TECNO HUMANISMO
https://tecnohumanismo.online

TecnoHumanismo. Revista Científica
Marzo - Mayo 2022
Volumen 2 / No.1
ISSN: 2710-2394
pp. 248-266
https://doi.org/10.53673/th.v1i5.38

Análisis clásico y bayesiano en la distribución beta rectangular

Classical and Bayesian analysis in the rectangular beta distribution

Análise clássica e bayesiana na distribuição beta retangular

ARTÍCULO GENERAL

Luis Humberto Chia Ramírez
https://orcid.org/0000-0002-5317-3656
luischia.r@gmail.com
lchia@pucp.pe
Pontifica Universidad Católica del Perú, Lima - Perú

Recibido 02 de Octubre 2021 | Arbitrado y aceptado 14 de Diciembre 2021 | Publicado en 04 Marzo 2022

RESUMEN

En el presente trabajo se aborda el problema de trabajar con datos expresados en proporciones que contengan valores extremos. El objetivo general del estudio fue estudiar las propiedades, estimar y aplicar a datos reales el modelo de distribución Beta Rectangular, que ha sido construido específicamente para llevar a cabo el análisis estadístico de datos expresados en proporciones que contengan valores extremos. El estudio se llevó a cabo desde el punto de vista clásico y bayesiano. Para la implementación de la inferencia Bayesiana se consideraron simulaciones de Montecarlo de Cadenas de Markov (MCMC). A fin de evaluar la robustez del modelo de distribución Beta Rectangular, se comparó con el modelo de distribución Beta tanto por inferencia clásica como por inferencia bayesiana, y se llevó a cabo estudios de simulación bajo diferentes escenarios generados por variaciones en el valor de los parámetros de la distribución. Los estudios de simulación demostraron, que el modelo de distribución Beta Rectangular es más robusto que el modelo de distribución Beta. En el caso complementario, es decir cuando los datos no incluyen valores extremos, se presenta una alternancia entre los modelos Beta Rectangular y Beta en relación a cuál de ellos es el que mejor se ajusta a los datos. Se concluye que el modelo de distribución Beta Rectangular presenta propiedades adecuadas para trabajar con conjuntos de datos expresados en proporciones, restringidos al intervalo [0, 1] de la recta real, y que presentan valores extremos. Cuando esta situación se da, el modelo de distribución Beta Rectangular tiene un mejor ajuste a los datos que el modelo de distribución Beta.

Palabras Claves: Análisis clásico bayesiano; Modelo de distribución beta rectangular; valores extremos; simulaciones de Montecarlo.

ABSTRACT

The problem of working with data expressed in proportions containing extreme values is addressed. The general objective of the study was to study the properties, estimate and apply to real data the Beta Rectangular distribution model, which has been specifically constructed to carry out the statistical analysis of data expressed in proportions containing extreme values. The study was carried out from the classical and Bayesian point of view. For the implementation of Bayesian inference, Markov Chain Monte Carlo (MCMC) simulations were considered. In order to evaluate the robustness of the Rectangular Beta distribution model, it was compared with the Beta distribution model by both classical and Bayesian inference, and simulation studies were carried out under different scenarios generated by variations in the value of the distribution parameters. The simulation studies showed that the Rectangular Beta distribution model is more robust than the Beta distribution model. In the complementary case, i.e. when the data do not include extreme values, there is an alternation between the Beta Rectangular and Beta models in relation to which of them is the best fit to the data. It is concluded that the Beta Rectangular distribution model presents adequate properties to work with data sets expressed in proportions, restricted to the interval [0, 1] of the real line, and that present extreme values. When this situation occurs, the Beta Rectangular distribution model has a better fit to the data than the Beta distribution model.

Keywords: Classical Bayesian analysis; Rectangular Beta distribution model; extreme values; Monte Carlo simulations.

RESUMO

O problema de trabalhar com dados expressos em proporções contendo valores extremos é abordado. O objetivo geral do estudo foi estudar as propriedades, estimar e aplicar a dados reais o modelo de distribuição Retangular Beta, que foi construído especificamente para realizar a análise estatística de dados expressos em proporções contendo valores extremos. O estudo foi realizado do ponto de vista clássico e bayesiano. Para a implementação da inferência Bayesiana, foram consideradas as simulações de Markov Chain Monte Carlo (MCMC). Para avaliar a robustez do modelo de distribuição Retangular Beta, foi comparado com o modelo de distribuição Beta por inferência clássica e Bayesiana, e estudos de simulação foram realizados em diferentes cenários gerados por variações no valor dos parâmetros de distribuição. Os estudos de simulação mostraram que o modelo de distribuição Retangular Beta é mais robusto do que o modelo de distribuição Beta. No caso complementar, ou seja, quando os dados não incluem valores extremos, ocorre uma alternância entre os modelos Beta Retangular e Beta em relação a qual deles se ajusta melhor aos dados. Conclui-se que o modelo de distribuição Retangular Beta apresenta propriedades adequadas para trabalhar com conjuntos de dados expressos em proporções, restritos ao intervalo [0, 1] da reta real, e que apresentam valores extremos. Quando esta situação ocorre, o modelo de distribuição Retangular Beta tem um melhor ajuste aos dados do que o modelo de distribuição Beta.

Palavras-chave: Análise Bayesiana Clássica; Modelo de distribuição retangular beta; Valores extremos; Simulações de Monte Carlo.

INTRODUCCIÓN

Al realizar procesos de investigación algunos expertos usan modelos para analizar variables cuya respuesta es restringida al intervalo (0,1), tales como la aprobación presidencial, la tasa de analfabetismo, la tasa de migración, PEA ocupada según ocupación o actividad económica, la proporción de accidentes de tránsito, la tasa de fatalidad en accidentes de tránsito, etc.

