sea lo más fácil posible para los expertos en la materia, en términos probabilísticos, al tiempo que se reduce la necesidad de un conocimiento acerca de la teoría de probabilidad.

# 3. Metodología propuesta

La propuesta de elicitación para determinar la apriori conjunta para el modelo de regresión logística se basa en el uso de la metodología indirecta de muestras hipotéticas. El modelo a elicitar es:

$$logit(\pi) = \beta_0 + \beta_1 x \tag{3-1}$$

Se pretende elicitar la distribución conjunta para $\beta_0$ y $\beta_1$, que finalmente será una distribución Normal bivarible.

## 3.1. Algoritmo propuesto

1. Se fijan los niveles de la covariable adecuados $x_1, x_2, ..., x_k$, estos puntos se deben elegir en consenso con el experto. Charloner y Larntz (1989) concluyen que para un modelo de regresión logística el número mínimo de puntos de diseño es igual número de parámetros a estimar.

   **Recomendación:** Los puntos de diseño deben ser tomados siguiendo las recomendaciones dadas a continuación, para evitar tomar puntos donde la probabilidad de éxito es muy cercana a 0 o muy cercana a 1.

   - El primer punto debe ser tomado de tal forma que corresponda a una probabilidad de éxito aproximadamente de 0.5, se debe seleccionar de tal manera que sea igualmente probable que el verdadero valor sea mayor o menor que este punto.
   - El segundo punto debe ser tomado de tal forma que la probabilidad de éxito corresponda aproximadamente al 0.25, se debe seleccionar de tal manera que si el valor verdadero está por debajo de la mediana, sea igualmente probable que sea por encima o por debajo de este valor.
   - El tercer punto debe ser tomado de tal forma que la probabilidad de éxito corresponga aproximadamente al 0.75, se debe seleccionar de tal manera que, si el valor verdadero está por encima de la mediana, es igualmente probable que sea por encima o por debajo de este valor.

2. Para cada nivel se procede así:

- Se fija un $n$ y se pide al experto dar el número de éxitos que el esperaría se den en una muestra hipotética de tamaño $n$, digamos $X_0$, calcule $E(\pi) = X_0/n$.
- Para el mismo $n$ se pide al experto dar el número máximo de éxitos que él esperaría aceptable, $X_L$, calcule $\pi_L = X_L/n$.
- Para el mismo $n$ se pide al experto dar el número mínimo aceptable, $X_I$, calcule $\pi_I = X_I/n$.

  Se puede repetir este paso las veces que se consideren necesarias, esto sirve para evaluar la consistencia del experto con las diferentes muestra hipotéticas.

3. A los valores elicitados en el punto 2. se ajusta una distribución Beta para estimar los parámetros $\alpha$ y $\beta$

   Sean:

   $E(\pi) = x_0/n$
   $P(\pi \geq X_L/n) = 0{,}05$
   $P(\pi \leq X_I/n) = 0{,}05$

   Los valores $\alpha$ y $\beta$ se obtienen de minimizar la siguiente función:

   $$f(\alpha,\beta) = (\pi_I - qbeta(0{,}05, \alpha, \beta))^2 + (\pi_L - qbeta(0{,}95, \alpha, \beta))^2 + (\pi + \alpha/(\alpha + \beta))^2$$

4. Calcule el N equivalente, esto permite cuantificar el conocimiento del experto en términos de tamaño muestral, este tamaño representa realmente el nivel de conocimiento sobre el parámetro que el experto tiene; donde un tamaño muestral pequeño indica un menor conocimiento y un tamaño muestral grande indica un mayor conocimiento (Sedlmeier, 1999). El N equivalente se halla usando la ecuación de un intervalo de confianza para la proporción basado en el teorema central del límite:

   $$\left(\hat{\pi} - Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}, \hat{\pi} + Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}\right) \tag{3-2}$$

   Donde $\hat{\pi}$ es dado por el experto como el número de éxitos más probable.

   Sean $a$ y $b$ el límite inferior y superior del intervalo de confianza respectivamente:

   $$a = \hat{\pi} - Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{3-3}$$

$$b = \hat{\pi} + Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{3-4}$$

Luego tomando la diferencia entre (4-3) y (4-4),

$$b - a = 2Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

$$N = \frac{4Z^2_{(\alpha/2)}\hat{\pi}(1-\hat{\pi})}{(b-a)^2} \tag{3-5}$$

Dado que se calcula un N equivalente por nivel de la covariable, el N equivalente del experto será el N del punto diseño que tenga asociado el menor valor.

5. Para cada nivel repita los siguientes pasos $m$ veces:

   - Genere un valor de la beta con $\alpha_i$ y $\beta_i$ hallados en el punto 3.

   $$\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{bmatrix}$$

   En este paso se obtiene un vector de tamaño k.

   - Genere una muestra de valores $y$ de la distribución binomial.

   $$\begin{bmatrix} y_1^{(1)}, & x_1 \\ y_2^{(1)}, & x_1 \\ \vdots & \\ y_{n_{eq}}^{(1)}, & x_1 \\ y_1^{(2)}, & x_2 \\ y_2^{(2)}, & x_2 \\ \vdots & \\ y_{n_{eq}}^{(2)}, & x_2 \\ \vdots & \\ y_1^{(k)}, & x_k \\ y_2^{(k)}, & x_k \\ \vdots & \\ y_{n_{eq}}^{(k)}, & x_k \end{bmatrix}$$

En este paso se obtiene una matriz de tamaño $(n_{eq} * k)$x2, Donde $n_{eq}$ es el N equivalente hallado en el punto 4 y $x_1, x_2, ...x_k$ son los niveles de la covarible.

