Un test de dos muestras multinomiales basado en modelos Bayesianos jerárquicos

Antonio Torres
Departamento de matemáticas
Universidad de Almería
Almería, España
atr159@ual.es

Andrés Masegosa
Departamento de matemáticas
Universidad de Almería
Almería, España
andresmasegosa@ual.es

Antonio Salmerón
Departamento de matemáticas
Universidad de Almería
Almería, España
antonio.salmeron@ual.es

***Resumen*—Decidir si dos muestras pertenecen o no a la misma distribución es un problema que aparece en muchas ramas de la ciencia y de la industria. La herramienta más usada para este problema es un contraste de hipótesis basado en el test Chi-cuadrado. El resultado de este procedimiento depende del llamado p-valor. El problema de este p-valor es que no recoge la incertidumbre asociada al tamaño de las muestras. Es decir, el proceso de decisión de si dos muestras multinomiales son o no iguales es el mismo tanto para tamaños muestrales pequeños como para tamaños muestrales muy grandes. En este trabajo presentamos una metodología Bayesiana que es muy fácil de implementar. En este caso devolvemos una distribución *a posteriori* sobre un índice de diferencia entre dos distribuciones. Al ser un esquema Bayesiano podemos cuantificar la incertidumbre en la que se basan nuestras decisiones, y como consecuencia más directa, podemos determinar si el tamaño de las muestras es suficientemente grande como para tomar decisiones robustas acerca de si dos muestras multinomiales pertenecen o no la misma distribución.**

***Index Terms*—Test de dos muestras, Test Chi-cuadrado, Test Bayesiano.**

I. Introducción

Los test de dos muestras (*two-sample tests*) son muy útiles tanto en industria como en ciencia. Estos se basan en comparar si dos conjuntos de datos están generados por la misma distribución o no [5]. Para el caso de la industria nos encontramos, por ejemplo, con los test A/B [4], que nos sirven para comparar dos versiones de un mismo diseño para ver cuál de los dos se comporta estadísticamente mejor. Ejemplos de esto podemos encontrarlos en las empresas que envían distintas versiones de un correo a diferentes destinatarios, o en la web, donde aparecen varias versiones de la misma página para los distintos visitantes [4]. En el caso de la ciencia, un claro ejemplo son los ensayos clínicos, en los que se dispone de dos grupos de individuos con una cierta enfermedad o accidente donde a uno de ellos se le administra un determinado tratamiento (grupo experimental) y el otro grupo (grupo de control) al que o bien no se le administra ningún tratamiento o bien se le administra otro en fase de pruebas y es comparado con el experimental con el objetivo de sacar conclusiones acerca de los tratamientos utilizados [1].

Mediante los test de hipótesis Chi-cuadrado podemos comparar si el conjunto de datos proviene o no de una misma distribución a través de una hipótesis nula y otra alternativa. En este caso, esto se mediría con el p-valor, que pasado de cierto umbral prefijado hace que se rechace o acepte la hipótesis nula. Sin embargo, el p-valor no mide la incertidumbre debido principalmente al tamaño de la muestra. Entonces nos surge la pregunta: ¿cuál es el tamaño de muestra necesario para estar seguros de que las conclusiones obtenidas en el test de hipótesis no dependen de dicho tamaño? El test Chi-cuadrado (y otros test frecuentistas) no puede dar respuesta a esta pregunta [3].

En este trabajo proponemos un método alternativo basado en estadística Bayesiana. Este método no da respuestas binarias como hace el test Chi-cuadrado o los test frecuentistas (se acepta o no se acepta la hipótesis nula) debido a que nos proporcionan una probabilidad a posteriori de que sea o no cierta la hipótesis nula, donde la incertidumbre se va reduciendo según se aumenta el tamaño de la muestra [2]. Este test se basa en una técnica llamada *a prioris jerárquicas de potencia*, o en inglés *Hierarchical Power Priors* (HPP) [2].

