Número 12

Universidad Mayor de San Andrés

# Varianza

Revista del Instituto de Estadística Teórica y Aplicada

*Estadística: Carrera acreditada por CUB - R. II/2014*

Varianza 2015

IETA

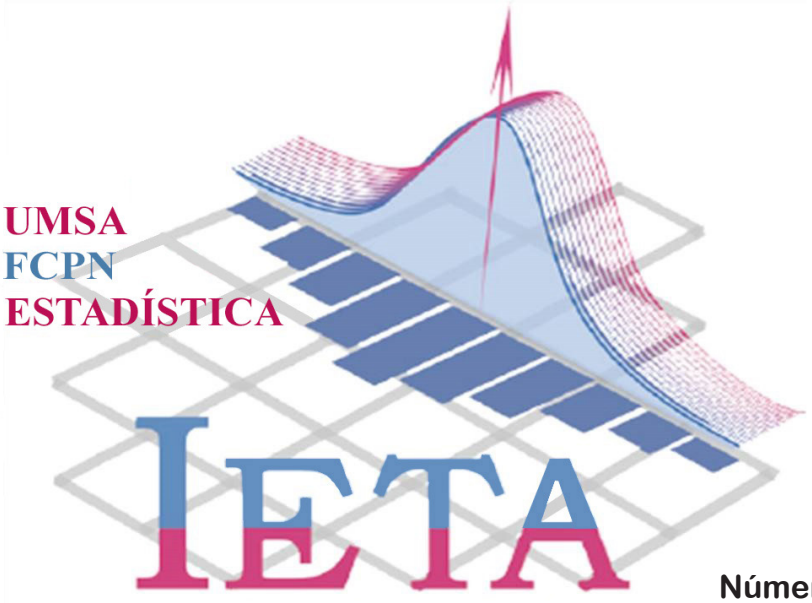
CARRERA DE ESTADISTICA
UMSA - FCPN

*Número 12*

# Varianza

## Revista de la Carrera de Estadística

*Publicación del Instituto de Estadística Teórica y Aplicada*




Instituto de Estadística
Teórica y Aplicada

Número 12

Septiembre, 2016

**ISSN 9876-6789**
**REVISTA VARIANZA**
Nº 12 - Septiembre, 2016



## PRESENTACIÓN

Las anteriores publicaciones de la revista VARIANZA, órgano oficial de difusión del Instituto de Estadística Teórica y Aplicada y por ende de la Carrera de Estadística, mostraban tanto el quehacer científico, como la difusión de actividades de la Carrera. La presente edición está orientada a mostrar la investigación realizada por nuestros docentes e investigadores asociados dando a nuestra revista un carácter científico.

La doceava publicación, muestra a la comunidad, no solo de la Carrera, sino Facultativa y de la UMSA; la madurez alcanzada en la publicación de artículos científicos de la ciencia Estadística.

Es, sin embargo el lector quien con sus críticas y sugerencias, hará posible la mejora constante de este material que se pone a consideración.

Lic. Raúl L. Delgado Álvarez
**DIRECTOR CARRERA DE ESTADÍSTICA**

Carrera de Estadística
Instituto de Estadística Teórica y Aplicada (I.E.T.A.)
Facultad de Ciencias Puras y Naturales
Universidad Mayor de San Andrés

La Paz - Bolivia
Edificio Antiguo - Planta Baja
Telefax: 2442100 -2612844
Correos:estadistica@umsa.bo - ieta@umsa.bo

*Dedicado a los Egresados*
*de la Carrera de Estadística,*
*futuros profesionales*

# Contenido

**Medición de la disparidad salarial en las regiones del departamento de La Paz**
***Autor**: Abalos Choque Melisa* ..... 1

**Tabla de vida de cohorte. Inferencia Estadística**
***Autor**: Coa Clemente Ramiro*..... 11

**Cálculo de las primas tipo pooling y primas diferenciadas en el sector de seguros de salud y la probabilidad de ocurrencia, a partir de la encuesta a hogares 2009**
***Autor**: Crespo Chuquimia Mercy*..... 17

**2 censos, 10 años de Encuestas a Hogares en Bolivia tiempo de reponderar y definir un diseño muestral comparable para los indicadores de bienestar**
***Autor**: Chirino Gutiérrez Álvaro*..... 25

**Análisis con Modelos Multinivel**
***Autor**: Flores López Juan Carlos*..... 34

**Modelación Poisson con enfoque Bayesiano para explicar el efecto de la educación y del área de residencia sobre la mortalidad durante los primeros años de vida**
***Autor**: Loza Cruz Patricia*..... 40

**Psychometric behaviour of the strengths and difficulties questionnaire (SDQ) in the Spanish national health survey 2006**
***Autor**: Pereyra Zamora Pamela et.al*..... 47

**Análisis Factorial**
***Autor**: Ruiz Aranibar Gustavo*..... 57

**Estimadores Robustos de Tendencia Central**
***Autor**: Valdez Blanco Dindo*..... 71

**Sobre la riqueza y la pobreza de una nación en términos de productividad "Una sencilla explicación estadística"**
***Autor**: Vargas Salazar F. Rodrigo*..... 75

# MEDICIÓN DE LA DISPARIDAD SALARIAL EN LAS REGIONES DEL DEPARTAMENTO DE LA PAZ

M. Sc. Abalos Choque, Melisa
✉ melisa.abalos@gmail.com

## RESUMEN

Las disparidades salariales han sido objeto de estudio en países europeos y americanos con el fin de aplicar políticas que permitan reducir dichas diferencias. El presente artículo desarrolla teóricamente los métodos de Propensity Score Matching (PSM) y Coarsened Exact Maching (CEM) a fin de mostrar al investigador una alternativa para medir el impacto en investigaciones de tipo social. En tal sentido, se utiliza la teoría de inferencia causal para medir la diferencia salarial entre la población que reside en la región Metropolitana versus las demás regiones del departamento de La Paz mediante el método CEM con datos de la encuesta Sociodemográfica 2010-2012. Estos resultados serán útiles para futuras investigaciones que busquen determinar las causas teóricas de los resultados aquí encontrados.

## PALABRAS CLAVE

*Disparidad, Inferencia Causal, CEM*

## 1. INTRODUCCIÓN

La literatura ha distinguido la existencia de diferencias espaciales de salario en países como; Chile, España y México, entre otros. Las disparidades salariales podrían deberse a la composición de habilidades de los trabajadores, dotaciones no humanas (como las características geográficas) e interacción entre trabajadores y empresas (Combes, Duranton & Gobillon, 2007).