- Con la tabla de datos construida en el punto anterior estime los parámetros de la regresión logística. Guarde los resultados.

  $betas^{(1)} < -glm(Y \sim X, family = "binomial")\$coef$

$$\begin{bmatrix} \beta_0^{(1)}, & \beta_1^{(1)} \\ \beta_{0,}^{(2)} & \beta_1^{(2)} \\ \vdots & \\ \beta_{0,}^{(m)} & \beta_1^{(m)} \end{bmatrix}$$

En este paso se obtiene una matriz de tamaño mx2.

6. Ajuste la normal bivariable a los betas hallados en el paso anterior.

## 3.2. Uso de la Distribución Normal trucada

En muchos casos donde la probabilidad de éxito es muy baja o muy alta, los parámetros de la distribución Beta están por debajo de 1 y en esta situación dicha distribución no es unimodal y tiene colas pesadas. En este caso se recomienda usar la distribución normal truncada entre [0,1] y aquí se garantizaría la unimodalidad. El algoritmo presentado a continuación es una modificación al presentado anteriormente y permitiría ajustar como distribución apriori para cada punto de diseño la Distribución Normal truncada.

1. Se fijan los niveles de la covariable adecuados $x_1, x_2, ..., x_k$. Charloner y Larntz (1986,1989) concluyen que para un modelo de regresión logística el número mínimo de puntos de diseño es igual número de parámetros a estimar.

   **Recomendación:** Los puntos de diseño deben ser tomados siguiendo las recomendaciones dadas a continuación, para evitar tomar puntos donde la probabilidad de éxito es muy cercana a 0 o muy cercana a 1.

   - El primer punto debe ser tomado de tal forma que corresponda a una probabilidad de éxito aproximadamente de 0.5.
   - El segundo punto debe ser tomado de tal forma que la probabilidad de éxito corresponda aproximadamente al 0.25
   - El tercer punto debe ser tomado de tal forma que la probabilidad de éxito corresponga aproximadamente al 0.75

2. Para cada nivel se procede así:

   - Se fija un $n$ y se pide al experto dar el número de éxitos que el esperaría se den en una muestra hipotética de tamaño $n$, digamos $X_0$, calcule $E(\pi) = X_0/n$.

- Para el mismo $n$ se pide al experto dar el número máximo de éxitos que él esperaría aceptable, $X_L$, calcule $\pi_L = X_L/n$.
- Para el mismo $n$ se pide al experto dar el número mínimo aceptable, $X_I$, calcule $\pi_I = X_I/n$.

  Se puede repetir este paso las veces que se consideren necesarias, esto sirve para evaluar la consistencia del experto con las diferentes muestra hipotéticas.

3. Los valores elicitados en el punto 2. permiten estimar los parámetros la media y la desviación típica de una distribución Normal:

   Sean:

   $E(\pi) = X_0/n$
   $P(\pi \leq X_L/n) = 0{,}95$
   $P(\pi \leq X_I/n) = 0{,}05$

   El valor de $\sigma$ se obtienen así:

   $dt1 = (\pi_I - \pi)/Q_{(0{,}05)}$
   $dt2 = (\pi_L - \pi)/Q_{(0{,}95)}$
   $\sigma = (dt1 + dt2)/2$

Donde $Q_{(0{,}05)}$ y $Q_{(0{,}95)}$ son cuantiles teóricos de la Distribución normal.

4. Calcule el N equivalente, esto permite cuantificar el conocimiento del experto en términos de tamaño muestral, este tamaño representa realmente el nivel de conocimiento sobre el parámetro que el experto tiene; donde un tamaño muestral pequeño indica un menor conocimiento y un tamaño muestral grande indica un mayor conocimiento (Sedlmeier, 1999). El N equivalente se halla usando la ecuación de un intervalo de confianza para la proporción basado en el teorema central del límite:

$$\left(\hat{\pi} - Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}, \hat{\pi} + Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}\right) \tag{3-6}$$

Donde $\hat{\pi}$ es dado por el experto como el número de éxitos más probable.

Sean $a$ y $b$ el límite inferior y superior del intervalo de confianza respectivamente:

$$a = \hat{\pi} - Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{3-7}$$

$$b = \hat{\pi} + Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \tag{3-8}$$

Luego tomando la diferencia entre (4-3) y (4-4),

$$b - a = 2Z_{(\alpha/2)}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}$$

$$N = \frac{4Z^2_{(\alpha/2)}\hat{\pi}(1-\hat{\pi})}{(b-a)^2} \tag{3-9}$$

Dado que se calcula un N equivalente por nivel de la covariable, el N equivalente del experto será el N del punto diseño que tenga asociado el menor valor.

5. Para cada nivel repita los siguientes pasos $m$ veces:

- Genere un valor de la Distribución Normal truncada en el intervalo $[0, 1]$ con $\mu_i$ y $\sigma_i$ hallados en el punto 3.

$$\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{bmatrix}$$

En este paso se obtiene un vector de tamaño k.