II. Conocimientos previos

En muchos problemas de inferencia Bayesiana no podemos calcular la distribución a posteriori directamente debido a que la constante de normalización no se puede calcular. Por ejemplo, sea x un conjunto de observaciones, y p un conjunto de variables latentes, si queremos calcular la distribución a posteriori $P(p|x)$, que sabemos que $P(p|x) = \frac{P(x|p)P(p)}{\int P(x|p)P(p)dp}$, en algunas ocasiones no podremos calcular el denominador, bien porque la dimensión sea muy grande o bien porque la integral tiene una expresión muy compleja. Para resolver esto se utiliza la estadística bayesiana, donde se define una $q(p|\omega)$ lo más cercana posible a $P(p|x)$ en distancia de Kullback-Leibler, esto es, buscamos $\arg\min_\omega \mathrm{KL}(q(p|\omega), P(p|x))$.

Es decir, cuando intentamos calcular esta probabilidad a posteriori lo que se hace es buscar una distribución que sea lo más cercana posible a esta, en términos de la distancia de Kullback-Leibler. Una vez hecho esto, se buscan los parámetros de la distribución $q(p|\omega)$ que mejor aproxima a la a posteriori verdadera $P(p|x)$.

III. A prioris jerárquicas de potencia

El modelo HPP [2] es un modelo probabilístico generativo que permite modelar cambios en los parámetros de una dis-

tribución. En la Fig. 1 podemos ver una descripción gráfica de este modelo y, por otro lado, el Algoritmo 1 nos muestra una descripción en pseudocódigo de este método: (1) en el primer paso, p_0 se muestrea de una distribución Dirichlet de hiper-parámetros $\alpha_u = (1, ..., 1)$, (2) en el segundo paso, se generan los datos de la primera muestra en base al parámetro p_0 previamente generado, (3) en el tercer paso se calcula la distribución a posteriori de p_0 dados los datos muestreados, que también sigue una distribución Dirichlet, donde α_0 son los parámetros de la distribución a posteriori de p_0, (4) en el cuarto paso, muestreamos ρ a partir de una distribución exponencial truncada en el intervalo $[0, 1]$ y de parámetro fijo $\gamma = 0.1$, (5) en un quinto paso, se muestrea el parámetro p_1 a partir de una Dirichlet con parámetro $\rho\alpha_0 + (1 - \rho)\alpha_u$, (6) y, por último, en el sexto paso se muestrea x_1 en base a una multinomial de parámetro p_1.

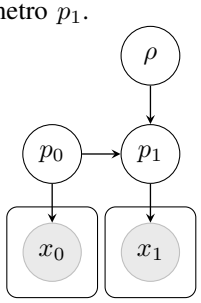

Figura 1: Modelo HPP.

Algoritmo 1 Modelo HPP.

1: $p_0 \sim \text{Dir}(1, ..., 1) \equiv \text{Dir}(\alpha_u)$
2: $x_0 \sim \text{Multinom}(p_0)$
3: $\text{Dir}(\alpha_0) = P(p_0|x_0)$
4: $\rho \sim \text{TruncExp}_{[0,1]}(\gamma)$
5: $p_1 \sim \text{Dir}(\rho\alpha_0 + (1 - \rho)\alpha_u)$
6: $x_1 \sim \text{Multinom}(p_1)$

En este modelo generativo de los datos, la variable ρ es la que define la transición entre los parámetros p_0 y p_1 que generan los datos, $p(p_1|p_0, \rho)$, de forma que la distribución a posteriori de p_1 dado x_0 se calcularía como:

$$p(p_1|x_0, \rho) = \int p(p_1|p_0, \rho)p(p_0|x_0)dp_0.$$

El problema es que para el caso de una distribución Dirichlet condicionada a otra distribución Dirichlet, como sucede en este caso para datos multinomiales, daría lugar a un modelo no-conjugado, lo que complica el cálculo de la distribución a posteriori. El modelo HPP define esta probabilidad de transición de manera implícita [2] bajo principios de máxima entropía. El resultado es un modelo de transición con las siguientes propiedades:

- Si $\rho = 0$, entonces $p(p_1|p_0, \rho = 0) = p(p_1|\alpha_u)$ y, denotamos $p(p_1|x_0, \rho = 0)$ como $p_u(p_1)$. Es decir, no existe relación entre p_0 y p_1.
- Si $\rho = 1$, entonces $p(p_1|p_0, \rho = 1) = \delta(p_1 - p_0)$ y, denotamos $p(p_1|x_0, \rho = 1)$ como $p_\delta(p_1|x_0)$. Es decir, existe una relación determinística de igualdad entre ambos, $p_0 = p_1$.
- Si $0 < \rho < 1$, entonces

  $$p(p_1|x_0, \rho) \propto p_\delta(p_1|x_0)^\rho p_u(p_1)^{(1-\rho)},$$

  es decir, es una mezcla de las dos situaciones extremas anteriores. En el caso de la distribución Dirichlet, esa operación se simplifica como la combinación convexa de sus hiper-parámetros [2].

El modelo HPP nos permite aplicar inferencia variacional para calcular la distribución a posteriori del parámetro ρ dados los datos, es decir, $p(\rho|x_0, x_1)$, siguiendo el Algoritmo 2. Siguiendo el método variacional, esta probabilidad a posteriori es aproximada por una distribución exponencial truncada, $q(\rho|\omega)$.

Como se ha comentado con anterioridad, ρ toma valores en $[0, 1]$. Si $\rho = 0$ significa que el cambio es absoluto, y por el contrario, si $\rho = 1$, entonces no hay cambio. La probabilidad a posteriori es la que nos da información sobre este cambio. Esto no se trata de un test estadístico, sino que es una cuestión de interpretación. Por otro lado, comentar que ω es el parámetro de la exponencial truncada que aproxima a la a posteriori.

La función de densidad, $q(\rho|\omega)$, así como el valor esperado, $\text{E}_\omega[\rho]$, y la varianza, $\text{Var}_\omega[\rho]$, se calculan de la siguiente manera:

$$q(\rho|\omega) = \frac{\gamma e^{\gamma\rho}}{1 - e^\gamma}, \quad (1)$$

$$\text{E}_\omega[\rho] = \frac{1}{1 - e^{-\omega}} - \frac{1}{\omega}, \quad (2)$$

$$\text{Var}_\omega[\rho] = \frac{1 + e^{2\omega} - e^\omega(2 + \omega^2)}{(e^\omega - 1)^2\omega^2}. \quad (3)$$

Además denotamos $\text{KL}(\lambda, \lambda')$ como la distancia de Kullback-Leibler entre dos distribuciones Dirichlet con parámetros λ y λ', que se puede expresar como

$$\begin{aligned}\text{KL}(\lambda, \lambda') = &\sum[(\lambda - \lambda')(\psi(\lambda) - \psi(\lambda_0))] \\ &+ \sum[\ln\Gamma(\lambda') - \ln\Gamma(\lambda)] \\ &+ \ln\Gamma(\lambda_0) - \ln\Gamma(\lambda'_0),\end{aligned} \quad (4)$$

siendo $\psi(x) = \frac{d}{dx}\ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ la conocida función *digamma*, $\lambda_0 = \sum\lambda$ y $\lambda'_0 = \sum\lambda'$.

Algoritmo 2 Calcular $q(\rho|\omega) \approx p(\rho|x_0, x_1)$.

Entrada: (x_0, x_1)
Salida: $p(\rho|\omega)$
1: $\lambda_u = (1, ..., 1)$
2: $\lambda_1 = x_0 + \lambda_u$
3: $\lambda_2 = x_1 + \lambda_1$
4: **Repetir**
5: $\quad\omega = \text{KL}(\lambda_2, \lambda_u) - \text{KL}(\lambda_2, \lambda_1) + \gamma$
6: $\quad\lambda_2 = x_1 + \text{E}_\omega[\rho]\lambda_1 + (1 - \text{E}_\omega[\rho])\lambda_u$
7: **hasta** convergencia
8: **return** $q(\rho|\omega)$