Para estos casos, tal como señala Ferrari y Cribari-Neto (2004) los modelos de regresión lineal no son apropiados ya que los valores estimados de la variable respuesta pueden exceder los límites inferior y superior del rango asignado.

Una de las alternativas para el modelamiento de este tipo de variables es la distribución Beta, la cual es bastante flexible, dado que su función de densidad puede tomar diferentes formas dependiendo del valor de los parámetros α y β.

Aún cuando la distribución esta es bastante flexible para modelar proporciones, Hahn (2008) nota que esta distribución no considera adecuadamente eventos que se encuentren en los extremos de la distribución. Esta es una característica limitante, pues en el mundo real necesitamos de distribuciones que sean capaces de incluir datos que se encuentren muy alejados en la muestra, tal es el caso de los datos de valor extremo. Es por ello que este autor propone usar una mixtura de distribuciones, ya que ello aumenta la robustez para la inferencia, y al mismo tiempo permite tener mayor flexibilidad y un mejor ajuste a los datos. Específicamente se propuso mezclar la distribución uniforme con la distribución Beta, añadiéndole un parámetro de mixtura θ, siendo $0 \leq \theta \leq 1$.

Ferrari y Cribari-Neto (2004) señalan que normalmente es más útil modelar a la media de la variable respuesta, sin embargo, Bayes, Bazán y García (2012) notan que la media de la distribución Beta Rectangular dada es una función de la mixtura de parámetros θ y μ, por lo que proponen una nueva parametrización de la distribución Beta Rectangular, que es sobre la cual se trabajará en este estudio.

La estimación de los parámetros se llevará a cabo a través de dos enfoques, por un lado, se desarrolla la inferencia estadística clásica para calcular los estimadores de Máxima Verosimilitud para los parámetros de la distribución Beta Rectangular reparametrizada. Para ello se hace uso del Algoritmo-EM (Expectation-Maximization) propuesto por Dempster, Laird y Rubin (1977), bajo la parametrización propuesta por Hahn (2008) lo cual es posible gracias a la propiedad de invarianza del Estimador de Máxima Verosimilitud. Para este caso los programas son desarrollados integramente sobre el software estadístico R (R Development Core Team (2011)).

Se estiman también los parámetros de la distribución Beta Rectangular reparametrizada bajo el enfoque bayesiano empleando simulaciones de Montecarlo de Cadenas de Markov. Para ello se hace uso de distribuciones a priori no informativas. Los programas que realizan esta tarea son desarrollados en lenguaje de los softwares estadísticos WinBUGS (Spiegelhalter, Thomas, Best y Lunn (2004)) y Open- BUGS (Sturtz, Ligges y Gelman (2005)), empleando la interface de los paquetes de BRugs y R2WinBUGS del software estadístico R para que los programas puedan correr sobre esta plataforma computacional.

Los métodos bayesianos introducen una nueva interpretación del concepto de probabilidad como una medida condicional de la incertidumbre asociada a la ocurrencia de un evento, dada la información disponible y las creencias previas. Congdon (2003) señala que en la inferencia clásica los datos correspondientes a un conjunto. Y son tomados como aleatorios, mientras que los parámetros poblacionales θ, de dimensión p son tomados como fijos. En el análisis bayesiano, en cambio, los parámetros siguen una distribución de probabilidad (sin haber considerado previamente el conjunto de datos disponible de Y), y esa información es resumida en una distribución a priori p(θ). En muchas situaciones puede ser beneficioso incluir en la densidad a priori la evidencia acumulada disponible acerca de un parámetro, proveniente de estudios científicos previos. Un ejemplo de esa información previa sería la proveniente de un ratio relativo al efecto de fumar más de cinco cigarrillos diariamente en el periodo de embarazo sobre el hecho que el peso del niño al nacer sea menor a 2.5 kg. Señala además que esta información puede ser obtenida de manera formal o informal a partir de estudios existentes.

Para entender mejor la diferencia entre la inferencia clásica y la bayesiana, suponga-mos que tenemos un conjunto de n observaciones Yn = y1, ..., yn. Siguiendo a Tomohiro (2010), para resumir la información podemos calcular fácilmente la media, varianza, curtosis, asimetría; sin embargo puede ser dificultoso obtener información más precisa sobre la estructura de un sistema o proceso proveniente de un número finito de datos observados. Por lo tanto, los investigadores usan familias de distribuciones paramétricas con densidades del tipo $f(y|\theta); \theta \in \Theta \subset Rp\}$ para explorar la naturaleza de la estructura de los datos y a partir de ello predecir el comportamiento futuro de la misma. Es decir, uno deriva conclusiones estadísticas basados en el modelo probabilístico asumido. La función de densidad predictiva $f(z|\theta)$ para observaciones futuras de z puede

ser construida reemplazando simplemente el vector de parámetros deconocidos por el estimado de máxima verosimilitud θEMV.

En el contexto bayesiano, en contraste a la aproximación clásica, el parámetro desconocido θ es tratado como una variable aleatoria; por lo tanto, para describir el conocimiento previo, la opinión experta la intuición o las creencias acerca del valor de θ, se prepara una distribución de probabilidad a priori p(θ) sobre el espacio paramétrico de Θ.