Para el caso del departamento de La Paz, aún no existe evidencia de estudios realizados para medir las diferencias de salario en espacios geográficos definidos. Medir la existencia o no de diferencias espaciales es un reto que podría ser afrontado con técnicas estadísticas.

La inferencia causal es una rama de la Estadística que estudia las técnicas de estimación de una variable de interés que puede variar gracias a la aplicación de un tratamiento específico y no a otros factores. Éstas técnicas podrían aplicarse mediante datos experimentales y no experimentales.

Los diseños experimentales son métodos estadísticos que nos permiten comparar dos grupos a fin de medir la causalidad de una variable específica. Estos métodos aseguran la equivalencia inicial la cual implica que los dos grupos son similares entre sí al momento de iniciarse el experimento. Sin embargo, en muchos casos, los diseños experimentales demandan costos elevados que no podrían ser asumidos por el investigador. Winship & Morgan 1999, sugieren utilizar datos observables de censos, encuestas, o registros administrativos para medir la causalidad cuando es imposible realizar experimentos aleatorios.

Los diseños transeccionales, que son parte de los diseños no experimentales, nos permiten obtener información en un momento dado. La debilidad de utilizar éste tipo de información sin un criterio previo es que las diferencias salariales podrían deberse a aspectos propios del individuo y no así al tratamiento.

En tal sentido, se debe garantizar la homogeneidad de los individuos estudiados antes de realizar las mediciones en la variable de interés. Rubin, Rosenbaum y Heckman, entre otros conceptualizan la utilización de un modelo contrafactual para resolver el problema de la heterogeneidad existente entre el grupo de tratamiento y el de control. Así los métodos contrafactuales se conciben como una solución factible para el fenómeno de no comparabilidad.

Un modelo contrafactual consiste en medir la diferencia de la variable de interés con un individuo que esté tanto en el grupo de tratamiento como en el de control al mismo tiempo, lo que es físicamente imposible. Sin embargo, podemos hacer esta comparación entre dos individuos, del grupo de tratamiento y del grupo de control que tengan características similares. En este contexto se utilizan los métodos matching los cuales consideran éstos aspectos.

Los métodos llamados matching o emparejamiento consisten en parear el resultado de la variable de interés del individuo tratado con el resultado de otro individuo perteneciente al grupo de control mismo que comparte características similares o cercanas al individuo tratado. El presente artículo desarrolla teóricamente los métodos de Propensity Score Matching (PSM) y Coarsened Exact Maching (CEM). Por otro lado, mide la diferencia salarial existente entre la población que reside en la región Metropolitana versus las demás regiones del departamento de La Paz mediante el CEM el cual garantiza un matching exacto. Estos resultados serán útiles para futuras investigaciones que busquen determinar las causas teóricas de los resultados aquí encontrados.

La sección 2 del artículo muestra la teoría de la inferencia causal incluyendo los métodos matching a desarrollar PSM y CEM. La sección 3 desarrolla las virtudes de la información utilizada. La sección 4 muestra los resultados obtenidos y finalmente la sección 5 presenta las conclusiones.

## 2. INFERENCIA CAUSAL

Los métodos de inferencia causal estudian como estimar el efecto en una variable de interés que se debe exclusivamente al tratamiento y no a otros factores. Entonces, se puede definir a la inferencia causal como una rama de la Estadística que estudia la obtención de efectos causales los cuales podrían realizarse mediante datos experimentales y no experimentales.

Los datos experimentales se obtienen mediante la realización de experimentos aleatorios. Un experimento aleatorio es controlado por el investigador, el cual asigna, de manera aleatoria, a los individuos a ser estudiados, al grupo de tratamiento o a un grupo de control[1]. La asignación aleatoria garantiza la equivalencia inicial que deben tener los dos grupos, o sea, que los grupos son similares al inicio del experimento. Sin embargo, la realización de experimentos aleatorios en muchos casos implica altos costos.

La estimación de efectos causales mediante datos observables, tales como, encuestas, censos y registros administrativos, es utilizada como una alternativa cuantitativa a la imposibilidad de diseñar experimentos aleatorios (Winship & Morgan, 1999). En tal sentido, consideremos que $y_i$ es el valor del salario para el individuo i.

1 El grupo de control esta formado por los individuos que no han recibido el tratamiento.

Se generan dos resultados potenciales, el primero en ausencia del tratamiento al individuo *i* denotado por $y_{0i}$ y el segundo en presencia del tratamiento para el individuo *i* denotado por $y_{1i}$. Los resultados potenciales se presentan en situaciones contrafactuales ya que $y_{1i}$ representa el salario hipotético que hubiera obtenido el individuo *i* (que no ha sido sometido al tratamiento) si hubiera sido sometido al tratamiento.

El efecto causal del tratamiento para el individuo *i* se define como la diferencia entre ambos resultados potenciales dado en (1).

$$\Delta = y_{1i} - y_{0i} \qquad (1)$$

En la realidad, no es posible observar ambos resultados potenciales, sino que únicamente se observamos el resultado $y_i$ que se puede expresar en (2), donde $D_i$ es una variable dummy que indica si el individuo *i* ha recibido el tratamiento o no.

$$y_i = D_i y_{1i} + (1 - D_i) y_{0i} \qquad (2)$$

En caso de que existiera homogeneidad[2] entre todos los individuos incluidos en el estudio, el problema podría estar resuelto. Sin embargo, generalmente existe alto grado de heterogeneidad en las características de los individuos y de sus respuestas. Por lo que no es garantizado que la diferencia refleje el efecto del tratamiento. Estudios realizados en esta área muestran que se puede calcular efectos individuales de tratamiento mediante el llamado efecto medio del tratamiento (ATE) y el efecto medio del tratamiento seleccionado (SATE).

