UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2015, Vol. 8, No. 2, pp. 225–243

Intervalos de confianza para valores propios en el análisis de correspondencias a partir de una muestra probabilística[1]

Confidence Interval for eigenvalues in Analysis of correspondence from a probabilistic sample

Javier Ramírez Montoya[a]
javier.ramirez@sunisucre.edu.co

Guillermo Martínez Flórez[b]
gmartinez@correo.unicordoba.edu.co

Stalyn Guerrero Gómez[c]
stalynguerrero@usantotomas.edu.co

Resumen

En este artículo se comparan los intervalos de confianza para los valores propios en el análisis de correspondencia a partir de una muestra probabilística propuesto en (Ramirez 2010), para lo cual se utiliza un diseño de muestreo probabilístico $p(.)$ y estrategias de remuestreo como es el caso de *Bootstrap-t* y *Jackknife-delete I*. Además se utilizan los intervalos de asintóticos de Anderson, ya que son utilizados frecuentemente en la práctica. Esta comparación se basa en la longitud y tasas de cobertura de los intervalos de confianza.

***Palabras clave*:** Análisis de correspondencias, *Bootstrap*, intervalos de confianza, *Jacknife*.

Abstract

This work compares the confidence intervals for the eingenvalues in the analysis of correspondence from a probabilistic sample proposed in (Ramirez 2010) using a probabilistic sampling design $p(.)$, and strategies of resampling as the case of *Bootstrap* and *Jackknife-delete I*, also used Anderson Asymptotics intervals, as they are used frequently in practice, this comparison is based on the length and coverage of the confidence intervals.

***Keywords*:** Analysis of correspondences,*Bootstrap*, confidence intervals,*Jacknife*.

1. Introducción

Un requisito fundamental en el análisis de correspondencia a partir de una muestra probabilística es la obtención de valores propios que permita, entre otras cosas, la obtención de las estimaciones de las coordenadas factoriales de los puntos filas y de los puntos columnas sobre los ejes factoriales. Con ello se obtiene la interpretación de las asociaciones entre las variables categóricas. Con relación a la estimación de los valores propios en un análisis de correspondencia bajo una muestra probabilística de diseño $p(.)$, el problema de la aplicación del análisis de correspondencia a partir de una muestra probabilística radica en la verificación de la calidad de las estimaciones de los valores propios.

En muchas ocasiones, resulta imposible expresar de manera explícita el estimador de algunos parámetros que intervienen en el modelo estadístico, debido al desconocimiento de la distribución de estos. Por tal razón, las medidas de la calidad

con los intervalos de confianza propuestos bajo esta metodología; en la sección 4 se describen los diseños de muestreo probabilístico MAS, πPT y πPT-MAS, en la sección 5 se hace una descripción de los escenarios de simulación y resultados; finalmente se presentan las conclusiones en la sección 6.

2. Análisis de correspondencias a partir de una muestra probabilística

(Ramirez 2010) propone una metodología para realizar el procedimiento de estimación de los elementos de base en el análisis de correspondencias, enfocado fundamentalmente en las estimación de los valores propios, mediante el π estimador de Horvitz-Thompson. Esto se hace con base en el muestreo de poblaciones finitas y el principio de expansión para estimar el total, reflejado en incrementar la importancia de los elementos de la muestra.

2.1. AC Simple

Dada la población $U = \{1, \ldots, N\}$. Suponga que a los elementos de U se le miden 2 variables, digamos Z_1 y Z_2, con p_1 y p_2 modalidades respectivamente. La matriz de datos, resultado de la medición de las variables sobre los N individuos es como sigue:

$$Z = \left[Z_1 \vdots Z_2 \right]$$

La matriz por diagonalizar es:

$$\widehat{S} = \widehat{F}' \widehat{D}_{p1}^{-1} \widehat{F} \widehat{D}_{p2}^{-1} \tag{1}$$

Con $\widehat{F}$ la matriz de frecuencias relativas estimada, $\widehat{D}_{p1}$ la matriz diagonal de perfiles fila y $\widehat{D}_{p2}$ la matriz diagonal de perfiles columna, para el caso simple, los valores propios estimados se encuentran resolviendo el polinomio característico:

$$|\widehat{S} - \widehat{\lambda} I_{p2}| = (-1)^{p2} \left(\widehat{\lambda}^{p2} + \widehat{b}_{p-1} \widehat{\lambda}^{p2-1} + \cdots + \widehat{b}_1 \widehat{\lambda} + \widehat{b}_0 \right) = 0 \tag{2}$$

Donde

$$\widehat{b}_r = f\left(\widehat{k}_{11}, \widehat{k}_{12}, \ldots, \widehat{k}_{1p2}, \widehat{k}_{21}, \widehat{k}_{22}, \ldots, \widehat{k}_{2p2}, \ldots, \widehat{k}_{p_1 1}, \widehat{k}_{p_1 2}, \ldots, \widehat{k}_{p_1 p_2} \right) \tag{3}$$

resultan ser funciones de totales o dominios estimados, los cuales se proponen mediante:

$$\widehat{k}_{ij\pi} = \sum_{l \in s} \frac{z_{ijl}}{\pi_l}$$

Donde, z_{ijl} corresponde a una variable dicótoma, si el individuo l selecciona la modalidad i de la pregunta p_1 y j de la pregunta p_2 corresponde el valor 1 y 0 cuando no sucede, bajo un diseño de muestreo probabilístico $p(.)$ con probabilidades de inclusión π_l.