La medida de probabilidad descrita anteriormente es una probabilidad bayesiana. Por lo tanto, se hace evidente que en el análisis bayesiano el teorema de Bayes es fundamental. Si denotamos dos eventos como A y B y P (A|B) a la probabilidad de ocurrencia del evento A luego de ocurrido el evento B, bajo la condición que la probabilidad de ocurrencia del evento B es positiva, es decir P (B) >, la probabilidad condicional del evento A dado B es dada por:

P (A B) = P (A ∩ B)

P (B) (1.1)

Transformando la ecuación obtenemos la regla del producto de probabilidades:

P (A ∩ B) = P (A|B)P (B)

Siguiendo a Tomohiro (2010) el teorema de Bayes puede ser derivado a partir de la Ley de Probabilidad Total. Si permitimos que A1, A2, ..., Am sean eventos disjuntos de modo que P (Ai ∩ Aj) = 0, i =/ j y P (A1 ∪ ... ∪ Am)=1 sea el evento seguro Ω, es decir P (Ω) = 1, luego tenemos que:

$$P(B) = P(B|\Omega) = \sum_{j=1}^{m} P(B|Aj)P(Aj)$$

Por lo tanto, el evento seguro es dividido en un evento disjunto de m piezas, y la probabilidad condicional de B dado cada evento dividido Am es añadido de forma conjunta. Bajo este marco, la probabilidad condicional del evento Ak dado el evento B, donde P (B) > 0, viene dada por:

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)}$$

$$P(A_k|B) = \frac{P(A_k \cap B)}{\sum_{j=1}^{m} P(B|A_j)P(A_j)}$$

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{m} P(B|A_j)P(A_j)}, \quad k = 1...m \qquad (1.2)$$

La ecuación anterior constituye el teorema de Bayes. Bajo este contexto permitamos ahora que el parámetro θ tome sólo m valores {θ1, ..., θm}, con probabilidades p(θ1), ..., p(θm)}.

Permitamos además que el evento Ak sea θ de modo que = θk y el evento B sea Xn, es decir el término observado en el teorema de Bayes de la ecuación (1.2), se sigue entonces a partir de este teorema que toda la información disponible acerca del valor θ está contenido en la distribución a posteriori correspondiente, es decir:

$$p(\vartheta = \vartheta_k|X_n) = \frac{f(X_n|\vartheta_k)p(\vartheta_k)}{\sum_{j=1}^{m} f(X_n|\vartheta_j)p(\vartheta_j)}$$

Donde f (Xn|θ) es la función de verosimilitud. Si el parámetro θ es una variable aleatoria continua, tendríamos:

$$p(\vartheta = \vartheta_k|X_n) = \frac{f(X_n|\vartheta_k)p(\vartheta_k)}{\int f(X_n|\vartheta_j)p(\vartheta_j)d\vartheta} \propto f(X_n|\vartheta)p(\vartheta) \qquad (1.3)$$

La ecuación (1.3) muestra que la influencia relativa del conocimiento previo y de los datos sobre la actualización de las creencias depende de cuánto peso se le de a la distribución a priori, lo que estará a su vez relacionado con la capacidad informativa de la misma, y con la importancia de los datos. Por ejemplo, un conjunto de datos grande puede tender a tener una influencia predominante sobre la actualización de las creencias a pesar que la priori sea informativa. En contraste, si la muestra es pequeña y esta es combinada con una priori informativa, entonces la distribución a priori tendrá una mayor influencia relativa sobre la actualización de las creencias. Este es el caso de muchos problemas reales, en los que la información disponible es sumamente limitada, por lo que emplear el método bayesiano puede tener una gran ventaja sobre el método clásico.

Integración de Monte Carlo

Suponiendo que se necesita evaluar la integral de una función f (θ) con respecto a una distribución p(θ), siguiendo a Mira (2005):

$$\mu_p(f) = \int \ominus f(\vartheta)p(\vartheta)d\vartheta$$

la autora señala que por lo general la distribución de interés es la posteriori del parámetro en estudio, mientras que la función puede ser la identidad (permitiendo recuperar la distribución posterior), la función indicadora o cualquier otra función integrable del parámetro de interés. Si se trabaja con una muestra aleatoria de observaciones (θ1, ..., θn) identica e independiemente disrtibuidas obtenidas de p, entonces podemos recurrir a la simulación de Monte Carlo y estimar μp (f) a partir de:

$$\mu\eta(f) = \frac{1}{n}\sum_{i=1}^{n} f(\vartheta i)$$

Asumiendo que f tiene varianza finita, la Ley de los Grandes Números garantiza que μn (f) es un estimador asintóticamente insesgado de μp(f) y el Teorema del Limite Central garantiza que la distribución límite del estimador de Monte Carlo, adecuadamente normalizado √ n (μn(f) − μp(f)), es Normal con varianza dada por σ(f). Por lo tanto, sin importar la dimensión de θ, el término de error es de orden √ n

Cadenas de Markov

Tal como define Mira (2005), una cadena de Markov es un proceso estocástico {X0, X1, ...} que evoluciona en el tiempo con la propiedad que el futuro el independiente del pasado dado el presente, es decir:

$$P\ X_t\ \in AX_0, X_1, \dots, X_{t-1}\} = PX_t \in AX_{t-1}$$

para cualquier A en el espacio E. Se identifica una Cadena de Markov con el Kernel de transición P definido por:

$$P(x, A) = P\ X_1 \in AX_0 = x\}$$

Mientras que el kernel de transición está dado por

$$P^n(x, A) = P\ X_n \in AX_0 = x\}$$

Al igual que la autora, denotamos como Px y como Pp a las probabilidades de que una Cadena de Markov inicie con X0 = x o con una distribución inicial X0 ∼ p respectivamente. Una cadena de Markov tiene distribución estacionaria p si:

$$p^P(A) = \int P(x, A)p(x)dx = p(A), \forall A \subset E$$

No todas las cadenas de Markov tienen distribuciones estacionarias y aun cuando exista una distribución estacionaria esta puede no ser única. Justamente el principio básico tras las simulaciones MCMC es que las Cadenas de Markov tengan convergencia a una única distribución y por lo tanto pueda ser usada para estimar expectativas con respecto a esa distribución. Mira (2005) señala cuáles son las propiedades que permiten identificar Cadenas de Markov que tengan una única distribución estacionaria, la misma que a su vez es la distribución límite del proceso. Estas propiedades son indicadas en el apartado siguiente.