IV. Experimentos

En esta sección vamos a ver una serie de experimentos (realizados en R) con el objetivo de comparar las diferencias entre usar el método Chi-cuadrado, y el modelo HPP. Para ello vamos a analizar cómo varían por un lado los p-valores, y por otro, la salida del modelo HPP. Para llevar esto a cabo, consideraremos una distribución multinomial p_0, muestrearemos y obtenemos una muestra x_0, donde el tamaño muestral irá variando. Con el mismo tamaño de la muestra, consideramos también una distribución p_1, muestreamos y obtenemos otra muestra x_1. Hecho esto es cuando aplicamos ambos test de hipótesis y comparamos los resultados. Este experimento lo repetiremos para diferentes tamaños de muestra (de menor a mayor). A su vez para cada tamaño muestral, muestrearemos un total de 1000 veces las muestras x_0 y x_1 y daremos un diagrama de cajas mostrando los resultados.

Como hemos dicho, el test Chi-cuadrado da como respuesta un valor, el p-valor, mientras que el método HPP da como respuesta una probabilidad a posteriori $p(\rho|x_0, x_1)$. Con el fin de comparar ambos, de la salida del método HPP mostraremos los siguientes valores: el *máximo a posteriori* (MAP) de ρ, $\arg\max_\rho p(\rho|x_0, x_1)$; el valor esperado de ρ, $\mathrm{E}_\omega[\rho]$; y la varianza de ρ, $\mathrm{Var}_\omega[\rho]$.

IV-A. Comparación de $q(\rho|\omega)$ y P-valores para dos probabilidades p_0 y p_1 iguales cuando aumenta el tamaño muestral.

En este caso consideramos que ambas distribuciones son iguales con $p_0 = p_1 = (0.5, 0.2, 0.1, 0.1, 0.1)^T$. De este experimento extraemos las siguientes conclusiones:

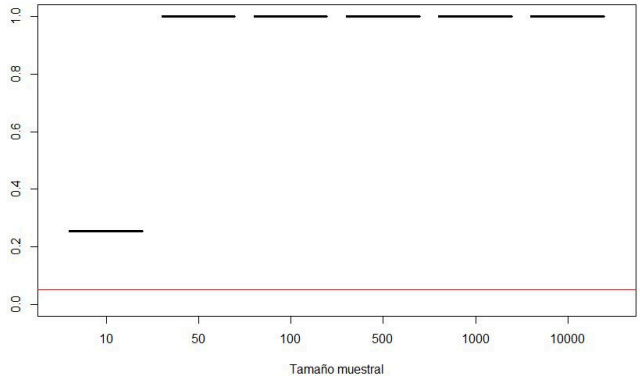

Figura 2: P-valor para dos distribuciones iguales cuando aumenta el tamaño muestral. La línea roja horizontal muestra el p-valor 0.05.

- Comenzaremos viendo qué ocurre con el p-valor. En la Fig. 2 podemos ver la representación gráfica de éstos según aumenta el tamaño muestral. Como puede verse con 10 muestras sale un p-valor menor que con el resto de muestras, pero aún así son valores suficientemente altos como para aceptar la hipótesis nula de igualdad entre las dos distribuciones al $95\,\%$ de confianza (es decir, con un nivel de significación de 0.05).
- En segundo lugar, representamos gráficamente el MAP de ρ en la Fig. 3. Como puede verse, la representación del

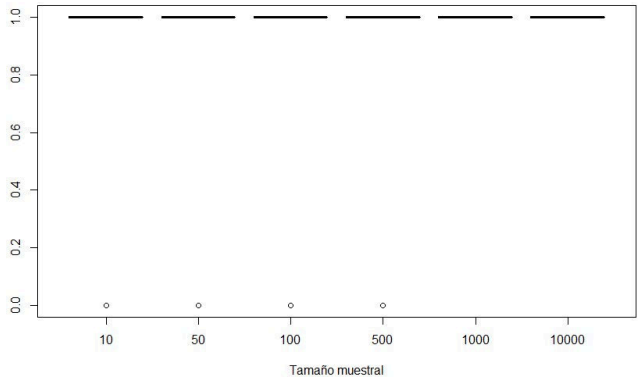

Figura 3: MAP para dos distribuciones iguales cuando aumenta el tamaño muestral.