2.2. AC múltiple

Para el caso múltiple, a través de un π estimador basado en una muestra probabilística s de tamaño n corresponden a m variables con $p_1, ..., p_m$ modalidades respectivamente, la matriz por diagonalizar es:

$$S = F' D_N^{-1} F D_p^{-1} = \frac{1}{m} Z' Z D^{-1} = \frac{1}{m} B D^{-1} \tag{4}$$

$$\widehat{t}_{zj'} = \sum_{i \in s} \frac{z_{ij'}}{\pi_i} \tag{5}$$

y

$$\widehat{t}_{zjj'} = \sum_{i \in s} \frac{z_{ij} z_{ij'}}{\pi_i}$$

se puede establecer que un estimador aproximado para λ es de la forma

$$\widehat{\lambda} = f\left(\widehat{t}_{z_1}, \ldots, \widehat{t}_{z_{j'}}, \ldots, \widehat{t}_{z_p}, \ldots, \widehat{t}_{z_{11}}, \ldots, \widehat{t}_{z_{jj'}}, \ldots, \widehat{t}_{z_{pp}}\right) \tag{6}$$

resultado de resolver el polinomio característico estimado a partir de la muestra s de la matriz

$$\widehat{S} = Z_n' \Pi^{-1} Z_n \widehat{D}^{-1} = \widehat{B} \widehat{D}^{-1} \tag{7}$$

dado por

$$p(\lambda) = \left|\widehat{S} - \widehat{\lambda} I\right| = (-1)^p \left(\widehat{\lambda}^p + \widehat{b}_{p-1} \widehat{\lambda}^{p-1} + \cdots \widehat{b}_1 \widehat{\lambda} + \widehat{b}_0\right) \tag{8}$$

En efecto, la construcción de elementos de base que componen el análisis de correspondencias a partir de una muestra probabilística, requieren una buena estimación de los valores propios y sus vectores asociados.

3. Intervalos de confianza

3.1. Intervalos de confianza Anderson a partir de una muestra probabilística

Sean $\lambda_1 > \lambda_2 > ... > \lambda_p$ los valores propios de la matriz por diagonalizar S en el análisis de correspondencia con vectores propios asociados $\beta_1, \beta_2, ..., \beta_p$ or-

tonormales. Sean $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p$ los valores propios en obtenidos en el análisis de correspondencia a partir de una muestra probabilística.

Se puede, entonces, mostrar que:

$$\hat{\lambda}_i \underset{n\to\infty}{\sim} N(\lambda_i, \frac{2\lambda_i^2}{n}) \tag{9}$$

Donde $\hat{\lambda}_i$ es un estimador consistente de λ_i. Así, mediante el teorema de Chebyshev, se tiene que:

$$\lim_{n\longrightarrow\infty} P(\mid \hat{\lambda}_i - \lambda_i \mid < \varepsilon) = 1, \ \forall \varepsilon > 0 \tag{10}$$

Teniendo en cuenta (9) obtenemos que:

$$P\left[-Z_{\alpha/2} < \sqrt{n}\frac{(\hat{\lambda}_i - \lambda_i)}{\sqrt{2}\lambda_i} < Z_{\alpha/2}\right] = 1 - \alpha \tag{11}$$

$$P\left[-\sqrt{2}Z_{\alpha/2}\lambda_i < \sqrt{n}(\hat{\lambda}_i - \lambda_i) < \sqrt{2}Z_{\alpha/2}\lambda_i\right] = 1 - \alpha$$

Entonces resolviendo tenemos:

$$P\left[\frac{\hat{\lambda}_i}{1 + \sqrt{\frac{2}{n}}Z_{\alpha/2}} < \lambda_i < \frac{\hat{\lambda}_i}{1 - \sqrt{\frac{2}{n}}Z_{\alpha/2}}\right] = 1 - \alpha \tag{12}$$

Por lo tanto, un intervalo de confianza de $100(1-\alpha)\,\%$ para los valores propios λ_i en el análisis de correspondencia es:

$$\left[\frac{\hat{\lambda}_i}{1 + \sqrt{\frac{2}{n}}Z_{\alpha/2}}, \frac{\hat{\lambda}_i}{1 - \sqrt{\frac{2}{n}}Z_{\alpha/2}}\right] \tag{13}$$

3.2. Intervalos de confianza Bootstrap a partir de una muestra probabilística

Teniendo en cuenta la importancia de utilizar el método de remuestreo Bootstrap para estimar valores propios, (Milan 1995), realiza una aplicación del Bootstrap paramétrico a modelos que incorporan valores singulares, donde se evidencian discusiones importantes sobre el efecto de la variación de muestreo en las estimaciones.

Para estimar los valores propios mediante el método el remuestreo Bootstrap paramétrico, se realizan los siguientes pasos:

1. Dada la muestra de tamaño n, calcular $\hat{\lambda}$. La distribución de esta muestra se considera equivalente a la distribución de la población y $\hat{\lambda}$ es el estimador muestral del parámetro poblacional λ.

2. Generar B muestras Bootstrap de tamaño n mediante muestreo con reemplazamiento de la muestra original y calcular los correspondientes valores $\hat{\lambda}^{*1}, \hat{\lambda}^{*2}, ..., \hat{\lambda}^{*B}$ para cada una de las B muestras Bootstrap.

3. Estimar el error estándar del parámetro estimado $\hat{\lambda}$ calculando la desviación estándar de las B réplicas Bootstrap.

Así, obtenemos que el error estándar es este

$$\sigma^*_{\lambda} = \sqrt{\frac{\sum_{b=1}^{B}\left(\lambda^{b^*} - \bar{\hat{\lambda}}^*\right)^2}{(B-1)}} = \sigma_{BOOT} \tag{14}$$

Donde

$$\bar{\hat{\lambda}}^* = \frac{1}{B}\sum_{b=1}^{B}\lambda^{b^*} \tag{15}$$

corresponde al promedio de los valores propios calculados en cada remuestra. Entonces un intervalo de confianza para los valores propios en el análisis de correspondencia está dado por:

$$\left[\hat{\lambda}^* - Z_{\alpha/2}\sigma^*_{\lambda}, \hat{\lambda}^* + Z_{\alpha/2}\sigma^*_{\lambda}\right] \tag{16}$$

Donde σ^*_{λ} corresponde al error estandar *Bootstrap* dado en (14)