Propiedades de las Cadenas de Markov

Se dice que una Cadena de Markov es $\varphi -$ irreducible para una distribución de probabilidad φ sobre E, si $\varphi(A) > 0$, lo cual implica:

$P_x \{\text{tiempo del primer retorno a } A < \infty\} > 0$

Entonces, una cadena es irreducible si esta es $\varphi -$ irreducible para algún φ. En otras palabras, una cadena irreducible tiene una probabilidad distinta de cero de pasar de una posición determinada en el espacio en el espacio de estados a cualquier otra posición, en un número finito de pasos. Ello garantiza que todas las porciones importantes del espacio de estados puedan ser visitadas. De otro lado, la recurrencia garantiza que todos los subcon- juntos de interés del espacio de estados puedan ser visitados un número infinito de veces, al menos desde casi todos los puntos de inicio. Si una cadena de Markov es irreductible y tiene una distribución estacionaria apropiada p, entonces esta debe ser positiva recurrente y p es también la única distribución estacionaria. Una condición suficiente para que una Cadena de Markov sea irreducible con respecto a la distribución φ es que el kernel de transición de $n -$ pasos tenga una densidad positiva con respecto a φ para algu´n $n \geq 1$.

Dada una Cadena de Markov irreducible con distribución estacionaria p y una función en los reales tal que $\int |f(x)|p(x)dx < \infty$, se tiene una Ley fuerte de los grandes nu´meros que es:

$$P_x \mu_n(f) \rightarrow \mu_p(f)\}$$

En otras palabras, el tiempo observado y esperado relacionado a un conjunto A converge a p(A). Adicionalmente, para obtener resultados más fuertes, se debe descartar el comportamiento periódico o cíclico.

Se dice que una Cadena de Markov es no-periódica si el máximo común divisor del número de pasos que le toma a la cadena al punto inicial, sin importar cual sea este, es

uno. Para descartar conjuntos nulos de puntos de partida iniciales en los que la Ley de los grandes números puede fallar, se deben considerar Cadenas de Markov Harris-recurrentes, es decir cadenas φ − irreductibles, donde φ sea la distribución irreductible máxima, y por lo tanto, para cada A ⊂ E con φ(A) > 0, tenemos:

Px {Xn ∈ Ainfinitamente} = 1

para todo x ∈ E. Una Cadena de Markov es ergódica si esta es irreductiblre, no-periódica y Harris-recurrente positiva. Sin embargo, la autora señala que en la mayoría de aplicaciones MCMC la ergodicidad es de poca importancia dado que típicamente los investigadores están interesados en resultados concernientes a trayectorias muestrales promedio.

Finalmente, las condiciones de ergodicidad uniforme y geométricas están relacionadas a la tasa a la cual la Cadena de Markov converge a la estacionariedad. En particular, una cadena con distribución estacionaria p es geométricamente ergódica si existe una función con valores reales extendidos a los no negativos M, tal que M (x)p(x)dx < ∞, y una constante positiva ρ < 1, tal que:

$$\ddot{P}^n(x,\cdot) - p(\cdot) \le M(x)\rho^n \quad \ddot{} \quad \forall x,\ n;$$

Si además M es constante finita y positiva, entonces la cadena es uniformemente ergódica.

METODOLOGÍA

Se hace una revisión exhaustiva de la literatura en relación a los modelos propuestos con distribución para proporciones, y posteriormente se implementa la inferencia estadística desde el punto de vista clásico y bayesiano del modelo de distribución Beta Rectangular. La inferencia clásica se lleva a cabo mediante el método de Máxima Verosimilitud, y dado que la estimación de los parámetros de la distribución Beta Rectangular de forma analítica se torna complicada se optar por emplear un método numérico iterativo, especificamente el algoritmo Expectation Maximitation generalmente conocido como Algoritmo-EM. Para la implementación de la inferencia Bayesiana se consideran simlaciones de Montecarlo de Cadenas de Markov (MCMC). A fin de evaluar la robustez del modelo de distribución Beta Rectangular, este es comparado con el modelo de distribución Beta tanto por inferencia clásica como por inferencia bayesiana, y se llevan a cabo estudios de simulación bajo diferentes escenarios generados por variaciones en el valor de los parámetros de la distribución

Objetivos de la investigación

General

Estudiar las propiedades, estimar y aplicar a datos reales el modelo de distribución Beta Rectangular desde el punto de vista clásico y bayesiano.

Específicos:

Revisar la literatura en relación a los modelos propuestos para proporciones.

Estudiar e implementar la inferencia estadística del modelo de Beta Rectangular mediante el método de áaxima Verosimilitud y mediante la inferencia bayesiana.

Estudiar e implementar la estimación del modelo Beta Rectangular desde el punto de vista clásico y bayesiano.

Implementar métodos de inferencia Bayesiana considerando simulación de Montecarlo de Cadenas de Markov (MCMC).

Implementar simulaciones del modelo de distribución Beta Rectangular sobre distintos escenarios.

Implementar el modelo Beta Rectangular a datos reales considerando que este no tiene variable respuesta.

