DISSERTATIO

Estadística

COLEGIO DE GRADUADOS EN CIENCIAS ECONÓMICAS DE ROSARIO
CONSEJO PROFESIONAL DE CIENCIAS ECONÓMICAS DE LA PROVINCIA DE SANTA FE CÁMARA II
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

TRABAJOS FINALES

RECENSIÓN DE TESINAS Y PRÁCTICAS PROFESIONALES DE LA CARRERA LICENCIATURA EN ESTADÍSTICA

CONSEJO PROFESIONAL DE CIENCIAS ECONOMICAS DE LA PROVINCIA DE SANTA FE
CAMARA II

UNR Universidad Nacional de Rosario

Colegio de Graduados en Ciencias Económicas de Rosario

Año III 2019, No 3, **ISSN 2591-5215**

ÍNDICE

CONFORMACIONES

Esta revista se pone a disposición de los profesionales matriculados al Consejo Profesional de Ciencias Económicas de la Provincia de Santa Fe Cámara II (CPCE), asociados del Colegio de Graduados en Ciencias Económicas de Rosario (CGCE), estudiantes y docentes de la Facultad de Ciencias Económicas y Estadística (FCEyE) de la Universidad Nacional de Rosario (UNR) y otras Instituciones vinculadas al quehacer profesional y académico.

El contenido de los trabajos finales no necesariamente refleja la opinión de los Comités responsables de esta publicación digital.

Las Instituciones no son responsables por el contenido de las informaciones y opiniones que viertan en esta revista quienes son identificados como autores de dichos trabajos finales, en todos los casos deberán ser cotejadas por los Profesionales y/o las fuentes.

DISEÑO: AXYOMA GLOBAL
Dir. Nahuel Gasparutti
Lic. DG. Agustín Roig Persig

Italia 1352, Planta Alta
S2000DFB, Rosario, Argentina
+54 (341) 5899241 / +54 (341) 2295698

ACOMPAÑANDO A LOS FUTUROS PROFESIONALES

Convencidos que la capacitación permanente es un camino indispensable para el ejercicio de la profesión es que revalidamos a través de este proyecto el trabajo conjunto entre el Consejo Profesional en Ciencias Económicas de la Provincia de Santa Fe Cámara II, el Colegio de Graduados en Ciencias Económicas de Rosario y la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario.

Iniciado en 2015 con la primera edición de Dissertatio Economía, en 2017 se incorporó Estadística y en 2018 Administración, logrando la participación de las tres escuelas de las licenciaturas que integran la facultad, y conformando una más de las tantas actividades que se realizan mancomunadamente en pos de agregar valor a las potencialidades de los futuros y recién graduados.

El desarrollo de estas revistas digitales, que compendian una selección de tesinas de grado y trabajos finales de carácter profesional, tiene además como objetivo propiciarles a los futuros graduados un marco de contención en sus carreras así como en el desarrollo profesional; incentivando la investigación, fortaleciendo el progreso intelectual, adaptado al cambio y a la innovación a través de la transformación digital, permitiéndoles transitar el camino que iniciaron al elegir esta profesión como proyecto de vida.

ESTRATEGIAS DE ALEATORIZACIÓN PARA LA EVALUACIÓN DE HIPÓTESIS MULTIVARIADAS COMPARATIVAS

LIC. ANA INÉS ALEMANNO

Directora: **DRA. MARTA QUAGLINO**
Codirectora: **LIC. PAULA MACAT**

El test de aleatorización MANOVA para un factor de clasificación, se utiliza para probar hipótesis multivariadas de igualdad entre poblaciones, con una alternativa no paramétrica más flexible que la clásica en cuanto a los supuestos requeridos. La estadística de prueba compara las distancias que existen entre observaciones de un mismo grupo, con las distancias entre observaciones de distintos grupos. La decisión se basa en la distribución de permutación de la estadística de prueba obtenida a partir de los propios datos de la muestra. En el presente trabajo, se presenta una breve síntesis de este test de aleatorización y se expone, a modo de ejemplo, una de las aplicaciones presentadas en la tesina, que corresponde a un problema arqueológico donde interesa evaluar si distintos sitios geográficos seleccionados, son similares en cuanto a mediciones de minerales en el suelo, que pueden brindar información sobre la presencia de diferentes culturas en el pasado.

INTRODUCCIÓN

En el campo de la investigación aplicada, frecuentemente el interés se centra en la comparación de efectos entre grupos, ya sea que la conformación de los mismos se dé en forma natural, como en un estudio observacional, o esté forzada o implícita por la variación de algún factor causal producida durante el desarrollo de un experimento. Tales efectos pueden ser expresados en términos de centros de gravedad, de variabilidades, o bien, de la forma misma de las distribuciones.

En el caso particular de hipótesis multivariadas de comparación entre vectores de promedios, la técnica clásicamente utilizada es el Análisis de la Variancia Multivariado (MANOVA por las siglas en inglés *Multivariate Analysisis of Variance),* el cual requiere para su correcta aplicación, de la verificación de supuestos distribucionales, así como de condiciones sobre las variabilidades y asociaciones entre las variables, representadas por las matrices de variancias y covariancias de cada grupo o población.

Las metodologías no paramétricas se presentan como alternativas flexibles para responder a la necesidad de realizar una comparación estadística de las poblaciones, cuando los datos limitan la aplicación de los tests estadísticos clásicos. Los tests de aleatorización constituyen una clase particular de tests no paramétricos, que permiten probar las hipótesis de interés sin realizar supuestos acerca de la distribución de las variables analizadas. Estos tests se denominan "autodocimantes", porque en lugar de adoptar un modelo distribucional, generan la distribución de las estadísticas de prueba a partir de la permutación de los datos de las muestras que representan la información empírica de las poblaciones. Otra ventaja de estas alternativas, es que es posible aplicarlas cuando el conjunto de datos tiene más variables que observaciones, situación en la cual la estimación del conjunto total de parámetros de la distribución multivariada, se vería comprometida (Anderson 2001).

Las hipótesis que plantean los tests de aleatorización en general consideran si es o no posible admitir que la "ubicación" de los individuos muestrales en cada grupo es al azar (Manly 2006). Si hay razones para considerar que no están distribuidos al azar, se rechaza la igualdad de los grupos.

La decisión respecto de la veracidad de la hipótesis nula se toma a partir de la comparación de la probabilidad asociada a dicha hipótesis (*p-value*), con el nivel de significación que asigna el investigador antes de comenzar el estudio (α). El *p-value* se estima como la proporción de valores de la estadística del test en la distribución de permutación, que resultan iguales o más extremos que el valor de la estadística que se obtiene en la muestra de datos originales.

Dado que estos tests se basan en la permutación de los datos, es posible, a partir de ellos, analizar conjuntos de datos pequeños. Además, contrariamente a las metodologías conocidas de tests estadísticos clásicos, los tests de aleatorización presentan la ventaja de no tener como requisito trabajar a partir de una muestra aleatoria de la población en estudio (Edgington y Onghena 2007). Incluso, en muchos casos, la misma muestra o conjunto de datos resultan ser la población de interés del investigador.

Una propiedad importante de estas estrategias inferenciales, es que los resultados que se obtienen a partir de los métodos paramétricos y de los tests de aleatorización, resultan similares cuando los supuestos requeridos por los tests paramétricos se verifican en el conjunto de datos (Manly 2006).

En general, para definir la estadística del test de aleatorización se comparan las distancias que existen entre las observaciones que pertenecen al mismo grupo con las de observaciones que pertenecen a grupos distintos. Las medidas de distancia no son únicas, y las propiedades matemáticas de las diferentes alternativas, pueden diferir. En este sentido es especialmente importante una propiedad llamada "métrica" que afecta a la valoración de las variabilidades intra grupo y entre grupos, conceptos que intervienen en la definición de las estadísticas del test, al igual que en el caso paramétrico clásico. De aquí que la elección de esta medida de distancia en la definición de los tests sea un paso muy importante en la aplicación del procedimiento.

Estos métodos autodocimantes, son costosos en cuanto a cálculos. Sin embargo, los avances computacionales producidos durante los últimos años, han permitido agilizar los procesos, facilitando su automatización. Actualmente, los tests de aleatorización, y en particular el denominado test permutacional MANOVA, han sido programados en diferentes entornos y están disponibles en programas de computación de libre acceso como R.

El objetivo propuesto en la tesina fue profundizar en el estudio del test no paramétrico MANOVA basado en permutaciones, revisando la alternativa clásica y sus supuestos, y presentar problemas de investigación, en los que la aplicación de estrategias de aleatorización permitiera inferir sobre la hipótesis comparativa de interés, aun cuando no fuera posible utilizar un test convencional.

ANÁLISIS DE LA VARIANCIA MULTIVARIADO (MANOVA) PARA UN CRITERIO DE CLASIFICACIÓN

El Análisis de la Variancia Multivariado se utiliza para probar hipótesis multivariadas de igualdad de promedios de grupos de observaciones definidos por uno o más factores de clasificación. Cada observación reúne las respuestas de p variables de interés, Para la aplicación del test se realiza la partición de la variabilidad total de los datos, diferenciando la variabilidad que existe entre los grupos de aquella que existe dentro de los grupos. La aplicación de esta técnica tiene como objetivo probar si variables categóricas independientes que definen factores de clasificación, influyen en el comportamiento promedio de variables continuas dependientes (Sharma 1996; Johnson y Wichern 2007).

Esta técnica realiza la prueba de hipótesis mediante la aplicación de un único test multivariado, en forma conjunta, sin recurrir a la prueba de las hipótesis individuales para cada una de las variables, lo que permite el control del error de tipo I. Además, a partir del test multivariado, la decisión respecto de la hipótesis nula planteada, tiene en cuenta la correlación que existe entre las variables (Manly y Navarro Alberto 2016).

El método MANOVA tiene una posible interpretación geométrica. Esta tiene base en el concepto de la similitud entre grupos de observaciones a partir de la distancia entre sus promedios, es decir, a menor distancia entre los promedios de los grupos, mayor similitud entre los mismos, y viceversa. El vector de promedios de las p variables a través de las observaciones, dentro de cada grupo, constituye la medida de localización central del grupo, denominado centroide (Anderson 2001). Los centroides son puntos que se representan en el espacio p-dimensional y la medida que se utiliza, preferentemente, para calcular la distancia entre ellos es la de Mahalanobis, ya que tiene en cuenta la correlación entre las variables dependientes o respuestas. A partir del cálculo de la distancia entre los centroides correspondientes a cada uno de los k grupos que identifica el factor de clasificación o variable independiente, puede probarse la existencia de diferencias entre los grupos, estableciendo como hipótesis nula que la distancia entre los promedios es nula y equivale a la igualdad de los promedios de los k grupos o a la no existencia de efecto del factor de clasificación.

Los supuestos que deben satisfacer los datos para que se verifiquen las propiedades distribucionales de la estadística del test MANOVA son:

- Distribución normal multivariada de los errores;
- Matriz de variancias y covariancias común para todos los grupos;
- Independencia de los errores.

Si alguno de los supuestos no se cumple, esto puede afectar tanto a la probabilidad de error de tipo I, como a la potencia del test.

Entre los criterios más utilizados para definir un test con buenas propiedades, se encuentran, el criterio de razón de verosimilitudes de Wilks, y el criterio de Roy o de la mayor raíz característica, que se basa en el criterio de Unión-Intersección.

TEST DE ALEATORIZACIÓN MANOVA PARA UN CRITERIO DE CLASIFICACIÓN

Los tests de aleatorización pueden aplicarse en diferentes contextos, inclusive para la comparación de promedios multivariados. A diferencia del MANOVA tradicional, en el cual se calcula la distancia de Mahalanobis entre los promedios de los distintos grupos, los tests de aleatorización utilizan, para la

construcción de las estadísticas, la distancia entre las observaciones. En estos casos, no necesariamente se debe considerar la medida de distancia de Mahalanobis, sino que se puede seleccionar aquella que resulte más adecuada según el tipo de variables bajo estudio.

La hipótesis nula que se plantea en el test de aleatorización MANOVA es similar a la descripta para el diseño MANOVA tradicional, pero sin utilizar vectores de promedios poblacionales en su planteamiento. Es decir, se pretende verificar si los grupos no difieren, a través de la comprobación que la clasificación de las observaciones es por azar.

Los tests de aleatorización MANOVA no realizan supuestos acerca de la distribución de las variables, ni sobre la naturaleza de las mismas, y permiten realizar el análisis, cualquiera sea el número y tipo de variables de interés definidas por el investigador (Anderson 2001). La única consideración que se debe hacer, es que las observaciones sean intercambiables bajo una hipótesis nula cierta. Esto equivale al supuesto que se realiza en el MANOVA tradicional, de independencia entre los errores e igual distribución (errores *üid*). Sin embargo, a diferencia del test paramétrico, la distribución de los errores no necesariamente se debe ajustar a la distribución multinormal.

Si se considera el enfoque geométrico del diseño MANOVA tradicional, el cálculo de las sumas de cuadrados dentro de los grupos representa la suma de las distancias Euclídeas entre cada observación y su centroide grupal. Sin embargo, el cálculo del centroide grupal puede resultar dificultoso en los análisis en los que no se trabaja con distancia Euclídea. Si se considera la equivalencia que existe entre la suma de los cuadrados de las distancias entre las observaciones respecto del centroide del grupo, y la suma de los cuadrados de las distancias entre los puntos, es posible conocer las sumas de cuadrados dentro de cada grupo y la suma de cuadrados total, sin necesidad de realizar el cálculo del centroide grupal (Anderson 2001). Esto permite obtener la partición de la variabilidad directamente a partir de una matriz simétrica de distancias o disimilitudes entre observaciones ($\mathbf{D}$), utilizando la medida de distancia que resulte más adecuada para el conjunto de datos. Como medidas de distancia se utilizan diferentes definiciones alternativas, de acuerdo al tipo de variables que describan a las poblaciones en estudio (cuantitativas, nominales con más de dos categorías, dicotómicas, o mixtas). En particular, en el ejemplo seleccionado para presentar en este trabajo, se utiliza el coeficiente de similitud de Gower que permite trabajar con variables de diferente naturaleza y, además, es posible aplicarla cuando en el conjunto de datos existen valores faltantes (Legendre y Legendre 1998).

La estadística de prueba del test se denomina $pseudo-F$, ya que su construcción se basa en el mismo criterio que se utiliza para la construcción de la estadística F del test paramétrico ANOVA, el cual consiste en la comparación de la variabilidad existente entre los grupos respecto de la variabilidad existente dentro de los grupos, mediante un cociente entre las mismas. La estadística compara las distancias de los puntos entre los grupos con las distancias de los puntos dentro de los grupos. Si los diferentes grupos tienen una localización central distinta, entonces la distancia entre las observaciones de diferentes grupos va a ser mayor que la distancia entre aquellas que pertenecen al mismo grupo, derivando en un valor grande de la estadística.

La distribución de la estadística $pseudo-F$ bajo la hipótesis nula, se construye intercambiando aleatoriamente los niveles del factor que identifican a las observaciones. Cada permutación asigna un nuevo valor de la F (F^{π}). Este procedimiento se repite para todos los posibles re-ordenamientos de las filas en relación con los niveles, generando la distribución completa de la $pseudo-F$ bajo la hipótesis nula para el conjunto de datos (Anderson 2001).

Para tomar una decisión en cuanto a la hipótesis de interés, es necesario conocer el *p-value* asociado a la hipótesis nula. Para ello se calcula, permutando las observaciones, la proporción de valores de F^{π} que son mayores o iguales al valor de la $pseudo-F$ obtenida con los datos originales. El valor de la estadística que corresponde a los datos originales es considerado como parte de la distribución para el cálculo del *p-value*, ya que, dentro de todos los posibles ordenamientos, se encuentra el obtenido originalmente (Anderson 2001). Si esta probabilidad asociada (P) resulta menor o igual que el nivel de significación prefijado, se rechaza la hipótesis nula de igualdad de los grupos.

El cálculo de la estadística F^{π} para todas las posibles permutaciones se puede realizar únicamente cuando la cantidad de observaciones es pequeña, ya que el número de permutaciones a realizar depende del tamaño de la muestra y, a medida que esta aumenta, se incrementan los costos y los tiempos de trabajo. Sin embargo, es posible calcular la probabilidad asociada a la hipótesis nula a partir de una importante muestra de la totalidad de las permutaciones posibles, esperando obtener iguales resultados, ya que, la ganancia que proporciona el trabajo con todas las posibles permutaciones respecto del trabajo con un subconjunto de las mismas, es muy pequeña. Si se trabaja con un nivel de significación del 5% son necesarias, como mínimo, 1000 permutaciones de las observaciones y, por lo menos, 5000 permutaciones para un nivel de significación del 1% (Manly 2006).

Cuando se lleva a cabo un test de aleatorización MANOVA con un criterio de clasificación, hay un acuerdo general en cuanto al procedimiento permutacional que resulta adecuado para lograr un test exacto. Este consiste en la permutación no restringida de las observaciones, lo que significa que, bajo una hipótesis nula cierta, los grupos no difieren entre sí y, por lo tanto, las observaciones son independientes de los grupos a los que pertenecen (Torres *et al* 2009).

ANÁLISIS DE UN CASO REAL PROVENIENTE DE DATOS ARQUEOLÓGICOS

La arqueología es la ciencia que estudia los cambios que se producen desde las sociedades antiguas hasta las actuales, a partir de restos materiales conservados a través del tiempo, que se encuentran dispersos en la geografía. En particular, el suelo es una importante fuente de información para el conocimiento de las sociedades humanas prehistóricas, ya que la interacción entre los humanos y el ambiente, tanto en el presente como en el pasado, tiene impacto en la composición del mismo.

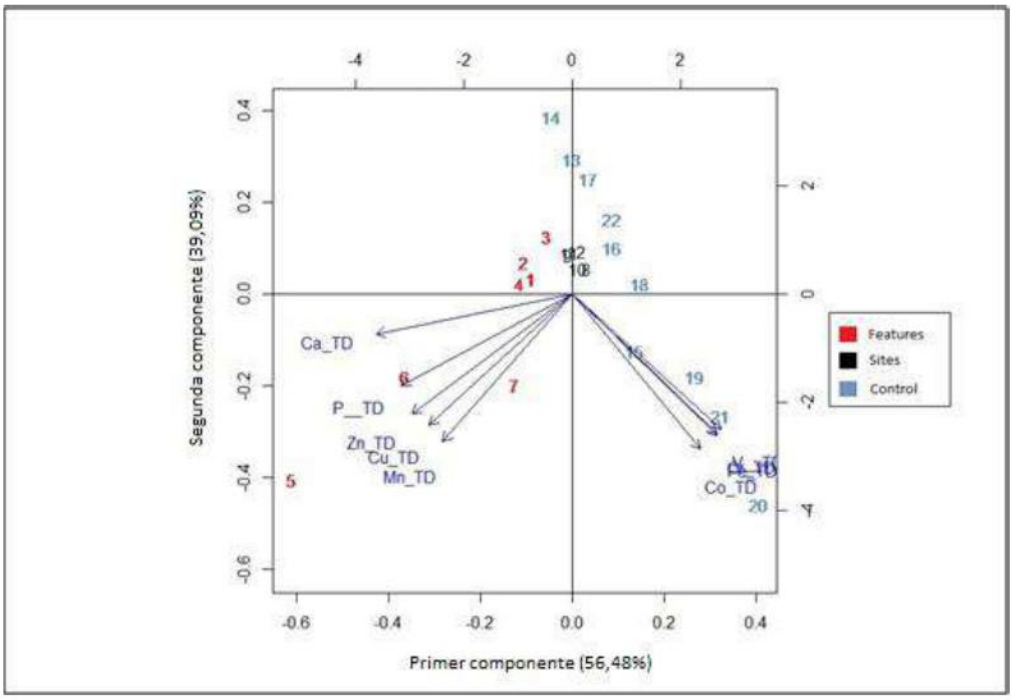
Según Linderholm (2010), existe una relación muy fuerte entre el fosfato del suelo y la vida humana. Sin embargo, el suelo se compone de varios elementos que también sirven para el estudio de la existencia de vida humana durante la prehistoria. Esto conlleva la necesidad de un análisis de forma conjunta de los principales componentes del suelo, es decir, la aplicación de un análisis multivariado de los elementos que lo componen. Este autor presenta un caso real de estudio en el que registró el nivel de nueve elementos en veintidós muestras de suelo, recogidas de una excavación arqueológica de un sitio fuera de Mjolby, Suecia. Los elementos considerados son: Fe (Hierro), Cu (Cobre), P (Fósforo), Mn (Manganeso), V (Vanadio), Co (Cobalto), Zn (Zinc), Cr (Cromo) y Ca (Calcio). Los niveles de los elementos en las muestras de suelo son variables del tipo cuantitativa continua, que no verifican el supuesto de normalidad univariada, por lo tanto, la distribución conjunta no se ajusta a la normal multivariada. Como característica adicional, que dificulta un análisis clásico, se presentan datos faltantes para las mediciones de Calcio, en seis muestras de suelo.

Las muestras recolectadas se clasifican en tres categorías, según el grado y la intensidad de influencia cultural humana. Siete muestras corresponden a la categoría "*Features*", que incluye aquellas con características reconocibles, cinco corresponden a la categoría "*Sites*", que son muestras de la capa de ocupación del sitio y 10 son muestras "*Control*", es decir, muestras recolectadas fuera del sitio.

El objetivo de la investigación es verificar si existe una diferencia significativa en la composición de la tierra, en los tres tipos de muestras recolectadas, aun cuando la información muestral es escasa para el contexto multivariado que se presenta, con nueve variables y tres poblaciones.

Como primer análisis exploratorio que permitiera visualizar gráficamente el conjunto total de las observaciones, identificadas según el sitio en el que habían sido recogidas, se aplica la técnica multivariada de Análisis de Componentes Principales (ACP), que permite proyectar sobre nuevos ejes las veintidós muestras del suelo recolectadas, manteniendo lo mejor posible los parecidos y diferencias conjuntas según los nueve elementos medidos. Las dos primeras componentes logran representar el 95,6% de la información original contenida en la variabilidad de las mediciones, lo cual se considera suficiente para estudiar la composición del suelo de los distintos sitios.

Gráfico 1. Proyección de los elementos (siglas) y las muestras de suelo (números), sobre las dos primeras componentes principales



Nota: Fe, Cr y V se superponen en el cuadrante inferior derecho.

En el Gráfico 1, las muestras pertenecientes a la categoría "*Sites*" quedan agrupadas muy cercanas al origen de coordenadas, las correspondientes a la categoría "*Features*" quedan un poco más dispersas sobre el semiplano izquierdo y las del sitio "*Control*" se observan incluso más dispersas y, en su mayoría, sobre el semiplano derecho. Es decir, ellas se muestran repartidas en lugares diferentes del gráfico, lo que induce a pensar que habría diferencia entre las poblaciones que representan, pero esto es una apreciación gráfica, no puede afirmarse que estas diferencias sean estadísticamente significativas. Para explicar cuáles son las características que diferencian a estas poblaciones, deben interpretarse el sentido de los ejes, el horizontal o primer componente principal y el vertical o segunda componente. La primer componente principal, que representa el 56,5% de la variabilidad total, representa muy bien a todas las variables medidas porque las proyecciones de los vectores que representan a cada elemento (variable) sobre dicho eje, son de módulo grande, están lejos del origen de coordenadas (algunas asociadas en forma negativa y otras positiva). Esta componente separa las muestras que tienen valores altos de Calcio, Fósforo, Zinc, Cobre y Manganeso (están hacia el lado negativo del eje), de aquellas que tienen valores altos de Hierro, Cromo, Vanadio y Cobalto (están hacia el lado positivo del eje). La segunda componente principal, que representa el 39,1% de la variabilidad total, representa bien a todos los elementos, excepto el Calcio que queda un poco más cerca del origen de coordenadas. Es decir, esta componente separa las muestras que presentan valores altos de los elementos, con excepción del Calcio, de aquellas que no.

Así, por ejemplo, se podría deducir que las características de la composición del suelo de las muestras clasificadas como grupo *"Features"*, por su posicionamiento en el gráfico, tienen mayor presencia de los componentes Calcio, Fósforo, Zinc, Cobre y Manganeso, mientras que las muestras del grupo *"Site"*, que están por encima del eje horizontal y más en el centro, se caracterizan por tener mediciones intermedias de todos los componentes y las del grupo "*Control*", están mezcladas, no muestran una composición uniforme.

Para verificar si estas diferencias que se observan gráficamente son estadísticamente significativas, es decir, si muestran evidencias de provenir de poblaciones que son diferentes, se aplica el test de aleatorización MANOVA. Se utiliza el coeficiente de similitud de Gower (Legendre y Legendre (1998)

ya que éste puede aplicarse con información faltante como se presenta en este caso, sin descartar las unidades incompletas.

Cuadro 1. Resultados del test de aleatorización MANOVA

Fuente de variabilidad	GL	SC	CM	Pseudo-F	P-asoc (1000 permut.)
Código del sitio	2	0,34329	0,17164	6,6012	0,00199(*)
Error	19	0,49403	0,02600		
Total	21	0,83732			

(*) Con 5000 permutaciones p=0,0002 y con 3000, p=0,00032.

El test muestra (Cuadro 1) que hay diferencias significativas, al menos dos sitios difieren en cuanto a la composición que presenta la tierra. Este resultado confirma que la dispersión de las muestras pertenecientes a distintos sitios a través de los semiplanos, observada en el análisis exploratorio, se debe a la existencia de distintas poblaciones.

COMENTARIOS FINALES

En la tesina se presenta un test alternativo al test clásico MANOVA a un criterio de clasificación, denominado test de aleatorización MANOVA, o MANOVA no paramétrico, que es adecuado en aquellos casos en los que no se verifican los supuestos distribucionales que requiere el test inferencial tradicional. La diferencia radica en que el efecto del factor de clasificación que estudia el test clásico es sobre vectores de promedios, mientras que el test no paramétrico se enfoca en un concepto de distancia entre observaciones multivariadas. De hecho, puede aplicarse aún cuando el cálculo de los “promedios” no sea apropiado. En particular, a partir del test de aleatorización MANOVA, se busca probar si la clasificación de las observaciones multivariadas en diferentes niveles de algún factor es por azar, o bien, si realmente existe un patrón en el conjunto de datos que provoca dicha distinción. El no cumplimiento de los supuestos no es un tema menor al momento de decidir qué análisis realizar, ya que los resultados que se obtienen pueden no ser válidos.