3.3. Intervalos de confianza Jacknife a partir de una muestra probabilística

Dada la cantidad de parámetros por calcular para estimar la varianza de Horvitz-Thompson se estudian los métodos de Jackknife y Bootstrap para la estimación de la varianza, los cuales son usados para este tipo de situaciones dada su simplicidad de cálculo y a los supuestos para su aplicación; por lo tanto, la estimación para la varianza de $\widehat{\lambda}$, según el método Jackknife presentado en (Wolter 1985) para parámetro de interés, en efecto para el análisis de correspondencias a partir de una muestra probabilística se propone como:

$$v_{jk} = \frac{n-1}{n}\sum_{l=1}^{n}\left(\widehat{\lambda}_{n-1,l} - \frac{1}{n}\sum_{l'=1}^{n}\widehat{\lambda}_{n-1,l'}\right)^2 \tag{17}$$

Luego este estimador se conoce como el *estimador Jackknife* (delete-1) de $V\left(\widehat{\lambda}_n\right)$, donde:

$$\widehat{\lambda}_{n-1,l} = f\left(\widehat{k}_{11\pi_{(n-1)}}, \widehat{k}_{12\pi_{(n-1)}}, \ldots, \widehat{k}_{p_1 1\pi_{(n-1)}}, \widehat{k}_{p_1 2\pi_{(n-1)}}, \ldots, \widehat{k}_{p_1 p_2 \pi_{(n-1)}}\right) \tag{18}$$

y

$$\widehat{k}_{ij\pi(n-1,l)} = \sum_{l' \in S-\{l\}} \frac{z_{ijl'}}{\pi_{l'}} \tag{19}$$

Es decir, $\widehat{\lambda}_{n-1,l}$ es el estimador del valor propio correspondiente, basado en la muestra de tamaño $n-1$ que resulta luego de eliminar el individuo l-ésimo de la muestra y $\widehat{k}_{ij\pi(n-1,l)}$ es la estimación de un dominio eliminado la misma observación.

Para determinar la consistencia del estimador de la varianza mediante el método Jackknife en el análisis de correspondencias a partir de una muestra probabilística. Suponemos una muestra aleatoria *iid* con distribución p-dimensional F. Además, sea $\lambda = \lambda_n(X_1, ..., X_n)$ la estadística de interés, es decir, el valor propio, bajo condiciones de regularidad y sea $\Sigma = var(\lambda)$ conocida:

$$(\lambda_n - \lambda)/\sigma_n \xrightarrow{d} N(0,1) \tag{20}$$

Donde $N(0,1)$ denota la distribución de una variable aleatoria normal estandar, para $\lambda = g(\lambda)$:

$$\sigma_n^2 = n^{-1} \nabla g(\lambda)' \Sigma \nabla g(\lambda) \tag{21}$$

y ∇g el gradiente de g. σ_n^2 es la varianza de λ_n, así:

$$(\lambda_n - \lambda)/\sqrt{v_{jk}} \rightarrow_d N(0,1) \tag{22}$$

Este es un resultado útil para inferencias estadísticas de muestras grandes. Entonces un intervalo de confianza para los valores propios mediante la técnica $Jackknife$ está dado por:

$$\left[\widehat{\lambda}_{n-1,l} - Z_{\alpha/2}\sqrt{v_{jk}}, \widehat{\lambda}_{n-1,l} + Z_{\alpha/2}\sqrt{v_{jk}}\right] \tag{23}$$

Donde v_{jk} esta dado en (17)

3.4. Estimación mediante muestreo aleatorio simple (MAS)

Bajo un diseño de muestreo probabilístico (MAS), con tamaño de muestra fijo en el que $\pi_k = \frac{n}{N}$ para todo $k = 1, 2, ..., N$ y $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ para $k \neq l$, $k, l = 1, 2, ..., N$ (Sarndal 2003). Luego, en el análisis de correspondencias a partir de una muestra probabilística del espacio $U = \{1, ..., N\}$, en el caso de m variables con $p_1, ..., p_m$ modalidades, respectivamente, se puede establecer un estimador aproximado para el valor propio λ como:

$$\widehat{\lambda} = f\left(\widehat{t}_{z_1}, \ldots, \widehat{t}_{z_{j'}}, \ldots, \widehat{t}_{z_p}, \ldots, \widehat{t}_{z_{11}}, \ldots, \widehat{t}_{z_{jj'}}, \ldots, \widehat{t}_{z_{pp}}\right) \tag{24}$$

Esto es resultado de resolver el polinomio característico estimado a partir de la muestra s de esta matriz:

$$\widehat{S} = Z_n' \Pi^{-1} Z_n \widehat{D}^{-1} = \widehat{B}\widehat{D}^{-1} \tag{25}$$

Donde la matriz de Burt estimada

$$\widehat{B} = Z_n' \Pi^{-1} Z_n \tag{26}$$

la matriz diagonal estimada obtenida a partir de la matriz de Burt, corresponde a una matriz de orden (p, p), dada por:

$$\widehat{D} = \text{diag}\{Z_n' \Pi^{-1} Z_n\} \tag{27}$$

y Π es la matriz diagonal de probabilidades de inclusión:

$$\Pi = \text{diag}\{\pi_1, \ldots, \pi_n\} \tag{28}$$

Donde los factores de expansión que afectan la estimación son $f_k = \frac{1}{\pi_k} = \frac{N}{n}$

3.5. Estimación mediante muestreo πPT

En este diseño se utiliza la información auxiliar disponible para todo $k \in U$ de una variable X altamente correlacionada con la variable de estudio Y. Con ayuda de esta información, se construyen probabilidades proporcionales al tamaño π_k, para todos los elementos de la población y fijando de antemano el tamaño de muestra n, donde:

$$\pi_k = \frac{n x_k}{\sum_U x_k}$$

con factores de expansión:

$$f_k = \frac{\sum_U x_k}{n x_k}$$

para todo $k = 1, ..., N$

3.6. Estimación mediante muestreo bietápico πPT-MAS

En el caso en el que la primera etapa se realiza un diseño πPT para seleccionar las UMP's tomando n_i de los N_i que componen la unidad, con:

$$\pi_I = \frac{n_I x_k}{\sum_U x_k}$$

Luego dentro de cada UMP se realiza un diseño MAS donde:

$$\pi_j = \frac{n_j}{N_j}$$

Se obtienen así los factores de expansión:

$$f_{exp} = f_1 * f_2 = \frac{1}{\pi_k} = \frac{(\sum_U x_k)(N_{i|I})}{(n_I x_k)(n_{i|I})}$$

para $k = 1, ..., N$.