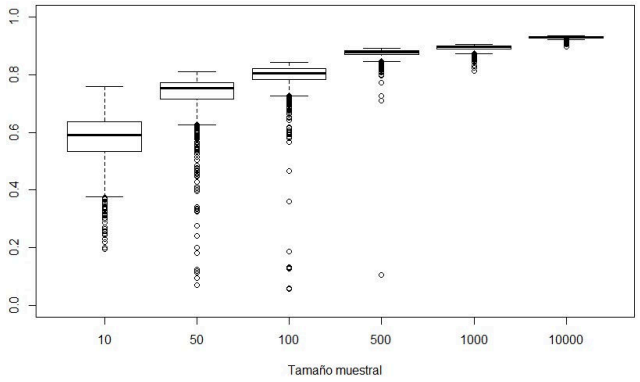

Figura 4: $\mathrm{E}_\omega[\rho]$ para dos distribuciones iguales cuando aumenta el tamaño muestral.

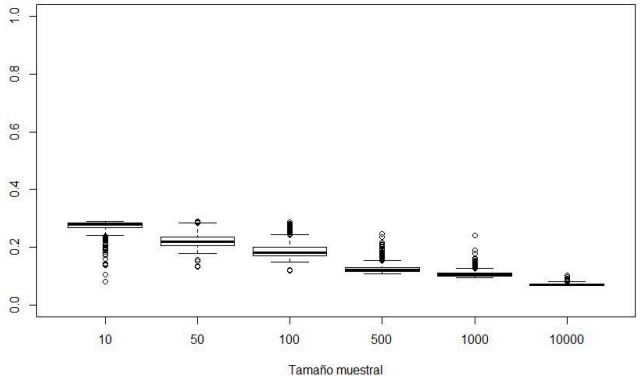

Figura 5: $\mathrm{Var}_\omega[\rho]$ para dos distribuciones iguales cuando aumenta el tamaño muestral.

MAP coincide bastante con la de los p-valores obtenidos mediante el método chi cuadrado. En este caso sin embargo, disponemos de una probabilidad a posteriori que nos permite saber también el valor esperado para ρ.

- En la Fig. 4 vemos la representación para $E_\omega[\rho]$ como comentábamos en el punto anterior. En esta imagen, y a diferencia de los p-valores, podemos ver que ésta va aumentando según aumenta el número de muestras y la variabilidad de la misma se va haciendo cada vez más pequeña, y por ello, más precisa. Esta precisión también puede verse a través de la varianza.
- En la Fig. 5 tenemos la representación gráfica de $\text{Var}_\omega[\rho]$. En esta puede apreciarse como la varianza se va reduciendo conforme aumenta el número de muestras, pudiendo comprobarse que, efectivamente, la incertidumbre se reduce según se aumenta el tamaño de la muestra.

IV-B. Comparación de $q(\rho|\omega)$ y P-valores para dos probabilidades p_0 y p_1 distintas cuando aumenta el tamaño muestral.

En este segundo experimento consideraremos que las dos distribuciones son diferentes con p_0 la misma que en el caso anterior y $p_1 = (0.1, 0.2, 0.5, 0.1, 0.1)$. De este experimento extraemos las siguientes conclusiones:

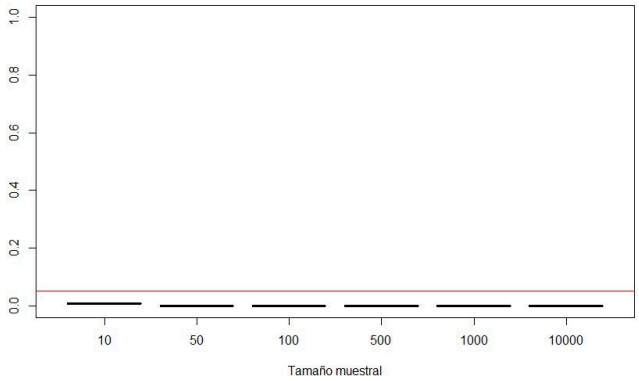

Figura 6: P-valor para dos distribuciones distintas cuando aumenta el tamaño muestral. La línea roja horizontal muestra el p-valor 0.05.