Se muestra un caso de aplicación, elegido entre varios de los presentados en la tesina, que considera una muestra de datos continuos que no se ajustan a la distribución normal multivariada y corresponde a un problema de investigación arqueológico que consiste en la comparación de la composición de la tierra de tres sitios arqueológicos con diferente registro de influencia cultural humana. Un análisis exploratorio multivariado evidenció gráficamente una disposición diferenciada de las muestras de suelo sobre un mapa factorial según el grupo de pertenencia, lo cual hacía suponer que los sitios explorados eran diferentes en cuanto a su composición química y por lo tanto habrían sido habitados por diferentes culturas. Este ordenamiento diferencial de las muestras fue confirmado por el test de aleatorización MANOVA que se realizó con una medida de distancia que admite la existencia de datos faltantes, ya que la base de datos, de pequeño tamaño, tiene algunas observaciones incompletas.

En la tesina se muestran otros ejemplos que buscan plasmar las flexibilidades que presenta el test para su aplicación en cuanto a la diversidad en el tipo de datos (discretos, presencia-ausencia, etc.), variedad de situaciones en las unidades muestrales y amplitud en las posibilidades de aplicación del test en diferentes ramas de la investigación. Los ejemplos desarrollados, también evidencian la riqueza que se obtiene sobre las conclusiones al aplicar conjuntamente los enfoques exploratorio y confirmatorio, ya que el test es capaz de confirmar, o negar, la significatividad de un efecto del factor que identifica las poblaciones pero no es capaz de describir las particularidades de cada grupo si ellos son diferentes. En

cambio, los análisis multivariados exploratorios describen las características de cada grupo, pueden sugerir la existencia de diferencias, pero no pueden confirmar su significatividad.

Para la aplicación del test de aleatorización MANOVA resulta indispensable contar con un software que realice cálculos con rapidez ya que la distribución de la estadística de prueba no es conocida en forma exacta, sino que se construye a partir de su medición en todas las posibles permutaciones de los datos. En la tesina que resume esta presentación, se utilizó la función "*adonis*" de R. La posibilidad de aplicar el test en un software libre y gratuito como R, es una gran ventaja ya que es de fácil acceso y amplio conocimiento. A pesar de que tiene ciertas limitaciones en la elección de medidas de distancia, éstas se pueden sortear fácilmente con mínimo esfuerzo de programación.

BIBLIOGRAFÍA

Anderson, M. J. (2001). "A new method for non-parametric multivariate analysis of variance", Austral Ecology, 26, 32-46.

Edgington, E. S.; Onghena, P. (2007). *Randomization tests*, 4º edición, Taylor & Francis Group, Boca Raton.

Johnson, R. A.; Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6º edición, Pearson education, Nueva Jersey.

Legendre, L.; Legendre, P. (1998). *Numerical Ecology*, 2º edición, Elsevier Science, Québec, Canada.

Linderholm, J. (2010). "The soil as a source material in archaeology: Theoretical considerations and pragmatic applications", Department of historical, philosophical and religious studies Umea University, Umea, Suecia.

Manly, B. F. J. (2006). *Randomization, Bootstrap and Monte Carlo Methods, in Biology*, 3º edición, Chapman & Hall/CRC, Boca Raton.

Manly, B. F. J.; Navarro Alberto, J. A. (2017). *Multivariate statistical methods*, 4º edición, Taylor & Francis Group, Boca Raton.

Sharma, S. (1996). *Applied multivariate techniques*, John Wiley & Sons Incorporated, Nueva York, USA.

Torres, P.; Quaglino, M.; Pillar, V. (2009). "Properties of a randomization test for multifactor comparisons of group", Journal of Statistical Computation and Simulation, 1-20.

ESTIMADORES ROBUSTOS PARA MUESTREO EN POBLACIONES FINITAS

LIC. EUGENIA BORTOLOTTO

Director: **MG. GONZALO MARÍ**

Entre los objetivos de las encuestas por muestreo se encuentra la estimación de parámetros de variables de interés. Una de las soluciones viene dada por los estimadores clásicos que gozan de buenas propiedades distribucionales como por ejemplo el insesgamiento. El problema surge cuando en la encuesta se observan, en algunas variables, valores alejados del común de los datos. Ante esta situación, los estimadores clásicos presentan dificultades que se ven traducidas en un desempeño pobre respecto a medidas relacionadas con la precisión. Se presentan algunos estimadores clásicos y sus versiones robustas para la estimación de totales poblacionales así como el estudio de sus propiedades a partir de simulaciones.

RESUMEN

Entre los objetivos de las encuestas por muestreo se encuentra la estimación de parámetros de variables de interés. Una de las soluciones viene dada por los estimadores clásicos que gozan de buenas propiedades distribucionales como por ejemplo el insesgamiento. El problema surge cuando en la encuesta se observan, en algunas variables, valores alejados del común de los datos. Ante esta situación, los estimadores clásicos presentan dificultades que se ven traducidas en un desempeño pobre respecto a medidas relacionadas con la precisión. Se presentan algunos estimadores clásicos y sus versiones robustas para la estimación de totales poblacionales así como el estudio de sus propiedades a partir de simulaciones.

INTRODUCCIÓN

Cuando se está interesado en conocer características de una determinada población, como totales, medias o proporciones, existen distintas formas de recolectar la información, pudiéndose mencionar entre las más importantes, a las encuestas. Las mismas consideran la recolección de datos a partir del muestreo. Las unidades sobre las cuales se recolectan los datos son un subconjunto de la población. El objetivo es, a partir de esta muestra, poder hacer inferencia a la población. Existen dos tipos de muestras, las probabilísticas y las no probabilísticas. Las primeras son las que aseguran su representatividad y permiten realizar inferencias válidas para la población considerada. Esto se logra asignando, a cada unidad de la población, una probabilidad no nula de ser seleccionada. Estas probabilidades son las que luego se utilizan en la etapa inferencial que debe contemplar el método de selección utilizado y diversas características como la no respuesta. En cambio, en el muestreo no probabilístico se desconoce la probabilidad de selección de las unidades, no se puede evaluar la precisión en términos probabilísticos y no se garantiza la representatividad de las muestras sobre la población.

Entre los estimadores clásicos más utilizados para estimar medias y totales de las variables de interés se pueden mencionar el estimador de Horvitz-Thompson (Horvitz y Thompson, 1952) y el de razón; para el caso de la media, además se puede mencionar el estimador de Hájek (Hájek ,1971). Con respecto a la estimación de la media, el estimador de Horvitz-Thompson (HT) se utiliza cuando el tamaño de la población sobre la cual se va a inferir es conocido, mientras que el estimador de Hájek se emplea generalmente cuando se desconoce esta cantidad.

Una de las dificultades que surgen en los datos recolectados en encuestas para una amplia gama de aplicaciones, es que los mismos contienen frecuentemente una o más observaciones atípicas llamadas *outliers*, que son observaciones que están separadas de la mayoría de los datos. En estos casos los estimadores clásicos de la media y el total, pueden estar muy influenciados por los *outliers* y no arrojar estimaciones precisas.

Una solución al problema planteado proviene de la estadística robusta. La misma contempla un conjunto de técnicas y herramientas que resultan menos sensibles a la aparición de estas observaciones atípicas. Existe una serie de estimadores considerados robustos que son más apropiados que los estimadores clásicos ante la aparición de observaciones extremas. Entre sus características se puede mencionar que los mismos son generalmente no lineales, y precisan de la definición de términos constantes para su aplicación.

Considerando la estimación del total poblacional de una variable, se presentan los estimadores clásicos de Horvitz-Thompson y de Razón, y se consideran alternativas robustas de ambos y un estimador que reemplaza valores extremos por el valor correspondiente a una cota, denominado de aquí en adelante winsorizado. También se considera el estimador de Hájek del total como un caso especial del estimador de razón. Luego, se introduce una versión robusta del estimador de Horvitz-Thompson a través de M-estimadores, los cuales forman una clase de estimadores robustos simples y flexibles. Considerando al estimador HT como un funcional de mínimos cuadrados, el proceso para convertirlo en robusto se realiza en forma análoga a la realizada para los estimadores de mínimos cuadrados en modelos lineales para poblaciones infinitas a través de los M-estimadores (Hampfel et al., 1986). Para este estimador, que se denomina Horvitz-Thompson Robusto (HTR), se consideran dos versiones del mismo, a un paso y M-estimador, según la cantidad de iteraciones que se realizan. Si además se cuenta con información auxiliar se puede considerar el estimador de razón robusta, para el cual también se consideran dos versiones, a un paso y M-estimador. Por otra parte, se presenta el estimador de Clark winsorizado, el cual utiliza un algoritmo para formar una región de detección y así tratar los valores *outliers*, otorgándoles un peso diferente.

ESTIMADORES CLÁSICOS

Se muestran tres estimadores clásicos, que son los más difundidos en la actualidad. Estos estimadores son sensibles ante la aparición de valores atípicos.

Estimador de Horvitz-Thompson

El estimador de Horvitz-Thompson surge al extender el concepto de Hansen y Hurwitz, quienes desarrollaron la teoría de muestreo con probabilidades proporcionales al tamaño con reemplazo, para muestras sin reemplazo de probabilidades desiguales. Horvitz y Thompson hicieron una contribución a la inferencia basada en diseños al formular tres clases de estimadores lineales y luego planteando la posibilidad que el mejor estimador (de mínima variancia) a través de todos los estimadores posibles lineales insesgados puede no existir para una muestra aleatoria simple.

El estimador de Horvitz-Thompson del total poblacional para una variable y es un estimador lineal e insesgado que emplea las probabilidades de inclusión de primer orden (π_k). El mismo es igual a $\hat{t}_{HT} = \sum_s \frac{y_k}{\pi_k}$. Para el caso de la media, este requiere conocer el tamaño de la población (N), y resulta igual a $\hat{t}_{HT}/N$.

Estimador de Hájek

Una alternativa para estimar parámetros poblacionales como la media cuando se desconoce N, es el estimador de Hájek (1971). La fórmula del estimador de Hájek del total es igual a $\hat{t}_H = N\tilde{y}_H = N\frac{\hat{t}_{HT}}{\hat{N}} = N\frac{\sum_s y_k/\pi_k}{\sum_s 1/\pi_k}$, donde $\tilde{y}_H$ es el estimador de Hájek de la media.

Este estimador es frecuentemente mejor que el estimador de Horvitz-Thompson, en el caso que N sea conocido. El estimador de Hájek es no lineal, y por lo tanto aproximadamente insesgado, además se puede obtener una expresión aproximada para la variancia a través de linealización por series de Taylor (Särndal et al., 1992).

Para algunos diseños los estimadores de Horvitz-Thompson y Hájek son idénticos, es decir, producen el mismo valor para todas las muestras como en los diseños simple al azar y muestreo estratificado simple al azar.

Cuando el tamaño de la población es desconocido no es posible utilizar el estimador de Horvitz-Thompson para la media con lo cual se debe optar por el estimador de Hájek. Sin embargo si N es conocido, y los estimadores difieren, se debe elegir uno. Como ya se ha indicado es preferible usar el estimador de Hájek por tres motivos. En primer lugar, por la forma de la variancia de Hájek, éste es preferible cuando los valores de la variable son todos cercanos a la media. En segundo lugar, tiene un mejor rendimiento en casos donde el diseño no es fijo, o sea, el tamaño de muestra es variable. Si el tamaño muestral es mayor que el promedio, la suma del numerador y la suma del denominador tendrán relativamente más términos. Análogamente, si el tamaño de la muestra es pequeño, ambas sumas tendrán pocos términos. La razón conserva de este modo una cierta estabilidad. Por el contrario, el estimador de Horvitz-Thompson tiene denominador fijo y por lo tanto carece de esta propiedad. Y finalmente, en casos donde π_k no está correlacionado o lo está en forma negativa con los valores y_k. Si se supone que la muestra contiene un elemento con un gran valor de y_k pero un pequeño valor de π_k, la suma del numerador será muy grande. Sin embargo, será compensada hasta cierto punto por grandes valores de $1/\pi_k$ en el denominador. En este sentido, el estimador de Hájek es mejor que el de HT, donde el denominador de N permanece fijo.

Estimador de Razón

El estimador de razón de totales poblacionales es uno de los estimadores que utilizan información auxiliar más antiguos dentro del muestreo de poblaciones finitas. Robinson (1987) considera una deducción basada en el diseño condicional de una muestra aleatoria simple cuando se cuenta con una variable auxiliar z positiva, de la cual se conoce el total poblacional t_z. La situación ideal ocurre cuando y_k/z_k es constante para todos los elementos k de la población, lo cual resulta en una mejor estimación.

Se define al estimador de razón del total t_y como $\hat{t}_r = t_z \frac{\hat{t}_{HTy}}{\hat{t}_{HTz}} = t_z \hat{R}$, donde $\hat{R}$ es el estimador de $R = \frac{\sum_U y_k}{\sum_U x_k}$, $\hat{t}_{HTy}$ es el estimador HT de t_y, $\hat{t}_{HTz}$ es el estimador HT de t_z y t_z es el total poblacional de z. En el caso que $z_k = 1$ para todo los elementos k de la población, estamos en presencia del estimador de Hájek.

Este es un estimador no lineal que requiere de información auxiliar para su estimación. Se asume que una medida positiva del tamaño z_k es conocida para los valores de la muestra y tiene correlación positiva con las variables de interés de la encuesta. Se denota con t_z el total poblacional de z_k, el cual debe ser conocido para utilizar el estimador. El mismo es frecuentemente más eficiente que el estimador de Horvitz-Thompson.

Para el estimador de razón, al igual que en el caso del estimador de Hájek, se aplica el método de linealización por series de Taylor para hallar una variancia aproximada. El estimador de razón es frecuentemente más eficiente que el estimador de Horvitz-Thompson para muestreo simple al azar, cuando el coeficiente de correlación entre z e y es mayor o igual a la mitad entre el coeficiente de variación de z sobre el coeficiente de variación de y.

El sesgo del estimador de razón es frecuentemente pequeño. Sin embargo, en muestras con pocas observaciones, el mismo puede ser lo suficientemente importante para ser ignorado, pero en muestras grandes el sesgo es despreciable.

ESTIMADORES ROBUSTOS

En las encuestas por muestreo, es común la aparición de valores que se alejan de la generalidad de los datos, denominados *outliers*. Los mismos pueden ser observaciones que se corresponden con valores observados y que resultan ser válidos. Estos valores afectan los estimadores tradicionales debido a que los mismos son sensibles ante la aparición de uno o más *outliers*. Ya que es frecuentemente difícil detectar *outliers* y decidir si estos son válidos o no, son necesarios los estimadores que se desempeñan bien en términos de sesgo y varianza con independencia de la naturaleza y la detección de posibles valores atípicos. Se presentan algunos de los estimadores robustos más difundidos desarrollados hasta la actualidad.

Estimador Horvitz-Thompson robusto

Es posible obtener una versión robusta del estimador de Horvitz-Thompson. Para ello se requiere conocer una medida positiva x del tamaño para toda la población y que debe tener correlación positiva con las variables de interés de la encuesta.

Al utilizar el estimador HT clásico, el razonamiento dado en la literatura de muestreo, es que tiene un error de muestreo o variancia igual a cero si las probabilidades de inclusión π_k son estrictamente proporcionales a y_k. El estimador tendrá sesgo robusto pero no variancia robusta con respecto a desviaciones de la proporcionalidad entre π_k e y_k (Rao, 1966).

Por lo tanto, si se está interesado en formular al estimador HT como un estimador de variancia, se debe expresar al mismo como un funcional de mínimos cuadrados de un estimador de la función de distribución poblacional, de forma tal que el diseño sea incorporado en el estimador de la función de distribución poblacional mientras que la proporcionalidad entre y y x es tenida en cuenta por el funcional de mínimos cuadrados.

El estimador de HT no depende del modelo de superpoblación. Sin embargo, en ese modelo la pendiente involucrada en el estimador HT es un estimador de mínimos cuadrados ponderado que incorpora la información del diseño a través de la función de probabilidad, así como la información en la variable auxiliar a través de la regresión.

Huber (1973a) extiende los resultados del estimador robusto de parámetros de posición al caso de regresión lineal al proponer un estimador a partir de un método iterativo que considera mínimos cuadrados ponderados. Estas ponderaciones no son fijas sino que dependen del estadístico.

Luego de separar el diseño y la información auxiliar y expresarlo como un funcional de mínimos cuadrados, el proceso para convertir en robusto al estimador HT es análogo al del estimador de mínimos

cuadrados en modelos lineales para una población finita a través del M-estimador (Hulliger, 1995).

El estimador HT robusto es un estimador no paramétrico. El modelo es simplemente utilizado para derivar la expresión del estimador HT como un funcional de mínimos cuadrados. Ni el estimador HT ni el estimador HT robusto necesitan del modelo para ser aplicado, por lo tanto no es necesario comprobar los supuestos del mismo.

El estimador de Horvitz-Thompson robusto, asume que las ponderaciones no poseen valores *outliers* y por lo tanto sólo los residuos deben convertirse en robustos. Este estimador puede ser expresado como uno ponderado, por lo tanto la solución puede ser obtenida con un algoritmo Reponderado Iterativo de Mínimos Cuadrados (IRLS, por sus siglas en inglés).

Una estimación para Horvitz-Thompson robusto viene dada por: $\hat{t}_{HTR} = t_x \beta^{(t+1)}$ en la $(t+1)$-ésima iteración, donde $\beta^{(t+1)} = \frac{\sum_s w_k u_k y_k}{\sum_s w_k u_k x_k}$ es la estimación de la pendiente, u_k son ponderaciones robustas que dependen de $\beta^{(t)}$ y $w_k = \frac{1}{\pi_k}$. El proceso iterativo es repetido hasta cumplirse los criterios de convergencia. Si a este proceso de iteración sólo lo realizamos una vez obtenemos el estimador Horvitz-Thompson robusto a un paso.

Estimador Hájek robusto

El estimador de Hájek robusto de la media a un paso se obtiene a partir del estimador HT robusto, reemplazando el total poblacional por su estimador $\hat{\bar{y}}_{HR} = \frac{\hat{t}_{HTR}}{\hat{N}} = \frac{\sum_s w_k u_k y_k}{\sum_s w_k \sum_s u_k / n}$.

Y a partir del mismo se obtiene el estimador de Hájek robusto un paso del total $\hat{t}_{HR} = N \hat{\bar{y}}_{HR}$. Un razonamiento similar se realiza para obtener el M-estimador de Hájek robusto.

Estimador de razón robusto

Se asume que existe una variable auxiliar $z_k > 0$ para todos los elementos k de la población, correlacionada con las variables de interés y que el total poblacional t_z es conocido.

Se construye un estimador de razón robusto de forma análoga al estimador HT robusto. Se parte de un estimador robusto de la pendiente. Luego, se estima la desviación estándar de los residuos a partir de la mediana de los residuos absolutos estandarizados. Se definen ponderaciones robustas.

Para obtener el estimador de razón robusto del total poblacional se realiza el proceso iterativo donde se obtiene: $\hat{t}_{RS} = t_z \frac{\sum_s w_k u_k y_k}{\sum_s w_k u_k z_k}$, con u_k ponderaciones robustas.

Con el estimador de razón robusto, al igual que pasa con en el estimador HT robusto, si sólo realizamos una iteración obtenemos el estimador de razón robusto un paso.

Total Clark winsorizado

Una técnica que permite detectar y tratar valores atípicos es llamada winzorisación. Este procedimiento reemplaza los valores extremos por otros, menos extremos, moviendo los valores originales hacia el centro de la distribución. El proceso de winsorización puede ser unilateral que trata sólo una cola de la distribución correspondientes a los valores más extremos o puede ser bilateral que trata simultáneamente ambos extremos. Los valores designados como atípicos son modificados reemplazándolos con valores que minimizan el Error Cuadrático Medio (ECM) de la estimación del total.

Se considera el estimador de Clark winzorizado, un método de una cola diseñado por Clark (1995). Este método utiliza un algoritmo para formar una región de detección y así tratar los valores *outliers*.

La efectividad de este método dependerá de la cantidad de *outliers* en la muestra, siendo muy conservador ante la presencia de los mismos, pero si no contiene valores atípicos, dejará de serlo.

Elección de la constante de robustez

El problema cuando se utilizan estimadores robustos univariados para distribuciones asimétricas es que, inevitablemente, los estimadores robustos tienen un sesgo. Afortunadamente, la menor varianza que estos presentan compensa, al menos parcialmente, el sesgo. Por lo tanto la cuestión de elegir la constante de robustez está relacionada a encontrar un equilibrio entre el sesgo y la variancia de los estimadores. Para ello se destacan dos aspectos (Hulliger, 2011b):

1- Lo deseable es estar protegidos contra un número mínimo de valores atípicos. Esto establece un límite superior a la constante de robustez. De hecho, si esta proporción mínima de valores atípicos que se esperan es α entonces existe una constante de robustez que determina una proporción α de las observaciones como valores atípicos. Para estar protegidos contra tal proporción de valores atípicos es que se acepta un cierto sesgo. Este es un precio a pagar por la seguridad de estar protegido al menos contra un pequeño número de valores atípicos.

2- Si existe un mayor número de valores atípicos, se desea estar más protegidos, lo cual implica una pérdida mayor en términos de sesgo. Sin embargo, esto provoca una ligera mejora con respecto a la variancia debido a la reducción de valores extremos.

A medida que se agranda esta constante, el estimador se asemeja a su estimador clásico, mientras que cuanto más chica sea, se consideran más valores como atípicos.

ESTIMADORES DISPONIBLES EN EL PROGRAMA ESTADÍSTICO R

El programa estadístico R constituye uno de los programas más utilizados en la actualidad debido, entre otros aspectos, a que su licencia es libre. Por otro lado, existe un gran desarrollo de códigos con metodologías actuales debido al empleo que tiene el mismo en el ámbito de la investigación. Se presentan a continuación dos paquetes del programa R que permiten estimar en el ámbito del muestreo en poblaciones finitas los estimadores que se presentan en esta revisión.

El paquete *survey* (Lumley, 2010) permite el cálculo de los estimadores de Horvitz-Thompson, de Hájek y de razón junto con sus correspondientes estimadores de variancia, para diversos diseños muestrales.

Con respecto a los estimadores robustos, no existe disponible, al día de la fecha, en la página de descarga *The Comprehensive R Archive Network* (CRAN) del programa R ningún paquete que los considere. El paquete *rhte* (Hulliger, 2011a) es uno de los que incluye los mismos y se obtiene a partir del pedido a sus autores. Permite el cálculo del estimador HT robusto, del estimador de Hájek y del estimador de razón robusto, para las dos versiones analizadas, el M estimador y el estimador a un paso, junto con el cálculo de sus respectivas variancias. Dado que funciona en forma conjunta con el paquete *survey*, comparte con éste la posibilidad de cálculo para los mismos diseños muestrales.

ESTUDIO DE LAS PROPIEDADES DE LOS ESTIMADORES A TRAVÉS DE SIMULACIONES

Para analizar cómo se comportan los distintos estimadores clásicos y robustos se lleva a cabo un estudio por simulación. El objetivo es investigar el desempeño de los estimadores robustos propuestos en términos de características propias de la distribución de los mismos como son la variancia, el sesgo y el ECM, para distintos escenarios. Para ello, se generaron seis poblaciones (Beaumont, Haziza y Ruiz-Gazen, 2013), cada una con una variable de interés y, y una variable auxiliar x, que tiene distribución Gamma con media 50 y variancia 500.

En las primeras tres poblaciones, de tamaños poblacionales 500, 1000 y 5000 respectivamente, se generan los valores de y considerando un modelo de razón $y_{1,k} = 2x_k + 3{,}7{x_k}^{1/2}\varepsilon_k$, donde el término de error ε_k se generó de una distribución normal estándar. Estas poblaciones no contienen ningún valor atípico. En las otras tres poblaciones, también de tamaños 500, 1000 y 5000 respectivamente, los valores de y se generan de acuerdo a un modelo de mezcla $y_{2,k} = \tau_k\left(2x_k + 3{,}7{x_k}^{1/2}\varepsilon_k\right) + (1-\tau_k)z_k$, donde los valores de z son generaciones independientes de una distribución normal con media 1200 y desvío estándar 200 y los valores de τ_k son generados en forma independientes de una distribución Bernoulli con probabilidad $p = 0.98$, o sea que estas poblaciones contienen aproximadamente 2% de valores *outliers*.

Para cada población se realizan simulaciones considerando dos escenarios distintos, incluyendo la variable auxiliar x en el diseño muestral a través de la definición de las probabilidades de inclusión, o incluyéndola en la etapa de estimación. En el primer escenario, se considera un diseño muestral proporcional al tamaño (PPT) mientras que en el segundo se utiliza un muestreo simple al azar. Esto se realiza con el objetivo de analizar cuál es la mejor forma de incluir información auxiliar. Además se consideran dos fracciones de muestreo $f = n/N$, 0,02 y 0,10.

En primer lugar se seleccionan $R = 1000$ muestras de acuerdo a un muestreo Poisson con probabilidad de inclusión π_k, proporcional a x_k , o sea $\pi_k = \tilde{n}x_k / \sum_U x_k$, donde ñ representa la esperanza del tamaño muestral.

La elección del muestreo Poisson en el estudio de simulación se justifica por el hecho de facilitar la coordinación de muestras en encuestas de panel, situación que se presenta en la mayoría de encuestas económicas.

En cada muestra simulada se calculan los estimadores de Horvitz-Thompson, tanto clásicos como robustos, y los estimadores de Hájek, también en sus versiones clásicas y robustas.

Luego se seleccionan las muestras de acuerdo a un muestreo simple al azar, en donde se calculan los estimadores que requieren de una variable auxiliar, como son los estimadores de razón clásicos y robustos con M y con una iteración y el estimador de Clark winzorizado. También se considera el estimador HT clásico para realizar comparaciones.

Para finalizar se comparan los estimadores obtenidos en cada población.