La homogeneidad entre los grupos se garantiza mediante un vector *X* de variables observables, tal que el tratamiento es independiente de los resultados potenciales condicionados en dichas variables. En tal sentido, por independencia condicional entre el tratamiento y los resultados potenciales, la esperanza matemática (E[ ]) de la diferencia del salario dado el vector X está dado por (3).

$$\begin{aligned} E\left[y_1 - \frac{y_0}{X}\right] &= E\left[y_1 - \frac{y_0}{X}, D = 1\right] \\ &= E\left[\frac{Y}{X}, D = 1\right] \\ &= 1 - E\left[\frac{Y}{X}, D = 0\right] \end{aligned} \qquad (3)$$

Los métodos llamados matching o emparejamiento son métodos no paramétricos que han sido estudiados recientemente y consisten en emparejar el resultado de cada individuo tratado con el resultado de otro individuo que no pertenece al grupo de tratamiento, pero que comparte características similares o cercanas al individuo tratado, dichas características se pueden incluir en el vector *X*. En el presente artículo se exponen dos estrategias econométricas relacionadas con los métodos matching los cuales son: Coarsened Exact Maching (CEM) y Propensity Score Matching (PSM) los cuales garantizan la comparabilidad entre grupos (Enriquez & Paredes, 2014).

### *2.1 PROPENSITY SCORE MATCHING (PSM)*

El Propensity Score Matching, *e(x)*, es la probabilidad condicional de recibir tratamiento (*D=1*) dado un vector de variables observables *X*, es decir, *e(x)=P(D=1/X)*. El propensity Score generalmente es desconocido, sin embargo, se puede estimar mediante una regresión logística binaria logit o probit (Winship & Morgan, 1999). El modelo logit asociado esta dado en (4), donde $D_i$ es la condición de tratamiento que toma un valor 1 si el individuo está en el grupo de tratamiento y 0 si el individuo está en el grupo de control para cada individuo *i=1,2,...,N,*

2 Esto es lo que se pretende en un experimento aleatorio.

el vector de variables observables es $X_i$ y el vector de regresión de parámetros es $\beta_i$.

$$P\left(\frac{D_i}{X_i} = x_i\right) = \frac{e^{x_i\beta_i}}{1 + e^{x_i\beta_i}} = \frac{1}{1 + e^{-x_i\beta_i}} \quad (4)$$

Aunque $D_i$ no es una función lineal de $X_i$ ésta es una variable transformada a través de una función Logit. La condición es que el tratamiento es independiente a los resultados potenciales, el cual se logra condicionando un vector $X$ de variables observables.

El proceso se puede resumir en tres pasos:

*Primero*, se estima el Propensity Score (PS) mediante el modelo de elección binaria. Se espera que la distribución conjunta entre todas las variables observables sea igual para cada grupo (tratamiento y control). Con ese propósito, los valores estimados del PS se dividen en estratos, contrastando las diferencias existentes entre los grupos. Si las diferencias no son significativas se acepta la distribución conjunta, caso contrario, se añaden potencias de orden superior de las variables observables e interacción entre las mismas hasta que se acepte la igualdad de la distribución conjunta de variables observables entre ambos grupos.

*Segundo*, las observaciones se ordenan en forma ascendente de acuerdo al valor estimado del PS. Los valores estimados extremos del PS del grupo de comparación son descartados. Con las observaciones restantes, se procede al emparejamiento de cada individuo del grupo de tratamiento con el grupo de control que tenga Propensity Score más próximo.

*Finalmente*, se calcula el efecto medio del tratamiento ATE mostrado en (5), donde $Y_i$ es el salario de los individuos pertenecientes al grupo de tratamiento y de control.

$$E[\Delta] = E\left[\frac{Y_1}{\hat{e}(x)}(x), D = 1\right] = E\left[\frac{Y_0}{\hat{e}(x)}(x), D = 0\right] \quad (5)$$

### *2.2 COARSENED EXACT MATCHING (CEM)*

CEM es parte de los métodos matching conocidos como "Monotonic Imbalance Bounding" (MIB) que muestra mejores rendimientos que los otros métodos matching (Iacus, King, Porro, 2011). El balance entre el grupo de control y tratamiento implica que la distribución conjunta de las covariantes $X$ en más similar entre los dos grupos. CEM trabaja en muestras y no requiere supuestos sobre el proceso generador de datos, es un método que mejora el balance entre el grupo de control y tratamiento, por tanto, el desequilibrio no será mayor a la que el investigador puede definir ex ante.

Para medir el imbalance se construye un histograma multidimensional de las celdas generadas por un producto cartesiano $H(X_1)x\ldots xH(X_k)=H(X)$. El imbalance multivariado se muestra en (6), donde, $f$ y $g$ son las distribuciones de frecuencia relativa para el grupo de tratamiento y control respectivamente.

$$L_1(f,g) = \frac{1}{2}\sum_{l_1\ldots l_k \epsilon H(X)} \left|f_{l_1\ldots l_k} - g_{l_1\ldots l_k}\right| \quad (6)$$

Si, $f^m$ y $g^m$ son las distribuciones de frecuencia relativa para emparejar el grupo de tratamiento y control respectivamente. Entonces, un buen método matching cumple que $\mathcal{L}_1(f^m, g^m) \leq \mathcal{L}_1(f,g)$ . Si $\mathcal{L}_1 = 1$ se dice que las dos distribuciones de datos son completamente diferentes y si $\mathcal{L}_1 = 0$, las distribuciones son exactamente iguales. Si por ejemplo $\mathcal{L}_1 = 0{,}7$, solamente el 30% de la densidad de los dos histogramas se solapan.

El algoritmo CEM puede resumirse en tres pasos. *Primero*, se recodifica cada variable de tal manera que los valores indistinguibles se agrupan y se asigna el mismo valor numérico (coarsar). *Segundo*, el algoritmo CEM crea un set de estratos, sϵS, cada uno con el mismo valor coarsado de $X$ y con por lo menos una unidad de tratamiento y una unidad de control. Se aplica el algoritmo matching exacto a los datos coarsados para determinar los emparejamientos. *Tercero*, los estratos solamente con unidades de control o tratamiento se eliminan y se toman en cuenta solamente los estratos que tienen al menos una unidad de tratamiento y una unidad de control. El algoritmo CEM asigna los pesos mostrados en (7), donde, $T^s$ son las unidades tratadas en el estrato $s$, $m_T^s$ es la cantidad de unidades tratadas en el estrato $s$, $C^S$ son las unidades de control en el estrato s, $m_C^s$ es la cantidad de unidades de control en el estrato $s$, $m_T, m_C$ son el número de unidades emparejadas en grupo de tratamiento y control respectivamente. Los valores no emparejados reciben un peso igual a cero, o sea, $w_i=0$, por tanto son eliminados. En tal sentido, CEM impone un matching exacto entre individuos evitando problemas con imposiciones arbitrarias de medidas de distancia.