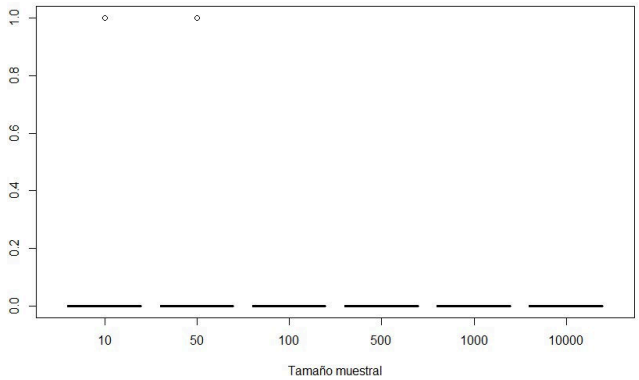

Figura 7: MAP para dos distribuciones distintas cuando aumenta el tamaño muestral.

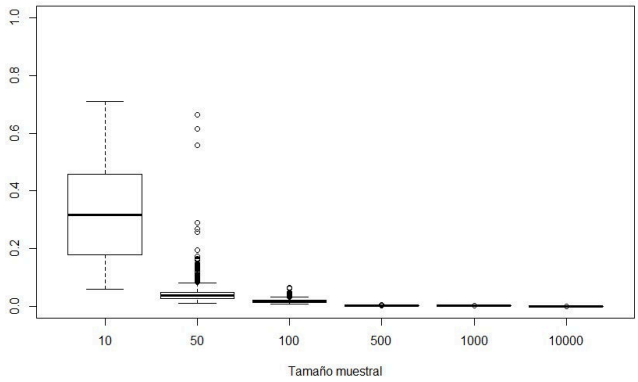

Figura 8: $E_\omega[\rho]$ para dos distribuciones distintas cuando aumenta el tamaño muestral.

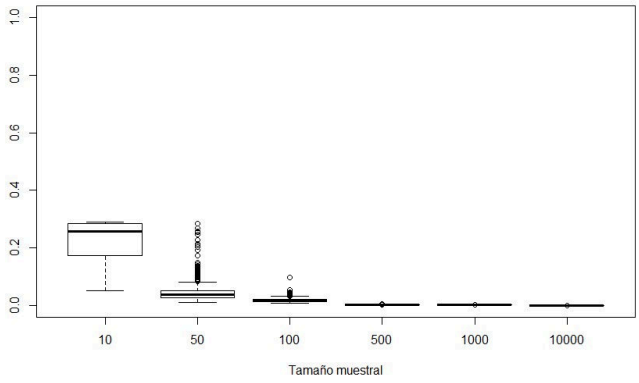

Figura 9: $\text{Var}_\omega[\rho]$ para dos distribuciones distintas cuando aumenta el tamaño muestral.

- Al igual que en el caso anterior, empezaremos por el p-valor. En la Fig. 6 podemos ver la representación gráfica de éstos. Desde el primer resultado con 10 muestras ya puede verse un p-valor muy pequeño, con lo que se rechazaría la hipótesis nula de igualdad de distribuciones.
- A continuación, representamos gráficamente el MAP de ρ en la Fig. 7. Como puede verse, la representación es prácticamente nula, como ocurría en el caso de los p-valores, lo que da como resultado el rechazo de la hipótesis nula.
- En cuanto a $E_\omega[\rho]$, podemos ver en la Fig. 8 la representación de éstas según el tamaño muestral. Al contrario que en la Fig. 4, en esta imagen tenemos que los valores esperados de ρ se van haciendo más pequeños a medida que aumenta el tamaño muestral, y, además, su variabilidad decrece de manera más rápida, por lo que es más exacta que en el caso anterior.
- Por último, vemos la representación gráfica de $\text{Var}_\omega[\rho]$ en la Fig. 9. Como comentábamos en el caso de la Fig. 8, se ve que la representación es más exacta, cosa que también puede verse reflejada en la varianza de ρ, dado que llega un momento que es prácticamente nula.