- Genere una muestra de valores $y$ de la distribución binomial.

$$\begin{bmatrix} y_1^{(1)}, & x_1 \\ y_2^{(1)}, & x_1 \\ \vdots & \\ y_{n_{eq}}^{(1)}, & x_1 \\ y_1^{(2)}, & x_2 \\ y_2^{(2)}, & x_2 \\ \vdots & \\ y_{n_{eq}}^{(2)}, & x_2 \\ \vdots & \\ y_1^{(k)}, & x_k \\ y_2^{(k)}, & x_k \\ \vdots & \\ y_{n_{eq}}^{(k)}, & x_k \end{bmatrix}$$

En este paso se obtiene una matriz de tamaño $(n_{eq} * k)$x2, Donde $n_{eq}$ es el N equivalente hallado en el punto 4 y $x_1, x_2, ...x_k$ son los niveles de la covarible.

- Con la tabla de datos construida en el punto anterior estime los parámetros de la regresión logística. Guarde los resultados.

  $betas^{(1)} < -glm(Y \sim X, family = "binomial")\$coef$

$$\begin{bmatrix} \beta_{0}^{(1)}, & \beta_{1}^{(1)} \\ \beta_{0,}^{(2)} & \beta_{1}^{(2)} \\ \vdots & \\ \beta_{0,}^{(m)} & \beta_{1}^{(m)} \end{bmatrix}$$

En este paso se obtiene una matriz de tamaño mx2.

6. Ajuste la normal bivariable a los betas hallados en el paso anterior.

# 4. Aplicación

## Modelo logístico para el cáncer de próstata en Colombia

## 4.1. Introducción

Desde 1990 el cáncer de próstata en Colombia ha venido en aumento; entre 1990 y 2013 el número de nuevos casos de tumores malignos de próstata al año pasó de 3.200 a 13.200, así lo reveló el estudio ´´La carga mundial del cáncer 2013" elaborado por el consorcio internacional de investigadores del instituto para medición y evalucación de la salud.

Es importante entender que este tipo de tumor está relacionado con el envejecimiento, es decir, a mayor edad, mayor probabilidad de desarrollarlo, por otra parte, el cáncer de próstata es más frecuente en los hombres afroamericanos que en los blancos; los hombres de raza negra presentan un mayor riesgo de padecer este tipo de cáncer que los de raza blanca. También tienen más probabilidades de desarrollar cáncer de próstata a una edad más temprana y de tener tumores agresivos, de crecimiento rápido. Se desconocen los motivos exactos de estas diferencias, los cuales pueden estar vinculados con factores socioeconómicos y de otros tipos. Los hombres hispanos tienen un menor riesgo de desarrollar cáncer de próstata y de morir por la enfermedad que los hombres de raza blanca. El cáncer de próstata se produce con más frecuencia en América del Norte y el norte de Europa. También parece que el cáncer de próstata está aumentando entre los asiáticos que viven en áreas urbanizadas, como Hong Kong, Singapur, y ciudades de América del Norte y de Europa, particularmente, entre aquellos que llevan un estilo de vida más occidental.

En este proyecto de tesis se aplica el método de elicitacón propuesto a la relación que existe entre el cáncer de próstata y la edad en hombres de raza negra, además se pretende determinar la prevalencia de este tipo de cáncer por edad. Conocer la prevalencia del cáncer de próstata por grupos de edad es importante, debido a que en personas de menor edad el diagnóstico puede ser mas tardío, puesto que se tiene la concepción de que este tipo de cáncer se presenta con mayor frecuencia sólo en hombres de edad avanzada y como se comentó anteriormente este supuesto puede no cumplirse para hombres de raza negra, además en algunos casos este tipo de cáncer no presenta síntomas muy evidentes

y un hombre aproximadamente por debajo de los 50 años probablemente no se haga chequeos rutinarios.

## 4.2. Metodología

En este proceso de elicitación se siguieron las etapas descritas en el capítulo 2 recopiladas por Jenkinson (2005) para lograr un proceso de elicitación éxitoso:

- Inicialmente se dío al experto una contextualización en el tema, donde se presentó el objetivo de la investigación, se le explicó que todas las preguntas estarían basadas en muestras hipotéticas de hombres de raza negra de diferentes edades, además se validó que estas fueran entendidas, que estuviera en capacidad y se sintiera cómodo dando este tipo de información.
- Como experto se tiene al Doctor Manuel García profesor de la Universidad Sur Colombiana que cuenta con investigaciones en biología prostática y biología de la reproducción y además un Posdoctorado de la universidad de Sao Paulo el cual titula "Analisis de microRNA que regula el Receptor de Androgeno en el cáncer de próstata".
- En la etapa de estructuración descomposición y entrenamiento en probabilidad, se le da al experto una introducción sobre la distribución binomial y su relación con el modelo logístico, información acerca de las heurísticas y sesgos a los cuales se tendría que enfrentar. Adicionalmente se le especifica al experto que la cantidad incierta que se quiere elicitar es el número de hombres de raza negra con cáncer de próstata en diferentes niveles de edad.
- Aplicación del método:
  1. Se pide al experto dar un intervalo de la edad en el cual sea de interés conocer la prevalencia del cáncer de próstata y además ubicarse en puntos cercanos al cuantil 25, 50 y 75. Luego de llegar a un consenso con el experto se eligieron los niveles de edad: 50, 60, 65, 70. El experto dío cuatro puntos y aunque para estimar los parámetros de una regresión logística con una sóla covarible son suficientes dos puntos, se decidío elicitar en estos cuatros puntos dado que el proceso de elicitación es de fácil aplicación.
  2. Para cada uno de los niveles de edad hallados en el punto anterior se da al experto diferentes muestras hipotéticas de hombres de raza negra y se le pide dar el número de casos de cáncer de prostata que el espera encontrar y el número mínimo y máximo de casos que el considera aceptable. Se repite este proceso 3 veces con 3 diferentes muestra hipotéticas, se calculan las proporciones en cada nivel de la edad y se verifica que el experto halla sido consistente con sus respuestas.
  3. Se calcula el N equivalente del experto, reemplazando (4-5) los valores elicitados en el punto anterior. Se obtuvo un N equivalente de 203.