$$w_i = \begin{cases} 1, & i \in T^s \\ \dfrac{m_C}{m_T}\dfrac{m_T^s}{m_C^s}, & i \in C^s \end{cases} \qquad (7)$$

El efecto medio de tratamiento de la muestra (SATT) y de la población (PATT) se presentan en (8), donde, $TE_i=Y_i(1)-Y_i(0)$, es el efecto de tratamiento de la unidad $i$, $n_T$ es la cantidad de elementos en el set de índices de unidades tratadas $T$ en la muestra, $T^*$ es el set de índices de unidades tratadas de la población y $N_T$ es la cantidad de elementos en $T^*$. *SATT* es un estimador insesgado de *PATT* de tal forma que $E(SATT)=PATT$.

$$SATT = \frac{1}{n_T}\sum_{i\in T} TE_i \ ,$$

**Tabla1**
**Departamento de La Paz: Distribución de municipios por región**

| REGIÓN | MUNICIPIO | | | |
|---|---|---|---|---|
| Amazonía | Ixiamas | San Buenaventura | Caranavi | Alto Beni |
| | Guanay | Tipuani | Mapiri | Teoponte |
| | Apolo | | | |
| Metropolitana | Viacha | Laja | La Paz | Palca |
| | Mecapaca | Achocalla | El Alto | |
| Yungas | Coroico | Coripata | Chulumani | Irupana |
| | Yanacachi | Palos Blancos | La Asunta | |
| Altiplano Sur | Sica | Umala | Ayo Ayo* | Calamarca |
| | Patacamaya | Colquencha | Collana | Papel Pampa* |
| | S.P. de Curahuara* | Santiago de Machaca* | San Andrés de Machaca* | Jesús de Machaca |
| | Chacarilla* | Catacora* | Coro Coro | Caquiaviri |
| | Calacoto | Comanche | Charaña* | Waldo Ballivián* |
| | Nazacara de Pacajes | Santiago de Callapa | | |
| Altiplano Norte | Puerto Acosta | Puerto Carabuco | Humanata* | Escoma |
| | Guaqui | Tiahuanacu | Desaguadero | Taraco |
| | Pucarani | Batallas | Puerto Pérez | Copacabana |
| | San Pedro de Tiquina | Tito Yupanqui | Achacachi | Ancoraimes |
| | Huatajata | Huarina | Santiago de Huata | Chua Cocani |
| Valles Norte | Charazani | Curva | Mocomoco | Pelechuco |
| | Sorata | Tacacoma | Quiabaya | Combaya |
| | Chuma | Ayata | Aucapata | |
| Valles Sur | Inquisivi | Quime | Cajuata | Colquiri |
| | Ichoca | Villa Libertad Licoma | Luribay | Sapahaqui |
| | Yaco | Malla | Cairoma | |

Fuente: Plan de Desarrollo del Departamento Autónomo de La Paz al 2020
*Municipios que no fueron encuestados
Elaboración Propia

$$PATT = \frac{1}{N_T} \sum_{i \in T^*} TE_i \qquad (8)$$

## 3. DATOS

El presente artículo utiliza los datos de la encuesta sociodemográfica realizada por la Universidad Mayor de San Andrés y el Gobierno Autónomo Departamental de La Paz entre las gestiones 2010 y 2012.

El objetivo de la encuesta fue conformar una línea base de información estadística que permita realizar un diagnóstico de la situación socio – demográfica y productiva para el planteamiento de políticas y la planificación de los municipios de las 7 regiones del Departamento de La Paz.

Según el Plan de Desarrollo del Departamento Autónomo de La Paz al 2020, el departamento de La Paz tiene 7 regiones de planificación las cuales son; Altiplano Norte, Altiplano Sur, Yungas, Amazonía, Valles Norte, Valles Sur y Metropolitana. La Tabla 1 muestra la conformación de los municipios incluidos en cada una de las regiones.

La información contenida en la Base de datos de la encuesta Sociodemográfica es representativa para hallar resultados de los municipios, provincias y regiones del Departamento de La Paz, en tal sentido proporciona estimadores robustos a los niveles mencionados. La muestra es de 33.011 personas del departamento de La Paz que forman parte de la población económicamente activa[3] y que declaran tener algún salario. De la muestra total del departamento de La Paz, 4.833 corresponden a la región Altiplano Sur, 5.341 a Altiplano Norte, 4.295 a Yungas, 4.154 a Amazonía, 3.657 a Valles Norte, 3.932 a Valles Sur y 6.799 a la región Metropolitana. El total de población incluida en la base de datos de la encuesta es de 103.054 personas.

## 4. RESULTADOS

Las disparidades salariales se remontan a diferencias en la composición de habilidades de la fuerza laboral (Combes, Duranton & Gobillon, 2008). En tal sentido, el vector de variables observables X está conformada por variables cuantitativas y cualitativas que suponemos que inciden en las variables salario las cuales son; edad, años promedio de escolaridad, años de experiencia[4] , genero, estado civil, área, relación de parentesco y categoría ocupacional. El grupo de tratamiento está conformado por los individuos que viven en la región Metropolitana y los grupos de control son los individuos que viven en otras regiones del departamento de La Paz.

La Tabla 2 muestra la diferencia de la media y la desviación estándar entre la región Metropolitana y el resto de las regiones, de las variables utilizadas para el emparejamiento. A primera vista el promedio del salario[5] es mayor en la región metropolitana al resto de las regiones. Las variables cuantitativas presentan diferencias aparentemente significativas. Las variables Dummy relacionadas con el género, estado civil, relación de parentesco con el jefe de hogar y la categoría ocupacional presentan, de la misma manera, muestran diferencias

3 Se considera a la Población Económicamente Activa a las personas mayores a 10 años que declararon trabajar al menos una hora la semana pasada a la encuesta y aquellos que no trabajaron por licencia o que atendieron o ayudaron en cultivos agrícolas o en algún negocio familiar o buscaron trabajo.

4 La experiencia (Exp) es igual a la edad (Edad) menos los años promedio de escolaridad (Esc) menos seis (Exp=Edad-Esc-6)..