4. Estudio de simulación

Con el objetivo de realizar la estimación de los valores propios a través de los intervalos de confianza en el análisis de componentes principales a partir de una muestra probabilística, se tomó como la población objetivo la base de datos Finanzas.SBD del *software* SPAD, la cual contiene 468 registros de clientes de un Banco. La tiene registrada las variables: el tipo de cliente (Tipo) con dos modalidades, donde el 50.64 % pertenece a la categoría 1 y el 49.36 % está en la categoría 2; Edad tiene cuatro modalidades que están distribuidas en la población así: 18.80 %, 32.05 %, 26.07 % y 23.08 % para las categorías 1, 2, 3 y 4, respectivamente; Estado Civil con las categorías 1, 2, 3 y 4, que cuentan con el 36.32 %, 47.22 % 13.03 % y 3.42 %, respectivamente; Antigüedad en el trabajo, que cuenta con 5 categorías están presentes en la población con un 42.52 % para la modalidad 1, 10.04 % para la 2, 14.74 % para la 3, 14.10 % para la 4 y 18.59 % en la categoría 5: finalmente la variable Profesión que tiene tres modalidades con el 16.45 % 50.64 % y 32.91 % en las modalidades 1, 2 y 3. Además, la base contiene la variable Cupo disponible, que se empleará como información auxiliar en los diseños de muestreo.

Para la realización del estudio de simulación mediante el *software* R, se generó una población artificial de tamaño $N = 20000$, mediante una distribución multinomial de tal forma que las proporciones descritas anteriormente se conservarán posteriormente se seleccionan muestras probabilísticas de tamaños $n = 25,\ 50\ y\ 100$, mediante los diseños MAS y πPT en una etapa. A su vez, en dos etapas se utilizó el diseño πPT-MAS, para emplear el diseño en dos etapas se definen 4 grupos aleatorios. La asignación de los elementos dentro de estos grupos se hizo de forma aleatoria, los tamaños de las muestras probabilística a seleccionar dentro de los grupos se hace mediante la asignación óptima. Se desea comparar la cobertura y la longitud promedio de los intervalos de confianzas construidos con la metodologías tradicionales y la propuesta; para ello, se calculan los intervalos de confianza para los valores propios λ_i a partir de las muestras probabilísticas, luego se determina si el valor propio real (calculado a partir de la población) se encuentra dentro del intervalo de confianza y se mide su longitud. Este procedimiento se repite 1000 veces, con lo que se determina la cobertura y la longitud promedio de los intervalos

de confianza construidos con estás metodologías. Para comparar los intervalos de confianza teniendo en cuenta de manera simultánea las cobertura y la amplitud se utiliza el índice de (Correa 2003), dado por:

$$I = \frac{2 - LPI}{2} \frac{NR}{NN}$$

Donde LPI: longitud promedio de los intervalos de confianza, NR: nivel real de simulación, NN: nivel nominal, cuyos resultados se presentan en los anexos, utilizando un nivel nominal del 95 %.

Resultados de simulación

Se seleccionan tamaños de muestra $n = 25,\ 50\ y\ 100$ los resultados para los intervalos de confianza (I.C) asintóticos de Anderson, Bootstrap y Jackknife para los valores propios en un análisis de correspondencias a partir de una muestra probabilística. Estas son sus siglas; VP: valor propio, LONGPA: longitud del Intervalo de Anderson, LONGPB: longitud del Intervalo Bootstrap, LONGPJ: longitud del Intervalo Jackknife, TCA: cobertura I.C Anderson, TCB: cobertura I.C Bootstrap, TCJ: cobertura I.C Jackknife, utilizando un nivel de confianza nominal del 95 %, se presentan a continuación:

4.1. Resultados diseño MAS

Tabla 1: *Resultados para* $n = 25$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,825	**0,986**	**0,154**	0,393	0,262	0,716
2	0,319	0,661	**0,989**	**0,103**	0,141	0,236	0,796
3	0,272	0,550	**0,998**	**0,079**	0,151	0,207	0,859
4	0,258	0,463	**1,000**	**0,060**	0,576	0,184	0,994
5	0,253	0,391	**1,000**	**0,048**	0,758	0,182	0,992
6	0,248	0,317	**0,997**	**0,060**	0,114	0,201	0,855
7	0,214	0,244	**0,979**	**0,058**	0,029	0,185	0,700
8	0,197	0,179	**0,806**	**0,048**	0,000	0,173	0,475
9	0,175	0,120	**0,454**	**0,036**	0,000	0,140	0,261
10	0,153	0,085	**0,175**	**0,023**	0,000	0,099	0,084

Es importante resaltar que los valores en negrilla corresponden a los mejores resultados, para cada valor propio.

Se nota que mediante el diseño de muestreo MAS, las tasas de cobertura de los intervalos de confianza de Anderson resultan ser mayores; sin embargo, resultan ser más amplios, por tener menor amplitud los intervalos de confianza Boostrap. Por otra parte, teniendo en cuenta los dos aspectos simultáneamente: tasas de co-

Tabla 2: *Resultados para* $n = 50$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,431	**0,989**	**0,122**	0,754	0,190	0,854
2	0,319	0,346	**0,996**	**0,079**	0,514	0,161	0,883
3	0,272	0,301	**0,996**	**0,061**	0,306	0,150	0,861
4	0,258	0,264	**1,000**	**0,047**	0,662	0,134	0,993
5	0,253	0,233	**1,000**	**0,037**	0,743	0,119	0,999
6	0,248	0,204	**1,000**	**0,043**	0,103	0,131	0,944
7	0,214	0,173	**0,996**	**0,045**	0,142	0,133	0,906
8	0,197	0,143	**0,971**	**0,042**	0,029	0,130	0,749
9	0,175	0,113	**0,814**	**0,038**	0,002	0,125	0,592
10	0,153	0,085	**0,481**	**0,032**	0,000	0,098	0,329