4. En esta etapa se realiza la simulación y se estiman los parámetros $\beta_0$, $\beta_1$ del modelo logístico para la prevalencia de cáncer de próstata en hombres de raza negra en relación con la edad.

- Una vez se estimaron los coeficientes del modelo, se mostraron al experto y se le explicó su significado e implicaciones, con el fin el de validar si dichos resultados reflejan sus creencias. En caso de que el experto estuviera de acuerdo con el modelo se daba por finalizado el proceso, en caso contrario, se ajustarían los casos de cáncer de próstata para las diferentes muestras hipotéticas.

## 4.3. Resultados

Las proporciones halladas para cada nivel de la edad son:

| Edad | $x_0$ | x | $x_L$ |
|---|---|---|---|
| 50 | 0.02 | 0.05 | 0.08 |
| 60 | 0.03 | 0.08 | 0.1 |
| 65 | 0.13 | 0.17 | 0.22 |
| 70 | 0.34 | 0.38 | 0.42 |

**Tabla 4-1**.: Proporciones elicitadas.

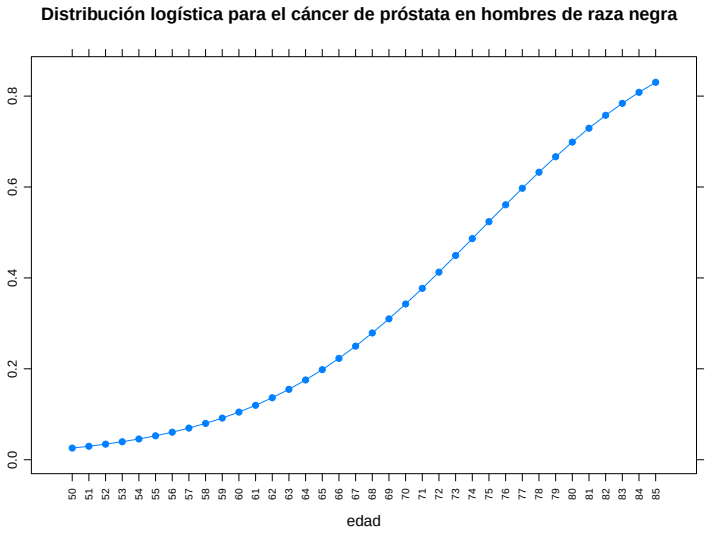
La estimación de los parámetros para el modelo logístico junto con su intervalo de probabilidad son:

| Parámetro | Estimación | LI | LS |
|---|---|---|---|
| $\beta_0$ | -11.104 | -12.449 | 7.430 |
| $\beta_1$ | 0.149 | 0.126 | 0.220 |

**Tabla 4-2**.: parámetros estimados

Así el modelo estimados es:

$$ln(\pi/1-\pi) = -11{,}104 + 0{,}149 * Edad$$




Del signo de $\beta_1$ se puede concluir que la edad es un factor de riesgo para el cáncer de próstata; al aumentar la edad aumenta la probabilidad de sufrir la enfermedad. Conocer como es el comportamiento de la prevalencia de la enfermedad por edad es importante a la hora de diseñar políticas de prevención, como se ve puede ver en el gráfico anterior a la edad de 60 años la prevalencia de cáncer de próstata empieza a incrementar más rápidamente, es por esto que a esta edad se deben incrementar los chequeos rutinarios para una detección temprana de la enfermedad. Existen diferentes estudios en los que se ha mostrado la relación de esta enfermedad con la edad pero para poblaciones en general, si bien se conoce que es más probable en hombres de raza negra no hay muchos datos que indiquen a que edad se deben incrementar los chequeos para una detección temprana.

# 5. Conclusiones

## 5.1. Conclusiones

- En esta tesis se estudiaron los diferentes puntos a considerar en el momento de afrontar un proceso de elicitación, con el fin de desarrollar un método que permitiera estimar los parámetros $\beta_0$, $\beta_1$ de la regresión logística con una sola covariable. El uso de un método indirecto es una alternativa amigable para el experto y aunque este tipo de metodología puede resultar un poco más complicada para el analista estadístico, el uso de muestras hipotéticas en particular disminuye este inconveniente, aunque el principal objetivo sea la fácilidad para el experto, estos dos puntos se traducen en un método de elicitación de fácil aplicación y de mayor facilidad para generar resultados. Desde este punto de vista es un aporte importante al área debido a que se hace uso de una distribucón apriori informativa y este método de elicitación no involucra apuestas.
- El uso de muestras hipotéticas como metodología de elicitación exige el cálculo de un N equivalente, para esta tesis se propuso el uso de la ecuación del intervalo de confianza para una proporción y de allí despejar el valor de n, en trabajos posteriores se podrían explorar diferentes aproximaciones para este cálculo.
- Los cálculos y resultados para el método de elicitación que se presenta en este proyecto se hicieron en el software estadístico R, la realimentación al experto se hizo mediante gráficos y resultados procesados en dicho software. Para posteriores desarrollos sería de gran ayuda implementar un aplicativo como el construido por Flórez (2015) y adaptarlo al método propuesto aquí. Esto permitiría capturar y almacenar información, además de realimentar en tiempo real a los expertos.
- Un trabajo posterior podría incluir la comparación de este método, se podría tomar en consideración el trabajo de tesis doctoral de Barrera (2015), en este trabajo se hace una comparación objetiva de dos métodos de elicitación garantizando que todos los expertos reciben la misma cantidad de información.