5 Se eliminaron valores atípicos encontrados por debajo del percentil 1 y por encima del percentil 99.

**Tabla 2**
**Departamento de La Paz: Media y Desviación estándar de las variables para emparejamiento por región**

| Variable | Metropolitana | | | | Resto Regiones | | | |
|---|---|---|---|---|---|---|---|---|
| | Media/ proporción | Std. Dev. | Min | Max | Media/ proporción | Std. Dev. | Min | Max |
| Salario | 2.399,44 | 1.843,55 | 170 | 10.500 | 1.477 | 1.719 | 17,32 | 21.217 |
| Edad | 39,79 | 13,54 | 12 | 93 | 42 | 16 | 10 | 98 |
| Escolaridad | 10,97 | 4,64 | 0 | 25 | 7 | 5 | 0 | 25 |
| Experiencia | 23,25 | 15,69 | 0 | 93 | 29 | 19 | 0 | 93 |
| Hombre | 56,1% | 0,496 | 0 | 1 | 62,3% | 0,485 | 0 | 1 |
| Mujer | 43,9% | 0,496 | 0 | 1 | 37,7% | 0,485 | 0 | 1 |
| Soltero | 24,7% | 0,432 | 0 | 1 | 17,4% | 0,379 | 0 | 1 |
| Casado | 56,1% | 0,496 | 0 | 1 | 59,4% | 0,491 | 0 | 1 |
| Jefe de Hogar | 50,5% | 0,500 | 0 | 1 | 55,4% | 0,497 | 0 | 1 |
| Esposa(o) | 22,2% | 0,416 | 0 | 1 | 22,8% | 0,420 | 0 | 1 |
| Hijo | 22,1% | 0,415 | 0 | 1 | 18,3% | 0,387 | 0 | 1 |
| Obrero/Empleado | 46,6% | 0,499 | 0 | 1 | 20,7% | 0,405 | 0 | 1 |
| Cuentapropista | 50,3% | 0,500 | 0 | 1 | 75,6% | 0,429 | 0 | 1 |
| Patrón Socio o Empleador | 1,5% | 0,123 | 0 | 1 | 2,2% | 0,148 | 0 | 1 |

Fuente: Elaboración propia con datos de la encuesta sociodemográfica (2010 - 2012)

entre ambos contextos geográficos. Las diferencias apreciadas en las variables para el emparejamiento son evidencia preliminar de que la diferencia del salario podría deberse a otros factores asociados a la productividad (Enriquez & Paredes, 2014) y no así a la condición de vivir en una región específica.

Las diferencias en media pueden ser encontradas con un análisis de regresión simple. La Tabla 3 muestra los resultados de la regresión lineal simple utilizando como variable dependiente al salario y como variable independiente a la variable dummy (Reg 7) que toma el valor 1 si la persona pertenece a

**Tabla 3**
**Diferencia de salario de la población entre la región metropolitana vs otras regiones sin emparejamiento**

| Source | SS | df | MS |
|---|---|---|---|
| Model | 1.1673e+09 | 1 | 1.1673e+09 |
| Residual | 6.4425e+11 | 33011 | 19516225.8 |
| Total | 6.4542e+11 | 33012 | 19550995.9 |

| | |
|---|---|
| Number of obs | = 33013 |
| F( 1, 33011) | = 59.81 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0018 |
| Adj R-squared | = 0.0018 |
| Root MSE | = 4417.7 |

| salario_mes | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Reg7 | 465.0006 | 60.12447 | 7.73 | 0.000 | 347.1545 | 582.8467 |
| _cons | 1894.962 | 27.28546 | 69.45 | 0.000 | 1841.481 | 1948.442 |

Fuente: Elaboración propia con datos de la encuesta sociodemográfica (2010 - 2012)

la región metropolitana y 0 si pertenece a otra región. Los resultados indican que el vivir en la región metropolitana tiene una prima positiva de 465 Bs a diferencia de vivir en otras regiones. Sin embargo, estos resultados no toman en cuenta las variables propias de cada individuo que podrían estar afectando la variable salario. El presente artículo utiliza un método de emparejamiento que considera a individuos con características similares con el propósito de eliminar las diferencias entre covariantes.

**Tabla 4**
**Cálculo del imbalance antes y después del emparejamiento**

| Covariantes | Antes de emparejar | | Después de emparejar | |
|---|---|---|---|---|
| | L1 | Diferencia en Media/ proporción | L1 | Diferencia en Media/ proporción |
| Edad | 0,08205 | -1,31760 | 0,03018 | -0,01409 |
| Escolaridad | 0,26994 | 2,49870 | 0,03925 | 0,0024 |
| Experiencia | 0,10223 | -3,81630 | 0,01595 | 0,01143 |
| Hombre | 0,03423 | -0,03423 | 1,4E-14 | -2E-14 |
| Mujer | 0,03423 | 0,03423 | 1,4E-14 | -7,4E-15 |
| Soltero | 0,04208 | 0,04208 | 2,2E-14 | -1,7E-14 |
| Casado | 0,01235 | 0,01235 | 2E-14 | -1,5E-14 |
| Jefe de Hogar | 0,01054 | -0,01054 | 1,9E-14 | -1,2E-14 |
| Esposa(o) | 0,00507 | -0,00507 | 3E-14 | 1,7E-15 |
| Hijo | 0,01298 | 0,01298 | 3,4E-14 | -5,3E-14 |
| Obrero/Empleado | 0,17113 | 0,17113 | 2,9E-14 | -2,4E-15 |
| Cuentapropista | 0,15518 | -0,15518 | 2,8E-14 | -1E-14 |
| Patrón Socio o Empleador | 0,01586 | -0,01586 | 3E-14 | 4,9E-17 |
| Distancia Multivariada | | **0,5573** | | **0,2576** |
| Estratos (Total-pareados) | | 9.222 | | 2.719 |

| | Metropolitana | Otra Región | Total | % |
|---|---|---|---|---|
| Total | 18.280 | 84.774 | 103.054 | 100% |
| Pareados | 17.080 | 70.070 | 87.150 | 85% |
| No pareados | 1.200 | 14.704 | 15.904 | 15% |

Fuente: Elaboración propia con datos de la encuesta sociodemográfica (2010 - 2012)

El estadístico $\mathcal{L}_1$ mide el imbalance entre la distribución conjunta de las covariantes entre el grupo de control y tratamiento. La Tabla 4 muestra el cálculo de los imbalances antes y después del emparejamiento, el valor de la distancia multivariada reduce 0,56 a 0,26 lo que muestra que un buen emparejamiento produce una reducción en el sesgo de la diferencia de medias. Por otro lado, los imbalances univariados reducen drásticamente después del emparejamiento. También se aprecia que el total de individuos pareados es de 87.150 que corresponde al 85% de la muestra de la población total. Una vez realizado el emparejamiento, utilizamos los pesos generados por el CEM para medir el SATT con los valores pareados. La tabla 5 muestra la diferencia de salario entre el