RESULTADOS

Como parte de la puesta en práctica la teoría estudiada se presentan los resultados de la misma en un estudio de caso:

Niveles de pobreza en los distritos del departamento de Ica

Descripción del caso

Uno de los mayores problemas que enfrentan los países de ingresos medios o bajos es la pobreza. La medición de la pobreza monetaria se calcula comparando los gastos de los hogares con la línea de pobreza. Dicha línea es aquella que permite adquirir una Canasta Básica de

Consumo suficiente para satisfacer requerimientos nutricionales y otras necesidades básicas de los hogares. Se define como pobre a la población que vive en hogares cuyo gasto, por persona, es inferior al monto establecido en la línea de pobreza, que para el año 2011 se estableció en 272 nuevos soles, tal como se menciona en el documento técnico "Mejoras Metodológicas para la Medición de la Pobreza" publicado por el Instituto Nacional de Estadística e Informática - INEI (2011a).

En el documento "Perú: Perfil de la Pobreza por departamentos, 2001-2010" INEI (2011b) se señala que la pobreza es medida de forma monetaria porque no considera las

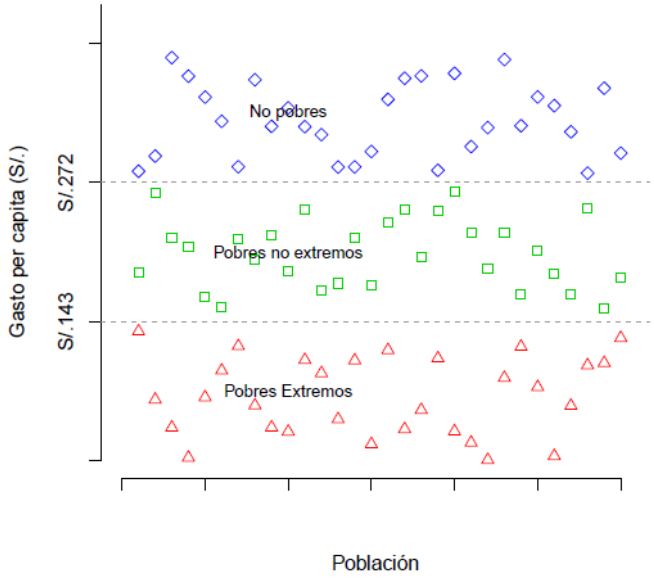
dimensiones no monetarias de la misma, como desnutrición, necesidades básicas insatisfechas, exclusión social, capacidades, entre otras.

La medición de la pobreza se refiere a una valoración absoluta pues esta se mide respecto a un valor de la línea que no depende de la distribuci´on relativa del bienestar de los hogares.

Para esta medición, señala el documento, se utilizan dos tipos de líneas, a saber, la línea de Pobreza Extrema y la línea de Pobreza Total. La Línea de Pobreza Extrema es un valor monetario necesario para la adquisición de una canasta de alimentos capaz de satisfacer un mínimo de necesidades nutricionales de las personas. La Línea de Pobreza Total es el valor de la línea de Pobreza Extrema más el valor monetario necesario para satisfacer un conjunto de necesidades no alimentarias consideradas esenciales (vestido y calzado, alquiler de la vivienda, combustible, muebles y enseres, cuidado de la salud, transportes y comunicaciones, esparcimiento, educación y cultura y otros gastos). En este sentido, se puede considerar que los pobres no extremos se encuentran en el rango que comprende la diferencia entre el valor de la línea de pobreza total y la línea de la pobreza extrema.

Si se mide en el eje de abscisas a la población y en el eje de ordenadas el Gasto per capita en nuevos soles, podría representarse gráficamente las clasificaciones de los niveles de pobreza.

Tal como se muestran en la figura1. Medición de la pobreza monetaria



Así una persona sumida en la pobreza extrema sería aquella cuya canasta básica de alimentos esté por debajo del umbral de los 143 nuevos soles mensuales. Los pobres no extremos serían aquellos cuya canasta alimentaria y no alimentaria, su canasta básica de consumo, fuera de menos de 272 nuevos soles, pero mayor a los 143 nuevos soles

mensuales. En general los pobres extremos y no extremos serán aquellos cuya canasta básica de consumo esté por debajo de los 272 nuevos soles mensuales; y en complemento, aquellos que gocen de una canasta cuyo valor sea superior a los 272 nuevos soles serán considerado como no pobres.

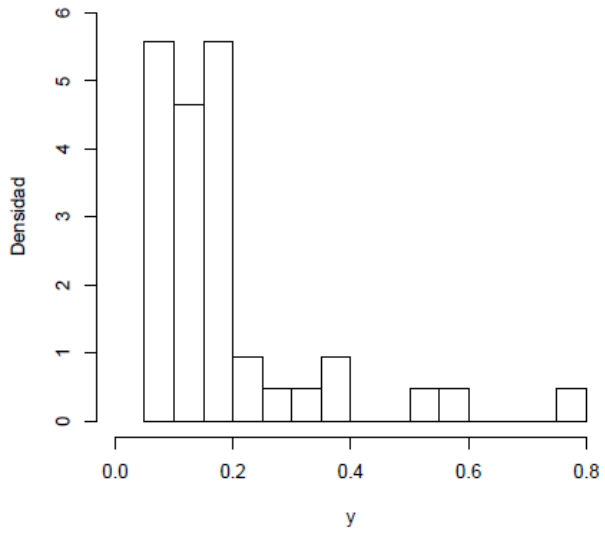
La aplicación que se lleva a cabo en la presente sección consiste en la estimación de los parámetros de las distribuciones Beta y Beta Rectangular por los métodos clásico y bayesiano. Para ello se ha empleado datos del total de la pobreza en los distritos del departamento de Ica. Los datos fueron extraídos del anexo estadístico del libro electrónico "Mapa de Pobreza Provincial y Distrital 2009" publicado en la dirección electrónica http://www.inei.gob.pe/ biblioineipub/bancopub/Est/Lib0952/index.htm. Estos datos pueden ser apreciados en el anexo A.