Tabla 3: *Resultados para* $n = 100$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,262	**0,995**	**0,091**	0,883	0,137	0,934
2	0,319	0,208	**1,000**	**0,059**	0,794	0,115	0,952
3	0,272	0,183	**0,995**	**0,045**	0,549	0,104	0,928
4	0,258	0,166	**1,000**	**0,035**	0,803	0,093	0,991
5	0,253	0,151	**1,000**	**0,028**	0,733	0,082	1,000
6	0,248	0,138	**1,000**	**0,032**	0,097	0,092	0,959
7	0,214	0,122	**1,000**	**0,034**	0,359	0,096	0,976
8	0,197	0,106	**0,997**	**0,033**	0,133	0,094	0,931
9	0,175	0,090	**0,961**	**0,031**	0,056	0,088	0,847
10	0,153	0,074	**0,809**	**0,030**	0,026	0,078	0,684

bertura y amplitud de los intervalos de confianza, el método Jacknnife resulta ser más eficiente. Esto, se cumple para los cinco primeros ejes factoriales (ver anexo). Es importante señalar que las tendencias de mayor cobertura y menor longitud promedio de los intervalos de confianza se presentan en todos los tamaños muestrales $n = 25,\ 50,\ y\ 100$.

Ahora, realizando la estimación mediante un diseño de muestreo probabilístico πPT, se obtienen los resultados que se muestran a continuación:

4.2. Resultados diseño πPT

De igual forma, se nota que las tasas de cobertura de los intervalos de confianza de Anderson resultan ser mayores; además, las longitudes promedios de los intervalos mediante remuestreo Boostrap son menores, lo que se cumple para todos los ejes factoriales y tamaños de muestra en este estudio.

Tabla 4: *Resultados para* $n = 25$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,841	**0,970**	**0,164**	0,375	0,278	0,713
2	0,319	0,663	**0,985**	**0,108**	0,138	0,237	0,758
3	0,272	0,549	**0,989**	**0,081**	0,138	0,214	0,885
4	0,258	0,460	**1,000**	**0,062**	0,531	0,183	0,996
5	0,253	0,383	**1,000**	**0,051**	0,711	0,191	0,981
6	0,248	0,306	**0,999**	**0,063**	0,098	0,197	0,803
7	0,214	0,235	**0,959**	**0,059**	0,042	0,185	0,657
8	0,197	0,169	**0,736**	**0,048**	0,000	0,171	0,422
9	0,175	0,112	**0,366**	**0,035**	0,000	0,135	0,209
10	0,153	0,081	**0,152**	**0,022**	0,000	0,101	0,079

Tabla 5: *Resultados para* $n = 50$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,438	**0,982**	**0,128**	0,738	0,200	0,852
2	0,319	0,350	**0,993**	**0,083**	0,464	0,167	0,843
3	0,272	0,303	**0,990**	**0,063**	0,279	0,155	0,857
4	0,258	0,265	**1,000**	**0,048**	0,642	0,137	0,994
5	0,253	0,232	**1,000**	**0,038**	0,746	0,122	0,998
6	0,248	0,201	**1,000**	**0,045**	0,106	0,137	0,932
7	0,214	0,169	**0,995**	**0,047**	0,117	0,139	0,901
8	0,197	0,138	**0,930**	**0,043**	0,010	0,131	0,684
9	0,175	0,109	**0,718**	**0,037**	0,002	0,124	0,521
10	0,153	0,083	**0,435**	**0,032**	0,000	0,098	0,280

Tabla 6: *Resultados para* $n = 100$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,264	**0,993**	**0,096**	0,876	0,144	0,930
2	0,319	0,210	**0,997**	**0,062**	0,778	0,120	0,937
3	0,272	0,185	**0,991**	**0,047**	0,493	0,110	0,901
4	0,258	0,166	**1,000**	**0,036**	0,785	0,097	0,993
5	0,253	0,151	**1,000**	**0,028**	0,724	0,085	1,000
6	0,248	0,137	**1,000**	**0,033**	0,115	0,097	0,972
7	0,214	0,121	**0,998**	**0,035**	0,298	0,101	0,979
8	0,197	0,104	**0,989**	**0,034**	0,084	0,097	0,897
9	0,175	0,088	**0,936**	**0,032**	0,037	0,093	0,818
10	0,153	0,072	**0,758**	**0,030**	0,011	0,079	0,638

4.3. Resultados diseño πPT-MAS

Finalmente, mediante el diseño de muestreo probabilístico bietápico πPT-MAS, se sigue confirmando las tendencias reflejadas en los demás diseños de muestreo

utilizados.

Tabla 7: *Resultados para* $n = 25$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,826	**0,966**	**0,158**	0,366	0,265	0,739
2	0,319	0,652	**0,991**	**0,106**	0,173	0,229	0,759
3	0,272	0,544	**0,997**	**0,081**	0,153	0,207	0,892
4	0,258	0,457	**1,000**	**0,060**	0,574	0,178	0,996
5	0,253	0,383	**1,000**	**0,047**	0,702	0,181	0,972
6	0,248	0,310	**0,994**	**0,061**	0,090	0,202	0,818
7	0,214	0,237	**0,959**	**0,059**	0,053	0,189	0,690
8	0,197	0,172	**0,782**	**0,048**	0,000	0,174	0,439
9	0,175	0,116	**0,396**	**0,036**	0,000	0,136	0,208
10	0,153	0,081	**0,140**	**0,023**	0,000	0,100	0,088