# A. Apéndice: Códigos

## A.1. Código: Uso de la distribución Beta como distribución apriori

```
# metodologia propuesta
library(XLConnect)
library(RODBC)
#
temp <- odbcConnectExcel2007("DatosElicitación.xlsx")
# Se leen los puntos de diseño dados por el experto
puntos <- sqlFetch(temp, "Puntos")
odbcCloseAll()
# Calculo n equivalente
n1 <- NULL
n_equ <- function(x){
for(i in 1:nrow(x)){
n <- round((4*1.96^2*x[i,3]*(1 - x[i,3]))/(x[i,4]-x[i,2])^2)
n1 <- rbind(n1,n)}
return(n1)
}
(n_exp <- min(n_equ(puntos)))
#
#
# ajuste de la distribución beta
ajuste.beta <- function(teta,valores= valores){
alfa<-teta[1]
beta<-teta[2]
cuantil0.05<-valores[1]
cuantil0.95<-valores[3]
media<-valores[2]
cuant1.teo<-qbeta(0.05,alfa,beta)
cuant2.teo<-qbeta(0.95,alfa,beta)
media.teo<-alfa/(alfa+beta)
res<-(cuantil0.05-cuant1.teo)^2
```

```
+(cuantil0.95-cuant2.teo)^2
+(media-media.teo)^2
return(res)
}
#
alfas1 <- NULL
par_betas <- function(x){
for(i in 1:nrow(x)){
valores <- x[i,2:4]
alfas <- optim(c(1,1),ajuste.beta,method="L-BFGS-B",
lower=c(1,1)/1000000,upper=c(10,10),
valores = valores)[[1]]
alfas1 <- rbind(alfas1,alfas)}
return(alfas1)
}
parametros_beta <- matrix(unlist(par_betas(x = puntos)), ncol = 2, byrow = F)
colnames(parametros_beta) <- c("alfa","beta")
#
puntos1 <- cbind(puntos,parametros_beta)
#
#
beta1 <- NULL
const1 <- NULL
for(i in 1:1000){
pi1 <- rbeta(n = 1, puntos1[1,5], puntos1[1,6])
pi2 <- rbeta(n = 1, puntos1[2,5], puntos1[2,6])
pi3 <- rbeta(n = 1, puntos1[3,5], puntos1[3,6])
pi4 <- rbeta(n = 1, puntos1[4,5], puntos1[4,6])

#
y1 <- sample(c(0,1), prob = c(pi1,1-pi1), n_exp, replace = T)
y2 <- sample(c(0,1), prob = c(pi2,1-pi2), n_exp, replace = T)
y3 <- sample(c(0,1), prob = c(pi3,1-pi3), n_exp, replace = T)
y4 <- sample(c(0,1), prob = c(pi4,1-pi4), n_exp, replace = T)

#
Y <- c(y1,y2,y3,y4) # se adicionan tanto y's como puntos de diseño
X <- c(rep(puntos1[1,1],n_exp),rep(puntos1[2,1],n_exp),
       rep(puntos1[3,1],n_exp),rep(puntos1[4,1],n_exp))
const <- glm(Y~X, family = "binomial")$coef[1]
beta <- glm(Y~X, family = "binomial")$coef[2]
#
```

```
#
const1 <- c(const1,const)
beta1 <- c(beta1,beta)
}
#
# Ajuste de la distribución normal multivariada
 a los parámetros Alfa y Beta hallados
library(MASS)
fitdistr(beta1,"normal")
fitdistr(const1,"normal")
#
histogram(beta1,xlab = "", ylab = "",
main = expression(paste('distribución del parámetro',sep = " ",beta,'1')))
#
histogram(const1, xlab = "", ylab = "" ,
main = expression(paste('distribución del parámetro',sep = " ",beta,'0')))
#
```