IV-C. Comparación de $q(\rho|\omega)$ y P-valores cuando p_1 se aleja de p_0, para un tamaño muestral fijo.

En este tercer y último experimento lo que hacemos es partir de dos distribuciones iniciales iguales y vamos modificando la segunda considerando un tamaño muestral fijo, en este caso, $p_0 = (0.5, 0.2, 0.1, 0.1, 0.1)^T$ y el tamaño muestral de 50. Estas modificaciones vienen dadas mediante la expresión

$$p_1 = (1 - \lambda_i)p_0 + \lambda_i \lambda_u, \tag{5}$$

donde $\lambda_i = \frac{i}{10}$ e $i = 0, 1, 2, ..., 10$. Las representaciones gráficas que veremos a continuación son el resultado de ir comparando las distribuciones p_0 y p_1.

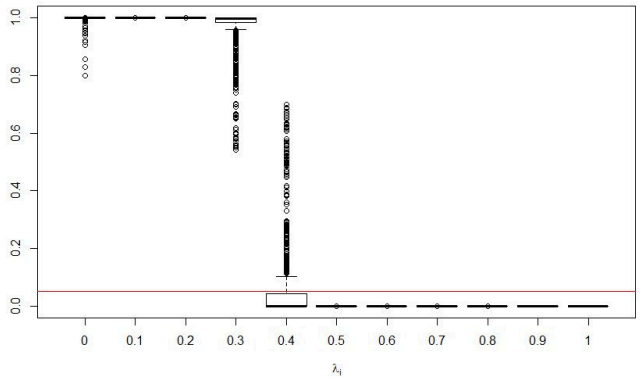

Figura 10: P-valor cuando una distribución se aleja de la otra según (5), para un tamaño muestral fijo. La línea roja horizontal muestra el p-valor 0.05.

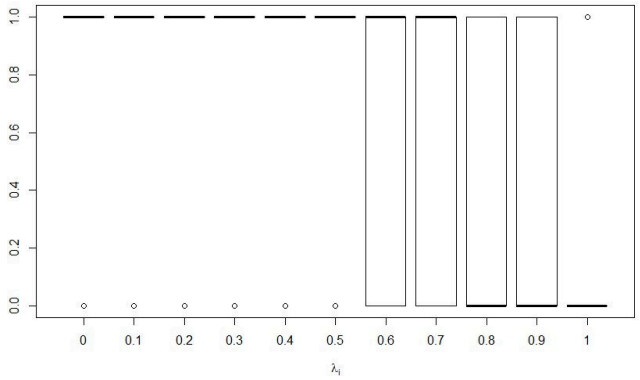

Figura 11: MAP de $q(\rho|\omega)$ cuando una distribución se aleja de la otra según (5), para un tamaño muestral fijo.

De este experimento extraemos las siguientes conclusiones:

- En la Fig. 10 tenemos la representación de los p-valores para las distintas distribuciones. Puede verse que con las tres primeras distribuciones, las clasifica como iguales, es decir, se acepta la hipótesis nula; para 0.4, comete bastantes errores a la hora de clasificarlas; y a partir de 0.5 no hay evidencia suficiente para aceptar la hipótesis nula, lo que quiere decir que las dos distribuciones no serán iguales.

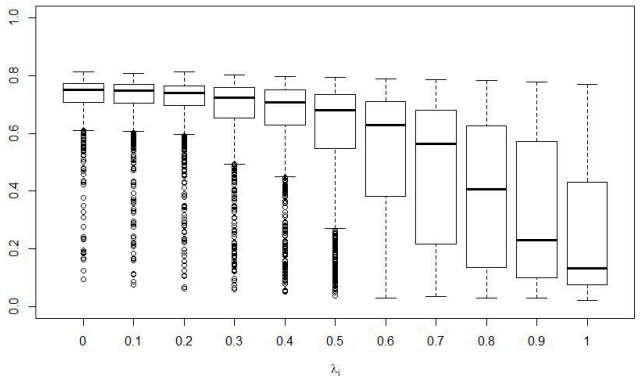

Figura 12: $\mathrm{E}_\omega[\rho]$ cuando una distribución se aleja de la otra según (5), para un tamaño muestral fijo.