Para el caso de muestreo Poisson, luego de analizar los resultados obtenidos, para las poblaciones de tamaño 500, en los escenarios donde se trabajó con la variable y_1 que no contiene *outliers* la eficiencia relativa es menor en los estimadores clásicos comparado con sus versiones robustas. A su vez el estimador de Hájek arroja mejores resultados que el estimador HT, esto se debe a las propiedades que se vieron anteriormente. Por otro lado, al analizar las poblaciones correspondientes a la variable y_2, los estadísticos robustos presentan eficiencias relativas menores a 1 lo cual implica un menor ECM para las versiones robustas. Para estos casos también el estimador de Hájek resulta mejor que el estimador HT.

A medida que aumenta el tamaño poblacional los estimadores robustos se vuelven más inestables. Cabe destacar que el comportamiento de los estimadores HTR y Hájek robusto son diferentes a medida que el tamaño de la población aumenta. En el primer caso, la ganancia respecto a los estimadores clásicos ante la presencia de *outliers* va disminuyendo, situación que referencia Beaumont et al (2013) en simulaciones con poblaciones similares. Para el estimador de Hájek, la situación es más extrema, llegándose a encontrar un comportamiento ineficiente de los estimadores robustos causado por un aumento del sesgo. Por otro lado, no se observan diferencias significativas entre los estimadores robustos M y un paso en ninguna de las poblaciones, si bien en todas las situaciones el estimador con más de una iteración se comporta levemente mejor que el que considera sólo un paso.

Se muestra a continuación, en los gráficos 1 y 2, cómo varía el ECM de los distintos estimadores analizados en el muestreo Poisson, tanto en las poblaciones que contienen valores atípicos como en las que no. Los gráficos se realizan para la fracción de muestreo f = 0,02 y en los tres tamaños poblacionales considerados.

Gráfico 1: ECM de los estimadores en poblaciones sin *outliers* según tamaño poblacional.

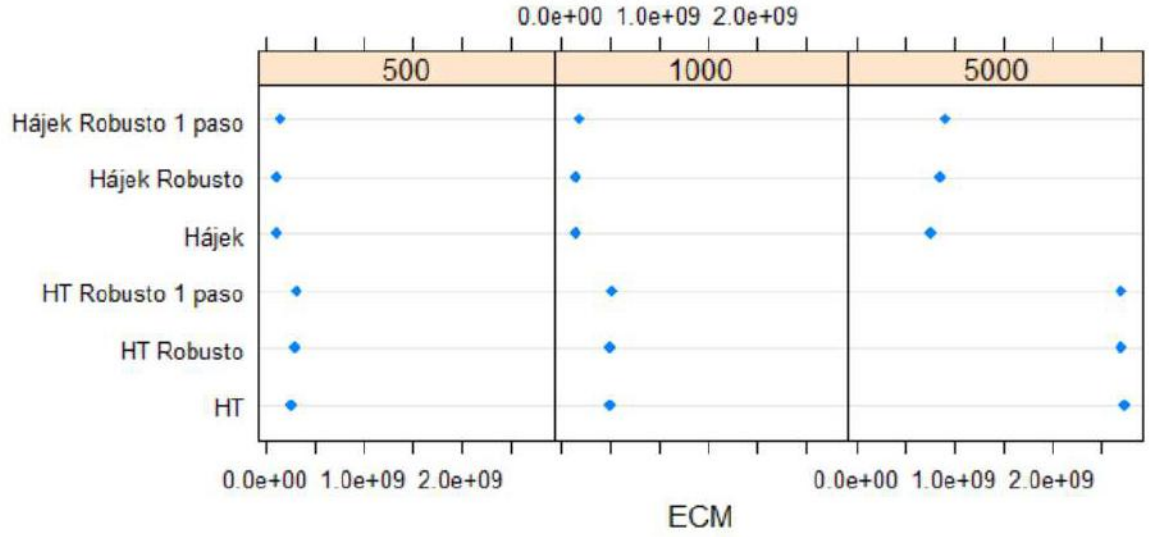
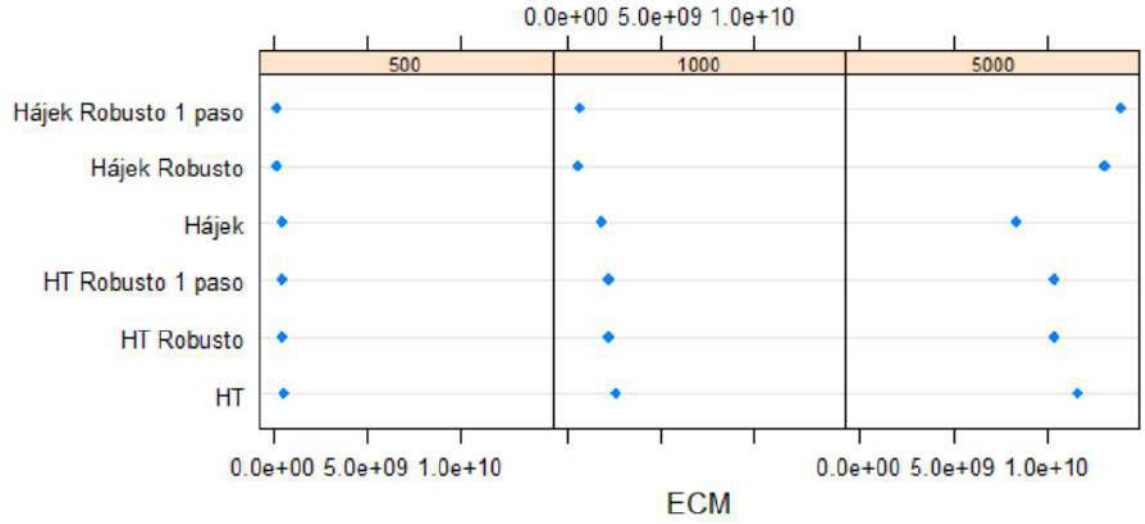


Gráfico 2: ECM de los estimadores en poblaciones con *outliers* según tamaño poblacional.



Se observa, en el gráfico 1, el aumento considerable del ECM para un tamaño poblacional de 5000 en los estimadores de Horvitz-Thompson tanto el clásico como los robustos. Mientras que en el gráfico 2, se muestra cómo incrementa el ECM en todos los estimadores analizados, destacándose los estimadores de Hájek robustos.

Se analizan las simulaciones realizadas utilizando muestreo simple al azar, en el cual se incorpora la información auxiliar a través de los estimadores. En estos casos, para tamaño poblacional N=500, fracción de muestreo 0,02 y cuando no se cuenta con *outliers* en la muestra, los estimadores de razón robustos arrojan una eficiencia similar al estimador de razón clásico. Estos a su vez, presentan una eficiencia relativa considerablemente menor al estimador HT, la cual se debe a una menor variancia en los estimadores de razón. Con el estimador de Clark winsorizado se obtiene una eficiencia relativa aproximadamente igual a uno, es decir que es igual de eficiente que el estimador HT. Al analizar el escenario donde contamos con valores atípicos, los estimadores de razón robustos mejoran considerablemente en cuanto a eficiencia relativa, ya que disminuye la variancia de estos estimadores, mientras que el estimador de razón clásico pasa a ser menos eficiente que en el escenario anterior, aunque lo sigue siendo más que el estimador HT. En cuanto al estimador de Clark también mejora considerablemente, disminuyendo su variancia, aunque no llega a ser tan eficiente como los estimadores de razón robustos.

Esta relación se mantiene al variar la fracción de muestreo a 0,1, observándose sólo un aumento en la eficiencia relativa para los estimadores de razón robustos en el escenario donde se presentan *outliers*.

Al aumentar el tamaño poblacional a $N = 1000$, en los escenarios con fracción de muestreo chica, los resultados obtenidos siguen siendo los mismos que los expuestos en párrafos anteriores, en donde se observa que para poblaciones sin valores atípicos, la eficiencia relativa de los estimadores de razón robustos son semejantes a su versión clásica y estos considerablemente mejores que el estimador HT. Para poblaciones con *outliers*, los estimadores de razón robustos se mantienen con eficiencias relativas bajas, mientras que la eficiencia relativa del estimador de razón clásico se acerca a uno. Y el estimador de Clark disminuye su eficiencia relativa en ambos casos. Sin embargo, al variar la fracción de muestreo, en poblaciones con valores atípicos, los estimadores de razón robustos incrementan su eficiencia relativa, pasando a ser menos eficientes que el estimador HT, esto se debe a un gran incremento del sesgo. En este escenario también el estimador de Clark empeora, aunque sigue presentando una eficiencia relativa mejor que el estimador HT.

Por último, al analizar los muestreos de las poblaciones con tamaño poblacional N=5000, en ambas fracciones de muestreo, se observa que en las muestras que no contienen *outliers* los estimadores de razón robustos, son tan eficientes como el estimador de razón clásico, arrojando eficiencias relativas cercanas a 0,2, mientras que el estimador de Clark arroja eficiencias relativas cercanas a 1. Por otro lado, en los escenarios que sí contienen *outliers*, los sesgos de los estimadores de razón robustos crecen de forma abrupta, arrojando eficiencias relativas mayores que uno. En estos casos el estimador de Clark sigue manteniendo su eficiencia relativa cercana a 1.

Estos resultados que se observan con el estimador de razón para tamaños poblacionales grandes en muestras que contienen *outliers*, son similares a los vistos en el muestreo Poisson para el estimador de Hájek, lo cual resulta lógico, ya que se mencionó que este último es un caso particular del estimador de razón.

Al igual que vimos con el muestreo Poisson, no se observan diferencias significativas entre los estimadores robustos M y su versión a un paso.

De igual manera que para el muestreo Poisson, se realizaron gráficos para ver cómo varía el ECM entre los estimadores, tanto en las poblaciones que contienen valores atípicos como en las que no, para la fracción de muestreo f = 0,02 y en los tres tamaños poblacionales considerados.

En el caso del muestreo simple al azar (gráfico 3), el ECM en poblaciones sin *outliers* crece para los estimadores de Clark winzorizado y el estimador HT. En cambio, en las poblaciones que sí contienen valores atípicos (gráfico 4), crece para los estimadores de razón robustos.

Gráfico 3: ECM de los estimadores en poblaciones sin *outliers* según tamaño poblacional.

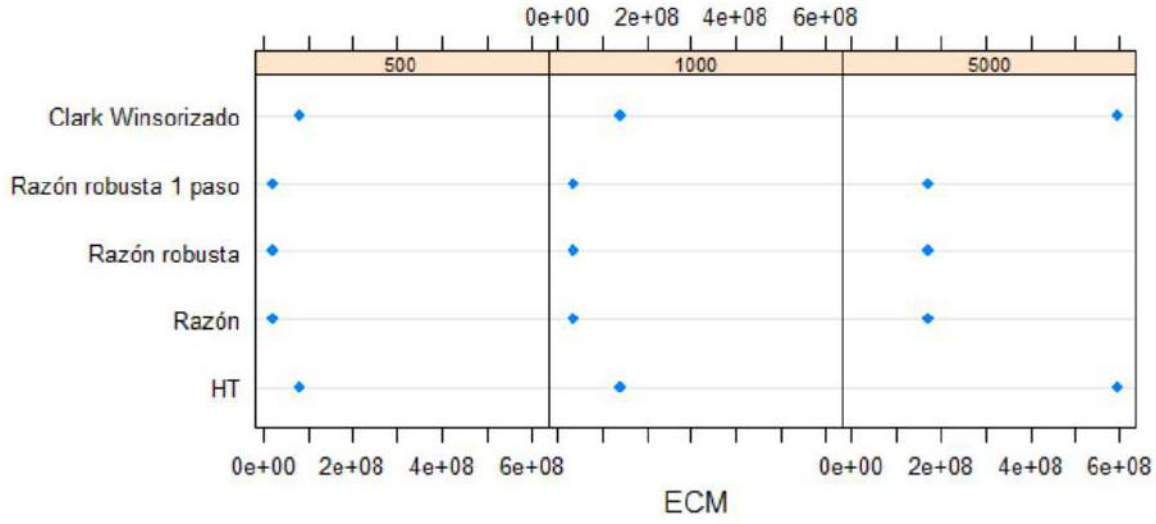
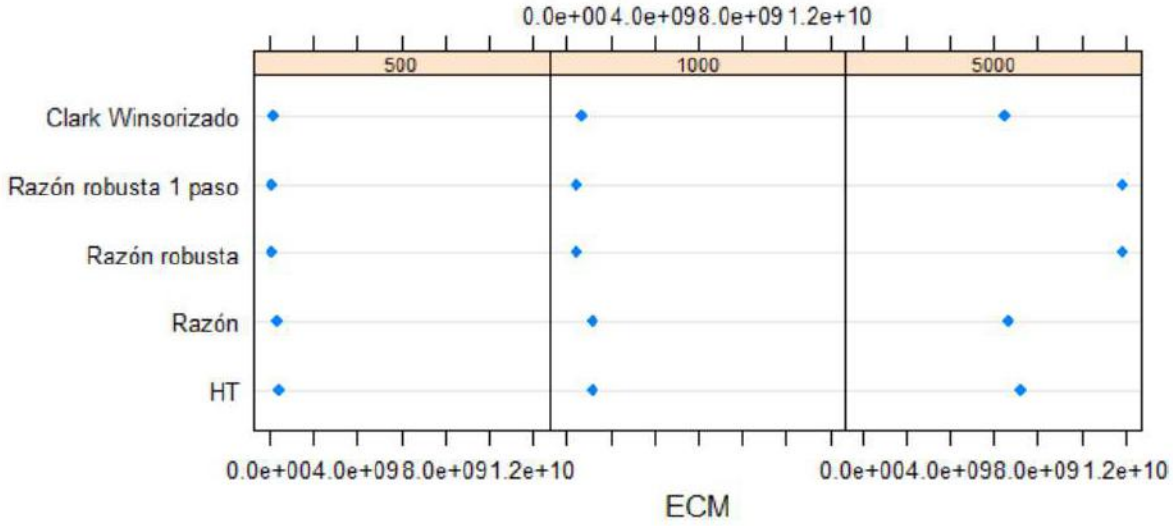


Gráfico 4: ECM de los estimadores en poblaciones con *outliers* según tamaño poblacional.



COMENTARIOS FINALES

Se presentan tres estimadores clásicos del total, el estimador HT, el estimador de Hájek y el estimador de razón y sus versiones robustas a M y un paso, y por último el estimador de Clark winsorizado para la estimación de parámetros de poblaciones finitas a partir de muestras seleccionadas en forma probabilística. Los primeros poseen el inconveniente de ser sensibles ante la aparición de valores atípicos, mientras que los estimadores robustos son sesgados.

Se plantea el uso de estimadores robustos como solución al problema de la presencia de valores atípicos en el muestreo. Para ello se realizan simulaciones en el programa estadístico R considerando diversos escenarios, donde varían los tamaños poblacionales, fracción de muestreo y diseño muestral. Además se simulan dos variables de interés, una que no contiene *oultiers* y la otra que contiene un 2% de valores atípicos.

Sobre las simulaciones realizadas, se calculan los estimadores planteados teóricamente y se comparan de acuerdo a su variancia, sesgo relativo, ECM y eficiencia relativa al estimador HT clásico.

A su vez, se analiza la incorporación de información auxiliar de dos formas distintas. En primer lugar incorporando una variable auxiliar x en el diseño muestral, para ello se optó por un muestreo Poisson, en el cual se calcularon los estimadores de Horvitz-Thompson, de Hájek y sus versiones robustas. Y por otro lado, se extrajeron muestras simple al azar y se incorporó la variable auxiliar en la etapa de estimación; para ello se tuvieron en cuenta los estimadores de razón, tanto clásico como robustos, y el estimador de Clark winsorizado.

Al evaluar los resultados obtenidos, se observa en primer lugar, que los estimadores de razón (y el de Hájek como un caso particular de razón) se vuelven muy inestables al aumentar el tamaño poblacional, ya que el sesgo incrementa considerablemente. Sin embargo, para tamaños poblacionales pequeños estos estimadores presentan buenos resultados, arrojando eficiencias relativas menores a las del estimador HT clásico.

Por otro lado, los estimadores HT robustos y de Clark, se mantienen estables al variar el tamaño poblacional, aunque para muestras chicas, su eficiencia relativa es mayor que en los estimadores de Hájek robustos.

Para todos los estimadores robustos, se consideraron dos versiones de los mismos, el M estimador y el estimador a un paso. En todos los casos, los resultados de estas dos versiones fueron similares, por lo tanto es indistinto el que se elija.

En cuanto a los estimadores clásicos, al analizar las poblaciones que contienen *outliers*, sus variancias se incrementan causando ECM mayores que los estimadores robustos. Además, como se analizó teóricamente, los estimadores de razón presentan eficiencias relativas menores que el estimador HT.

Se concluye que utilizar estimadores clásicos en muestras con *outliers* no resulta conveniente. Además, si se cuenta con tamaños poblacionales grandes (N = 5000), tampoco es apropiado utilizar estimadores de razón robusto o de Hájek robusto. Se recomienda utilizar los estimadores de HT robusto en caso de contar con muestras en donde el diseño muestral incorpore información auxiliar, o el estimador de Clark en muestras que no se incorpore información auxiliar en el diseño.

Este trabajo da lugar a posibles investigaciones futuras, ya sea investigar el comportamiento de algún otro estimador robusto o bien plantear alguna otra posible solución al problema de los *outliers*. Por otra parte, se deberá seguir estudiando el comportamiento de los estimadores de razón en poblaciones grandes con *outliers*.

BIBLIOGRAFÍA

Beaumont, J.-F., Haziza, D., Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100 (3), 555-569.

Chambers, R., Kokic, P., Smith, P. y Cruddas, M. (2000). Winsorization for identifying and treating outliers in economic surveys. *ICES II, The Second International Conference on Establishment Surveys, Survey Methods for Businesses, Farm, and Intitutions*, American Statistical Association, 717-726.

Clark, R. (1995). *Winsorization Methods in Sample Surveys*. Master's Thesis. Department of Statistics. Australia National University.

Hajek, J. (1971) Discussion of *An essay on the logical foundations of survey sampling* by Basu, D. in Foundations of Statistical Inference (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., y Stahel, W.A. (1986). *Robust Statistics*. Ney York: Wiley.

Horvitz-D.G., y Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35, 73-101.

Huber, P. J. (1973a). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Statistics*, 1, 799-821.

Huber, P. J. (1973b). The use of Choquet capacities in statistics. *Proceedings of the 39th Session of the ISI*, Vol. 45, pp. 181-188.

Hulliger, B. (1995). Outlier Robust Horvitz-Thompson Estimators. *Survey Methodology*, 21, 79-87.

Hulliger, B. (1999). *Simple and Robust Estimators for Sampling. Proceedings of the Survey Research Methods Section, American Statistical Association*, 1999, 54-63

Hulliger, B. Alfons, A., Filzmoser, P., Meraner, A., Schoch, T., Templ, M. (2011a) *R Programmes for Robust Procedures Including Manual.* AMELI Deliverable D4.1. AMELI Project.

Hulliger, B. Alfons, A., Filzmoser, P., Meraner, A., Schoch, T., Templ, M. (2011b) *Robust Methodology for Laeken Indicators.* AMELI Deliverable D4.2. AMELI Project.

Lohr, S. (1999). *Sampling: Design and Analysis, 2nd Edition*. Cengage Lerning.

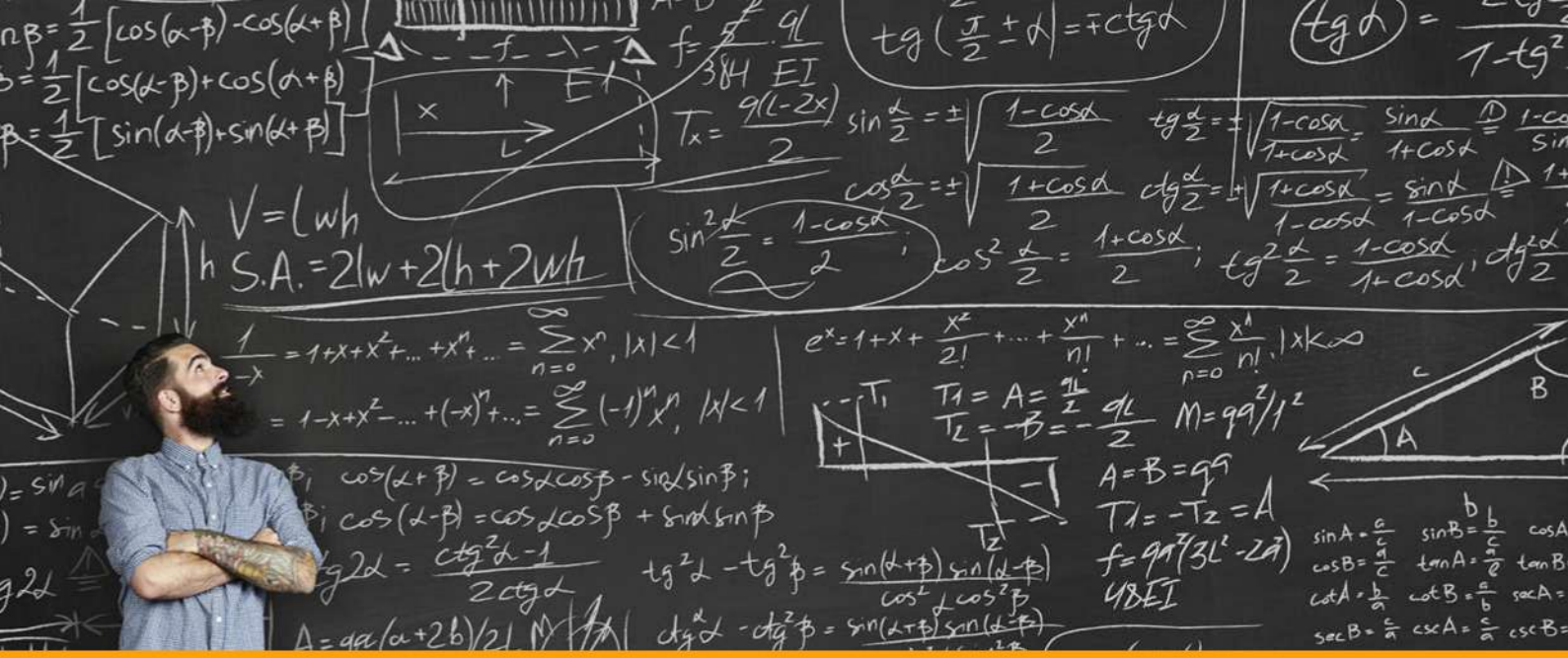
Lumley, T. (2010). *Complex Surveys: A guide to Analysis Using R*. New Jersey: Wiley & Sons.

Maronna, R.A., Martin, R.D., Yohai, V.J. (2006). *Robust Statistics.* New York: John Wiley & Sons.

Rao, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya A*, 28, 47-60.

Robinson, J. (1987). Conditioning ratio estimates under simple random sampling. *Journal of the American Statistical Association* 82: 826-831.

Särndal, C.E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer & Verlag.



LEY DE LOS GRANDES NÚMEROS Y TEOREMA CENTRAL DEL LÍMITE. UN ESTUDIO DE SIMULACIÓN CONSIDERANDO DISTINTOS ESCENARIOS

LIC. MARÍA CELESTE CARBONE

Directora: **MG. FERNANDA MÉNDEZ**

Este trabajo aborda el estudio de la Ley de los Grandes Números y el Teorema Central del Límite. Se presenta una breve reseña histórica. Se incluyen algunas aplicaciones que se le dieron al Teorema Central del Límite y se ejemplifica, mediante diferentes escenarios de simulación, los resultados de dicho teorema no sólo para el caso de variables aleatorias independientes e idénticamente distribuidas sino también bajo condiciones menos restrictivas. Se verifica empíricamente que el Teorema Central del Límite continúa cumpliéndose aun cuando se suprime la hipótesis de que las distribuciones sean idénticas y se debilita la hipótesis de independencia.

INTRODUCCIÓN

La Teoría de la Probabilidad surgió en el siglo XVII en Francia, cuando los reconocidos matemáticos Pierre de Fermat y Blaise Pascal comenzaron a interesarse en los juegos de azar, a raíz de lo cual muchos matemáticos de la época empezaron a introducirse en el tema, entre ellos Abraham De Moivre, Christiaan Huygens y Jacob Bernoulli. Así fue como se fundó definitivamente lo que hoy se conoce como Teoría de la Probabilidad (Aronna et al. (2004)).

Como en toda área de la matemática, era necesario establecer una estructura adecuada para la investigación de esas "experiencias aleatorias". Para tal fin se crearon las llamadas variables aleatorias, objetos fundamentales dentro de la Teoría de la Probabilidad. Éstas representan resultados numéricos obtenidos a partir de la realización de un experimento; y matemáticamente se modelizan como una función a valores reales definida sobre el conjunto de posibles resultados del experimento (espacio muestral). Mientras se fue avanzando en el estudio de estos sucesos azarosos, resultó cada vez más frecuente la necesidad de relacionar muchas variables para poder analizarlos. Surgió naturalmente de esta manera el problema de estudiar la distribución de la suma de una infinidad de variables aleatorias (Aronna et al. (2004)).

Los dos resultados fundamentales vinculados a este problema son la Ley de los Grandes Números (LGN) y el Teorema Central del Límite (TCL).

La LGN, en sus versiones débil y fuerte, permite dar fundamento matemático a la argumentación heurística que interpreta la esperanza de una variable aleatoria como el valor al cual tiende el promedio de varias realizaciones de la variable correspondientes a la repetición de experimentos independientes. También permite fundamentar la noción heurística de la probabilidad de un evento como el valor límite de las frecuencias relativas con que ocurre el evento cuando se repiten muchos experimentos independientes. La ley débil expresa estos resultados en términos de convergencia en probabilidad y la ley fuerte en términos de convergencia casi segura.

El TCL establece que la distribución de la suma de una gran cantidad de variables aleatorias independientes, bajo ciertas condiciones adicionales, se aproxima a una distribución Normal (Aronna et al. (2004)).

Se comienza por aclarar que la denominación Teorema Central del Límite, es relativamente reciente. Fue utilizada por primera vez en 1920 por George Polya. El término "central" significa "fundamental" o de "importancia central".