## A.2. Código: Uso de la distribución Normal truncada como distribución apriori

```
# metodologia propuesta
#
library(XLConnect)
library(RODBC)
# se leen los puntos de diseño dados por el experto
temp <- odbcConnectExcel2007("DatosElicitación.xlsx")
puntos <- sqlFetch(temp, "Puntos")
odbcCloseAll()
#
# Calculo n equivalente
n1 <- NULL
n_equ <- function(x){
for(i in 1:nrow(x)){
n <- round((4*1.96^2*x[i,3]*(1 - x[i,3]))/(x[i,4]-x[i,2])^2)
n1 <- rbind(n1,n)}
return(n1)
}
(n_exp <- min(n_equ(puntos)))
#
#
```

```
ajuste.normal <- function(valores){
media<-valores[2]
cuantil0.05<-valores[1]
cuantil0.95<-valores[3]
#
# necesitamos calcular la desviación típica
dt1<-(cuantil0.05-media)/qnorm(0.05)
dt2<-(cuantil0.95-media)/qnorm(0.95)

desvi.tip<-(dt1+dt2)/2

return(c(media,desvi.tip))
}
param1 <- NULL
for(i in 1:nrow(puntos)){
valores <- puntos[i,2:4]
param <- ajuste.normal(valores)
param1 <- c(param1,param)
}
param1 <- matrix(unlist(param1), ncol = 2, byrow = T)
colnames(param1) <- c("media","desv")
puntos1 <- cbind(puntos,param1)
#
#
beta1 <- NULL
const1 <- NULL
for(i in 1:10000){
pi1 <- qnorm(runif(1,pnorm(0,mean = puntos1[1,5],sd = puntos1[1,6]),
pnorm(1, mean =puntos1[1,5] ,sd = puntos1[1,6])),
mean = puntos1[1,5],sd = puntos1[1,6])
#
pi2 <- qnorm(runif(1,pnorm(0,mean = puntos1[2,5],sd = puntos1[2,6]),
pnorm(1, mean =puntos1[2,5] ,sd = puntos1[2,6])),
mean = puntos1[2,5],sd = puntos1[2,6])
#
pi3 <- qnorm(runif(1,pnorm(0,mean = puntos1[3,5],sd = puntos1[3,6]),
pnorm(1, mean =puntos1[3,5] ,sd = puntos1[3,6])),
mean = puntos1[3,5],sd = puntos1[3,6])
#
pi4 <- qnorm(runif(1,pnorm(0,mean = puntos1[4,5],sd = puntos1[4,6]),
pnorm(1, mean =puntos1[4,5] ,sd = puntos1[4,6])),
mean = puntos1[4,5],sd = puntos1[4,6])
```

```
#
y1 <- sample(c(0,1), prob = c(1-pi1,pi1), n_exp, replace = T)
y2 <- sample(c(0,1), prob = c(1-pi2,pi2), n_exp, replace = T)
y3 <- sample(c(0,1), prob = c(1-pi3,pi3), n_exp, replace = T)
y4 <- sample(c(0,1), prob = c(1-pi4,pi4), n_exp, replace = T)
#
Y <- c(y1,y2,y3,y4) # se adicionan tanto y's como puntos de diseño
X <- c(rep(puntos1[1,1],n_exp),rep(puntos1[2,1],n_exp),
      rep(puntos1[3,1],n_exp),rep(puntos1[4,1],n_exp))
#
const <- glm(Y~X, family = "binomial")$coef[1]
beta <- glm(Y~X, family = "binomial")$coef[2]
#
#
const1 <- c(const1,const)
beta1 <- c(beta1,beta)
}
#
quantile(beta1, pr = c(0.25,0.975))
quantile(const1, pr = c(0.25,0.975))
#
library(MASS)
fitdistr(beta1,"normal")
fitdistr(const1,"normal")
#
histogram(beta1,  xlab = "", ylab = "",
main = expression(paste('distribución del parámetro',sep = " ",beta,'1')))
#
histogram(const1, xlab = "", ylab = "" ,
main = expression(paste('distribución del parámetro',sep = " ",beta,'0')))
#
0.1171922 + 2*0.01337853
# intervalo de confianza para b1
0.1493276076 + 1.96 *(0.0328542429/sqrt(n_exp))
0.1493276076 - 1.96 *(0.0328542429/sqrt(n_exp))
#
# para b0
pii <- NULL
edad <- 50:85
for(i in edad) {
```

```
pi <- exp(-11.10430093 + 0.1493276076*i)/(1+exp(-11.10430093 + 0.1493276076*i))
pii <- c(pii,pi)
}
pronosticos <-  unlist(pii)
pronosticos <- cbind(edad, pi = pronosticos)
plot(pronosticos[,2], ylab = "",xlab = "Edad",type = "l",
main = "Probabilidad para el cáncer de próstata \n en hombres de raza negra",
axes = "F")
axis(1, 1:36, 50:85, cex =0.6 )
axis(2)
box()
xyplot(pronosticos[,2] ~ edad, type = c("b"),
main = "Distribución logística para el cáncer de próstata
        en hombres de raza negra",
ylab = "" ,pch = 19,
scale = list(rot = 90,x = list(at = 50:85, labels = 50:85,cex = .7)) )
```

# Bibliografía


[1] Barrera C., 2015, Analysis of the elicited prior distributions using tools of functional data analysis *Tesis Doctoral,* Universidad Nacional de Colombia.

[2] Bedrick E., Christensen R., Johnson W., 1996, A new perspective on priors for generalized linear models, *Journal of the American Statistical Association,* Vol 91, 1450-1460.

[3] Bedrick E., Christensen R., Johnson W., 1997, Bayesian Binomial Regression: Predicting Survival at a Trauma Center, *Journal of the American Statistical Association,* Vol 51, 211-218 No. 3.

[4] Bonano E., Hora S., Keeney R., von Winterfeldt D., 1989, Elicitation and Use of Expert Judgment in Performance Assessment for High-Level Radioactive Waste Repositories., *Sandia Report,* NUREG/CR-5411, No. SAND89-1821.