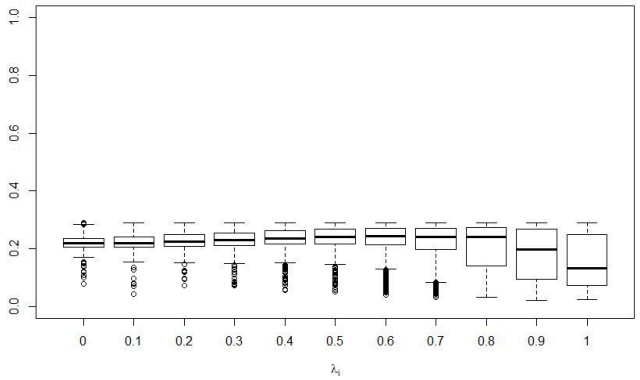

Figura 13: $\mathrm{Var}_\omega[\rho]$ cuando una distribución se aleja de la otra según (5), para un tamaño muestral fijo.

- En la figura Fig. 11 tenemos la representación gráfica del MAP. En este caso, podemos ver que se acepta la hipótesis nula hasta 0.7, pero para los valores entre 0.6 y 0.9 hay bastante error a la hora de tomar la decisión.
- La Fig. 12 está representada el valor esperado de ρ para cada una de las distintas probabilidades. Podemos ver que comenzamos con valores bastante altos debido a que las dos distribuciones son iguales en 0, y, a medida que vamos aumentando p_1, va tomando valores cada vez más pequeños. A su vez, puede verse que la variabilidad al comienzo es bastante más pequeña, y conforme se va avanzando en la gráfica, esta se va haciendo más grande.
- En cuanto a $\mathrm{Var}_\omega[\rho]$, podemos ver en la Fig. 13 que hasta el valor 0.7 va creciendo un poco (lo que explica que la variabilidad de la gráfica anterior sea un poco mayor entre estos valores), pero decrece a partir de este hasta el final, tomando incluso un valor más bajo que en inicial.

V. Conclusiones y Trabajos Futuros

En este artículo hemos explorado un test Bayesiano de dos muestras basado en el modelo HPP [2]. Este test devuelve una probabilidad *a posteriori* sobre la tasa de cambio de la distribución, $p(\rho|x_0, x_1)$. Hemos visto que se comporta de forma similar al test Chi-cuadrado cuando comparamos con el

valor MAP. Pero tiene la ventaja de que es capaz de cuantificar la incertidumbre asociada al tamaño de la muestra, al contratio que el test Chi-cuadrado. Como trabajos futuros, pretendemos extender el test para datos continuos y compararlo con otros test Bayesianos presentes en en la literatura.

Referencias

[1] A. Martín, J. D. Luna, "Bioestadística para las ciencias de la salud (+)", Capitel Editores, 2004.

[2] A. Masegosa, T. D. Nielsen, H. Langseth, D. López-Ramos, A. Salmerón, A. L. Madsen, "Bayesian Models of Data Streams with Hierarchical Power Priors," PMLR 70, Sydney, Australia, 2017.

[3] G. M. Sullivan, R. Feinn, "Using effect size—or why the P value is not enough", J Grad Med Educ, vol. 4, pp 279–282, 2012.

[4] J. Nielsen, "Putting A/B testing in its place", Useit.com Alertbox, 2005.

[5] K. M. Borgwardt, Z. Ghahramani, "Bayesian two-sample tests", *arXiv preprint arXiv:0906.4032*, Jun 2009.

[6] M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, "Stochastic variational inference", The Journal of Machine Learning Research, vol. 14, pp. 1303-1347, 2013.

[7] W. L. Hu, P. C. Wu, L. Y. Pan, H. J. Yu, C. C. Pan, Y. C. Hung, "Effect of laser acupuncture on dry eye: A study protocol for a 2-center randomized controlled trial", Medicine, vol. 97, Jun 2018.