El TCL estudia el comportamiento de la suma de variables aleatorias, cuando crece el número de sumandos, asegurando su convergencia hacia una distribución Normal en condiciones muy generales. Este teorema, del cual existen diferentes versiones que se han demostrado a lo largo de la historia, tiene una gran aplicación en inferencia estadística, pues muchos parámetros de diferentes distribuciones de probabilidad, como la media, pueden expresarse en función de una suma de variables. También permite aproximar muchas distribuciones de uso frecuente: Binomial, Poisson, Chi-cuadrado, t-Student, Gamma, etc., cuando los valores de sus parámetros crecen y el cálculo se hace difícil. Por otro lado, la suma de variables aleatorias aparece en forma natural en muchas aplicaciones (Aronna et al. (2004)).

En este trabajo se estudian las características y propiedades de la LGN y del TCL. Se intenta, mediante simulaciones, comprobar los resultados del TCL no sólo para el caso de variables aleatorias independientes e idénticamente distribuidas sino también bajo condiciones menos restrictivas. Se extienden los resultados suprimiendo la hipótesis de que las distribuciones sean idénticas y debilitando la hipótesis de la independencia.

METODOLOGÍA

Ley de los Grandes Números

La LGN describe el comportamiento del promedio de una sucesión de variables aleatorias conforme aumenta el número de ensayos. Las diferentes formulaciones de la LGN (y sus condiciones asociadas) especifican la convergencia de formas distintas.

Teorema 1 (Ley Débil de los Grandes Números): Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias no correlacionadas, es decir $Cov(X_i, X_j) = 0$ si $i \neq j$, tal que $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i^2$ para cada $i \in \mathbb{N}$. Se considera la sucesión de variables aleatorias $(\bar{X}_n)_{n\geq 1}$ donde $\bar{X}_n$ es el promedio de las primeras n variables y sea $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\mu_i$, entonces si $\lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\right) = 0$, se tiene que $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$.

Teorema 2 (Ley Fuerte de los Grandes Números): Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias independientes tal que $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i^2$ para cada $i \in \mathbb{N}$. Se considera la sucesión de variables aleatorias $(\bar{X}_n)_{n\geq 1}$ definida por $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i$ y sus respectivas medias $\bar{\mu}_n = E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mu_i$. Entonces si $\sum_{i=1}^{\infty}\frac{\sigma_i^2}{i^2} < \infty$, se tiene que $\bar{X}_n - \bar{\mu}_n \xrightarrow{c.t.p.} 0$.

Una aplicación estadística importante de la convergencia en probabilidad es la consistencia de una secuencia de estimadores. La teoría asintótica estudia las propiedades de los procedimientos de inferencia estadística cuando el tamaño de la muestra n que se utiliza es grande, más precisamente, en el límite cuando n tiende a infinito. Una propiedad deseable para un estimador, es que cuando n es grande la sucesión de estimadores $\delta_n(X_1, \dots, X_n)$ se aproxime en algún sentido al valor que se quiere estimar (Lehmann, (2013)).

Teorema Central de Límite

El TCL no es un único teorema, sino que consiste en un conjunto de resultados acerca del comportamiento de la distribución de la suma (o promedio) de variables aleatorias. Se refiere, con Teorema Central del Límite, a todo teorema en el que se afirma, bajo ciertas hipótesis, que la distribución de la suma de un número muy grande de variables aleatorias se aproxima a una distribución Normal. La siguiente es la versión más simple del TCL para el caso de variables independientes idénticamente distribuidas (Blaiotta et al. (2004)).

Teorema 3 (Teorema Central del Límite): Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias independientes idénticamente distribuidas con variancia finita. Se llama $\mu = E(X_i)$ y $\sigma^2 = Var(X_i) > 0$. Sean las sumas parciales $S_n = \sum_{i=1}^{n}X_i$ y $Z_n = \frac{S_n - \mathrm{E}(S_n)}{\sqrt{Var(S_n)}}$, entonces $Z_n \xrightarrow{D} N(0,1)$.

La demostración puede consultarse en James, B. (2008).

El TCL sigue valiendo bajo condiciones menos restrictivas. El Teorema de Lindeberg da una condición suficiente para que una sucesión de variables aleatorias independientes no necesariamente idénticamente distribuidas converja en distribución a la Normal estandarizada (Yohai (2008)).

Teorema 4 (Teorema Central de Lindeberg): Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias independientes con $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i^2$ para todo $i \in \mathbb{N}$, donde $\sigma_i^2 < \infty$. Sea $S_n = \sum_{i=1}^{n}X_i$ y se llama a $s_n^2 = \sum_{i=1}^{n}\sigma_i^2 = Var(S_n)$. Se definen las variables aleatorias centradas $Y_i = X_i - \mu_i$.

Una condición suficiente para que $Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \xrightarrow{D} N(0,1)$, es que para todo $\varepsilon > 0$, $\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\int_{\{|y|\geq s_n\varepsilon\}} y^2 dF_{Y_i}}{s_n^2} = 0$.

Del Teorema Central del Límite de Lindeberg se deduce la siguiente versión:

Teorema 5 (Teorema Central del Límite de Lyapunov): Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias independientes con $E(X_i)=\mu_i$ y $Var(X_i)=\sigma_i^2<\infty$. Se llama $Y_i=X_i-\mu_i$ a las variables aleatorias centradas. Una condición suficiente para que $Z_n=\frac{S_n-E(S_n)}{\sqrt{Var(S_n)}}\xrightarrow{D}N(0,1)$, es que exista $\delta>0$ tal que $\lim_{n\to\infty}\frac{\sum_{i=1}^{n}E(|Y_i|^{2+\delta})}{{S_n}^{2+\delta}}=0$.

Esta condición es útil cuando las variables tienen momentos finitos de orden mayor que dos.

Las pruebas del Teorema de Lindeberg y Lyapunov pueden consultarse en James, B. (2008).

También se considera el caso de sucesiones de variables aleatorias que están distribuidas de forma idéntica pero no son independientes. Se enuncian teoremas que nos permiten mostrar la normalidad asintótica de promedios o sumas de variables aleatorias para ciertos problemas estadísticos con una cantidad limitada de dependencia entre las variables.

Se supone que $X_1,\dots,X_n$ tienen una distribución conjunta con media común $E(X_i)=\mu$ y covariancias $Cov(X_i,X_j)=\gamma_{ij},\ i,j=1,\dots,n$. Se verifica entonces que: $E(\bar{X}_n)=E\left(\frac{\sum_{i=1}^{n}X_i}{n}\right)=\mu$ y $V(\bar{X}_n)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}$. Una condición suficiente, para que $\bar{X}_n$ sea un estimador consistente de μ es que $V(\bar{X}_n)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}\to 0$.

Entonces, para series estacionarias (Peña, (2010)), la variancia de la media muestral puede escribirse como: $V(\bar{X}_n)=\frac{1}{n}\left[\sigma^2+2\sum_{i=1}^{n-1}\left(1-\frac{i}{n}\right)\gamma_i\right]$.

Se puede observar que, cuando las covariancias no sean cero, la variancia de la media muestral para procesos estacionarios puede ser considerablemente mayor que para observaciones independientes. En efecto, si las γ_i son positivas, el sumatorio puede ser muy grande. La condición para que la variancia de la media muestral tienda a cero al aumentar n es que el sumatorio converja a una constante al aumentar n. Una condición necesaria (aunque no suficiente) para que la serie asociada converja es $\lim_{i\to\infty}\gamma_i=0$. Lo que supone que la dependencia de las observaciones tiende a cero al aumentar el retardo (Peña, (2010)).

Teorema 6 (Proceso media móvil): Sea $X_1,X_2,\dots$ una secuencia de un proceso de media móvil de orden q (Peña, (2010)) con $E(X_i)=\mu$ y $Var(X_i)=\sigma^2<\infty$, entonces $\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{D}N(0,\tau^2)$, donde $\tau^2=\sigma^2+2\sum_{i=1}^{q}\gamma_i$.

La demostración puede consultarse en Ferguson (1996).

Se concluye este tema considerando un caso en el que la normalidad asintótica de la media muestral se mantiene, aunque $\gamma_i\neq 0$ para todo i.

Teorema 7 (Proceso autorregresivo de orden uno): Sea un proceso autorregresivo de orden uno (Peña, (2010)) definido como: $X_t=\mu+\phi(X_{t-1}-\mu)+a_t$, con $|\phi|<1$ y a_t ruido blanco de media cero y variancia ${\sigma_a}^2$. Se supone además que $X_1\sim N(\mu,\sigma^2)$, luego esto implica que $\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{D}N\left(0,\frac{1+\phi}{1-\phi}\sigma^2\right)$.

La demostración puede consultarse en (Ferguson (1996), Hunter (2014)).

APLICACIÓN

Se realiza un estudio de simulación con el fin de verificar empíricamente las propiedades estudiadas, utilizando simulaciones a partir de distintos escenarios de distribución de probabilidad que permiten estudiar el comportamiento de la sucesión de promedios de variables aleatorias. El proceso de simulación utilizado en este trabajo consiste en generar conjuntos de valores aleatorios que respondan a un modelo teórico específico, contemplando diferentes escenarios. En todos los casos se generan 1000 muestras de distintos tamaños (n = 20, 50, 100, 1000) y en cada una de ellas se calcula la media muestral. Luego, se resume y presenta gráficamente la distribución simulada a través de las 1000 muestras. La programación para las simulaciones se realiza en el *software* estadístico R.

Escenario 1: Variables aleatorias independientes e igualmente distribuidas, con esperanza y variancia finitas.

Distribución Normal Estándar

Las medias muestrales están centradas en cero (valor esperado bajo la distribución Normal Estándar) cualquiera sea el tamaño de muestra (gráfico 1). Se ve una notable reducción de la variabilidad a medida que aumenta el n. Para cualquier n, se observa que la forma de la distribución es aproximadamente simétrica con forma de campana y centrada en cero (gráfico 2).

Gráfico 1. Boxplots de las medias muestrales para 1000 muestras tomadas de una distribución Normal Estándar

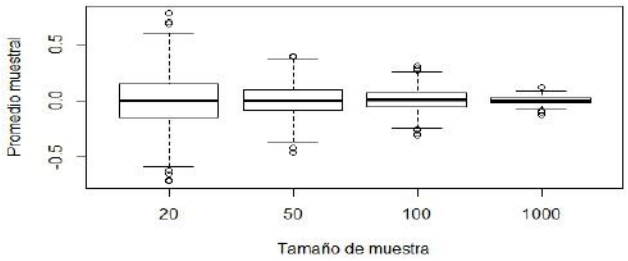
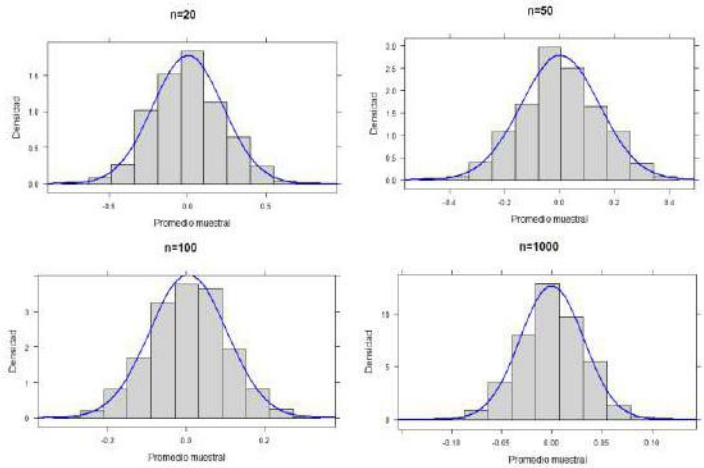


Gráfico 2. Histogramas de frecuencia de las medias muestrales para 1000 muestras de tamaño n tomadas de una distribución Normal Estándar



Escenario 2: Variables aleatorias independientes e igualmente distribuidas, con esperanza y variancia no definidas.

Distribución Cauchy

Este ejemplo muestra que la tendencia Normal de la media muestral afirmada por el TCL requiere algunas suposiciones más allá de que las variables sean independientes e idénticamente distribuidas. En el grafico 3 se ve un outlier que llama mucho la atención para el tamaño de muestra igual a 1000. Además, se observa que la media no converge a ningún valor en particular. En el gráfico 4 se observa que la forma de la distribución para cualquier tamaño de muestra no es simétrica y se ve claramente que no se está cumpliendo el supuesto de normalidad. Se puede demostrar, mediante función característica, que la media muestral tiene la misma distribución de Cauchy para cualquier tamaño de muestra.

Gráfico 3: Boxplots de las medias muestrales para 1000 muestras tomadas de una distribución Cauchy $C(0,1)$

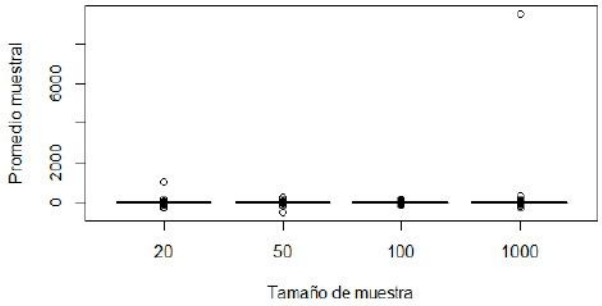
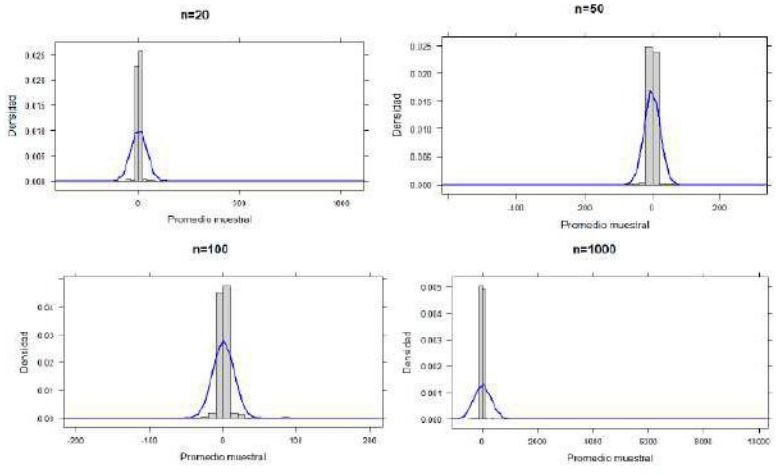


Gráfico 4. Histogramas de frecuencia de las medias muestrales para 1000 muestras de tamaño n tomadas de una distribución Cauchy $C(0,1)$



Escenario 3: Variables aleatorias independientes e igualmente distribuidas, con esperanza finita y variancia no definida.

Distribución t-Student

En el gráfico 5 se observa que las medias muestrales se centran en cero a medida que el tamaño de muestra aumenta con una menor variabilidad. Además, se puede ver la presencia de varios valores extremos que parecen disminuir cuando n aumenta. En el gráfico 6 se ve claramente la presencia de observaciones atípicas. También se observa que la media muestral está centrada en cero y además disminuye la variabilidad a medida que aumenta el tamaño muestral. En cuanto a la forma de la distribución, se observa que no es simétrica en ninguno de los tamaños de muestra calculados. Se concluye que, como es de esperar, el Teorema Central del Límite no se verifica en este caso, dado que por más que el tamaño de muestra sea grande, la distribución de la media no tiende a la distribución Normal.

Gráfico 5. Boxplots de las medias muestrales para 1000 muestras tomadas de una distribución t-Student t_2

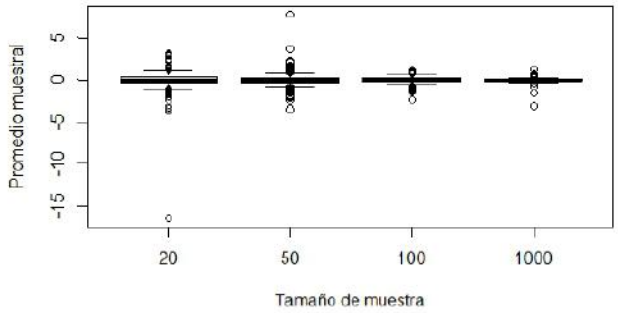
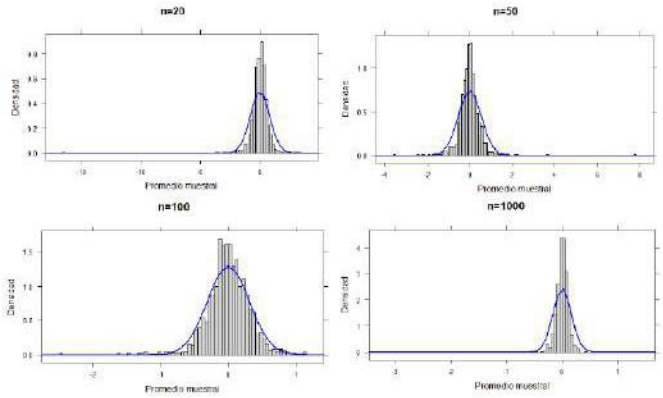


Gráfico 6. Histogramas de frecuencia de las medias muestrales para 1000 muestras de tamaño n tomadas de una distribución t-Student t_2



La media muestral, en el límite, tendrá distribución Normal cualquiera sea la función de distribución siempre que dicha función tenga momento de segundo orden. El grado de aproximación depende del tamaño de la muestra y además de la función particular con la que se quiere trabajar.

Escenario 4: Variables aleatorias igualmente distribuidas, con esperanza y variancia finita, y tales que al menos un par de ellas están correlacionadas.

Proceso de media móvil de orden uno

Particularmente, en este escenario de simulación, los parámetros de la distribución que se utilizaron fueron $\theta = -0.5$, $\mu = 5$ y $\sigma_a{}^2 = 10$.

Gráfico 7. Boxplots de las medias muestrales para 1000 muestras tomadas de un proceso de media móvil de orden uno

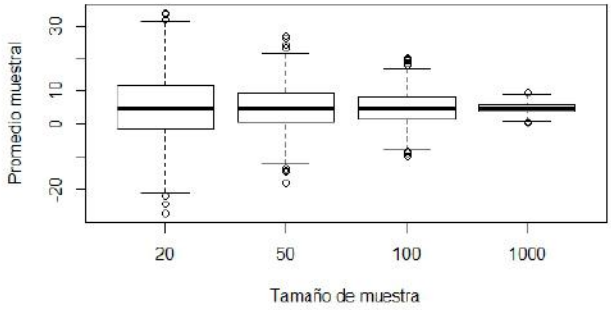
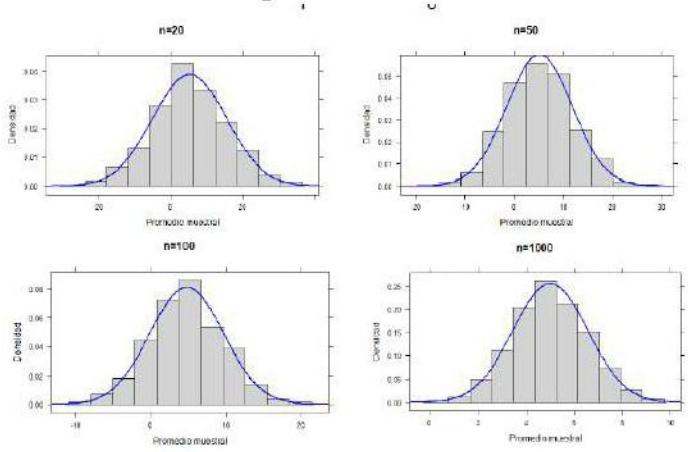


Gráfico 8. Histogramas de frecuencia de las medias muestrales para 1000 muestras de tamaño n tomadas de un proceso de media móvil de orden uno



En el gráfico 7 se observa que las medias muestrales tienden a cinco (valor de la media poblacional del proceso). También se observa una disminución de la variabilidad cuando se incrementa n. En el gráfico 8 se observa que las distribuciones de las medias muestrales se aproximan a la Normal para cualquier tamaño de muestra.

DISCUSIÓN FINAL

La presente tesina abordó el estudio de la Ley de los Grandes Números y del Teorema Central del Límite. Se intentó ejemplificar, mediante diferentes escenarios de simulación, los resultados de dicho teorema. Más precisamente, se verificaron los resultados suprimiendo la hipótesis de que las distribuciones sean idénticas y debilitando la hipótesis de independencia.

El estudio por simulación, realizado en esta tesina, permitió verificar empíricamente algunos de los siguientes resultados:

- El TCL no dice nada acerca de la forma de la función de densidad dada. Cualquiera que sea la distribución, siempre que dicha función tenga momento de segundo orden finito, la media muestral tendrá aproximadamente, para muestras grandes, distribución Normal.
- El TCL se aplica tanto a distribuciones continuas como a distribuciones discretas.
- El grado de aproximación a la distribución Normal depende del tamaño de la muestra y además de la función particular con la que se trabaja.
- Como la variancia de la media muestral es igual a la variancia poblacional dividida por el tamaño de muestra, se deduce que cuanto mayor sea la muestra, más cierto se puede estar de que la media muestral sea una buena estimación de la media poblacional.
- El TCL clásico se refiere al caso en el que las variables son independientes e idénticamente distribuidas. Las extensiones de Lyapunov y Lindeberg proporcionan condiciones bajo las cuales la suma de variables aleatorias independientes que no son necesariamente idénticamente distribuidas tiene una distribución asintótica Normal.
- El TCL continúa cumpliéndose para una secuencia de variables aleatorias dependientes si la dependencia es suficientemente débil. Esto se ilustró en particular para procesos promedio móvil de orden q y para el proceso autorregresivo de orden uno.

BIBLIOGRAFÍA

Aronna, M.; Del Barco, V; Tolonei, P.; Vittone, F. (2004). Teorema Central del Límite. Universidad Nacional de Rosario. Secretaria de Ciencia y Técnica. Facultad de Ciencias Exactas, Ingeniería y Agrimensura.

Blaiotta, J.; Delieutraz, P. (2004). Teorema Central del Límite. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.

Casella, G.; Berger, R. (2008). *"Statistical Inference"*. Segunda Edición. Thompson Press.

Ferguson, T. (1996). *"A course in Large Sample Theory"*. Chapman and Hall.

Hunter, D. (2014). *"Notes for a graduate-level course in asymptotics for statisticians"*. Chapman and Hall.

James, B. (2008). *"Probabilidade: um curso em nível intermediário"*. Associaçao Instituto Nacional de Matemática Pura e Aplicada, IMPA, Tercera edición.

Lehmann, E.L. (2013). *"Elements of Large-Sample Theory"*. Springer Texts in Statistics. Springer.

Peña, D. (2010). *"Análisis de series temporales"*. Alianza Editorial.

Robert, C.P.; Casella, G. (2010). "*Monte Carlo Statistical Methods*". Springer Texts in Statistics. Springer. Segunda Edición.

Yohai, V. (2008). Notas de Probabilidades y Estadística. Basadas en apuntes de clase tomados por Déboli, Alberto durante el año 2003. Versión corregida durante 2004 y 2005 con la colaboración de Szretter, María Eugenia. Departamento de Matemática, UBA.

HERIDOS POR ARMAS DE FUEGO EN LA CIUDAD DE ROSARIO EN EL 2012. SU COMPORTAMIENTO ESPACIAL

LIC. MARTÍN CASTRO

Directora: **MG. VIRGINIA BORRA**
Codirector: **DR. JOSÉ ALBERTO PAGURA**

Los hechos de violencia armada en la ciudad de Rosario, crecieron en el último tiempo. Es por esto que resulta de interés analizar la distribución espacial de los casos de violencia armada, y qué factores explican dicha distribución, con el fin de analizar si existen evidencias de que la ubicación de los hechos es un factor fundamental al momento de tomar medidas de seguridad, o simplemente los casos se distribuyen de forma aleatoria dentro del territorio.

En este trabajo se presentan distintas técnicas para el análisis de datos espaciales, aplicados al conjunto de datos de “Heridos por armas de fuego en la ciudad de Rosario durante el año 2012”.

INTRODUCCIÓN

La violencia armada es uno de los principales obstáculos para el desarrollo humano. Su impacto sobre el bienestar de las personas y el efecto nocivo que tiene sobre las instituciones y la economía, hacen que la reducción de la violencia armada sea una prioridad para las políticas públicas de desarrollo. Este tipo de violencia, presente en todas las sociedades, consiste en el uso o la amenaza de uso de armas de fuego para causar heridas, muerte o daño psicológico. Según la Organización Mundial de la Salud, la violencia armada está entre las primeras cinco causas de muerte entre los adultos.

Se entiende por arma de fuego a cualquier instrumento de defensa y ataque que utiliza la combustión de pólvoras de distintos tipos, en un espacio confinado, para la proyección a distancia de un agente ofensivo (Medicina Legal, 2017).

Las armas de fuego han cobrado un creciente protagonismo en la violencia social en la ciudad de Rosario en la última década. La cantidad de heridos por arma de fuego en la ciudad ha estado en ascenso; entre los años 2003 y 2015 el uso de armas de fuego en el total de homicidios ha crecido un 25,4% en la ciudad. En 2003, 5 de cada 10 casos de homicidios eran causados por un arma de fuego y en 2015 esta cifra asciende a 7 de cada 10 (Observatorio de convivencia y seguridad ciudadana, 2016).

Se define como “herida por arma de fuego” a toda lesión contusa ocasionada por el paso del proyectil a través de los tejidos del cuerpo humano (Medicina Legal, 2017). Es de interés analizar cómo se distribuyen espacialmente los casos de violencia en un territorio, y qué factores explican dicha distribución. Esta información permitirá detectar si existen evidencias de que la ubicación de los hechos es un factor fundamental al momento de tomar medidas de seguridad, o simplemente los casos se distribuyen de forma aleatoria dentro del territorio.

Si bien el número de casos de heridos con arma de fuego puede considerarse como una serie de datos sobre la que puede aplicarse las técnicas estadísticas habituales, cada uno de estos datos tiene asociada coordenadas geográficas o cartesianas y éstas aporta una información adicional que puede emplearse para obtener resultados estadísticos de diversa índole. Más aún, el análisis en exclusiva de los valores de la variable sin considerar la componente espacial asociada puede no ser adecuado por no cumplir algunos de los supuestos de la estadística clásica. Por lo tanto, debido a la naturaleza de este tipo de datos, denominados “Datos Espaciales”, es necesario recurrir a técnicas estadísticas que tengan en cuenta la información geográfica. Al conjunto de métodos y técnicas estadísticas que se han desarrollado para analizar este tipo de datos se lo conoce con el nombre de “Estadística Espacial o Geoestadística”.