[5] Burgman M., Fidler F., McBride M., Walshe T., Wintle B., 2007, Eliciting Expert Judgments: Literature Review, *University of Melbourne,* Round 1, Project 11.

[6] Cannell C., 1977, A summary of studies of interviewing methodology, *Vital and Health Statistics,* Vol. 2, NO. 69

[7] Chaloner K., Duncan T., 1983, Assessment of a Beta Prior Distribution: PM Elicitation, *The Statistician*, Vol 27, 174-180.

[8] Chaloner K., Larntz K., 1989, Optimal Bayesian Design Applied to Logistic Regression Experiments, *Journal of Statistical Planning and Inference*, Vol 21, 191-208.

[9] Chesley G., 1975, Elicitation of Sub active Probabilities: A Review, *The Accounting Review*, Vol 50, 325-337, No. 2.

[10] Choy L., James, A., Mengersen, K., 2009, Expert elicitation and its interface with technology: a review with a view to designing Elicitator, *The Accounting Review*, 18th World IMACS / MODSIM Congress, Cairns, Australia, 13-17 julio del 2009.

[11] Clemen R., Reilly T., 2004, Making Hard Decisions with DecisionTools, *Duxberry*, 18th World IMACS. /

[12] Correa J., 2011, Elementos de Estadística Bayesiana, *Notas de clase*

[13] De Finetti B., 1937,). La Prevision: ses lois logigues, ses sources subjectives, *Annal es de l'Institut Henri Poincard* , vol 7, 1-68.

[14] De Finetti B., 1974, Theory of Probability, *Wiley*, vol 1.

[15] Denham R., Mengersen K., 2007, Geographically assisted elicitation of expert opinion for regression models, *Bayesian Analysis,* Vol 2, 99-136, No. 1.

[16] Devilee J., Knol A., 2011, Software to support expert elicitation: An exploratory study of existing software packages, *National Intitute for Public Health and the Environment, Ministry of Health, Welfare and Sport,*RIVM Letter Report 630003001.

[17] Duran B., Booker J., 1988, A Bayes Sensitivity Analysis when Using the Beta Distribution as a Prior, *IEEE Transactions on Reliability*, Vol 37, 239-247, No. 2.

[18] Elfadaly F., Garthwaite P., 2012, On Eliciting Some Prior Distributions for Multinomial Models, *Department of Mathematics and Statistics,* The Open University, UK.

[19] Flórez A., 2015, Eliciación de una distribución subjetiva del vector de parámetros $\pi$ de la distribución Multinomial, *Tesis de maestría,* Universidad Nacional de Colombia.

[20] Fox B., 1966, A Bayesian Approach to Reliability Assessment, *Memorandum RM-5084-NASA,* The Rand Corporation, Santa Monica, CA, 23 pp.

[21] Garthwaite P., Dickey J., 1988, Quantifying expert opinion in linear regression problems, *Journal of the Royal Statistical Society,* Vol 29, 462-474.

[22] Garthwaite P., Kadane ., O'Hagan A., 2005, Statistical Methods for Eliciting Probability Distributions, *Journal of the American Statistical Association,* Vol 100, 680-700, No. 470.

[23] Garthwaite P., Al-Awadhi S., 2006, Quantifying Opinion About a Logistic Regression Using Interactive Graphics, *Statistics Group,* Vol 6.

[24] Gavasakar U., 1988, A comparison of two elicitation methods for a prior distribution for a Binomial parameter, *Management Science,* Vol 134, 784-79.

[25] Goossens L., Cooke R., Hale A., Rodic-Wiersma., 2007, Fifteen Years of Expert Judgement at TUDelft, *Safety Science*, Vol. 46, 234-244.

[26] Gross A., 1971, The Application of Exponential Smoothing to Reliability, *Technometrics,* Vol 13, 877-883, No. 4.

[27] Hamada M., Martz H. F., Reese C. S., Wilson A. G., 2001, Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms, *The American Statistician,* Vol. 55, 175-181, No. 3

[28] Hampton M., Moore P., Thomas H., 1973, Subjective Probability and Its Measurement, *Journal of the Royal Statistical Society,* Vol. 136, 21-42, No.1

[29] Hogarth RM., 1975, Cognitive process and the assessment of subjective probability distributions, *Journal of the American statistical association,* Vol. 70.

[30] Hora S., Winterfeldt D., 1997, Nuclear waste and future societies: A look into the deep future, *Technological Forecasting and Social Change,* Vol. 56, 155-170.

[31] Hora S., 2007, Advances in Desicion Analysis: From Foundations to Applications, *Cambridge University Press,*129-153.

[32] Hosmer D., Lemeshow S., Sturdivant R., 2013, Applied Logistic Regression, *Wiley*, vol 3.

[33] Huson LW., Kinnersley N, 2008, Bayesian fitting of a logistic dose-response curve with numerically derived priors, *John Wiley  Sons Inc,*

[34] James A., Low Choy S. y Mengersen K., 2010, Elicitator: an expert elicitation tool for regression in ecology, *Environmental Modelling  Software,* Vol. 25, 129-145, No. 1.