**Tabla 5**
**Diferencia de salario de la población entre la región metropolitana vs otras regiones con emparejamiento**

| Source | SS | df | MS |
|---|---|---|---|
| Model | 389821150 | 1 | 389821150 |
| Residual | 3.1226e+11 | 29171 | 10704334 |
| Total | 3.1265e+11 | 29172 | 10717329.9 |

| | |
|---|---|
| Number of obs | = 29173 |
| F( 1, 29171) | = 36.42 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0012 |
| Adj R-squared | = 0.0012 |
| Root MSE | = 3271.7 |

| salario_mes | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| Reg7 | 282.7978 | 46.8618 | 6.03 | 0.000 | 190.9465 | 374.649 |
| _cons | 2032.655 | 21.57889 | 94.20 | 0.000 | 1990.36 | 2074.951 |

Fuente: Elaboración propia con datos de la encuesta sociodemográfica (2010 - 2012)

grupo de tratamiento y control después del emparejamiento. El SATT es de Bs. 283, mismo que es significativo, lo cual expresa que el individuo que vive en la región metropolitana gana más que un individuo que vive en otras regiones del departamento de La Paz.

## 5. CONCLUSIONES

La inferencia causal estudia la estimación de efectos causales de una variable de interés. La medición de efectos causales podría realizarse mediante datos experimentales y no experimentales. La estimación de efectos causales mediante datos observables, tales como, encuestas, censos y registros administrativos, es utilizada como una alternativa cuantitativa a la imposibilidad de diseñar experimentos aleatorios (Winship & Morgan, 1999). Los métodos matching son una alternativa para medir efectos causales mediante datos observables.

El presente artículo aplica el método de Coarsened Exact Matching (CEM) que son parte de los métodos matching para medir la diferencia del salario entre los habitantes de la Región Metropolitana (Tratamiento) versus los habitantes de otras regiones (Control) del departamento de La Paz. Los resultados descritos en la sección 4 muestran que el CEM es un método de emparejamiento que en primer lugar disminuye el desbalance entre covariantes propias del individuo, lo que asegura una homogeneidad entre grupos, mostrando una mejora en el estadístico $L_1$. Lo cual implica que utilizar este método puede reducir los sesgos de la diferencia de medias.

En segundo lugar, nos permite utilizar los datos de la encuesta Socio demográfica para medir la diferencia de medias que no sean afectadas por variables particulares de cada individuo y que pueden influir en la variable salario por tanto se usa datos transeccionales y no se depende de un experimento aleatorio. En tercer lugar, se verifica que existe una diferencia en el salario de la población que vive en la región Metropolitana versus otras regiones. El vivir en la región metropolitana, tiene un premiun de Bs. 283 más que vivir en otras regiones.

La Región Metropolitana contiene a las ciudades de La Paz y El Alto, que según el Censo Nacional de Población y Vivienda 2012 juntos tienen una población de 1.605.725 habitantes y representa 59% de la población total del departamento. Esta aglomeración de los habitantes podría estar generando economías a escala, que podrían explicar el premio de vivir en la región metropolitana, hecho que puede ser estudiado posteriormente mediante las teorías de economías de aglomeración.

## BIBLIOGRAFÍA

*Plan de Desarrollo del Departamento Autónomo de La Paz al 2020*.

Blackwell, Iacus, King & Porro (2010), "*Coarsened Exact Matching in Stata"*.

Castro (2005), "*Salarios y desigualdad territorial en las áreas urbanas de México, 1992-2002"*, Tesis de doctorado en economía, Universidad Autónomo de Barcelona.

Combes, Duranton & Gobillon, (2008), "*Spatial wage disparities: Sorting matters*", Journal of Urban Economics 63, pp. 723-742.

Enriguez & Paredes (2014), "*Migración Interna y Diferenciales de Ingreso: Evidencia para Bogotá (Colombia) a Partir de Métodos de Emparejamiento*", Revista de economía & Administración, Vol. 11, N°1, pp. 65-83.

García, I (2009), "*Metodología y diseño de estudios para la evaluación de políticas públicas"*, Barcelona – España.

Garza & Quintana (2014), "*Determinantes de la desigualdad salarial en las regiones de México 2005-2010, una visión alternativa a la teoría del capital humano*", paradigma económico, año 6, núm. 1, pp. 33-48.

Guo & Fraser (), "A*dvanced Quantitative Techniques in the Social Sciences Series"*, Propensity Score Marching, pp. 127-208.

Iacus, King & Porro, 2011, "*Causal Inference without Balance Checking: Coarsened Exact Matching"*, Oxford University Press on behalf of the society for political methodology.

Iacus, King & Porro, 2011, "*Multivariate Matching Methods That Are Monotonic Imbalance Bounding"*, Journal of the American Statistical Association, Vol. 106, No. 493, Theory and Methods, pp. 345-361

Khandker, Koolwal & Samad, 2010, "*Handbook on Impact Evaluation-Quantitative Methods and Practices*", The International Bank for Reconstruction and Development / The World Bank, Washington DC.

King, Nielsen, Coberley, Pope & Wells, (2011), "*Comparative effectiveness of Marching Methods for causal inference*".

Larraz & Pavia (2014), "*Concentración salarial en España, un análisis espacial*", International Conference On Regional Sciences.

Rosembaum & Rubin (1985), "*Constructing a Control Group Using Multivariate Marched Sampling Methods That Incorporate de Propensity Score*", American Stadistical Association, Vol. 39, N°1, pp. 33-38.

Winship & Morgan (1999), "*The estimation of causal effects from observational data"*, Rev. Social 25, pp. 659-706.

Winship & Morgan (2007), "*Conterfactuals and Causal Inference – Methods and Principles for Social Research"*, Cambridge University Press, United States of America.