Tabla 8: *Resultados para* $n = 50$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,429	**0,972**	**0,123**	0,738	0,188	0,845
2	0,319	0,345	**0,994**	**0,081**	0,521	0,161	0,857
3	0,272	0,298	**0,993**	**0,062**	0,341	0,148	0,881
4	0,258	0,262	**1,000**	**0,046**	0,691	0,128	0,994
5	0,253	0,232	**1,000**	**0,036**	0,738	0,118	0,999
6	0,248	0,202	**1,000**	**0,043**	0,108	0,133	0,919
7	0,214	0,171	**0,993**	**0,046**	0,160	0,135	0,891
8	0,197	0,140	**0,950**	**0,043**	0,030	0,131	0,712
9	0,175	0,110	**0,750**	**0,037**	0,003	0,120	0,528
10	0,153	0,086	**0,519**	**0,032**	0,000	0,100	0,354

Tabla 9: *Resultados para* $n = 100$

Eje	VP	LONGPA	TCA	LONGPB	TCB	LONGPJ	TCJ
1	0,410	0,258	**0,971**	**0,089**	0,721	0,133	0,861
2	0,319	0,207	**0,998**	**0,059**	0,764	0,115	0,949
3	0,272	0,182	**0,999**	**0,045**	0,552	0,105	0,945
4	0,258	0,165	**1,000**	**0,034**	0,800	0,092	0,993
5	0,253	0,151	**1,000**	**0,027**	0,749	0,082	1,000
6	0,248	0,137	**1,000**	**0,031**	0,120	0,091	0,928
7	0,214	0,122	**1,000**	**0,034**	0,438	0,101	0,952
8	0,197	0,105	**0,981**	**0,034**	0,172	0,097	0,858
9	0,175	0,089	**0,927**	**0,032**	0,077	0,093	0,790
10	0,153	0,075	**0,827**	**0,030**	0,042	0,079	0,716

5. Conclusiones

- Cuando se realiza el procedimiento de estimación de los valores propios mediante muestreo probabilístico en el análisis de correspondencias, es recomendable realizar la complementación mediante el calculo de los intervalos de confianza, teniendo en cuenta la estrategia más eficiente en este estudio, logrando una mejor decisión en la escogencia de planos factoriales.
- Los intervalos de confianza para los valores propios en el análisis de correspondencias a partir de una muestra probabilística mediante el método de remuestreo Bootstrap resultan tener menor longitud, pero sesgados en sus estimaciones, reflejados en su cobertura.
- Los resultados de los intervalos de confianza mediante remuestreo Jackknife para los diseños de muestreo probabilístico MAS, πPT y πPT-MAS, tienen mejores resultados teniendo en cuenta la longitud y las tasas de coberturas Se resalta que en el diseño de muestreo πPT-MAS se debe tener cuidado en la escogencia de la variable auxiliar para el calculo de los factores de expansión.
- El índice de Correa et al. (2003) puede ser utilizado para comparar los intervalos de confianza de los valores propios en el análisis de correspondencias a partir de una muestra probabilística.

Recibido: 20 de agosto del 2015
Aceptado: 13 de octubre del 2015

Referencias

Correa, J. & Sierra, E. (2003), 'Intervalos de confianza para la comparación de dos proporciones', *Revista Colombiana de Estadística* **26**, 61–75.

Milan, L. & Whittaker, R. (1995), 'Application of the parametric bootstrap to models that incorporate a singular value decomposition', *Applied Statistics* **44**, 31–49.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *`http://www.R-project.org/`

Ramirez, J. & Martinez, G. (2010), 'Análisis de Correspondencias a partir de una muestra probabilística', *Revista Colombiana de Estadística* **33**, 273–293.

Sarndal, C., S. B. . W. J. (2003), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Wolter, K. (1985), *Introduction to Variance Estimation*, Springer-Verlag, Berlin.

A. Tablas

Tabla 10: *Resultados Indices para MAS*

n=25			n=50			n=100		
And.	Boot.	Jack.	And.	Boot.	Jack.	And.	Boot.	Jack.
0,610	0,382	0,655	0,817	0,745	0,814	0,963	0,920	0,988
0,697	0,141	0,739	0,867	0,520	0,855	0,981	0,889	0,991
0,762	0,153	0,811	0,891	0,312	0,838	0,991	0,856	0,990
0,809	0,588	0,950	0,914	0,680	0,975	0,995	0,927	1,017
0,847	0,779	0,949	0,930	0,768	0,989	0,998	0,769	1,024
0,883	0,116	0,810	0,945	0,106	0,929	1,002	0,159	0,997
0,905	0,030	0,669	0,958	0,146	0,890	1,007	0,603	1,012
0,772	0,000	0,457	0,949	0,030	0,737	1,012	0,347	1,001
0,449	0,000	0,256	0,808	0,002	0,584	1,013	0,300	0,995
0,176	0,000	0,084	0,485	0,000	0,329	0,989	0,182	0,951

Tabla 11: *Resultados Indices para πPT*

n=25			n=50			n=100		
And.	Boot.	Jack.	And.	Boot.	Jack.	And.	Boot.	Jack.
0,592	0,362	0,646	0,807	0,727	0,807	0,907	0,878	0,908
0,693	0,137	0,703	0,862	0,468	0,813	0,939	0,794	0,927
0,755	0,139	0,832	0,884	0,284	0,832	0,947	0,507	0,896
0,811	0,542	0,952	0,913	0,660	0,975	0,965	0,811	0,995
0,851	0,729	0,934	0,931	0,770	0,986	0,973	0,751	1,008
0,891	0,100	0,762	0,947	0,109	0,914	0,981	0,119	0,974
0,891	0,043	0,628	0,959	0,120	0,883	0,987	0,308	0,978
0,709	0,000	0,406	0,911	0,010	0,673	0,987	0,087	0,898
0,364	0,000	0,205	0,715	0,002	0,514	0,942	0,038	0,821
0,154	0,000	0,079	0,439	0,000	0,280	0,769	0,011	0,645

B. Códigos

```
require(TeachingSampling)
require("ade4")
require(boot)
# Funciones previas
## tasa de cobertura
TSC<-function(LIM,ValoR){
x<-cbind(LIM,ValoR)[, c(1, 3, 2)]
# se ordena
LIF<=ValoR<=LIS
```