[35] Jekinson D., 2005, The Elicitation of Probabilities - A Review of the Statistical Literature

[36] Kadane J., 1980, Predictive and structural method for elicitating prior distributions. In Bayesian Analysis in Econometrics and Statistics: Essays in Honour of Harold Jeffreys (A. Zellner, ed.) 89-93

[37] Kadane J., Dickey J., Winkler R., Smith W., Peters S., 1980, Interactive Elicitation of opinion for a Normal Linear Model, *Journal of the American Statistical Association,* Vol. 75, 845-854, No.372

[38] Kadane J., Wolfson L., 1998, Experiencies in Elicitation, *he Statistician,* Vol. 47, 3-19, No. 1

[39] Kadane J., Winkler R., 1988, Separating Probability Elicitation From Utilities, *Journal of the American Statistical Association,* Vol. 83, 357-363, No.402

[40] Koehler A., 2006,Comment: Expert Elicitation for Reliable System Design, *JStatistical Science,* Vol. 21, 454-455, No. 4

[41] Kynn M, 2005, Designing ELICITOR: Software to graphically elicit expert priors for logistic regression models in ecology,*Department of Mathematics and Statistics, Fylde College, Lancaster University.*

[42] Kynn M., 2008, The Heuristics and Biases Bias in Expert Elicitation,*Journal of the Royal Statistical Society,* Vol 171, 239-264, No 1.

[43] León C., Vázquez F., León C., 2003, Elicitation of Expert Opinion in Benefit Transfer of EnvironmentalGoods,*Environmental and Resource Economics,* Vol 26, 199-210

[44] Lichtenstein S., Fischhoff B., Phillips L., 1980, Calibration of Probabilities: The State of the Art to 1980, *Decision Research a Branch of Perceptronics,* Vol 62, No. 776-800.

[45] Madden LV., Hughes G., 2002,Plants epidemic, models and analysis, *Wiley.*

[46] Martin T., Kuhnert P., Mengersen K, Possingham H., 2005, The power of expert opinion in ecological models: a Bayesian approach examining the impact of livestock grazing on birds, *Ecological Applications,* Vol 15, 266-280.

[47] Oakley J., Daneshkhah A., O'Hagan A., 2010, Nonparametric Prior Elicitation using the RouletteMethod, *School of Mathematics and Statistics, University of Sheffeld, UK*

[48] O'Leary R., Low Choy S., Mengersen K., Kynn M., Kuhnert P., Denham R., Martin T., Murray J., Jarman P., 2008, Comparison of expert elicitation methods for logistic regression for presence of endangered brush-tailed rock-wallaby Petrolage penicillata , *Environmetrics*

[49] Phillips LD., 1999, Group elicitation of probability distributions: are many heads better than one?, *in Decision Science and Technology: Reflections on the contributions of Ward Edwards*ed. Shanteau, J., Mellers, B. and Schum, D.

[50] Raiffa H., Schlaifer R., 1964, Applied Statistical Theory, *Harvard University Press: Boston.*

[51] Ramsey F., 1931, General propositions and causality. In The Foundations of Mathematics and Other Logical Essays, *Canadian Journal of Philosophy,* vol 10, 497-511.

[52] Savage L., 1971, Elicitation of Personal Probabilities and Expectations, *Journal of the American Statistical Association,* 783-801.

[53] Sedlmeier P.,1999, Improving statistical reasoning: theoretical models and practical textifimplications, Lawrence Erlbaum, Mahwah, NJ.

[54] Shephard G., Kirkwood C., 1994, Managing the Judgmental Probability Elicitation Process: A CaseStudy of Analyst/Manager Interaction, *IEEE Transactions on Engineering Management,* Vol 41 414-425.

[55] Slovic P., 1972, From Shakespeare to Simon: Speculations and some evidence about man's ability to process information, *Oregon Research Institute Monograph,* Vol 2 No. 2.

[56] Tversky A., 1974, Assessing uncertainty , *J.R. Statis. Soc.* Vol 36, 148-159.

[57] Tversky A., Kahneman D., 1974, Judgment under Uncertainty: Heuristics and Biases, *Sciense, New Series,* Vol 185, No. 4157 1124-1131.

[58] Umesh G., 1988, comparison of Two Elicitation Methods for a Prior for a Binomial Parameter, *Management Science* Vol. 34, 784-790.

[59] Walker KD., Evans JS., MacIntosh D., 2001, Use of expert judgment in exposure assessment - Part I. Characterization of personal exposure to benzene, *Journal ofExposure Analysis and Environmental Epidemiology* Vol. 11, 308-322.

[60] Walss L., Quigley J., 2001, Building prior distributions to support Bayesian reability growth modelling using expert judgement, *Reability Engineering and system safety* Vol. 74, 117-128.

[61] Waterman M., Martz H., Walter R., 1976, Fitting Beta Prior Distributions in Bayesian ReliabilityAnalysis A Sat of TabSss *Los alamos scientific laboratory of the University of California,* LA-6395-MS.

[62] Weiler H., 1965, The Use of Incomplete Beta Functions for Prior Distributions in Binomial Sampling, *Institute of Statistics and Decision Sciences,* Vol 7, 335-347, No. 3.

[63] Wilson A., 1994, Cognitive factors affecting subjective probability assessment, *Technometrics,* Vol 7, 335-347, No. 3.

[64] Winkler R., 1967, The Assessment of Prior Distributions in Bayesian Analysis, *Journal of the American Statistical Association* Vol 62, No. 776-800.

[65] Wood L., Ford J., 1993, Structuring interviews with experts during knowledge elicitation, *International Journal of Intelligent Systems* Vol 8, No. 71-90.