Índice

# TABLA DE VIDA DE COHORTE. INFERENCIA ESTADÍSTICA

Dr. (c) Coa, Ramiro
✉ clementecoa@gmail.com

## RESUMEN

Tablas de vida para cohortes de personas son raramente elaboradas en el país debido principalmente a limitaciones de la información. Las encuestas de demografía y salud, sin embargo, son probablemente la única fuente valiosa de información en nuestro medio para este propósito. El presente trabajo tiene como propósito construir tablas de vida de cohorte para los primeros doce años de vida, incorporando elementos de inferencia estadística. Para el tramo de edad de 0 a 12 años, los niños bolivianos viven en promedio 10.6 años de vida. En áreas urbanas este promedio alcanza a 11.0 años mientras en áreas rurales se reduce a 10.1 años. Estadísticamente, las diferencias de mortalidad entre las cohortes de niños urbanos y rurales están en el primer año de vida. No se registran diferencias estadísticamente significativas en el intervalo de edad de 1 a 12 años. En consecuencia, se recomienda que las políticas de salud, nutricionales y otras, orientadas a reducir la mortalidad en los primeros años de vida estén focalizadas principalmente en áreas rurales del país y en el tramo de edad de 0 a 1 año.

## PALABRAS CLAVE

*Tabla de vida de cohorte, Tasa de mortalidad infantil, Esperanza de vida.*

## 1. INTRODUCCIÓN

La tabla de vida proporciona varios indicadores importantes del estado de salud de la población, como las probabilidades de muerte y las esperanzas de vida a distintas edades. Su utilidad, sin embargo, no solo se limita al campo de la salud, también tiene importantes aplicaciones en áreas de la educación, empleo y migración, entre otros temas.

Cuando la información con la que se construyen las tablas de vida proviene de encuestas por muestreo, estos indicadores normalmente son construidos sin hacer referencia a su precisión. Actualmente, sin embargo, se han desarrollado metodologías bastante robustas para hacer inferencia estadística sobre las funciones de una tabla de vida. El propósito de este trabajo está relacionado con este aspecto.

## 2. OBJETIVO

Construir una tabla de vida de cohorte para el país y las áreas de residencia urbana y rural, para los primeros doce años de vida, incorporando elementos de inferencia estadística.

La fuente de información adecuada en el país para elaborar una tabla de vida de cohorte para los primeros 12 años de vida es la Encuesta Nacional de Demografía y Salud. Para el presente caso se usa la del año 2008 (ENDSA 2008). La citada encuesta incluye un módulo relacionado con la historia de nacimientos. Es esta historia que permite construir la tabla para los primeros doce años de vida.

Para la construcción de la tabla de vida de cohorte se consideró los nacimientos ocurridos en el año 1996, grupo que en el año 2008 tendría alrededor de los 12 años.

## 3. METODOLOGÍA

Una tabla de vida de cohorte describe la experiencia de mortalidad de un grupo de personas desde el nacimiento de la primera persona hasta la muerte del último miembro del grupo.

En lo que sigue, la edad $x$ se trata como una variable continua. La tasa de mortalidad instantánea, también conocida como la tasa de riesgo o fuerza de mortalidad a la edad $x$, es denotada por $h(x)$ y es definida como

$$h(x) = \lim_{\Delta x \to 0} \frac{l(x) - l(x+\Delta x)}{l(x)\Delta x}$$

o equivalentemente

$$h(x) = -\frac{l'(x)}{l(x)} \qquad (1)$$

donde $l(x)$ es la función de sobrevivencia a la edad exacta $x$. Tratando la relación (1) como una ecuación diferencial se obtiene

$$l(x) = l(0)e^{-\int_0^x h(x)dx},$$

donde $l(0)$ es la raíz de una tabla de vida.

Si $l(0)=1$, la función densidad de probabilidad de morir a la edad exacta $x$ está dada por:

$$f(x) = h(x)l(x)$$

A partir de esta función de densidad se puede obtener una serie de funciones importantes. Por ejemplo, la probabilidad condicional de que una persona de la cohorte muera en el intervalo de edad $[x, x+n)$ dado que llegó con vida hasta la edad $x$, representada por ${}_nq_x$, queda expresada como

$${}_nq_x = \frac{\int_x^{x+n} f(x)dx}{\int_x^{\infty} f(x)dx} = \frac{l(x) - l(x+n)}{l(x)}.$$

De manera complementaria, la probabilidad condicional de que la persona de la cohorte viva hasta la edad $x+n$ dado que llegó con vida hasta la edad $x$ se la obtiene como

$${}_np_x = 1 - {}_nq_x = \frac{l(x+n)}{l(x)} = \frac{e^{-\int_0^{x+n} h(x)dx}}{e^{-\int_0^{x} h(x)dx}}$$

Ahora, si la variable aleatoria $X$ representa la edad a la que va a morir una persona de la cohorte, su valor esperado condicional, condicionado a que la persona va a morir después de alcanzar la edad $x$, es dado por

$$E(X \Big/ X \geq x) = \frac{\int_x^{\infty} xf(x)dx}{\int_x^{\infty} f(x)dx} = \frac{\int_x^{\infty} x[-l'(x)]dx}{\int_x^{\infty}[-l'(x)]dx} = x + \frac{T(x)}{l(x)}, \qquad (2)$$

donde $T(x)$ representa el tiempo de vida que queda a partir de la edad $x$ y se la expresa como

$$T(x) = \int_x^{\infty} l(x)dx.$$

De la relación (2) se tiene que la esperanza de vida a la edad $x$, $e_x$, queda definida como

$$e_x = \frac{T(x)}{l(x)}.$$

Ahora, si la edad discretizada toma los valores $i=0,1,\ldots,w$, siendo $w$ la edad máxima, la probabilidad de sobrevivencia se la puede estimar como

$$\hat{p}_i = \frac{l_{i+1}}{l_i}, \qquad (3)$$

cuya varianza está dada por

$$S^2_{\hat{p}_i} = S^2_{\hat{q}_i} = \frac{1}{l_i}\hat{p}_i\hat{q}_i \qquad (4)$$

donde $\hat{q}_i = 1 - \hat{p}_i$

La probabilidad de que una persona de edad *i* llegue con vida a la edad *j* es un indicador importante tanto en demografía como en el área de la salud. El estimador de la probabilidad de estar vivo en el intervalo de edad *(i,j)* se lo puede expresar como

$$\hat{p}_{ij} = \hat{p}_i \hat{p}_{i+1} \cdots \hat{p}_{j-1}$$

$$\hat{p}_{ij} = (1-\hat{q}_i)(1-\hat{q}_{i+1}) \ldots (1-\hat{q}_{j-1}) \quad (5)$$

puesto que para que una persona de la cohorte sobreviva al intervalo *(i,j)* también debe sobrevivir cada intervalo intermedio. Remplazando (3) en (5) se tiene