Descripción de los datos

Se seleccionó para esta aplicación los datos de la población en situación de pobreza total en los distritos del departamento de Ica debido a la adecuación de este conjunto de observaciones a los requerimientos para el ajuste de los modelos de distribución Beta y Beta Rectangular. Los datos de pobreza tienen un rango que pertenece al intervalo (0,1), es decir, está medido en proporciones. Al no ser la pobreza un valor homogéneo a lo largo de toda la extensión del país, y tampoco dentro del territorio de un departamento, se pueden presentar situaciones en los que los datos registrados contengan valores extremos, caso en el que nos encontramos si analizamos los niveles de pobreza en los distritos del departamento de Ica.

En la figura 2 se muestra el histograma de los datos registrados de la población en situación de pobreza. Se puede apreciar que esta variable presenta valores que podrían ser considerados atípicos o extremos. Hay distritos que presentan una proporción de la poblacion pobre cercana al 80 %, o de forma complementaria distritos en el departamento de Ica que cuentan con población no pobre de sólo 20 %.

Figura 2: Panel (a): histograma de la población en situación de pobreza en los distritos del departamento de Ica. Panel (b): Histograma de la población en situación de no pobreza en los distritos del departamento de Ica.



El cuadro 1 muestra las estadísticas resumen de las variable "población en situación de pobreza" en los departamentos del distrito de Ica. Como puede apreciarse, las variables est´an medidas en porcentajes, por ello, para tenerlas en proporciones son divididas entre 100. Se tiene que el distrito con menor pobreza tiene un 5.1 % de pobres y corresponde espec´ıficamente al distrito de Tambo de Mora; el distrito con mayor cantidad de población pobre es el distrito de Chavín con 79.1 % de la población en esta situación; la media de esta variable es de 18.1 %, mientras que la mediana es de 14.7 %; y el rango intercuantil va de 9.85 % a 19.05 %.

Cuadro.1: Población en situación de pobreza en los distritos del departamento de Ica

Min.	Q1	Mediana	Media	Q3	Max.
5.10	9.85	14.70	18.11	19.05	79.10

Resultados de la aplicación: Estimación por máxima verosimilitud

En la presente sección se ajustan los modelos Beta y Beta Rectangular a un conjunto de datos reales, tal como han sido descritos en la sección anterior.

El modelo Beta a estimar sigue la distribución Beta Reparametrizada dada en la ecuación, es decir

$$y_i \sim Beta(\mu, \varphi)$$

Del mismo modo, el modelo Beta Rectangular a estimar sigue la distribución dada en la ecuación, es decir:

$$y_i \sim BR(\mu, \varphi, \vartheta)$$

Los resultados de la aplicación mediante la estimación de máxima verosimilitud son presentados en la tabla 2. Como puede apreciarse el valor calculado para el Criterio de Información Bayesiano (BIC) es menor en el caso del modelo Beta Rectangular, por lo que este modelo sería escogido bajo este criterio.

Analizando los parámetros estimados observamos que en el caso del modelo Beta Rectangular este presenta un valor para la media del modelo, μBR = 0,1363, menor a la reportada por el modelo Beta μBeta = 0,1908. De otro lado, se muestra un mayor valor del parámetro de precisión en el modelo Beta Rectangular φBR = 29,3309 frente al estimado mediante el modelo Beta φ = 8,3527.

Cuadro 2: Estimación por máxima verosimilitud de los modelos Beta y Beta Rectangular

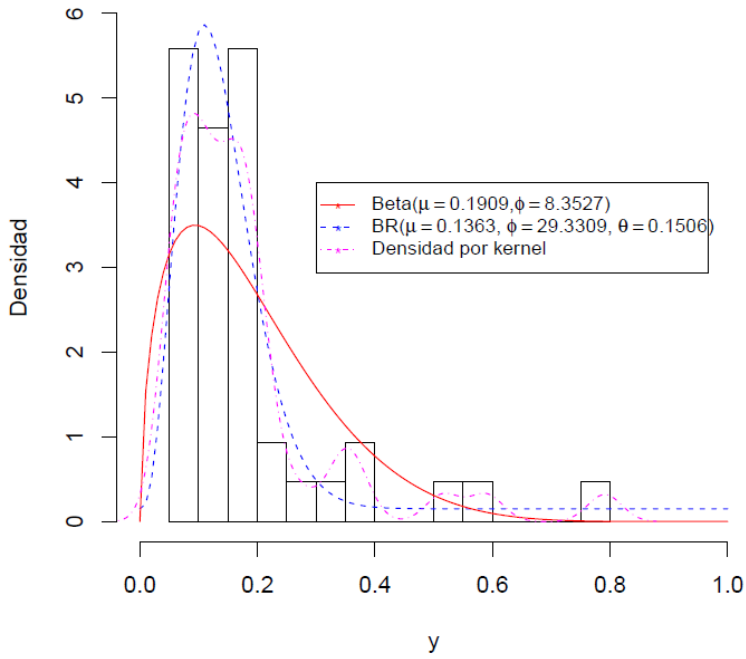
Modelo	Beta			Beta Rectangular		
Parámetro	Media	AIC	BIC	Media	AIC	BIC
μ	0.1908			0.1363		
φ	8.3527	-62.09	-51.05	29.3309	-76.92	-70.60
θ	–			0.1506		

Analizando el histograma de los datos podemos observar claramente que la densidadestimada por el modelo Beta Rectangular (linea discontinua) ajusta mucho mejor que la estimada por el modelo Beta (linea continua), tal como muestra la figura 3.