Una característica que tienen estos datos es la autocorrelación espacial, considerada como un proceso intrínsecamente geográfico que puede decir mucho acerca del comportamiento de la información georreferenciada a diferentes escalas (Celemín, 2009).

Por lo tanto, el objetivo de este trabajo es analizar la ocurrencia de hechos delictivos con armas de fuego en la ciudad de Rosario en el año 2012, teniendo en cuenta el lugar del hecho, para poder contribuir a la toma de decisiones municipales.

MATERIALES Y MÉTODOS

Fuente de datos

Los datos de los heridos de arma de fuego en la ciudad de Rosario se obtuvieron del Observatorio de Convivencia y Seguridad Ciudadana, los cuales surgen al cruzar información de distintos medios. Por un lado a través de informes de Salud Pública de la Municipalidad de Rosario, que notifica los distintos ingresos de los heridos; por otro lado, a través de partes policiales y los distintos medios de comunicación.

Cada herido por arma de fuego que se notifica, se registra agregándole datos personales como edad, sexo, historia clínica; y datos del hecho, como la ubicación georreferenciada, la cantidad de personas que participaron del mismo, la fecha y hora del evento, motivo, entre otros.

Como variable en estudio se considera el número de personas heridas por arma de fuego, en cada radio censal de la ciudad de Rosario, registradas entre el 1 de enero y 31 de diciembre de 2012. Para analizar dicha variable es necesario tener en cuenta dos características fundamentales: es una variable de tipo conteo y tiene asociada coordenadas del lugar de cada hecho.

Datos espaciales

Se denomina dato espacial a todo aquella unidad que tiene asociada una referencia geográfica y ésta es relevante o informativa, de tal modo que se puede representar la ubicación, el tamaño y la forma de un objeto en el planeta Tierra (Haining, 2003).

Estos tipos de datos poseen características que constituyen a los denominados efectos espaciales, los cuales pueden ser divididos en dos tipos: autocorrelación espacial y heterogeneidad espacial. El primer efecto que se plantea en el análisis de estos datos es que las unidades tomadas en una región específica no son independientes. Así, se supone que si se encuentra una determinada unidad en un punto de este área es más fácil (o inversamente más difícil) encontrar unidades semejantes en puntos próximos a éste que en puntos alejados. Es decir, la variable en estudio presenta mayor correlación cuanta más cercanía hay entre las unidades. Esta dependencia puede ser expresada según la primera ley de la geografía de Tobler (1970), en la cual "todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes". El segundo tipo de efecto, la heterogeneidad espacial, está relacionado con la ausencia de estabilidad en el comportamiento de la variable en el espacio. Más precisamente, esto implica que los parámetros y formas funcionales varían con la ubicación y no son homogéneos en los conjuntos de datos.

Métodos descriptivos

El primero de los métodos presentados es el *Box Map*, el cual es una versión cartográfica del *Box Plot* y muestra la ubicación de los cuartiles representados en el diagrama de caja, sobre el mapa. Además permite identificar valores atípicos de la variable en estudio. En primer lugar, se ordenan los valores de la variable y se divide la distribución de la variable en cuatro grupos correspondientes a los cuantiles [0,25], [25,50], [50,75] y [75,100]. Los cuartiles dividen a una variable en cuatro partes, de forma que cada una contenga aproximadamente igual número de observaciones. Estos cuartiles son representados en el mapa con distintos colores y dependiendo la tonalidad del color varían los valores de la variable en estudio; colores más claros están asociados a valores bajos de la variable y colores más oscuros a valores altos.

Existen distintos índices para medir el grado de autocorrelación espacial global. El más utilizado es el índice I de Moran. Este índice varía entre -1 y 1, pero su interpretación es diferente de los coeficientes de correlación convencionales, los cuales también varían en ese rango. La escala numérica de I está relacionada con su valor esperado, $E(I)$, bajo un patrón espacial aleatorio. Valores menores que la $E(I)$ están asociados con un patrón uniforme/disperso (autocorrelación negativa) y valores más grandes que la $E(I)$ están típicamente asociados con patrones de agrupamientos (autocorrelación positiva). Un valor cercano a $E(I)$ (la cual tiende a 0 cuando n crece) indica ausencia de autocorrelación. Algunas veces el estadístico I de Moran no detecta ciertas estructuras locales, ya que supone que el grado de autocorrelación espacial es igual para todas las unidades espaciales. Sin embargo puede que existan pequeños conglomerados en los que la variable en estudio presenta una concentración importante. Es por esto que se utiliza el índice de asociación espacial local de Moran (I_i). Una característica particular es que se puede llegar al estadístico global de Moran si se suman todos los estadísticos locales I_i ponderados. Esto resulta de gran utilidad, ya que si se detectan valores extremos en la distribución de los índices locales se puede determinar cuáles son las unidades que más aportan al índice global. Este índice (I_i) puede tomar valores entre -1 y 1, indicando autocorrelación espacial para valores alejados a su esperanza y ausencia de correlación para valores cercanos a la $E(I_i)$. En base a dicho índice se pueden determinar distintos tipos de grupos: unidades con valores altos de la variable en estudio y que estén rodeadas de unidades con similares características o viceversa; también se pueden encontrar unidades con valores altos (bajos) rodeados de unidades con valores bajos (altos), a este tipo de agrupación se los conoce como "valores atípicos espaciales"; por último, se pueden encontrar unidades con ausencia de autocorrelación espacial significativa. La representación cartográfica de los índices locales, se realiza a través de mapas coropléticos que permiten estudiar el patrón espacial de la autocorrelación local, conocidos como mapas LISA. En estos mapas se representan las localizaciones de los indicadores de Moran locales con valores significativos. Hay dos clases: los mapas de significación y los mapas de conglomerados.

Conocer la ubicación de los eventos y poder establecer si los conglomerados de eventos resultan significativos en una región en estudio determinada resulta de gran importancia al momento de estudiar la distribución espacial de los datos. Es por eso, que existe un método conocido como detección de conglomerados significativos para corroborar si los conglomerados de puntos son estadísticamente

significativos, basado en un test robusto que puede detectar conglomerados de cualquier tamaño, localizados en cualquier región del mapa y determinar su localización (Estevez, 2011).

Métodos de modelado de la variabilidad espacial

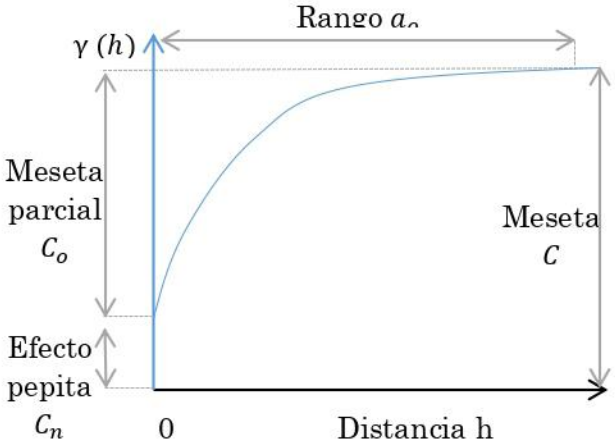
Se plantea, mediante un modelo matemático, la relación existente entre la distancia que separa a las observaciones con los valores de la variable en estudio. Dichos modelos se denominan semivariograma, los cuales responden a la siguiente pregunta ¿Cuán parecidos son los puntos en el espacio a medida que estos se encuentran más alejados? (Guiraldo, 2006).

La semivariancia se define en forma general como:

$$\gamma(h) = \frac{1}{2} Var(y_i - y_{i'}) \qquad (1)$$

En la Figura 1 se ejemplifican las componentes básicas de un modelo de semivariograma y a continuación del mismo se define cada una de las componentes.

Figura 1. Semivariograma teórico



Meseta o silla (C)**:** hace referencia a la máxima semivariancia encontrada entre pares de puntos, se conoce como *sill* y coincide con la variancia de la población.

Rango (a_0): es la distancia a la que la semivariancia deja de aumentar. El rango, por lo tanto, indica la distancia a partir de la cual las unidades son espacialmente independientes unas de otras.

Efecto pepita o nugget (C_n)**:** es la variancia no explicada por el modelo y se calcula como la intersección con el eje Y. Se conoce también como variancia error, puesto que la variancia de dos puntos separados por 0 metros (la intercepción con el eje Y) debería ser cero. Es por ello que esta variancia está normalmente indicando variabilidad a una escala inferior a la muestreada.

Basándose en la tendencia observada en los puntos del semivariograma experimental se deduce el modelo matemático o modelo de semivariograma que mejor se ajusta a ellos, llamado modelo teórico (Ambrosio Flores, 1999). Dependiendo de la forma del semivariograma teórico, se distinguen algunos modelos, como ser el esférico, exponencial, efecto agujero o lineal, entre otros.

Ajustes semivariograma teórico

Si las unidades observadas presentan correlación espacial, en una primera instancia se debe identificar el modelo de semivariograma y estimar sus parámetros. Para ello, el método de mínimos cuadrados ponderados puede ser utilizado. En el ajuste basado en mínimos cuadrados, se desea estimar el vector de parámetros $\boldsymbol{\theta}_q = (a_0,\ C,\ C_n)$ del semivariograma teórico $\gamma(h)$, de modo tal que se minimice

la suma de cuadrados de las diferencias ponderadas entre el semivariograma empírico y teórico, $R(\boldsymbol{\theta}_q)$, dado por la siguiente expresión:

$$R(\boldsymbol{\theta}_q) = \sum_{l=1}^{L} p_i^2 (\hat{\gamma}(h_l) - \gamma(h_l; \boldsymbol{\theta}_q))^2, \quad (2)$$

donde los pesos son $p_i^2 = \frac{1}{Var(\hat{\gamma}(h_l))}$ para mínimos cuadrados ponderados y $p_i^2 = 1$ en el caso de mínimos cuadrados ordinarios.

Existen medidas de bondad de ajuste que permiten la comparación y elección del mejor modelo de semivariograma. Cuando se utiliza mínimos cuadrados ponderados, aquel modelo que presenta el menor valor de $R(\boldsymbol{\theta}_q)$, tiene el mejor ajuste. Otra medida de bondad de ajuste es el Criterio de Información de Akaike (AIC), el cual asume que los errores de los modelos se distribuyen de forma normal e independiente. Esta suposición no es correcta en el ajuste de la semivariancia, sin embargo, AIC también se puede definir de manera operativa sobre $R(\boldsymbol{\theta}_q)$ como:

$$AIC = L\, ln\left(\frac{R(\boldsymbol{\theta}_q)}{L}\right) + 2q \quad (3)$$

donde L es la cantidad de intervalos que distan a una distancia h y q son los parámetros del modelo de semivariograma.

RESULTADOS

Se presentan los resultados de las técnicas para datos espaciales recientemente mencionadas, aplicadas al conjunto de personas heridas por arma de fuego en la ciudad de Rosario durante el año 2012. Dichos datos se obtuvieron por medio del Observatorio de Convivencia y Seguridad Ciudadana, dedicado a generar un sistema de indicadores e índices en seguridad y convivencia con el objetivo de sistematizar, analizar y comunicar información georreferenciada sobre el delito. En este marco, el observatorio registra todos los hechos de violencia armada agregándole datos personales como edad, sexo, historia clínica; y datos del hecho: ubicación georreferenciada, cantidad de personas que participaron del mismo, fecha y hora del evento, motivo, entre otros.

Se registran 643 heridos por arma de fuego durante el año 2012 en la ciudad de Rosario. De estos se logra obtener información del lugar del hecho en 564 casos. En la Figura 2 se presenta el mapa de la ciudad de Rosario dividida según distrito y se muestra la ubicación exacta del lugar del hecho registrado en el periodo en estudio. Teniendo en cuenta los porcentajes de heridos en cada distrito, se destaca que la mayoría de los casos ocurrieron en el distrito oeste, donde se registraron más del 30,00% de los hechos. Los distritos sur, sudoeste y noroeste, son seguidos con cifras cercanas al 18,00%. Los distritos norte y centro son en los que se registraron menor cantidad de hechos, con cifras inferiores al 10,00%.

Por otro lado, en el mapa de la Figura 3 se observa la distribución geográfica de los radios censales en función de la cantidad de heridos. Hay una mayor concentración de los hechos en los radios de los distritos oeste, sur y sudoeste de la ciudad, los cuales se representan en el mapa con un color azul oscuro. Mientras que en los distritos centro y norte se observan gran cantidad de radios en los que no se registraron hechos de violencia armada durante el año 2012.

Figura 2. Distribución de los heridos de arma de fuego según distrito (n=564)

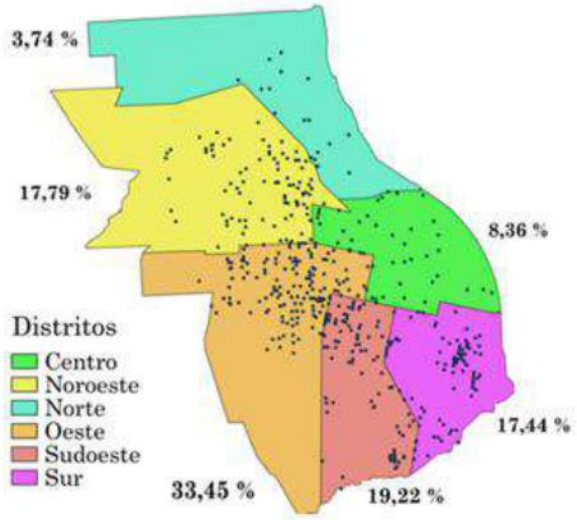
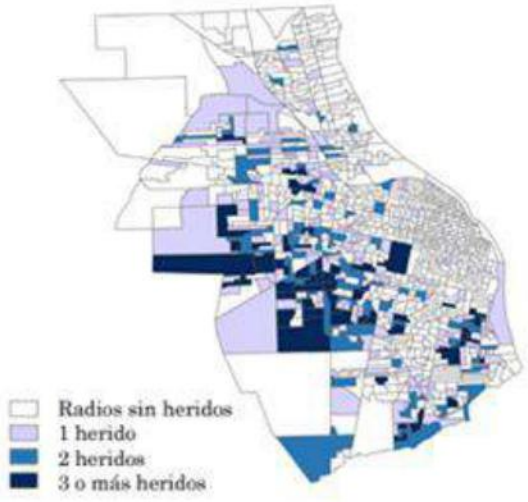


Figura 3. Distribución geográfica del porcentaje de personas heridas por arma de fuego según radio censal



El Índice de Moran alcanza un valor de 0,33 y resulta ser significativo, ya que supera ampliamente su valor esperado ($E(I)$=-0,0009) y su probabilidad asociada es igual a 0,01, es decir, los hechos de violencia no ocurrieron aleatoriamente en la ciudad de Rosario, en cuanto a su ubicación geográfica.

Como se mencionó, el estadístico *I* de Moran puede no detectar la autocorrelación espacial local, es por eso que se presentan en las Figuras 4 y 5 los mapas LISA para evaluar la significación de autocorrelación espacial local, con el fin de mostrar algún agrupamiento de radios censales en diversas regiones de la ciudad.

Figura 4. Mapa de significación LISA

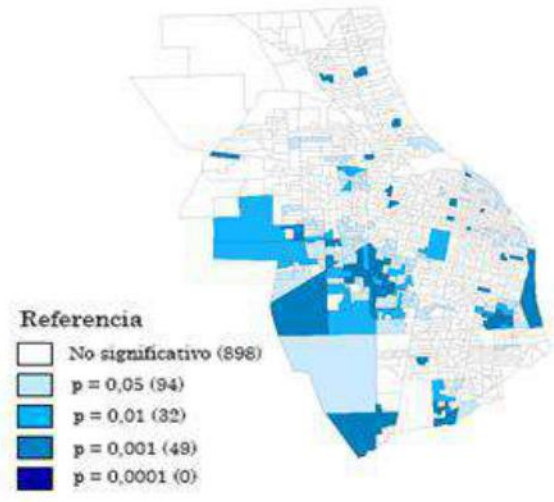
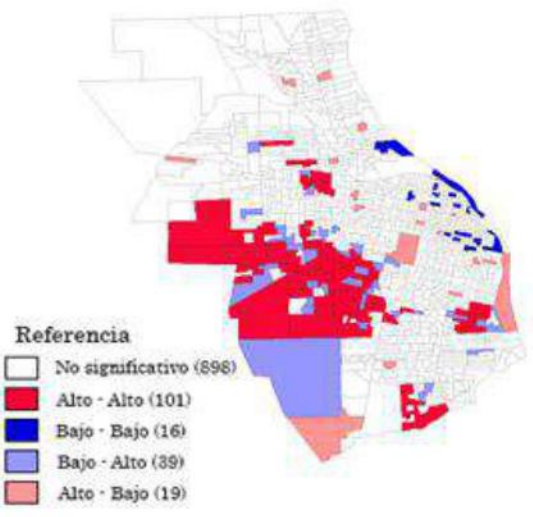


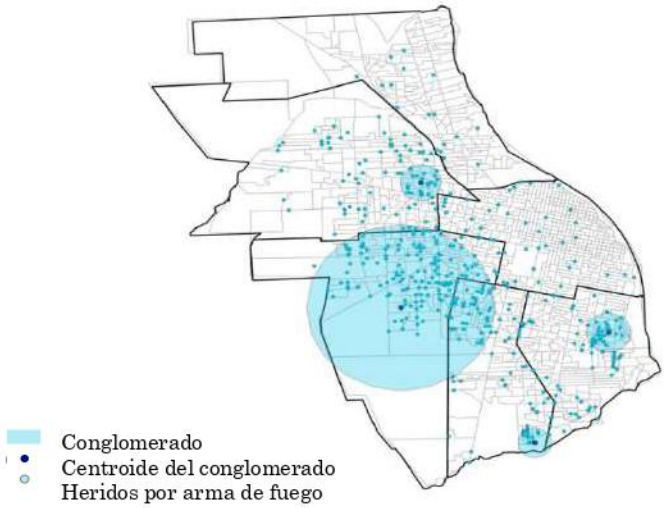
Figura 5. Mapa de conglomerados LISA



Debido a la falta de aleatoriedad de la distribución de los hechos y la concentración de puntos en determinados lugares de la ciudad, se busca la detección específica de conglomerados. Para esto, se trabaja con la cantidad de hechos y con la población en cada radio censal utilizando los datos recolectados en el Censo de Población, Hogares y Viviendas del año 2010 del INDEC. Se busca obtener conglomerados de radios que presenten valores altos y que resulten significativos a un valor α prefijado.

Se encontraron cuatro conglomerados significativos, al 5,00%, en toda la ciudad. De los 4 conglomerados, uno abarca 126 radios censales y contiene 233 hechos de violencia armada; otro incluye 21 radios y contiene 43 casos; el tercero de los conglomerados significativos comprende 13 radios y contiene 30 lesionados y el restante, abarca 4 radios censales donde ocurrieron en su interior 18 heridos por arma de fuego.

Figura 6. Ubicación de los conglomerados significativos



Se modeló la variabilidad de la variable número de heridos por arma de fuego por radio censal en la ciudad de Rosario según el modelo esférico.

En base a las estimaciones obtenidas, la expresión analítica del modelo que mejor ajusta los datos es la siguiente:

$$\hat{\gamma}_{esférico} = \begin{cases} 0{,}2815 + 1{,}2637 \left(\frac{3}{2} \frac{h}{5.946{,}08} - \frac{1}{2} \frac{h^3}{5.946{,}08^3} \right) & 0 < h \leq 5.946{,}08 \\ 0{,}2815 + 1{,}2637 & h > 5.946{,}08. \end{cases}$$

CONSIDERACIONES FINALES

Por medio del análisis espacial se registraron radios censales con alto número de heridos por arma de fuego rodeados por radios de las mismas características. Esto indica el agrupamiento de los eventos estudiados. También se observaron radios censales con pocos hechos rodeados con radios de igual característica. Se obtuvieron, además, por medio de un test robusto, la detección de 4 conglomerados de radios censales significativos.

Se modela la variabilidad espacial de los datos y se obtiene que el modelo que mejor ajusta los datos es el esférico, obteniendo también las estimaciones correspondientes de los parámetros.

Como continuación de la línea de trabajo se considera útil replicar este análisis considerando los heridos por arma de fuego durante diferentes años, junto con investigaciones de tipo cualitativa y/o criminal para la comprensión profunda de los factores que explican este tipo de delito.

BIBLIOGRAFÍA

Ambrosio Flores, L. (1999). Estadística Espacial. Universidad politécnica de Madrid, España. ISBN: 84-7401- 156-6.

Celemín, J. (2009). Autocorrelación espacial e indicadores locales de asociación espacial. Importancia, estructura y aplicación.

Estevez, J. (2011). Análisis espacial del brote de dengue en la ciudad de Rosario durante el año 2009.

Guiraldo, A (2006). Geoestadística. Disponible en: https://fjferrer.webs.ull.es/Bibliog/Biblio/Geoestadistica.pdf [Último acceso: agosto 2017].

Haining, R. (2003). Spatial Data Analysis: Theory and Practice.

Medicina Legal. Jurisprudencia médica. Disponible en: https://medicinalegalaldia.blogspot.com.ar/2008/03/heridas-por-armas-de-fuego.html. [Último acceso: agosto 2017].

Observatorio de convivencia y seguridad ciudadana. Municipalidad de Rosario. (2016)

ESTUDIO DE LA SITUACIÓN DE ALQUILERES EN LA CIUDAD DE ROSARIO

LIC. FRANCO COMETTO

Tutora académica: **LIC. GUILLERMINA ISERN**
Tutora institucional: **LIC. NORA VENTRONI**

Este trabajo es producto de la Práctica Profesional llevada a cabo en el Centro de Asesoramiento Social en Alquileres (CASA) -dependiente del Servicio Público de la Vivienda y el Hábitat- y la Dirección General de Estadística, ambas dependencias de la Municipalidad de Rosario.

La oficina solicitante deseaba contar con un perfil de la población que alquila en la ciudad de Rosario como también reconocer la dinámica del mercado de alquileres. Para ello se consultaron tres fuentes. En primer lugar, datos oficiales provistos por el INDEC, para lograr una caracterización de los hogares de la ciudad. En un segundo momento se trabajó con información de avisos clasificados para describir el mercado de alquileres. Por último, se procesó la información de los registros internos de la institución.

El trabajo realizado constituye un avance hacia un mayor entendimiento de la problemática de los alquileres en la ciudad, posibilita adecuar las herramientas de trabajo para ofrecer un servicio más eficiente ante las necesidades de los inquilinos.

INTRODUCCIÓN

Desde el Centro de Asesoramiento Social en Alquileres (CASA) -dependiente del Servicio Público de la Vivienda y el Hábitat- de la Municipalidad de Rosario se planteó la necesidad de obtener información actualizada y confiable en cuanto a la situación de alquileres de la ciudad. Esta temática es tratada con frecuencia en diferentes medios de comunicación, los cuales citan cifras al respecto que son muy disímiles según su fuente, o bien no tienen un origen claro.

La oficina demandante de colaboración y asesoramiento, opera en la ciudad de Rosario desde septiembre de 2016. Su objetivo es proveer al inquilino de asistencia para el acceso a un inmueble de alquiler.

El CASA ofrece puntualmente, a través del programa *Hoy Alquilo* de la Municipalidad de Rosario, tres servicios para los inquilinos:

- Asesoramiento legal gratuito con el fin de brindar información y evacuar dudas de acuerdo a los compromisos asumidos a la hora de formalizar un contrato de alquiler.

- Acceso a una línea de créditos, a través del Banco Municipal, con el objetivo de afrontar los gastos iniciales derivados de un contrato de alquiler. El mismo cuenta con una tasa fija, subsidiada por la Municipalidad de Rosario.

- Posibilidad de una garantía o fianza bancaria a través del Banco Municipal que podrá ser utilizada en los contratos de alquiler de vivienda dentro de la ciudad.

A partir de las necesidades de la oficina se procedió a obtener datos confiables y actualizados respecto a la situación de alquileres en la ciudad, para lo cual se decidió consultar fuentes oficiales de información. En particular se trabajó con información proveniente de dos fuentes, ambas provistas por el Instituto Nacional de Estadística y Censos (INDEC). Se tomaron datos correspondientes al Censo Nacional de Población, Hogares y Viviendas, el cual fue realizado por última vez en el año 2010 y a la Encuesta Permanente de Hogares, la cual se realiza de manera continua en el país. De esta última se utilizó, particularmente, información respecto a los años 2005, 2010 y 2015.

En segundo lugar, para describir el mercado de alquileres se recurrió a la observación de los avisos clasificados publicados por un medio virtual en dos períodos considerados. El portal seleccionado corresponde a la edición web de un diario de amplia cobertura y con gran tradición en la publicación de clasificados. Por cuestiones de acceso y conformación de bases de datos con la información, se trabajó con su formato web en vez de utilizar formato papel. Otro de los motivos por los que se seleccionó este medio es que, a diferencia de otros portales disponibles en internet, incluye la publicación de alquileres ofrecidos directamente por propietarios.

Para finalizar se evaluó la información disponible en los registros de asistentes al CASA. Se recolectaron datos de las planillas archivadas en la oficina, se conformó una base de datos con las atenciones realizadas en los últimos meses disponibles (marzo y abril de 2017), para luego generar un informe.

REVISIÓN DE FUENTES OFICIALES

Para el análisis de datos censales se utilizó la base de datos Redatam. Esta herramienta permite el procesamiento de los datos correspondientes al Censo 2010 en su cuestionario básico de manera libre y gratuita. La máxima desagregación que permite esta herramienta es de radios censales. Estos radios definidos previamente al relevamiento por el IPEC/INDEC, cubren un espacio territorial con límites geográficos y una determinada cantidad de viviendas a relevar. En zonas urbanas, cada radio tiene, en promedio, 300 viviendas. Los radios son el máximo nivel de desagregación disponible para realizar análisis específicos de acuerdo a la problemática que se desee estudiar.

El interés en este primer análisis consistió en caracterizar los hogares pertenecientes a viviendas alquiladas en comparación al total de hogares de la ciudad. Por este motivo se analizaron las variables como el total de habitaciones, el número de dormitorios y el total de personas en el hogar, que refieren a las características de los hogares, y las variables como edad, nivel educativo y condición de actividad del jefe de hogar, que caracterizan a el/la jefe/a de hogar.