Tabla 12: *Resultados Indices para πPT-MAS*

n=25			n=50			n=100		
And.	Boot.	Jack.	And.	Boot.	Jack.	And.	Boot.	Jack.
0,597	0,355	0,675	0,804	0,729	0,806	0,890	0,725	0,846
0,703	0,172	0,707	0,866	0,526	0,829	0,942	0,780	0,942
0,764	0,155	0,842	0,890	0,348	0,859	0,956	0,568	0,943
0,812	0,586	0,955	0,915	0,711	0,979	0,966	0,828	0,997
0,851	0,722	0,931	0,931	0,763	0,990	0,973	0,778	1,009
0,884	0,092	0,774	0,946	0,111	0,903	0,981	0,124	0,932
0,890	0,054	0,658	0,956	0,165	0,875	0,988	0,453	0,951
0,752	0,000	0,422	0,930	0,031	0,700	0,978	0,178	0,859
0,393	0,000	0,204	0,746	0,003	0,522	0,932	0,080	0,793
0,141	0,000	0,088	0,523	0,000	0,354	0,838	0,044	0,724

```
(apply(x,MARG=1,function(x)
ifelse(x[1]<=x[2]&x[2]<=x[3],1,0)))
}
## Estimación Boot
hat.V.p<- function(z.s,i){
z.sb <- z.s[i,]
(p <-  dudi.coa(z.sb, scann = FALSE)$eig[1:10])
}
## Estimación jackknife
jackknife<-function(z.s,conf=0.95)
{
thetahat <- dudi.coa(z.s, scann = FALSE)$eig[1:10]
n <- length(thetahat)
u <- matrix(NA,ncol=n,nrow=nrow(z.s))
for(i in 1:nrow(z.s)) {
u[i,] <- dudi.coa(z.s[-i,], scann = FALSE)$eig[1:10]
}
#u<-na.omit(u)
mJ<-colMeans(u,na.rm = T)
vjk<-sqrt(((n-1)/(n))*rowSums((t(u)-mJ)^2,na.rm = T))
ICJ<-cbind(LIF=thetahat-qnorm((1-conf)/2,lower.tail=F)*vjk,
LIS=thetahat+qnorm((1-conf)/2,lower.tail=F)*vjk,
Valor.P = mJ)
return(ICJ)
}
-----------------------------------------------------------------
Sim.MAS <- function(datos,n,conf = 0.95){
N<-nrow(datos)
x<-datos[,-c(2,6)]
# obtencion de la matrix z
```

```
z <- acm.disjonctif(x)
# Valores propios reales
(Valor.P.V <- dudi.coa(z, scann = FALSE)$eig)
# para seleccionar de estan la poblacion mediante un MASS
el <- S.SI(N, n)[,1]
x.s <-data.frame(x[Sel,])
# Calculo del factor de expancion
F.exp <- rep(N/n,n)
# Matriz Z ponderada
z.s <- acm.disjonctif(x.s)*F.exp
# Interalo de confianza tradicional mediante Anderson
V.pro <- dudi.coa(z.s, scann = FALSE)$eig[1:10]
ICnorm2 <- data.frame(LI1=V.pro/(1+sqrt(2/n)*
qnorm((1-conf)/2,lower.tail=F)),
LS1=V.pro/(1-sqrt(2/n)*qnorm((1-conf)/2,lower.tail=F)))
TCnorm<-TSC(ICnorm2,Valor.P.V)
# Valores propios estimados mediante Bootstrap
rb<-boot(data=z.s, statistic=hat.V.p, R=1000)
# Intervalo de confianza Bootstrap
MRB<-rb$t0
SD<-sqrt(diag(var(rb$t,na.rm = T)))
ICBoot <- data.frame(LI1=MRB-SD*qnorm((1-conf)/2,
lower.tail=F),
LS1=MRB/(1-SD*qnorm((1-conf)/2,lower.tail=F)))
# Tasa de cobertura
TCBoot<-TSC(ICBoot,Valor.P.V)
#### Intervalo de confianza Bootstrap
ICJacn <- jackknife(z.s,conf)
TCJacn<-TSC(ICJacn[,-3],Valor.P.V)
## Salida
c(Anderson=c(Valor.P = V.pro,LI = ICnorm2[,1],
  LS = ICnorm2[,2],TC = TCnorm),
  Boot =   c(Valor.P = MRB,LI   = ICBoot[,1],
  LS  = ICBoot[,2], TC = TCBoot),
  Jacn =   c(Valor.P = ICJacn[,3],LI   = ICJacn[,1],
  LS  = ICJacn[,2], TC = TCJacn))
}

-----------------------------------------------------------------
#Haciendo el diseño pipt para seleccionar una muestra
#de estan la poblacion.
# Tomando como variable auxiliar el cupo disponible
Sim.pips<-function(datos,n,conf=0.95){
N<-nrow(datos)
x<-datos[,-c(2,6)]
# obtencion de la matrix z
```

```
z <- acm.disjonctif(x)
# Valores propios reales
(Valor.P.V <- dudi.coa(z, scann = FALSE)$eig)
V.aux <- datos$cupo.dissponible
Sel <- S.piPS(n, V.aux)
x.s <-data.frame(x[Sel[,1],])
# Calculo del factor de expancion
F.exp <- 1/Sel[,2]
# Matriz Z ponderada
z.s <- acm.disjonctif(x.s)*F.exp
# Interalo de confianza tradicional mediante Anderson
V.pro <- dudi.coa(z.s, scann = FALSE)$eig[1:10]
ICnorm2 <- data.frame(LI1=V.pro/(1+sqrt(2/n)*
qnorm((1-conf)/2,lower.tail=F)),
LS1=V.pro/(1-sqrt(2/n)*qnorm((1-conf)/2,lower.tail=F)))
TCnorm<-TSC(ICnorm2,Valor.P.V)
# Valores propios estimados mediante Bootstrap
rb<-boot(data=z.s, statistic=hat.V.p, R=1000)
# Intervalo de confianza Bootstrap
MRB<-rb$t0
SD<-sqrt(diag(var(rb$t,na.rm = T)))
ICBoot <- data.frame(LI1=MRB-SD*qnorm((1-conf)/2,lower.tail=F),
LS1=MRB/(1-SD*qnorm((1-conf)/2,lower.tail=F)))