Podemos también observar en la figura 3 que la cola derecha de la densidad Beta no incluye los valores extremos cercanos al límite superior del rango de la distribución, es decir los valores que se encuentran cercanos a la unidad. En oposición la densidad Beta Rectangular al presentar una cola derecha más pesada que la distribución Beta, permite incluir los valores extremos, por lo tanto se considera que esta es mejor para modelar este caso específico de datos cuyo rango se encuentra en el intervalo (0,1) y que además presenta valores extremos.

En relación a los criterios de selección, tanto el AIC como el BIC sugieren que se debe elegir el modelo de distribución Beta Rectangular. Se debe escoger aquel modelo que tenga menor valor de AIC y BIC.

Figura 3: Población pobre en los distritos del departamento de Ica



Resultados de la aplicación: Estimación por inferencia bayesiana

Se lleva a cabo el ajuste de los datos de la población pobre en los distritos del departamento de Ica mediante los modelos Beta y Beta Rectangular desde la perspectiva bayesiana. El modelo Beta a estimar sigue la distribución Beta Reparametrizada dada en la ecuación, es decir:

$$y_i \sim Beta\ (\mu, \varphi)$$

Con distribuciones a priori para μ y φ, es decir:

$$\mu \sim U(0, 1)$$
$$\varphi \sim Gamma\ (a, b)$$

Del mismo modo, el modelo Beta Rectangular a estimar sigue la distribución dada en la ecuación:

$$y_i \sim BR\ (\mu, \varphi, \theta),$$

y distribuciones a priori para μ, φ y θ:

$$\mu \sim U(0, 1)$$
$$\varphi \sim Gamma(a, b)$$
$$\theta \sim U(0, 1)$$

En ambos casos los parámetros de la distribución Gamma son a = b = 0,01, tal como se realiza en el trabajo de Bayes, Bazan y García (2012). El cuadro 3 muestra los resultados de los parámetros estimados.

Cuadro 3: Estimación bayesiana de los modelos Beta y Beta Rectangular

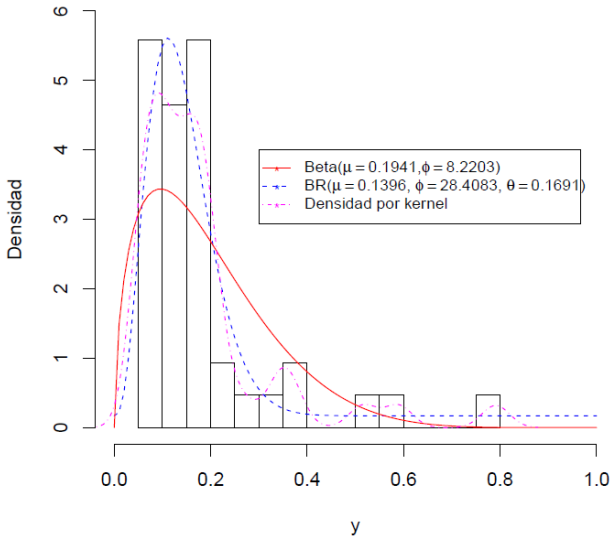
Modelo	Beta						B ta Rectangular					
Parámetro	Media	2.5 %	97.5 %	EAIC	EBIC	DIC	Media	2.5 %	97.5 % DIC	EAIC	EBIC	
μ	0.1941	0.1589	0.2371				0.1396	0.1145	0.1672			
φ	8.2203	5.1069	12.2413	-60.1	-56.6	-62.1	28.4082	13.7443	50.8065	-73.9	-68.7	-77.2
θ	–	–	–				0.1691	0.0482	0.3359			

El modelo Beta Rectangular estima los parámetros μ = 0,1395, φ = 28,4082 y θ = 0,1691, mientras que el modelo Beta estima los parámetros μ = 0,1941 y φ = 8,2203. De este modo, si comparamos los modelos para los datos según las dos distribuciones, es decir y ~ BR (μ = 0,1395, φ = 28,4082, θ = 0,1691) y y ~ Beta (μ = 0,1941, φ = 8,2203), se encuentra que los criterios de selección en el cuadro 3 sugieren que se elija en todos los casos el modelo de distribución Beta Rectangular. Asimismo, se observa también que los parámetros estimados para el modelo Beta son muy similares ya sea que sean estimados por máxima verosimilitud o de forma bayesiana, y del mismo modo para el modelo Beta Rectangular, es decir ambas estimaciones guardan coherencia, pero con la ventaja del modelo Beta Rectangular de generar el parámetro de mixtura, en este caso θ Bayesiano = 0,1691, lo que permite incluir los valores extremos del conjunto de datos.

La figura 4 muestra, al igual que en el caso de la inferencia clasica, que la densidad Beta Rectangular evaluada en los parámetros estimados se ajusta mejor a los datos estudiados. Ello confirma lo que se había sido ya anticipado por el valor de los criterios de información EAIC, EBIC y DIC, es decir que se debe escoger el modelo Beta Rectangular para la estimación de los parámetros de los datos.

Por lo tanto, mediante esta aplicación se ha podido ilustrar lo que ya se había explicado a lo largo de este trabajo de tesis, es decir la gran utilidad de contar con una distribución que permita modelar adecuadamente datos cuyo rango se encuentra en el intervalo (0,1) y que además toma valores extremos.