Con el objetivo de complementar y actualizar la información obtenida en base al Censo Nacional de Población, Hogares y Viviendas, se analizaron los resultados de la Encuesta Permanente de Hogares (EPH) en el periodo correspondiente al año del último censo (2010). En este caso se comparan las características de la población y los hogares con el fin de detectar diferencias entre las mediciones realizadas en cada fuente, dado que la EPH tiene una cobertura que excede a la ciudad, es decir, la información de ésta corresponde al Gran Rosario, y además los datos provienen de una muestra.

Además se utilizó información de la EPH correspondiente a los años 2005 y 2015 con el fin de observar variaciones a través del tiempo de ciertas variables de interés.

En el Censo de 2010, se registró para la provincia de Santa Fe un total de 1023777 hogares. De estos, 157265 (15.4%) se encontraban en situación de alquiler.

En Rosario se censaron 320532 hogares, de los cuales 60388 residían en viviendas alquiladas (18.8%). En estos vive el 15.2% del total de población de la ciudad (144599 de los 948312 habitantes).

La Tabla 1 resume los principales resultados censales, mientras que la Tabla 2 muestra información para 3 períodos seleccionados de la EPH.

Tabla 1. Características de los hogares en la ciudad de Rosario. Año 2010

		Hogares en viviendas alquiladas		Total de hogares en Rosario	
		Cantidad	Porcentaje	Cantidad	Porcentaje
Cantidad de ambientes	1	9976	16.5%	31032	9.7%
	2	19645	32.5%	74926	23.4%
	3	18502	30.7%	101280	31.6%
	4	7664	12.7%	63170	19.7%
	5 o más	4601	7.6%	50124	15.6%
Cantidad de dormitorios	1	25780	42.7%	73346	22.9%
	2	27139	44.9%	159466	49.8%
	3	6430	10.7%	73508	22.9%
	4	777	1.3%	12020	3.7%
	5 o más	262	0.4%	2192	0.7%
Total de personas en el hogar	1	18587	30.8%	68302	21.3%
	2	18758	31.1%	83115	26.0%
	3	11161	18.5%	62905	19.6%
	4	7268	12.0%	55787	17.4%
	5 o más	4614	7.6%	50423	15.7%
Condición de actividad del jefe de hogar	Ocupado	49312	81.6%	228255	71.2%
	Desocupado	1725	2.9%	8407	2.6%
	Inactivo	9351	15.5%	83939	26.2%
Edad del jefe de hogar	Menos de 20	1265	2.1%	3209	1.0%
	De 20 a 39	34275	56.7%	99767	31.1%
	De 40 a 64	18996	31.5%	141024	44.0%
	65 o más	5852	9.7%	76601	23.9%
Nivel educativo del jefe de hogar	Primario Incompleto	2385	4.0%	33769	10.7%
	Primario Completo	7312	12.2%	76845	24.4%
	Secundario Incompleto	6517	10.9%	40059	12.7%
	Secundario Completo	12574	21.0%	65033	20.6%
	Superior Incompleto	17428	29.0%	42145	13.4%
	Superior Completo	12240	20.4%	52014	16.5%
	Post Universitario Incompleto	774	1.3%	2006	0.6%
	Post Universitario Completo	786	1.3%	3611	1.1%
Composición del hogar	Solo jefe/a de hogar	18587	30.8%	68302	21.3%
	Solo jefe/a y cónyuge/pareja	10920	18.1%	49153	15.3%
	Otros	30881	51.1%	203068	63.4%

Fuente: Censo Nacional de Población, Hogares y Viviendas 2010 – INDEC.

Tabla 2. Distribución porcentual de los hogares según relación de parentesco de los miembros. Aglomerado Gran Rosario. Años 2005, 2010 y 2015

Composición del hogar	2005		2010		2015	
	Hogares en viviendas alquiladas	Total de hogares	Hogares en viviendas alquiladas	Total de hogares	Hogares en viviendas alquiladas	Total de hogares
Solo jefe de hogar	18.1%	13.8%	29.7%	21.7%	21.0%	21.5%
Solo jefe y cónyuge/pareja	20.0%	16.0%	21.2%	16.0%	16.3%	14.2%
Solo jefe e hijos	8.1%	11.1%	7.7%	11.0%	9.4%	11.4%
Jefe con cónyuge/pareja e hijos	33.9%	40.6%	28.3%	34.3%	37.9%	36.6%
Otros	19.8%	18.5%	13.0%	17.0%	15.4%	16.3%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Fuente: Encuesta Permanente de Hogares – INDEC.

ANÁLISIS DE LOS AVISOS CLASIFICADOS DE ALQUILER DE VIVIENDAS

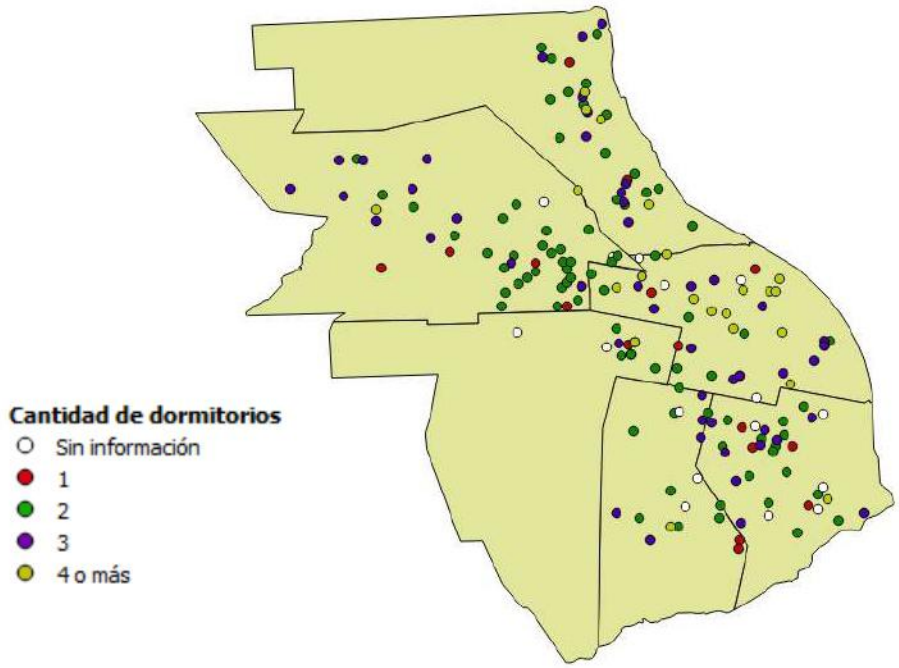
Con el objetivo de caracterizar la oferta de alquileres en la ciudad se analizaron los avisos publicados en un diario local en su versión digital. Se procesó información de los meses de mayo y junio de 2017, a modo de establecer una propuesta de seguimiento mensual de variables relacionadas con la oferta de alquileres. Entre ellas se evaluaron características de tamaño, ubicación y precio de las viviendas en alquiler.

Se optó por el Diario La Capital como fuente de datos por sobre otros portales de clasificados vía web, debido tanto a su volumen como a su contenido. La cantidad de avisos en este portal en el periodo considerado fue de aproximadamente 1500, mientras que otras páginas web disponibles no superaban las 400 publicaciones. Por otro lado, esta fuente incluye no solo ofertas realizadas por inmobiliarias sino también aquellas con trato directo con el propietario.

La conformación de la base de datos se realizó registrando en forma manual los avisos, extrayendo información referida a la cantidad de dormitorios del inmueble, origen de la publicación (inmobiliaria o propietario), monto de alquiler y dirección, así como también, identificando aquellos inmuebles que tuvieran características particulares que pudieran afectar el costo del alquiler (amueblados, con cochera y/o impuestos incluidos en el monto publicado).

A modo de ejemplo, en mayo de 2017 la mayor parte de la oferta de casas ubicadas en el distrito Centro se caracterizó por tener 3 o más dormitorios (Gráfico 1). En tanto que, la oferta de departamentos se concentró principalmente en el distrito Centro (Tabla 3) y casi un 58 % corresponde a monoambientes o con un dormitorio (Tabla 4).

Gráfico 1. Ubicación de la oferta de casas según cantidad de dormitorios



Fuente: Avisos clasificados del diario La Capital, mayo de 2017.

Tabla 3. Distribución por distritos de las propiedades en alquiler

Distrito	Departamentos		Casas	
	Cantidad	Porcentaje	Cantidad	Porcentaje
Centro	1264	86.0%	42	22.3%
Norte	86	5.9%	32	17.0%
Noroeste	20	1.3%	49	26.1%
Sur	48	3.3%	33	17.6%
Sudoeste	14	1.0%	20	10.6%
Oeste	37	2.5%	12	6.4%
Total	**1469**	**100%**	**188**	**100%**

Fuente: Avisos clasificados del diario La Capital, mayo de 2017.

Tabla 4. Inmuebles ofertados según cantidad de dormitorios

Dormitorios	Departamentos		Casas	
	Cantidad	Porcentaje	Cantidad	Porcentaje
Monoambiente	244	16.3%	-	-
1	623	41.6%	17	8.7%
2	343	22.9%	88	45.1%
3	87	5.8%	50	25.6%
4 o más	5	0.3%	23	11.8%
Sin información	196	13.1%	17	8.7%
Total	**1498**	**100%**	**195**	**100%**

Fuente: Avisos clasificados del diario La Capital, mayo de 2017.

ANÁLISIS DE LAS CONSULTAS EN EL CENTRO DE ASESORAMIENTO SOCIAL EN ALQUILERES

El CASA lleva un registro de las consultas recibidas mediante un formulario que se completa de manera manuscrita por personal de la oficina. En dicho formulario se registran características demográficas de los consultantes, su situación de alquiler, situación laboral y además, detalla el motivo de la consulta.

Con el objetivo de describir las características de las personas que asisten al CASA, así como conocer las necesidades e inquietudes que presentan más frecuentemente, se analizaron dichos registros durante los meses de marzo y abril de 2017. En tales meses se produjeron 103 y 88 consultas, respectivamente (Tablas 5, 6 y 7).

Tabla 5. Distribución por edades de los asistentes al CASA entre marzo y abril de 2017

Grupo etáreo	Cantidad	Porcentaje
De 20 a 39	122	63.9%
De 40 a 64	52	27.2%
65 o más	9	4.7%
Se desconoce	8	4.2%
Total	**191**	**100%**

Fuente. Registro de consultas en el CASA.

Tabla 6. Consultas mensuales de los asistentes al CASA según nivel educativo

Nivel educativo	Marzo		Abril	
	Cantidad	Porcentaje	Cantidad	Porcentaje
Primario	0	0.0%	2	2.3%
Secundario	5	4.9%	22	25.0%
Terciario o Universitario	13	12.6%	55	62.5%
Se desconoce	85	82.5%	9	10.2%
Total	**103**	**100%**	**88**	**100%**

Fuente. Registro de consultas en el CASA, marzo y abril de 2017.

Tabla 7. Consultas mensuales de los asistentes al CASA según servicio requerido

Servicio	Marzo		Abril	
	Cantidad	Porcentaje	Cantidad	Porcentaje
Asesoramiento Legal	63	61.2%	61	69.3%
Crédito	28	27.2%	22	25.0%
Garantía	24	23.3%	19	21.6%

Fuente. Registro de consultas en el CASA, marzo y abril de 2017.

Nota: los porcentajes fueron calculados sobre el total de personas que asistieron al CASA. Puede suceder que una misma persona consulte acerca de varios servicios. Por este motivo la suma de los porcentajes excede el 100%.

CONCLUSIONES

Analizando datos de fuentes oficiales (Censo de Población y Encuesta Permanente de Hogares) se logró hacer una caracterización de la población de la ciudad y/o el Aglomerado. En particular, de aquellos hogares residentes en viviendas alquiladas. Aproximadamente el 19% del total de hogares de la ciudad, se encontraban en situación de alquiler en el año 2010, según datos censales.

Los hogares, tanto en el caso de viviendas alquiladas como en general, están conformados mayormente (en un 60%) por grupos familiares. En el caso de los jefes de hogar que pertenecen a viviendas alquiladas, a diferencia de la población general, es más frecuente que presenten edades menores (50% por debajo de los 40 años) y hayan cursado algún tipo de educación superior (aproximadamente 50% de los hogares).

Respecto a la información provista por la EPH, se observó un cambio en los diferentes períodos considerados, encontrando que el porcentaje de hogares en los que habitan sólo una o dos personas en el año 2015 es menor que en los años 2010 y 2005. Se debe mencionar que, si bien se utilizaron diferentes períodos para comparar las características de interés en diferentes años, los cambios registrados pueden deberse a múltiples factores y no deben ser extrapolados más allá de lo observado en el presente análisis.

Al proceder al análisis de la oferta de alquileres de casas (en base a avisos clasificados), aquellas que cuentan con dos dormitorios cubren entre el 45% y el 50% del total de avisos y están ubicadas por fuera de la zona centro. Mientras que en ésta se ofrecen casi exclusivamente casas con 3 dormitorios o más.

En el caso de los departamentos, la oferta se concentra principalmente en la zona centro donde se ubica más del 85% de los avisos publicados. En su mayoría las publicaciones son de departamentos de pequeñas dimensiones. Alrededor del 60% del total corresponde a monoambientes o departamentos con un dormitorio.

En los dos meses en que fueron realizadas las mediciones se encontraron resultados similares. Por lo que es posible suponer que, sucesivas mediciones utilizando la metodología propuesta, permiten evaluar las características de los inmuebles ofrecidos por este medio. Sin embargo, al tomar períodos tan cercanos en el tiempo podría haber propiedades que estén presentes en ambos meses produciendo un solapamiento entre las muestras analizadas. La continuidad de esta metodología a largo plazo, tomando en cuenta las consideraciones previas, permitiría realizar análisis acerca del comportamiento de las variables de interés a través del tiempo observando variaciones interanuales, cambios estacionales, etc.

A través del análisis del Registro de consultas a la oficina del CASA se observó que aproximadamente dos de cada tres asistentes tenían menos de 40 años, dato que se corresponde con la información obtenida de otras fuentes. Respecto al nivel educativo, en función de los datos en el mes de abril, la mayor parte de los mismos cursa o cursó estudios superiores.

De los servicios ofrecidos a los inquilinos, el asesoramiento legal fue requerido por más del 60% de las personas que asistieron en los meses considerados. Entre las consultas realizadas, la mayor parte era referida al pago de impuestos o expensas, arreglos que debían realizarse o fueron realizados en la propiedad y acerca del contrato de alquiler, particularmente por rescisión del mismo o consultas al momento de renovarlo.

En los casos en que el consultante ya se encontraba alquilando al momento de la consulta y se conocían los montos de alquiler, se calculó el incremento del mismo entre el primer y segundo año del contrato. En promedio, éste resultó de un 28.9%.

Si bien el objetivo de este análisis resulta puramente descriptivo, basado en la información disponible en cuanto a los registros de atención de los últimos meses, es importante destacar la necesidad de contar con información completa, confiable y actualizada acerca de la atención en el CASA. En el futuro, el Centro contará con un sistema informático de registro de las asistencias, lo cual facilitará el procesamiento de esta información. Esto permitirá evaluar el funcionamiento del servicio, con el fin de brindar una mejor atención a los inquilinos y ajustarse a las necesidades e inquietudes que estos presenten.

Es innegable la importancia de contar con datos confiables, completos y actualizados para hacer frente a una problemática tan vasta y compleja como lo son los alquileres en la ciudad. El trabajo realizado en este sentido, partiendo de las fuentes disponibles y de los registros propios del CASA, constituye un avance hacia un mayor entendimiento de la misma. Esto posibilita adecuar las herramientas de trabajo para ofrecer un servicio más eficiente que esté a la altura de las necesidades de los inquilinos.

BIBLIOGRAFÍA

Dirección de Cartografía - Dirección General de Topografía y Catastro - Secretaría de Hacienda y Economía - Municipalidad de Rosario. (22 de Diciembre de 2016). *Distritos Descentralizados.* Recuperado en Abril de 2017, de Rosario Datos: http://datos.rosario.gob.ar/dataset/distritos-descentralizados

Dirección de Cartografía - Dirección General de Topografía y Catastro - Secretaría de Hacienda y Economía - Municipalidad de Rosario. (22 de Diciembre de 2016). *Barrios.* Recuperado en Abril de 2017, de Rosario Datos: http://datos.rosario.gob.ar/dataset/barrios

Instituto Nacional de Estadística y Censos. (8 de Junio de 2015). *Base de datos Redatam*. Recuperado en Abril de 2017, de http://200.51.91.245/argbin/RpWebEngine.exe/PortalAction?BASE=CPV2010B

Centro Latinoamericano y Caribeño de Demografía - División de Población de la Comisión Económica para América Latina y el Caribe - Naciones Unidas. (2017). *Download Redatam*. Recuperado en Abril de 2017 de https://www.cepal.org/es/temas/redatam/download-redatam

Quantum GIS Development Team. (2017). *Quantum GIS Development Team.* Recuperado en Mayo de 2017, de Open Source Geospatial Foundation Project: https://www.qgis.org/es/site/forusers/download.html

Instituto Provincial de Estadística y Censos. (2017). *Plantas Urbanas por localidad. Provincia Santa Fe*. Recuperado en Mayo de 2017 de http://www.ipec.santafe.gov.ar/descarga/index.php

Dirección de Cartografía - Dirección General de Topografía y Catastro - Secretaría de Hacienda y Economía - Municipalidad de Rosario. (5 de Octubre de 2017). *Nomenclador de calles*. Recuperado en Mayo de 2017, de Rosario Datos: http://datos.rosario.gob.ar/dataset/nomenclador-de-calles

INTRODUCCIÓN AL ANÁLISIS DE COSTO-EFECTIVIDAD

LIC. PABLO COTTET

Directora: **MG. NORA ARNESI**

En las últimas décadas, el razonamiento económico se fue incorporando al campo de la salud ya que sus premisas son perfectamente aplicables a lo que ocurre en los sistemas sanitarios. En especial en el área de la salud pública donde los recursos son escasos, es necesario tomar decisiones racionales sobre cuál es la mejor forma de asignarlos. Si bien la guía fundamental debe ser la eficiencia, la seguridad y la efectividad clínica, la evaluación económica incorpora al proceso de la toma de decisiones los costos asociados a cada uno de los tratamientos disponibles. Es decir, se elige aquél tratamiento que sea costo-efectivo.

INTRODUCCIÓN

La evaluación económica en salud es un término genérico que engloba diversas técnicas o procedimientos que pueden usarse para recabar información sobre la relación que existe entre el costo y el resultado de las intervenciones en salud (Cardozo, 2007).

Si bien la guía fundamental de las decisiones tomadas en el sistema sanitario deben ser la eficiencia, la seguridad y la efectividad clínica, la evaluación económica de las intervenciones sanitarias incorpora al proceso de toma de decisiones, además del análisis de los resultados, los costos asociados a cada una de las intervenciones.

Los términos eficacia, efectividad y eficiencia son ampliamente utilizados en este contexto y sus definiciones suelen confundirse. Por este motivo resulta de interés para este trabajo clarificar brevemente cada una de sus definiciones.

La eficacia mide la probabilidad de que un individuo, en una población definida, se beneficie de la aplicación de una tecnología médica a la resolución de un problema de salud determinado, bajo condiciones ideales de actuación. Se establece, habitualmente, en forma experimental y tiene validez universal. La efectividad también pretende medir la probabilidad de que un individuo, en una población definida, se beneficie de la aplicación de una tecnología médica, pero en condiciones reales de aplicación. Su establecimiento, por lo tanto, no tendrá validez universal. La diferencia entre eficacia y efectividad será mayor cuanto más se alejen las condiciones reales de las ideales. La eficiencia relaciona los beneficios medidos por la efectividad, con los costos que supone obtenerlos. Se trata de un concepto relativo que requerirá por tanto, formas de comparación (tanto de la efectividad, como del costo).

Una manera racional de practicar la toma de decisiones por parte del profesional de la salud puede orientarse de la siguiente manera: entre todas las posibles alternativas a su alcance para diagnosticar o tratar una enfermedad se deben elegir primero las más eficaces. Entre estas, aquellas que más beneficios médicos rindan en la práctica, en condiciones reales de aplicación, es decir, las más efectivas. Por último entre las alternativas efectivas seleccionadas deberá buscar la que rinda mayores beneficios económicos, es decir, la más eficiente.

A nivel mundial, en especial en el área de salud pública donde los recursos que se destinan son limitados, existe una creciente demanda de estudios que acrediten el costo-efectividad de un nuevo medicamento o tratamiento médico, los cuales deben incluir datos tanto de la eficiencia como de la seguridad de los mismos (Willan y Briggs, 2006).

La estadística de mayor interés en las evaluaciones económicas en salud es la Razón Costo-Efectividad Incremental, generalmente conocida por su sigla ICER (del inglés *Incremental Cost-Effectiveness Ratio*). Resulta natural intentar cuantificar la incertidumbre en la estimación de dicha razón a través de la construcción de intervalos de confianza. Sin embargo, uno de los problemas que se presenta es la necesidad de realizar supuestos acerca de la distribución muestral del estimador de la ICER sumado a complejidad de la obtención de su variancia. Para subsanar este inconveniente, gran parte de la bibliografía reciente sobre el tema se ha focalizado en el uso de diversos métodos, basados en un enfoque no paramétrico.

El objetivo principal de este trabajo consiste en presentar los fundamentos básicos del análisis de costo- efectividad, con principal énfasis en las distintas alternativas de construcción de los intervalos de confianza para la ICER.

CONCEPTOS BÁSICOS DEL ANÁLISIS DE COSTO- EFECTIVIDAD

Existen dos enfoques generales para interpretar una evaluación económica de una intervención de cuidados médicos. Un enfoque combina en un modelo de análisis de decisión los datos de eficacia y seguridad, provenientes de ensayos clínicos aleatorizados o estudios observacionales, con datos de fuentes secundarias del costo, en general no disponibles a nivel paciente en dichos estudios. El segundo enfoque se basa en datos recogidos en pacientes de forma individual y prospectivamente como parte de ensayos clínicos, en los cuales se recopila información sobre la efectividad y los recursos que conlleva el uso de

cuidados médicos. Estos datos combinados con la apropiada ponderación del precio dan una medida del costo por paciente. La medición de la efectividad y costo a nivel detallado del paciente permite el uso de métodos convencionales de inferencia estadística lo que posibilita cuantificar la incertidumbre debido al muestreo y obtener medidas del error (Willan y Briggs, 2006).

Los recursos incluidos dependen de la perspectiva del sistema de cuidados médicos utilizada: en un caso se incluyen solo aquellos cubiertos bajo el sistema, mientras que si se utiliza una perspectiva más "social", se deben incluir además los costos no cubiertos por el sistema, tales como el tiempo de trabajo perdido y cuidados por parte de los miembros de la familia.

Una componente fundamental para obtener medidas de efectividad es la medición de la calidad de vida relacionada con la salud (conocida como QALY sigla del inglés de *Quality Adjusted Life Year*). La calidad de vida es un término genérico que abarca aspectos relacionados a las capacidades físicas y mentales de un individuo, agregando componentes emocionales, sociales, económicas y circunstanciales. Desde el punto de vista práctico, se está más interesado en los aspectos directamente relacionados con la salud.

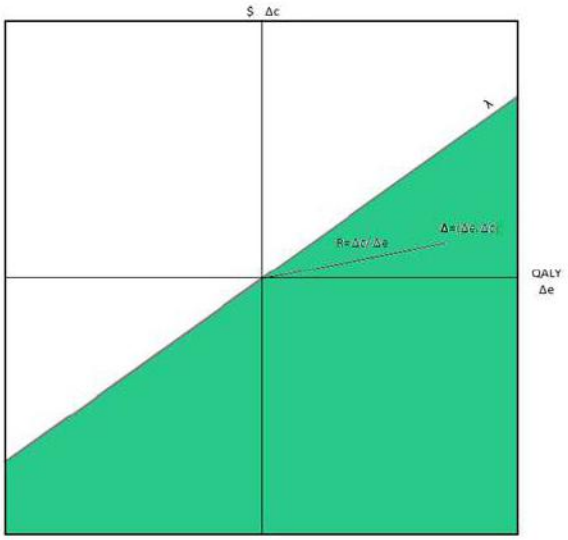
PLANO DE COSTO-EFECTIVIDAD E ICER

El plano de costo-efectividad (CE), es una herramienta gráfica ampliamente utilizada para realizar una correcta interpretación en este tipo de análisis. Este plano es una representación gráfica en la cual la abscisa y la ordenada representan las diferencias de medias de efectividad y costo entre los tratamientos, Δe y Δc respectivamente. Un punto en el plano se identifica con $\boldsymbol{\Delta} = (\Delta e, \Delta c)$.

Si para una comparación en particular de un nuevo tratamiento Alternativo (A) vs. un tratamiento Standard (S), el punto $\boldsymbol{\Delta}$ está localizado en el IV cuadrante (es decir, $\Delta e > 0$ y $\Delta c < 0$), el tratamiento Alternativo "domina" al Standard, ya que es más efectivo y más barato, y este argumento ya es suficiente para reemplazarlo. En cambio, si $\boldsymbol{\Delta}$ está localizado en el II cuadrante (es decir, $\Delta e < 0$ y $\Delta c > 0$), el tratamiento Alternativo es "dominado" por el Standard, y la elección racional es su rechazo para reemplazarlo. Los cuadrantes I y III, refieren a los cuadrantes de compensación, donde se necesitan considerar las magnitudes de Δe y Δc, para saber si el tratamiento es costo-efectivo.