# Tasa de cobertura
TCBoot<-TSC(ICBoot,Valor.P.V)
#### Intervalo de confianza Bootstrap
ICJacn <- jackknife(z.s,conf)
TCJacn<-TSC(ICJacn[,-3],Valor.P.V)
## Salida
c(Anderson=c(Valor.P = V.pro,LI = ICnorm2[,1],
  LS = ICnorm2[,2],TC = TCnorm),
  Boot =   c(Valor.P = MRB,LI   = ICBoot[,1],
  LS  = ICBoot[,2], TC = TCBoot),
  Jacn =   c(Valor.P = ICJacn[,3],LI   = ICJacn[,1],
  LS  = ICJacn[,2], TC = TCJacn))
}
-------------------------------------------------------------
#Diseño piPS - MAS
sim.pips.MAS<- function(datos,n,conf=0.95){
N<-nrow(datos)
x<-datos[,-c(2,6)]
# obtencion de la matrix z
z <- acm.disjonctif(x)
# Valores propios reales
(Valor.P.V <- dudi.coa(z, scann = FALSE)$eig)
# Variable indica de las UPM
```

```
UPM <- datos$edad
# La muestra de observaciones a tomar es de n = 98,
# para lo cual se selecionaran 2 de los 4 grupos de
#edad que estan la poblacion
nI = 2
# en cada uno de ellos se hara una afijacion de neyman
#para distribuir el total de
# individuos a encuestar.
# Nuemro de UPM
UPM.xk.aux<-as.vector(tapply(datos$cupo.dissponible,UPM,mean))
NI <- length(unique(UPM))
# selecion de UPM
Sel.UPM <- S.piPS(n = nI,x = UPM.xk.aux)
PikI <- Sel.UPM[,2]
Sel.UPM <- Sel.UPM[,1]
# Filtro de UPM selecionadas
G.UPM <-as.numeric(UPM[UPM %in% Sel.UPM])
x.UPM<-subset(x,UPM %in% Sel.UPM)
# asignacion de la variable auxiliar para realizar la
#afijacion de Neyman
V.aux <- subset(datos,select =cupo.dissponible, UPM %in% Sel.UPM)
# estimacion de la varianza por UPM
Si <-do.call("c",as.list(by(V.aux,G.UPM,function(X)sd(X[,1]))))
# Numero de individuo por UPM seleccionada
NII <- as.vector(table(G.UPM))
# Numero de individuos a seleccionar por UPM, por
#afijacion de Neyman
nII <- round(n*NII*Si/sum(NII*Si),0)
# Probabilidades de inclusion de primer orden para la UPM
piiI <- rep(PikI,nII)
# Seleccion en las USM
Sel.USM <- S.STSI(G.UPM,Nh=NII,nh = nII)
# Factor de expancion
piiII <- rep(nII/NII,nII)
F.exp <- 1/(piiI*piiII)
x.s <-data.frame(x.UPM[Sel.USM[,1],])
# Matriz Z ponderada
z.s <- acm.disjonctif(x.s)*F.exp
# Interalo de confianza tradicional mediante Anderson
V.pro <- dudi.coa(z.s, scann = FALSE)$eig[1:10]
ICnorm2 <- data.frame(LI1=V.pro/(1+sqrt(2/n)*
qnorm((1-conf)/2,lower.tail=F)),
LS1=V.pro/(1-sqrt(2/n)*qnorm((1-conf)/2,lower.tail=F)))
TCnorm<-TSC(ICnorm2,Valor.P.V)
# Valores propios estimados mediante Bootstrap
rb<-boot(data=z.s, statistic=hat.V.p, R=1000)
```

```
# Intervalo de confianza Bootstrap
MRB<-rb$t0
SD<-sqrt(diag(var(rb$t,na.rm = T)))
ICBoot <- data.frame(LI1=MRB-SD*qnorm((1-conf)/2,lower.tail=F),
LS1=MRB/(1-SD*qnorm((1-conf)/2,lower.tail=F)))
# Tasa de cobertura
TCBoot<-TSC(ICBoot,Valor.P.V)

#### Intervalo de confianza Bootstrap
ICJacn <- jackknife(z.s,conf)
TCJacn<-TSC(ICJacn[,-3],Valor.P.V)
## Salida
c(Anderson=c(Valor.P = V.pro,LI = ICnorm2[,1],
  LS = ICnorm2[,2],TC = TCnorm),
  Boot =   c(Valor.P = MRB,LI   = ICBoot[,1],
  LS  = ICBoot[,2], TC = TCBoot),
  Jacn =   c(Valor.P = ICJacn[,3],LI   = ICJacn[,1],
  LS  = ICJacn[,2], TC = TCJacn))
}

-----------------------------------------------------------------
#SIMULACION
# Fijar semilla
set.seed(10)
# Lectura y adecuacion de la base
datos<-read.delim("basesin.csv",header=T,sep=",")
Nom.Col <- paste(rep(c("Anderson","Boot","Janc"),c(4,4,4)),
c("Valor.P","Li","Ls","Tc"),sep="_")
Nom.fila <- paste("Valor.p",1:10,sep="_")
# Simulacion para un MAS con n = 25
MAS.98<-replicate(2000,Sim.MAS(datos,25,0.95))
round(matrix(rowMeans(MAS.98,na.rm = T),nrow = 10,
dimnames =list(Nom.fila,Nom.Col)),3)
# Simulacion para un pips con n = 25
pips.98<-replicate(2000,Sim.pips(datos,25,0.95))
round(matrix(rowMeans(pips.98,na.rm = T),nrow = 10,
dimnames =list(Nom.fila,Nom.Col)),3)
# Simulacion para un pips MAS con n = 25
pips.MAS.188<-replicate(2000,sim.pips.MAS(datos,25,0.95))
round(matrix(rowMeans(biMAS.188,na.rm = T),nrow = 10,
dimnames =list(Nom.fila,Nom.Col)),3)
```