Figura 4: Población pobre en los departamentos de Ica



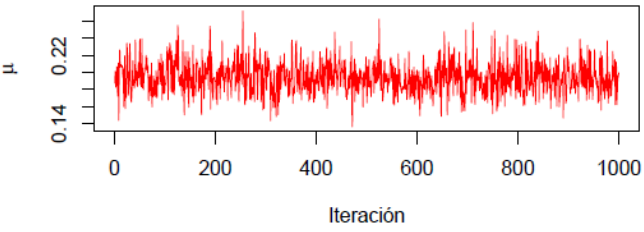
Del conjunto de datos mediante el modelo Beta la estimación podría haber estado sesgada respecto a las estimaciones efectuadas mediante el modelo Beta Rectangular, y mucho más aun si el proceso de estimación hubiera sido realizado mediante el método clásico en comparación con el método bayesiano.

En el gráfico 5 se muestran las últimas 1000 observaciones de las Cadenas de Markov simuladas para los parámetros μ y φ en el modelo Beta. Como puede apreciarse claramente ambos parámetros han convergido a su valor estimado.

Del mismo modo, el gráfico 6 se muestran las últimas 1000 observaciones de las Cadenas de Markov simuladas para los parámetros μ, φ y θ en el modelo Beta Rectangular. Al igual, que en el caso del modelo Beta, puede apreciarse que los tres parámetros han convergido a sus valores estimados.

Considerando estos resultados, se puede concluir que la estimación de los parámetros del modelo de distribución Beta Rectangular para este conjunto de datos específicos es superior a la estimación mediante el modelo Beta.

Figura 5: Cadenas de Markov para μ y φ



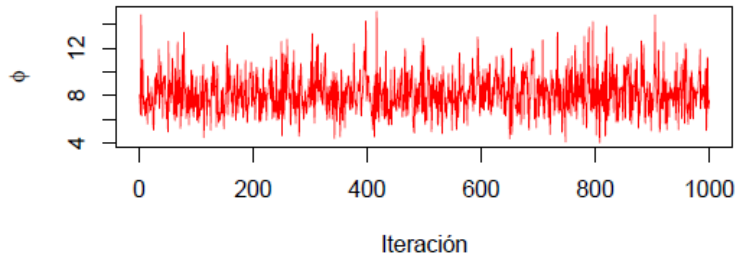
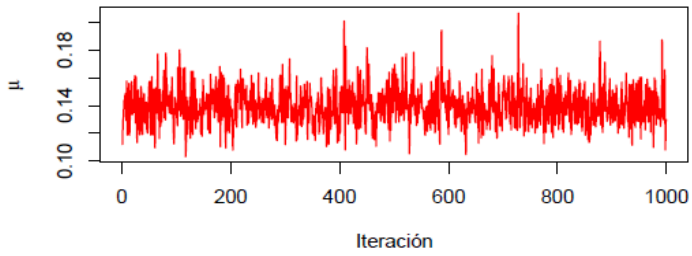
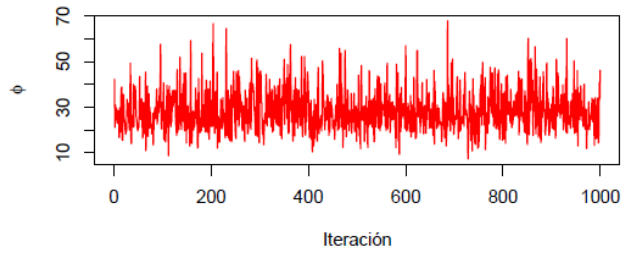


Figura 6: Cadenas de Markov para μ, φ y θ





CONCLUSIONES

Preliminarmente se puede apreciar que el modelo Beta Rectangular ajusta mejor cuando los datos presentan valores outlier como se puede observar en el estudio de simulación y en la aplicación.

Asimismo, podemos observar que mediante la estimación bayesiana se logra obtener en los casos estudiados un mejor ajuste de los datos, y por lo tanto una mejor estimación de los parámetros, en relación al ajuste y parámetros obtenidos mediante la inferencia clásica mediante el método de máxima verosimilitud.

Sugerencias para investigaciones futuras

Realizar estudios de simulación considerando otras distribuciones a priori para los parámetros.

Estudiar modelos de regresión considerando la distribución Beta Rectángular como el proceso generador de los datos.

Considerar nuevas aplicaciones del modelo a otro tipo de datos reales que también puedan ser expresados en proporciones.

REFERENCIAS

Bayes, C., Bazán, J. y García, C. (2012). A new regression model for proportions, Bayesian Analysis

Congdon, P. (2003). Applied-Bayesian-Modelling, John Wiley & Sons.

Ferrari, S. y Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, Journal of Applied Statistics 31: 799–815.

Hahn, E. (2008). Mixture densities for project managment activity times: A robust approac to pert, European Journal of Operational Research 188: 450–459.

INEI (2011a). Mejoras metodológicas para la medici´on de la pobreza, Reporte técnico, Instituto Nacional de Estadística e Informática, Perú.

Mira, A. (2005). Mcmc methods to estimate bayesian parametric models, Handbook of Statistic 25: 415–436.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Spiegelhalter, D. J., Thomas, A., Best, N. G. y Lunn, D. (2004). WinBUGS User Manual Version 1.4.1. http://www.mrc-bsu.cam.ac.uk/bugs.

Sturtz, S., Ligges, U. y Gelman, A. (2005). R2winbugs: A package for running winBUGS from R, Journal of Statistical Software 12(3): 1–16. http://www.jstatsoft.org.

Tomohiro, A. (2010). Bayesian Model Selection and Statistical Modeling, CRC Press.