Como se mencionó anteriormente, la ICER es la medida usada tradicionalmente por los investigadores para la toma de decisiones. Esta medida, se define como $R = \Delta_c/\Delta_e$, o de forma equivalente $R = \Delta_c(1/\Delta_e) = \Delta_c \times \text{NNT}$. La sigla NNT, representa el "número necesario para tratar", y se define como el número de pacientes que se estima que es necesario tratar con el nuevo tratamiento, en lugar de con el tratamiento control, para prevenir un evento. Es decir, la ICER es el producto entre el número de pacientes que necesitan ser asignados al tratamiento Alternativo para alcanzar una unidad extra de efectividad y el costo incremental de tratar cada uno de los pacientes con dicho tratamiento. Es por lo tanto, el incremento del costo para lograr una unidad de efectividad por usar el tratamiento Alternativo en vez del Standard. En el plano CE, la ICER es la pendiente de la recta que pasa por el origen y el punto $\boldsymbol{\Delta}$. Por ejemplo, si la medida de la efectividad es la probabilidad de sobrevivir, entonces la ICER es el costo de salvar una vida (o de prevenir una muerte). En cambio, si la medida de efectividad es la supervivencia media o la supervivencia ajustada por la calidad media, la ICER es el costo de lograr un año extra o un año de vida ajustado por la calidad, respectivamente. Entonces, la ICER es el costo de una unidad adicional de efectividad si el tratamiento Alternativo es adoptado sobre el Standard. Este valor por lo tanto necesita ser comparado con el monto que el tomador de decisiones esté dispuesto a pagar, denominado "voluntad a pagar" WTP (sigla del inglés *Willingness-To-Pay)*, la cual se simboliza con la letra λ. Al trazar una línea que pasa por el origen con pendiente λ, la cual se denomina Umbral, el plano CE se divide en 2 regiones (Gráfico 1).

Gráfico 1. Plano de costo-efectividad



Si un punto está debajo y a la derecha del Umbral, el tratamiento Alternativo es considerado costo-efectivo, pero si está arriba y a la izquierda no lo es. Dado que λ es siempre positivo, puntos del cuadrante IV están siempre debajo del umbral y entonces corresponden a comparaciones en el cual el tratamiento Alternativo es costo-efectivo. En cambio, puntos del cuadrante II están siempre sobre el umbral y corresponden a comparaciones en el cual el tratamiento Alternativo no es costo-efectivo. En los cuadrantes I y III el concepto de WTP permite consensuar entre efectividad y costo:

- En el cuadrante I la pendiente de cualquier punto debajo de la recta es menor que λ, si $\Delta c/\Delta e < \lambda$, lo cual que $\Delta c < \Delta e\lambda$. Entonces, el incremento en valor $(\Delta e\lambda)$ es mayor que el incremento en costo, haciendo al tratamiento Alternativo costo-efectivo.

- En el cuadrante III la pendiente de cualquier punto debajo de la recta es mayor que λ, entonces Δc y Δe son ambos negativos (es decir menos efectivo el tratamiento Alternativo y menos costoso). Siendo $\Delta c/\Delta e = |\Delta c|/|\Delta e| > \lambda$, esto implica que $|\Delta c| > |\Delta e\lambda|$. Es decir, la pérdida de valor ($|\Delta e\lambda|$) es menor que el monto salvado ($|\Delta c|$), haciendo el tratamiento Alternativo costo-efectivo.

Curva de aceptabilidad del costo-efectividad

La curva de aceptabilidad del costo-efectividad, CEAC (del inglés *Cost-Effectiveness Acceptability Curve*), se deriva de la distribución conjunta del incremento del costo y de la efectividad. La técnica más utilizada para estimar la distribución conjunta de los datos observados es el método *bootstrap*.

La CEAC se construye graficando la proporción de pares de $(\Delta e, \Delta c)$ que son costo-efectivos para un rango de valores de λ. Esta proporción es fácil de identificar del plano costo-efectividad como la proporción de puntos que se encuentran debajo del umbral que pasa por el origen y tiene pendiente λ. El proceso de construcción del CEAC comienza calculando la proporción de puntos debajo de un umbral de pendiente nula (equivalente al eje X). El proceso es repetido numerosas veces para umbrales con distintas pendientes, hasta un valor máximo de λ de infinito (equivalente al eje Y). Puntos del cuadrante II nunca serán considerados costo-efectivos y entonces tampoco serán contados. Puntos en el IV siempre serán considerados costo-efectivos y siempre serán contados. Como la pendiente del Umbral aumenta de 0 a infinito, puntos en los cuadrantes I y III pueden ser (o no) considerados costo-efectivos dependiendo de si están debajo del valor λ.

A partir de la CEAC se puede obtener la probabilidad de que el tratamiento Alternativo (A) sea costo-efectivo comparado con el tratamiento Standard (S), dada la información y para una voluntad de pago máxima aceptable λ (Fenwick & Byford, 2005).

Estimación de la ICER

El cálculo de la estimación puntual de la ICER y sus límites de confianza, se obtienen a partir de la distribución obtenida con el método *bootstrap*. En este método, se realizan remuestreos de los datos de forma aleatoria con reemplazo y se repite el proceso hasta que el conjunto de datos haya sido remuestreado B veces. Si se realizaran muestreos sin reemplazos de la muestra inicial, simplemente se obtendría esa misma muestra. En cambio al tomar muestras con reemplazo se obtiene variabilidad a través de la chance de que algunas observaciones aparezcan en la muestra remuestreada al menos una vez mientras que otras pueden no aparecer nunca.

En consecuencia, algunos pacientes están en al menos una de las muestras, mientras que otros no lo están. La validación de este método depende de dos propiedades asintóticas:

I. Dado que el tamaño muestral inicial tiende al tamaño poblacional, entonces la distribución muestral tiende a la distribución poblacional.
II. Si el número de replicaciones de la muestra original, tiende a infinito entonces la estimación por *bootstrap* de la distribución muestral de "la estadística" se aproxima a la verdadera distribución muestral.

A partir de la distribución obtenida por este método se puede calcular las estimaciones de Δ_e, Δ_c y la ICER para los datos remuestreados, denominados como Δ_{ei}^*, Δ_{ci}^* y R_i^*, $i = 1,2,\dots,B$, respectivamente. El conjunto de valores de R_i^* provee un estimador de la distribución de $\hat{R}$.

Algunos autores sugieren que la estimación por *bootstrap* 'ideal' corresponde a un número infinito de remuestreos. En la práctica, no obstante, no hay reglas formales sobre el número de replicaciones requeridas para una estimación confiable. Se sugiere que 50 replicaciones son usualmente adecuadas para proveer una estimación de la variancia y muy pocas veces más de 200 replicaciones. Para la estimación de los percentiles, se requieren mayor cantidad de replicaciones a fin de obtener estimaciones de las colas de la distribución. Generalmente se recomienda realizar entre 2000 y 5000 remuestreos, los cuales serían suficientes para lograr estabilidad para estimar los límites de confianza. Existen variantes, basadas en el enfoque de *bootstrap*, tales como la aproximación normal, de percentiles o correcciones del sesgo o tendencia.

Se describen brevemente los dos primeros métodos que son los más utilizados en la práctica.

✓ Aproximación Normal

Este método utiliza la estimación del desvío estándar obtenido a través del *bootstrap* a fin de construir el intervalo de confianza correspondiente. Para ello se asume que la distribución muestral de los $\bar{R}_i^*$ es Normal:

$$\hat{\sigma}^* = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\bar{R}_i^* - \bar{R}^*)^2} \; ; \qquad \bar{R}^* = \frac{1}{B}\sum_{i=1}^{B}\bar{R}_i^*$$

Como se mencionó anteriormente, el número de réplicas necesarias para obtener una buena estimación del desvió estándar no es elevado, por lo general con unas 50 es suficiente, y con escasa frecuencia se necesitan más de 200 (Gil Abreu, 2014).

El intervalo de confianza resultante para el $100\ (1-\alpha/2)\%$ es $\{\hat{R}^* \mp Z_{\alpha/2} \times \hat{\sigma}^*\}$. Este método, puede tener serios problemas si la distribución no es Normal. Ignora la importancia de la información obtenida de la distribución muestral, que es claramente alejada de la Normal.

Este método de aproximación no es recomendado porque los valores de R_i^* no están normalmente distribuidos y ocurren valores extremos cuando Δ_{ei}^* se aproxima a cero.

✓ Método de los percentiles

El método de los percentiles evita este problema al hacer uso de la distribución muestra empírica. Los valores de los percentiles $100(\alpha/2)$ y $100(1-\alpha/2)$ de la distribución muestral estimada por *bootstrap* se usan como los limites inferior y superior del intervalo de confianza de la ICER. El atractivo de este método reside en su simpleza y en el hecho de no requerir hacer supuestos de normalidad para la ICER.

Cada punto del plano CE representa a pares $\Delta^* = (\Delta_{ei}^*, \Delta_{ei}^*)$ de una de las re-muestras. Para construir el intervalo de confianza $100\ (1-\alpha/2)\%$ se deben encontrar dos "rayos", que pasen por el origen, que encierren $\hat{\Delta}$ y el $100\ (1-\alpha/2)\%$ de los puntos remuestreados Δ^*. El método de percentiles está basado en un ordenamiento de los puntos Δ^*. Este orden no puede basarse en valores de R_i^*, porque dos remuestreos pueden encontrarse en diferentes cuadrantes y tener el mismo valor de R_i^*, y en el caso de valores negativos tienen interpretaciones completamente diferentes. Este punto es pasado por alto en las derivaciones teóricas y aplicaciones del método de *bootstrap* en un análisis de costo-efectividad (Willan y Briggs, 2006).

Intervalo de confianza basado en el CEAC

La CEAC fue introducida como una alternativa a la elaboración de intervalos de confianza alrededor de la ICER, los cuales pueden ser estadísticamente desafiantes

Para los métodos de inferencia discutidos anteriormente, el nivel de significación se fija en un valor α. Esto no siempre resulta apropiado, por lo cual la CEAC aparece como un método más flexible y natural para expresar la incertidumbre de la estimación de los parámetros, más cercana a un pensamiento "bayesiano".

La CEAC es un gráfico que muestra la probabilidad de que un tratamiento Alternativo sea costo-efectivo siendo una función de la voluntad de pago. En términos bayesianos, es la probabilidad de que el Beneficio Neto Incremental (o conocido por su sigla, INB) para un determinado valor del WTP sea mayor que 0, $\left(P(\lambda\hat{\Delta}_e - \hat{\Delta}_c) > 0\right)$, y es la probabilidad de que el punto $\boldsymbol{\Delta}$ caiga debajo del Umbral en el plano CE, recta que pasa por el origen y tiene pendiente λ. La CEAC es la $P(\Delta_c < 0)$ para $\lambda = 0$, es la $P(\Delta_e < 0)$ cuando $\lambda \to -\infty$ y es la $P(\Delta_e > 0)$ cuando $\lambda \to +\infty$. Las probabilidades están determinadas por la distribución posterior de $\boldsymbol{\Delta}$, la cual es una distribución a priori no informativa, por lo tanto no es necesario especificar la distribución de $\widehat{\Delta}$.

Considerando a $f_{\hat{\boldsymbol{\Delta}}}$ la correspondiente función de densidad, entonces el CEAC para un valor particular de λ esta dado por

$$\mathcal{A}(\lambda) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\lambda_E} f(E,C)\, dC\, dE\ = \Phi\left(\hat{b}_\lambda/\sqrt{v_\lambda}\right)$$

Donde $\Phi(\cdot)$ es la función de distribución acumulada de una variable aleatoria Normal estándar. Se puede simplificar el cálculo de $\mathcal{A}(\lambda)$ al pensarlo como la proporción de puntos que son costo-efectivos para un cierto valor de λ. Lo cual se representa de la siguiente manera:

$$\mathcal{A}(\lambda) \equiv \sum_{i=1}^{B} I\{\lambda\Delta_{Ei}^{*} - \Delta_{Ci}^{*} > 0\}/B,$$

y el intervalo de confianza estará dado por $\mathcal{A}^{B}(\alpha/2)$ y $\mathcal{A}^{B}(1-\alpha/2)$. De esta forma se definen los límites las regiones, las cuales tienen forma de "moño" (Willan y Briggs, 2006).

La CEAC tiene dos importantes fortalezas, por un lado es una medida de ambas magnitudes y de la incertidumbre, y por otro expresa la incertidumbre en términos probabilísticos acerca de si el tratamiento Alternativo es costo-efectivo, lo cual es a menudo considerado más sencillo para quienes deben tomar decisiones.

Si bien la CEAC tiene un aspecto similar a la función de distribución de probabilidad acumulada, en la práctica una CEAC puede tomar una gran de variedad de formas, incluyendo aquellas que tengan pendiente negativa (Willan y Briggs, 2006).

APLICACIÓN A UN PROBLEMA DE SALUD POR MEDIO DE SIMULACIÓN

Se pone a prueba la metodología expuesta a partir de la simulación de dos muestras de 200 pacientes, bajo el esquema de un árbol de decisión, de una situación hipotética donde se comparan dos técnicas quirúrgicas. Los datos necesarios para la construcción de un árbol (probabilidades de transición, costos y efectividades medias) son ficticias, pero los mismos pueden obtenerse a partir de estudios previos.

A partir de los resultados obtenidos, se puede concluir que el método no paramétrico de los percentiles proporcionó buenos resultados con la ventaja de no imponer supuestos distribucionales. En el caso de realizar supuesto distribucionales de los estimadores, el que mejor desempeño tuvo es calcular la variancia del estimador a través de *bootstrap*.

CONSIDERACIONES FINALES

En las últimas décadas, los razonamientos económicos se fueron incorporando al campo de la salud, ya que sus premisas son perfectamente aplicables a lo que ocurre en los sistemas sanitarios, en especial porque los recursos son escasos. Suele ocurrir que cuanto más sana es la sociedad mayor es la demanda de asistencia médica y cuanto mayor es el progreso médico alcanzado mayor es el costo de obtener mejoras. Dado que los recursos son escasos, es necesario tomar decisiones sobre cuál es la mejor forma de gastarlos. Esto se debe a que cuando los recursos se gastan de forma determinada, se pierde la opción de usarlos de otra. Por esta razón, la incorporación de la economía en la toma de decisiones trata de asegurar que los beneficios obtenidos al seleccionar una opción sean mayores que los que se habrían obtenido con otras.

Las técnicas de evaluación usan la teoría económica para facilitar la elección de las intervenciones alternativas cuando los recursos son escasos, es decir ayudan a priorizar. El criterio que se usa, la eficiencia, constituye la base teórica de las evaluaciones económicas. En particular, en el ámbito sanitario una técnica o tratamiento es eficiente cuando se logra el máximo nivel de salud a partir de los recursos dados, es decir es costo-efectivo.

El análisis de costo-efectividad es el tipo de evaluación económica preferido en la actualidad por los tomadores de decisiones. Es frecuente la publicación de listas de tratamientos y programas sanitarios ordenados según su costo por efectividad. La idea es establecer prioridades para financiar en primer lugar, aquellos tratamientos con un menor costo por efectividad y los menos prioritarios son aquellos con un mayor costo por unidad de efectividad (Birnbaum & Greenberg, 2017).

Para saber si una intervención es eficiente, hay que considerar cuál es el resultado del estudio basal del análisis económico. Este resultado debe expresarse en términos incrementales, es decir, cuál es

el costo adicional de conseguir una unidad extra de efectividad. Además, el médico debe evaluar si la ICER es diferente en distintos grupos, como suele ocurrir. Una vez comprobado si una intervención evaluada es o no eficiente en el subgrupo de interés debe asegurarse de que los costos y los resultados sanitarios del paciente sean similares a los de los pacientes del estudio.

REFERENCIAS

Birnbaum H., Greenberg P. (2017). *Decision making in a wold of comparative effectiveness research: A practical guide*. Páginas: 41-44.

Cardozo, E. (2007). Monografía: Medicina basada en la eficiencia. Curso de auditoría médica 2007.

Combescure, Castelli, Daurès (2005). *Méthodologie statistique des évaluations médico-économiques*.

EuroQol (2017). Disponible en www.euroqol.org. Página consultada en Septiembre de 2017.

Fenwick E., Byford S. (2005). *A guide to cost-efectiveness acceptability curves*. Páginas: 106-108.

Herdman, M., Badia X., Berra S. (2001). El EuroQol-5D: Una alternativa sencilla para la medición de la calidad de vida relacionada con la salud en atención primaria. Aten. Primaria 2001 28(6):425 – 429.

Willan A., Briggs A. (2006). *Statistical Analysis of Costo-Efectiveness Data*. John Witley & Sons, Ltd. Páginas: 1-25, 43-57.

REGRESIÓN LINEAL MÚLTIPLE EN GRANDES DIMENSIONES

LIC. IVÁN MILLANES

Directora: **DRA. MARTA BEATRIZ QUAGLINO**
Codirectora: **LIC. MARÍA BELÉN ALLASIA**

En la actualidad es común encontrarse con bases de datos donde el número de variables supera ampliamente la cantidad de observaciones. En este contexto, la estimación mínimo-cuadrática de los parámetros de un modelo de regresión lineal pierde sentido, debido a que las estimaciones que se obtienen no son únicas. Por este motivo, es necesario recurrir a métodos alternativos de estimación que funcionen en este contexto de "grandes dimensiones", como las regresiones Ridge y LASSO. Estos métodos de estimación minimizan una suma de cuadrados penalizada y producen estimaciones únicas de los parámetros para cada valor de un parámetro de suavizado, dando lugar a un camino de soluciones. En esta tesina se estudian sus propiedades y se compara el desempeño de sus estimadores haciendo estudios por simulación que evidencien la bondad de predicción del modelo y características distribucionales. También se aborda el tema de implementación de los algoritmos de estimación en la regresión LASSO, caso en el cual no existen soluciones explicitas.

INTRODUCCIÓN

El análisis de regresión es una técnica estadística utilizada para investigar y modelar la relación entre variables. Esta técnica estudia la relación entre una variable respuesta o dependiente y una o más variables explicativas o predictores. Se puede usar con un fin descriptivo, es decir, para conocer la función que describe la relación entre las variables, detectando cuáles de las variables explicativas están relacionadas con la respuesta y explorando la forma e intensidad de esa relación, o bien, una vez conocida esta relación, con un fin predictivo, para conocer el valor probable de la respuesta a partir del valor conocido de los predictores.

Son numerosas las aplicaciones de la regresión, y las hay en casi cualquier campo, incluyendo ingeniería, ciencias físicas y químicas, economía, administración, ciencias biológicas y ciencias sociales. De hecho, puede ser que el análisis de regresión sea la técnica estadística más usada (Montgomery et al., 2012). En el mundo de los negocios, la predicción de oportunidades y riesgos es uno de los usos más comunes de los análisis de regresión. Por ejemplo, en los análisis de demanda se puede predecir el número de productos que una persona va a consumir en un determinado momento, basándose en características sociodemográficas, hábitos de consumo, nivel educativo, capacidad de ahorro, entre otros. Este tipo de análisis también puede ser utilizado para predecir la cantidad de clientes que van a pasar delante de un pasillo específico, con el fin de determinar el costo de un anuncio publicitario. Por otro lado, las compañías de seguro se basan en análisis de regresión para estimar la solvencia crediticia de los asegurados y el posible número de reclamos en un período de tiempo determinado.

El método clásico de estimación de los parámetros de un modelo de regresión es el de mínimos cuadrados, pero éste falla en “grandes dimensiones” porque los estimadores que se obtienen no son únicos y la interpretación de las soluciones pierde sentido. Cuando el número de variables explicativas es mayor que el de observaciones, usualmente se recurre a las regresiones Ridge y LASSO. La tesina está orientada al estudio de métodos de estimación en modelos de regresión que se adecuen al contexto de grandes dimensiones de datos, haciendo estudios comparativos por simulación que evidencien sus propiedades en cuanto a bondad de predicción y características distribucionales de los estimadores de los parámetros. También se aborda el tema de implementación de los algoritmos de estimación en la regresión LASSO, caso en el cual no existen soluciones explicitas.

ANTECEDENTES DE LOS MODELOS DE REGRESIÓN

La regresión lineal fue el primer tipo de análisis de regresión en ser estudiado con rigurosidad y utilizado ampliamente en aplicaciones prácticas (Yan and Su, 2009). En este enfoque, las relaciones se modelan usando funciones lineales en los parámetros, los cuales son constantes desconocidas en el modelo que identifican o definen la vinculación entre las variables que se piensan influyentes y la respuesta de interés. Por lo general, estos modelos se ajustan usando el método de estimación denominado mínimos cuadrados. Las estimaciones obtenidas con este método minimizan la suma de cuadrados de los residuos de todas las observaciones, los cuales se definen como la diferencia entre la respuesta observada y la respuesta esperada según el modelo por el modelo.

El método de mínimos cuadrados surgió de los campos de la astronomía y la geodesia cuando científicos y matemáticos intentaban proporcionar soluciones a los desafíos de navegar los océanos de la Tierra durante la llamada *era del descubrimiento*, la cual tuvo lugar en Europa desde finales del siglo XV hasta finales del siglo XVIII y se caracterizó por una extensa exploración de ultramar. En esa época, la descripción precisa del comportamiento de los cuerpos celestes era la clave para permitir a los buques navegar en mar abierto, donde los marineros ya no podían confiar en avistamientos de tierra para la navegación. La primera exposición clara y concisa del método de mínimos cuadrados fue publicada por Legendre en 1805. En su publicación, la técnica se describe como un procedimiento algebraico para ajustar ecuaciones lineales a los datos. Legendre utilizó el método para determinar, a partir de observaciones astronómicas, la órbita de cuerpos celestes alrededor del Sol. El valor del método de mínimos cuadrados presentado por Legendre fue reconocido inmediatamente por los principales astrónomos y geodestas de la época.

En 1809, Carl Friedrich Gauss publicó su propio método para calcular la órbita de cuerpos celestes. En su trabajo afirmó haber descubierto el método de mínimos cuadrados en 1795, lo que condujo naturalmente a una disputa de prioridad con Legendre conocida como tal en la historia de la estadística (Stigler, 1981). El desarrollo de Gauss sobre este método fue superior al de Legendre, logrando su conexión con los principios de probabilidad y distribución normal.

La estimación mínimo-cuadrática fue la primera forma de “regresión”, término introducido por Galton a finales del siglo XIX. Sir Francis Galton fue un naturalista, antropólogo, astrónomo y estadístico autodidacta. Al estudiar la altura relativa de padres e hijos, Galton observó que aquellos hijos cuyos padres tenían una estatura muy superior al promedio tendían a ser altos, pero con estaturas más cercanas al valor medio. Para aquellos hijos cuyos padres tenían una estatura muy inferior al promedio ocurría algo similar, es decir, también eran bajos pero tendían a reducir su diferencia respecto a la estatura media. A este fenómeno Galton lo llamó “regresión hacia la mediocridad", lo que en términos modernos se conoce como “regresión a la media" (Galton, 1886).

El método de mínimos cuadrados es una técnica tan popular que usualmente cuando las personas hablan de regresión lineal, en realidad se están refiriendo a la regresión mínimo-cuadrática. Esta popularidad se debe a que su aplicación es fácil de entender, provee estimaciones de parámetros que son fácilmente interpretables y su implementación en computadoras es sencilla, pudiendo resolver rápidamente problemas con cientos de predictores y cientos de miles de observaciones.

MÉTODOS DE ESTIMACIÓN ALTERNATIVOS

Sin embargo, existen dos motivos por los cuales el analista de los datos puede no estar satisfecho con los estimadores mínimo-cuadráticos. Uno de los ellos está relacionado con la precisión en la predicción, es decir, con el margen de error de los resultados obtenidos. Los estimadores mínimo-cuadráticos no tienen sesgo pero su variancia suele ser grande en comparación con otros estimadores que no necesariamente son insesgados (Tibshirani, 1996). La precisión en la predicción puede mejorarse reduciendo o fijando en cero alguno de los coeficientes. Si bien esto trae aparejado un incremento en el sesgo, también reduce la variancia de los valores predichos, y esta compensación disminuye el margen de error de los resultados obtenidos. El otro motivo tiene que ver con la interpretación y el uso de los modelos. Cuando se tiene una gran cantidad de predictores, generalmente se desea determinar un subconjunto más pequeño de variables que estén muy relacionadas con la respuesta, con el objetivo de evitar sobreajustes y entender mejor el fenómeno a explicar. Sin embargo, al aplicar mínimos cuadrados, puede ser que un gran número de predictores resulten significativos, siendo débil su relación con la respuesta. Esto puede deberse a la existencia de predictores altamente correlacionados, fenómeno que recibe el nombre de multicolinealidad. Cuando un modelo tiene muchos predictores, es más difícil de interpretar y generalmente no resulta bueno para predecir la respuesta en nuevas observaciones (Hastieet al., 2009).

REGRESIÓN EN GRANDES DIMENSIONES

En la actualidad, los grandes avances tecnológicos y la capacidad de almacenamiento creciente de los medios informáticos permite disponer de grandes bases de datos que hacen más compleja la tarea de extraer información en forma comprensible para interpretar los fenómenos investigados a través del planteo de modelos estadísticos (Nisbet et al., 2009; Han et al., 2011; Leskovec et al., 2014; Larose and Larose, 2015). Esta característica de modelos donde la cantidad de variables explicativas es muy numerosa es cada vez más frecuente. Google, redes sociales, Netflix, Mercado Libre y los supermercados son todos ejemplos de fuentes de información masiva que dan recursos para plantear modelos de interés.

El análisis de regresión en este escenario recibe el nombre de regresión en “grandes dimensiones". El método de mínimos cuadrados falla en este contexto porque los estimadores que se obtienen no son únicos. Esta falta de unicidad de los estimadores hace que la interpretación de las soluciones pierda sentido, ya que para una solución el coeficiente estimado para un predictor puede ser positivo, mientras que para otra, puede

ser negativo, es decir, el efecto de ese predictor sobre la respuesta depende de la solución elegida (Hastie et al., 2009).

A través de los años, diferentes técnicas fueron desarrolladas para mejorar la estimación mínimo-cuadrática, entre ellas: selección de un subconjunto de variables, regresiones *Ridge* y LASSO (*Least Absolute Shrinkage and Selection Operator*).

Beale et al. (1967) y Hocking and Leslie (1967) fueron los primeros autores que publicaron trabajos relacionados con los procedimientos de selección de un subconjunto de variables. Este método provee modelos interpretables, pero extremadamente inestables debido a que se trata de un proceso discreto, es decir, los predictores se retienen o se excluyen del modelo, y pequeños cambios en los datos pueden resultar en la selección de modelos muy diferentes, lo que reduce la precisión de las predicciones.

REGRESIONES RIDGE Y LASSO

La regresión *Ridge* fue presentada por Hoerl and Kennard (1970) como una alternativa a los estimadores mínimo-cuadráticos en presencia de multicolinealidad. Se trata de un proceso continuo que contrae los coeficientes, es decir, reduce su magnitud en valor absoluto, y por lo tanto es más estable. Sin embargo, no fija los coeficientes de variables muy poco asociadas con la respuesta exactamente en cero, razón por la cual no provee modelos fácilmente interpretables en presencia de muchas variables explicativas (Tibshirani, 1996).

En 1996, Tibshirani, intentando retener lo mejor de la selección de un subconjunto de variables y de la regresión *Ridge*, propuso la técnica denominada LASSO, la cual contrae algunos coeficientes y fija en cero a otros (Tibshirani, 1996).

Los métodos Ridge y LASSO permiten tratar con la presencia de multicolinealidad de las diferentes variables económicas y evitar el uso de modelos sobreajustados, los cuales resultan inestables al momento de predecir nuevas observaciones (Pereira et al., 2015).

La regresión LASSO ha sido utilizada por ejemplo, para estudiar la pobreza en América Latina, seleccionando aquellas variables que tienen mayor influencia sobre este fenómeno (Gonzalez Ferraro, 2015). Podría emplearse para predecir la quiebra de empresas, fenómeno de interés para inversores, empresas otorgadoras de préstamos y hasta gobiernos. Se utiliza también para procesar la información generada en redes sociales o mediante la navegación por internet, para conocer preferencias y así definir estrategias de marketing personalizadas, o sugerir diseños de productos o de publicidades, llevar a cabo estudios de fidelización de clientes, a través de ofertas personalizadas según los gustos y consumos del cliente, sistemas de recomendación (como utilizan Netflix o Mercado Libre). Los registros de páginas visitadas por internet, de los clics en pantalla y hasta de los movimientos típicos del cursor pueden ser utilizados para ajustar modelos capaces de detectar patrones de comportamiento, determinar costos de publicidades de acuerdo a la zona en la que se ubiquen u otras características de interés.

En este contexto de grandes dimensiones, los métodos usuales de selección de subconjuntos de variables como los denominados “selección del mejor subconjunto”, “selección paso a paso hacia adelante” o “selección paso a paso hacia atrás”, son impracticables, debido a la gran cantidad de cálculos que requieren. Por este motivo, cuando el número de variables explicativas es mayor que el número de observaciones, usualmente se recurre a las regresiones *Ridge* y LASSO.

Estas técnicas, denominadas métodos de regularización, imponen restricciones adicionales para la estimación de los parámetros y las soluciones dependen de la elección de un parámetro de suavizado, dando lugar a un camino de soluciones. De todas las soluciones calculadas para algunas variantes de este parámetro previamente establecidas, se elige la mejor de acuerdo a algún criterio, como por ejemplo, minimización del Error Cuadrático Medio, recomendándose el uso de validación cruzada.

ESTUDIO POR SIMULACIÓN

En el campo de estudio de la ciencia estadística, resulta de interés conocer el comportamiento de estas técnicas bajo diferentes escenarios que puedan presentarse en los conjuntos de datos reales con el objeto de poder anticipar el sesgo y precisión que pueda conseguirse según el método elegido, para realizar inferencias confiables. A tal fin, la tesina abarca el estudio teórico y por simulación de las propiedades de los estimadores obtenidos a partir de las regresiones Ridge y LASSO, en cuanto a bondad de predicción del modelo y características distribucionales de los estimadores de los parámetros, concluyendo que el método LASSO funciona mejor cuanto más disperso es el problema, es decir, cuanto mayor sea el conjunto de variables explicativas con respecto a la cantidad de observaciones. El estudio de simulación realizado considera tres situaciones, una donde el número de observaciones es igual al de parámetros del modelo, y dos en las cuales hay el doble y el cuádruple de parámetros. En cada caso se construyen distintos modelos variando el grado de esparcimiento o cantidad de parámetros no nulos. Los resultados muestran que la capacidad predictiva de los estimadores mínimo-cuadráticos resulta peor que la de los métodos de regularización. Si bien el ajuste mínimo-cuadrático provee estimadores insesgados tanto para los parámetros nulos como los no nulos, la variabilidad de los mismos es mucho mayor que la de las regresiones penalizadas. En todas las situaciones, al comparar los métodos de regularización, se observa que la regresión LASSO tiene mejor desempeño que Ridge en los modelos más esparcidos y que esta relación se revierte en modelos densos. Con respecto a la estimación de parámetros no nulos, los estimadores Ridge resultan sesgados, con sesgo constante para los distintos niveles de esparcimiento. En cada caso, LASSO presenta menor sesgo que Ridge en los modelos más esparcidos. Con respecto a la estimación de parámetros nulos, los estimadores de los métodos de regularización resultan insesgados en todas las situaciones. La variabilidad de los estimadores LASSO es siempre menor que la de los estimadores Ridge.

COMENTARIOS FINALES

La habilidad de LASSO para estimar con cero a parámetros nulos es muy buena en todas las situaciones, presentando mejor desempeño cuando el número de variables explicativas es mayor. En general, todos los métodos empeoran su desempeño cuando el número de parámetros no nulos es grande. A modo de recomendación final, y a partir de las propiedades estudiadas, puede enunciarse que la regresión LASSO sería preferible en lugar de Ridge en contextos de grandes dimensiones de datos, mientras que mínimos cuadrados no es una alternativa admisible en estos escenarios.

BIBLIOGRAFÍA

Beale, E., Kendall, M., and Mann, D. (1967). The discarding of variables in multivariate analysis. *Biometrika*, 54(3-4):357-366.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246-263.

Gauss, C. F. (1809). *Theoria motus corporum coelestum*. Hamburgi: Sumtibus F. Perthes et I.H. Besser.

Gonzalez Ferraro, Alejandro-Matias. (2015). Poverti in Latin America: theory and statistical application with lasso regression.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning, 2nd edition*. New York, NY: Springer.

Hocking, R. and Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4):531-540.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55-67.

Laplace, P. S. (1810). *Mémoire sur les approximations des formules qui sont fonctions de tr`es-grands nombres, et sur leur application aux probabilités*. París: Baudouin.

Larose, D. T. and Larose, C. D. (2015). *Data mining and predictive analytics*. Hoboken, NJ: John Wiley & Sons.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. París: F. Didot.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge, UK: Cambridge University Press.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis, 5th edition*. Hoboken, NJ: John Wiley & Sons.

Nisbet, R., Miner, G., and Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*. London: Academic Press.

Pereira, J. M., Basto, Mario, Ferreira da Silva, Amelia (2016). The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance. Vol 39, pages 634-641*. Elsevier.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465-474.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267-288.

Yan, X. and Su, X. (2009). *Linear regression analysis: theory and computing*. Singapore: World Scientific.

USO DE MODELOS LINEALES GENERALIZADOS PARA RESPUESTA ORDINAL EN EL ANÁLISIS DE TIEMPOS DE SUPERVIVENCIA AGRUPADOS

LIC. BRENDA NICCOLAI

Directora: **MG. LETICIA HACHUEL**

En ciertas oportunidades resulta de interés estudiar el tiempo hasta que ocurre un evento específico. Existen situaciones en las que dicho evento se observa en intervalos o en determinados puntos del tiempo a pesar de que éste pueda ocurrir en cualquier momento. En este trabajo se muestra cómo describir adecuadamente factores que afecten el tiempo hasta que se presenta el evento de interés adaptando un Modelo Lineal Generalizado (MLG) para respuesta ordinal, abarcando así dentro de un enfoque tradicional al análisis de datos de supervivencia sin ignorar ni obviar sus peculiares características.

La aplicación a datos sobre rendimiento de alumnos universitarios intenta ilustrar sobre sus alcances y limitaciones.

INTRODUCCIÓN

El objetivo de un modelo estadístico es capturar las principales características de un proceso empírico de investigación. Generalmente, el primer paso consiste en centrarse en un conjunto restringido de variables respuestas y considerar los datos como provenientes de un proceso de generación de esas variables dado un conjunto de variables explicativas o covariables.

Las variables, tanto las de respuesta como las explicativas, pueden ser clasificadas de acuerdo a sus niveles de medida. Pero en el caso de la variable respuesta resulta crucial, para formular un buen modelo estadístico, considerar también el proceso que la ha generado. Por ejemplo, una variable ordinal que mide el nivel de dolor (leve, moderado, alto) puede estar representada mediante los valores 1, 2 y 3. Otro ejemplo puede ser el tiempo, en meses, entre el diagnóstico de una enfermedad y su cura luego de un tratamiento. Los procesos que generan estos valores son de naturaleza variada y por lo tanto requieren modelos estadísticos diferentes.

Un enfoque usual es utilizar los denominados Modelos Lineales Generalizados (MLG) ya que permiten abarcar dentro de un mismo esquema metodológico un grupo importante de variables respuesta, a saber: continuas, binarias, multicategóricas y de conteo. Para variables binarias o multicategóricas, sobre todo ordinales, suelen plantearse modelos en términos de una variable latente continua no observada a la cual se le establecen umbrales que determinan intervalos dentro de los cuales realmente se efectúa la observación.

Por otro lado, suele ser de interés estudiar el tiempo hasta que se produce un evento en particular, lo que constituye el denominado análisis de datos de supervivencia o de duración. Un aspecto distintivo de estos estudios es que las duraciones son siempre no negativas y puede ocurrir que el evento en estudio no se presente antes de la finalización del período de observación, fenómeno que se conoce como censura a la derecha. Fundamentalmente estos aspectos hacen que a datos con estas características no se les puedan aplicar los MLG estándares para variable continua.

En este trabajo se muestra cómo adaptar un MLG para respuesta ordinal a fin de describir adecuadamente factores que afecten el tiempo hasta que se presente un evento. Para ello es necesario considerar que los eventos se observan en determinados puntos o que el tiempo se agrupa en intervalos aunque el evento pueda ocurrir en cualquier momento del tiempo. Este enfoque resulta relevante ya que la aplicación de los usuales modelos de supervivencia, cuando se cuenta con este tipo de datos, generalmente no se recomienda dado que ocasiona un número elevado de ligas o tiempos empatados.

El objetivo general de este trabajo es mostrar cómo analizar datos de supervivencia agrupados considerando al tiempo hasta la ocurrencia de un evento como una respuesta medida en escala ordinal, mediante el empleo de un MLG.

Para ilustrar el enfoque se aplica el mismo a datos sobre rendimiento académico de los alumnos de la cohorte de ingresantes de la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario (UNR), en el año 2008.

METODOLOGÍA

Los MLG representan un enfoque unificador de una gran cantidad de modelos, entre ellos el modelo clásico de regresión lineal con error normal y el modelo de regresión logística. La unificación de la estructura algebraica y el método de inferencia se logran considerando que la variable respuesta pertenece a la denominada familia exponencial a un parámetro (Agresti, A., 2010; Dobson, D., 2001).

Estos modelos se definen a través de tres componentes: 1) una componente aleatoria que tiene que ver con la distribución de probabilidad de la variable respuesta, 2) una componente sistemática o predictor lineal, es decir, una combinación lineal de variables explicativas y parámetros y 3) una función de enlace $g(.)$ que permite conectar el predictor lineal con la media de la variable respuesta. Combinando las tres componentes se define un determinado MLG.

Modelo lineal generalizado para respuesta multicategórica ordinal

En el caso de que la variable respuesta sea ordinal, es decir cuando hay un ordenamiento de las categorías de respuesta, se pretende que los modelos tengan términos que reflejen esa característica ordinal de la componente aleatoria.

Para ello, los modelos se expresan en términos de probabilidades acumuladas:

$$P_j = P(Y \leq j) = p_1 + \cdots + p_j \qquad j = 1,2,\dots,c$$

Sea una función $g(.)$ de enlace arbitraria, entonces un MLG se explicita:

$$g(P_j) = g[P(Y \leq j)] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x} \qquad j = 1,2,\dots,c-1$$

Este modelo se compone de $(c-1)$ ecuaciones que presentan el mismo efecto $\boldsymbol{\beta}$ acompañando a las variables explicativas. Una manera de justificarlo es a través del concepto de variable latente. Esto es, se piensa una variable Y^* continua no observada y se suponen puntos de corte en esa escala continua para definir la variable observada Y que satisface:

$$Y = j \qquad si \qquad \alpha_{j-1} < Y^* \leq \alpha_j\,,$$

donde α_j representan los puntos de corte y $j = 1,2,\dots,c$.

Para definir un MLG se aplican diferentes enlaces $g(.)$ a las probabilidades acumuladas, por ejemplo los enlaces logit o log-log del complemento.

Conceptos básicos del análisis de supervivencia

El análisis de supervivencia suele ser utilizado para examinar datos sobre el tiempo desde un origen prefijado hasta la ocurrencia de un evento en particular o de interés. Este tiempo es siempre no negativo y el evento generalmente es una categoría de una variable cualitativa, por lo que lo que se pretende medir es el tiempo hasta el cambio de categoría. En algunas ocasiones este tiempo de supervivencia se considera censurado, es decir, el evento de interés no ocurre en el tiempo de estudio o período de observación. Se denomina censura por derecha cuando el tiempo de supervivencia de un individuo es mayor que el tiempo de duración del estudio. Además, los datos de supervivencia pueden provenir de tiempos medidos en escala continua o discreta (Collet, D., 2003).

En el análisis de los datos de supervivencia existen dos funciones de importancia: la denominada función de supervivencia S_t la cual representa la probabilidad de no presentar el evento más allá del tiempo t, es decir la probabilidad de que el tiempo de supervivencia sea mayor que t y la función "hazard", h_t , que representa el riesgo de presentar el evento en el tiempo t. Otra función muy utilizada es H_t, la función de riesgo acumulada.

Un modelo usual para el análisis de datos de supervivencia es el modelo de regresión de Cox o modelo semiparamétrico de riesgos proporcionales. Éste se basa en el supuesto de que el "hazard" o riesgo de que un individuo presente el evento en un momento dado es proporcional al riesgo de otro individuo en ese mismo momento. Este modelo se denomina semiparamétrico porque no es necesario hacer ningún supuesto acerca de la distribución de probabilidad de los tiempos de supervivencia.

Análisis de datos de supervivencia agrupados mediante el uso de Modelos Lineales Generalizados

En el caso en el que se cuenta con datos agrupados de supervivencia, generalmente se produce un elevado número de empates (ligas), por lo cual utilizar el modelo de Cox no es aconsejado. Uno de los enfoques que permite abordar esta situación es considerar al tiempo de supervivencia como una variable

ordinal pero considerando que puede ser censurada a la derecha. En este caso, es apropiado el ajuste de un MLG para respuesta multicategórica ordinal, es decir un ajuste ordinal en el que cada observación consta de dos partes: el tiempo hasta el evento medido en escala ordinal y si es o no censurado.

Para ello se supone una variable aleatoria continua que indica el tiempo hasta la ocurrencia de un evento (T) pero se asume que este tiempo solamente puede tomar valores discretos positivos *t=1,2…,j…,c.*

Se define a $\boldsymbol{P_t}$ como la probabilidad de ocurrencia del evento o censura antes o en el tiempo $\boldsymbol{t}$ como:

$$P_t = P(T \leq t)$$

Se asume una transformación de las probabilidades de ocurrencia de un evento o censura antes o en el tiempo t como una función lineal de los predictores de la forma:

$$g(P_t) = \alpha_t + \boldsymbol{\beta' x}$$

Esta expresión, coincide con el argumento que considera una variable continua "tiempo" latente. Dicha variable representa el tiempo subyacente hasta el evento pero del cual se observan sólo determinados umbrales. Esto es, se supone que la variable respuesta ordinal T resulta del agrupamiento de una variable continua "tiempo" T^* usando diversos puntos de corte $\alpha_1 < \alpha_2 < \cdots < \alpha_{t-1}$; entonces se puede decir que T=1 si T^* es menor que α_1, T=2 si $\alpha_1 < T^* \leq \alpha_2$ y T=*c-1* si $T^* > \alpha_{t-1}$.

Una función de enlace usual para modelar probabilidades acumuladas de tiempo hasta el evento es la función log-log del complemento, la cual es útil cuando el tiempo es de naturaleza continua pero los sucesos son observados en intervalos discretos.

El modelo así conformado representa una versión para tiempos discretizados del modelo de riesgos proporcionales en tiempo continuo, o más sencillamente modelo de regresión de Cox:

$$log[-log(1 - P_t)] = \alpha_t + \boldsymbol{\beta' x} \qquad t = 1,2 \ldots, c-1$$

En él:

- $1 - P_t$ representa la probabilidad de supervivencia más allá del intervalo de tiempo t (función de supervivencia).
- $\boldsymbol{x}$ es un vector de $p \times 1$ covariables las cuales no varían a través del tiempo. Sin embargo, pueden representar el promedio de la variable a través del tiempo o el valor de la covariable en el tiempo de ocurrencia del evento.
- $\{\alpha_t\}$ es un conjunto de $c - 1$ constantes que representan el logaritmo del riesgo basal integrado cuando $\boldsymbol{x} = \boldsymbol{0}$.

Los parámetros α_t se relacionan con los "puntos de corte" o "umbrales" que se definen para generar la variable ordinal a partir de la variable latente.

Läärä,E. y Matthews,J. (1985) demostraron la relación que justifica la interpretación en términos de razones de "hazards" o razones de riesgo del modelo que aplica el enlace log-log del complemento a las probabilidades acumuladas.

Estimación Máximo Verosímil

El método de estimación de los parámetros del MLG para respuesta ordinal es el Máximo Verosímil, pero para el tratamiento ordinal de los tiempos de sobrevida se debe tener en cuenta las censuras por derecha, las cuales se presentan cuando al final del período de observación el individuo no presenta el evento de interés. Además se asume que las censuras de los tiempos son no informativas. Es necesario, entonces, utilizar paquetes computacionales que permitan realizar estas adaptaciones, como por ejemplo el procedimiento NLMIXED de SAS, el cual fue utilizado en este trabajo.

APLICACIÓN

Se aplica el enfoque MLG para datos de sobrevida o duración a una base de datos sobre el rendimiento de los alumnos de la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario (UNR). El plan de estudio de los ingresantes a esta facultad comienza con un Ciclo Introductorio a las carreras de Contador Público, Licenciatura en Economía y Licenciatura en Administración. La elección de la carrera se efectúa a partir de la aprobación de cuatro asignaturas de este ciclo.

Bajo estas condiciones resulta de interés analizar el tiempo que emplea el alumno hasta estar en condiciones de elegir carrera y la posible asociación de la demora en conseguirlo con variables demográficas y socioeconómicas que caracterizan a los estudiantes.

Se dispone de información relacionada con el seguimiento de los 1723 alumnos de la cohorte real ingresante en 2008 hasta marzo de 2013. Los datos considerados son obtenidos del Programa de Seguimiento de Planes de Estudios que lleva adelante la Secretaría Académica de esta Facultad. Se analizan diversas variables referidas al alumno inscripto (datos personales, antecedentes educativos, ocupación) y a la familia del alumno al momento de la inscripción.

Ajuste del MLG para los datos de supervivencia agrupados

Se define la variable respuesta como “tiempo hasta que el alumno cumple el requisito para elegir carrera”. El cumplimiento del requisito se puede dar en cualquier momento de los años académicos pero sólo se usa el registro que se realiza al momento de la reinscripción en el mes de marzo de cada año. Por lo cual, se puede suponer que existe una variable continua Y^* “tiempo hasta cumplir el requisito”, pero sólo se observa en determinados momentos, o sea la variable Y observada, toma los valores 1, 2, 3, 4 o 5 según la cantidad de años posteriores al 2008 que transcurren hasta aprobar las 4 materias del Ciclo Introductorio. De esta manera los datos se adecúan para aplicar el enfoque MLG ordinal a datos de supervivencia o duración. Se utiliza el enlace log-log del complemento y se interpretan los coeficientes del modelo en término de razones de “hazards”:

$$log[-log(1-P_t)] = \alpha_t + \boldsymbol{\beta}'\boldsymbol{x} \qquad t = 1,2 \ldots ,5$$

En este caso, dado que el cumplimiento del requisito de aprobar 4 materias es un suceso al que aspiran los estudiantes, las razones de riesgo- o razones de “hazards”- se denominan “Razones de oportunidad”.

En la Tabla 1 se presentan dichas razones de oportunidad las cuales se obtienen a partir de calcular e^{β}, donde β es el coeficiente asociado a cada variable explicativa, luego de haber obtenido un modelo de ajuste satisfactorio.

Tabla 1. Razones de oportunidad estimadas a partir del MLG

Variable		Razón de oportunidad
Sexo Femenino vs. Masculino		1,216
Año de egreso de secundario		-
Clase de escuela secundaria Privada vs. Pública		1,300
Estado ocupacional del alumno No trabaja vs. Trabaja		0,735
Estado ocupacional de la madre		-
Nivel educativo del padre Aumento unitario		1,094
Localidad Rosario vs. Otras provincias Rosario vs. Resto de Santa Fe		 0,819 0,808
Rendimiento académico promedio Aumento en dos unidades Aumento en tres unidades		 3,611 5,417
Año de egreso de secundario* Estado ocupacional de la madre	No trabaja	2007 vs. 2006 y anteriores= 1,456
	Trabaja	2007 vs. 2006 y anteriores= 2,332

A partir de los valores de la Tabla 1 se puede concluir que la oportunidad de alcanzar la condición necesaria para poder elegir carrera en forma más temprana resulta:

- el 21,6 % mayor para las mujeres que para los hombres.
- el 30% mayor para los alumnos que egresaron de una escuela secundaria privada que para los que egresaron de una pública.
- el 26,5% menor para los alumnos que no trabajan que para los que sí lo hacen.
- el 9,4% mayor a medida que aumenta en una unidad el nivel de instrucción del padre, es decir por ejemplo, al pasar de la categoría "No hizo estudios y Escuela primaria incompleta" a "Escuela primaria completa y/o Escuela secundaria incompleta".
- el 18,1% menor para los alumnos cuya localidad de procedencia es la ciudad de Rosario que para aquellos que provienen de otras provincias.
- el 19,2% menor para los alumnos cuya localidad de procedencia es la ciudad de Rosario que para aquellos que lo hacen de cualquier otra localidad de la provincia de Santa Fe.
- casi 4 veces mayor para los que presentan un rendimiento académico promedio "Medio" con respecto a uno "Bajo" y casi 6 veces mayor para los que presentan un rendimiento promedio "Alto" con respecto al "Bajo".
- si la madre no trabaja: 45,6% mayor para los alumnos que terminaron la escuela secundaria en el año 2007 que para aquellos que la finalizaron antes. En cambio, si la madre trabaja: 2 veces mayor para los alumnos que terminaron la escuela secundaria en el año 2007 que para aquellos que la finalizaron antes.

Además, utilizando el modelo estimado, se pueden calcular las probabilidades de demora en cumplir el requisito, es decir las $\hat{P}(T > t)$, para diferentes perfiles de alumnos definidos de acuerdo a sus características. A modo de ejemplo: una alumna egresada en el año 2007 de la escuela secundaria, siendo ésta de gestión privada, que trabaja y su madre también, el nivel educativo alcanzado por el padre es secundario completo o terciario incompleto, su lugar de procedencia es la ciudad de Rosario y su

rendimiento académico promedio es “Medio”, tiene una probabilidad de demorar más de un año en cumplir el requisito para elegir carrera de 0,05 y de 0,01 en demorar más de dos. En cambio, para una alumna con las mismas características pero que haya egresado de la escuela secundaria antes del año 2007, la probabilidad de demorar más de un año en cumplir el requisito se estima en 0,28, es decir se incrementa notablemente la demora.

CONSIDERACIONES FINALES

En este trabajo se presenta una forma de analizar datos de sobrevida agrupados, utilizando un modelo de regresión para una variable respuesta medida en escala ordinal que pertenece a la clase de los MLG. De esta manera se logra abarcar dentro de un enfoque más convencional al análisis de datos de sobrevida o duración sin ignorar ni obviar sus peculiares características lo cual permite, de forma sencilla, dar respuesta a interrogantes que puedan surgir de un problema particular.

En este enfoque, el tiempo de sobrevida refiere a una variable ordinal, la cual puede ser observada o bien censurada. Si se aplica el enlace log –log del complemento a la función de supervivencia se obtiene una versión para tiempos agrupados del modelo de “hazards” proporcionales para tiempo continuo o modelo de Cox.

Un enfoque alternativo, también dentro del contexto de los MLG, representa el tiempo de supervivencia asociado a cada individuo como un conjunto de variables binarias, el cual fue desarrollado en la tesina de grado de la Lic. Julia Angelini. El ajuste de modelos con ambos enfoques, bajo el enlace log-log del complemento, y el cumplimiento del supuesto de “hazards” proporcionales, brindan resultados idénticos para los parámetros de las variables que no dependen del tiempo. El enfoque ordinal presentado en esta oportunidad es conveniente de implementar cuando se tiene un gran número de observaciones dado que no modifica el tamaño de la base de datos como si lo hace el enfoque dicotómico (Hedeker,D. y Mermelstein, R., 2000).

La aplicación muestra cómo explicar la demora de los alumnos de la Facultad de Ciencias Económicas y Estadística de UNR en cumplir el requisito indispensable para poder elegir carrera en términos de factores demográficos y socioeconómicos de los estudiantes. Los resultados hallados permiten detectar subpoblaciones de alumnos cuyas características inducen a presumir que demorarán más tiempo en alcanzar el requisito necesario para elegir la carrera universitaria deseada. Esta información podría ser utilizada para diseñar estrategias educativas orientadas a estos grupos tendientes a disminuir esta demora y mejorar la eficiencia del cursado.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2010). “Analysis of Ordinal Categorical Data”– Second Edition.

Collet, D. (2003). “Modelling Survival Data in Medical Research” – Second Edition.

Dobson, A. J. (2001). “An introduction to Generalized Linear Models”- Second Edition.

Fahrmeir, L.; Tutz, G. (1994). “Multivariate Statistical Modelling Based on Generalized Linear Models”.

Hedeker,D., Mermelstein, R. (2000). “Analysis of longitudinal substance use outcomes using ordinal random-effects regression models”.

Hedeker, D., Siddiqui,O., B Hu,F. (2000). “Random-effects regression analysis of correlates grouped-time survival data”.

Läärä,E. y Matthews,J. (1985). “The equivalence of two models for ordinal data”.

McCullagh, P. (1980). “Regression models for ordinal data”.

Schmid, M., Tutz,G. (2015). “Modelling discrete time to event data”.

Skrondal, A., Rabe-Hesketch, S. (2004). “Generalized latent variable modelling”.