$$\hat{p}_{ij} = \frac{l_{i+1}}{l_i} * \frac{l_{i+2}}{l_{i+1}} * \ldots * \frac{l_j}{l_{j-1}}$$

$$= \frac{l_j}{l_i}, \quad i < j; \;\; i,j = 0,1,\ldots,w.$$

Esta relación permite derivar la varianza muestral de $\hat{p}_{ij}$ como

$$S^2_{\hat{p}_{ij}} = \frac{1}{l_i}\hat{p}_{ij}(1-\hat{p}_{ij})$$

Por otra parte, la varianza muestral de la esperanza de vida estimada a la edad α, $\hat{e}_\alpha$, se la puede obtener a partir del hecho de que la $\hat{e}_\alpha$ es la media muestral de futuros tiempos de vida $l_\alpha$. Esto es,

$$\hat{e}_\alpha = \overline{Y}_\alpha.$$

Usando este hecho se tiene que la varianza muestral de la esperanza de vida está dada por:

$$S^2_{\hat{e}_\alpha} = \frac{1}{l^2_\alpha}\sum_{i=\alpha}^{w}[(i-\alpha+in_i)-\hat{e}_\alpha]^2\, d_i$$

donde $d_i$ representa el número de muertes a la edad *i*.

## 4. RESULTADOS

Para la construcción de las tablas de vida de cohorte se usó la información de la ENDSA 2008. La cohorte estudiada comprende los nacimientos ocurridos en el año 1996, grupo que en el año 2008 tiene alrededor de los 12 años.

En el Cuadro 1 se exhiben tanto las probabilidades de muerte como las esperanzas de vida a determinadas edades de los niños. Estos mismos resultados son expuestos en los *Gráficos 1.1 y 1.2.*

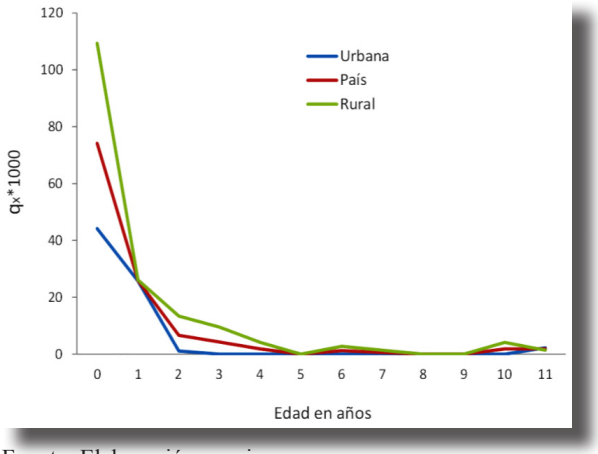
**Cuadro 1**

**Bolivia: Probabilidades de morir y esperanzas de vida a determinadas edades para la cohorte de nacimientos de 1996, por área de residencia, a partir de la ENDSA de 2008**

| Edad (años) | Probabilidades de morir ($q_x$) | | | Esperanzas de vida ($e_x$) | | |
|---|---|---|---|---|---|---|
| x | Urbana | Rural | País | Urbana | Rural | País |
| 0 | 0.044 | 0.109 | 0.074 | 11.0 | 10.1 | 10.6 |
| 1 | 0.026 | 0.026 | 0.026 | 10.5 | 10.3 | 10.4 |
| 2 | 0.001 | 0.013 | 0.007 | 9.7 | 9.6 | 9.7 |
| 3 | 0.000 | 0.009 | 0.004 | 8.8 | 8.7 | 8.7 |
| 4 | 0.000 | 0.004 | 0.002 | 7.8 | 7.8 | 7.8 |
| 5 | 0.000 | 0.000 | 0.000 | 6.8 | 6.8 | 6.8 |
| 6 | 0.000 | 0.003 | 0.001 | 5.8 | 5.8 | 5.8 |
| 7 | 0.000 | 0.001 | 0.001 | 4.8 | 4.8 | 4.8 |
| 8 | 0.000 | 0.000 | 0.000 | 3.7 | 3.8 | 3.8 |
| 9 | 0.000 | 0.000 | 0.000 | 2.7 | 2.8 | 2.8 |
| 10 | 0.000 | 0.004 | 0.002 | 1.8 | 1.8 | 1.8 |
| 11 | 0.002 | 0.001 | 0.002 | 0.8 | 0.8 | 0.8 |

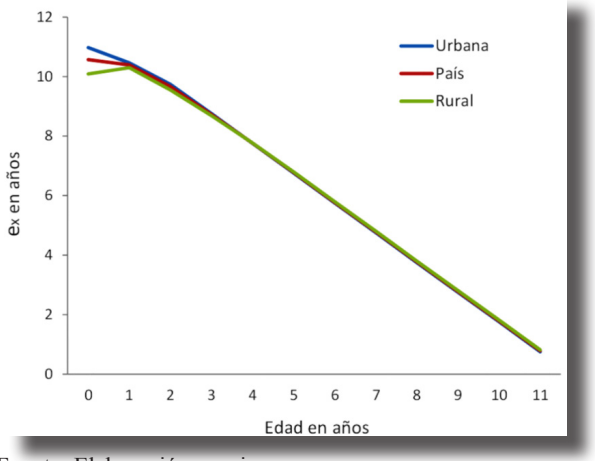
Fuente: Elaboración propia

**Gráfico 1.1**
**Bolivia: Probabilidades de morir (por mil) a determinadas edades para la cohorte de nacimientos de 1996, por área de residencia, a partir de la ENDSA de 2008**




Fuente: Elaboración propia

**Gráfico 1.2**
**Bolivia: Esperanzas de vida a determinadas edades para la cohorte de nacimientos de 1996, por área de residencia, a partir de la ENDSA de 2008**




Fuente: Elaboración propia

Una tabla de vida de cohorte permite estudiar la experiencia de mortalidad, en el tiempo, de una determinada generación de personas. De los resultados expuestos en el Cuadro 1, la probabilidad de morir en el primer año de vida para los nacidos en 1996 es 74 por cada mil nacidos vivos. Este valor es coherente con la estimación de la tasa de mortalidad infantil para el periodo 1993-1998, obtenida con otro método y publicada oficialmente por el Ministerio de Salud y el INE (76 por mil nacidos vivos). A partir del segundo año de vida la probabilidad de morir de la cohorte desciende fuertemente y de manera sistemática. También puede observarse que las principales diferencias en los niveles de mortalidad entre áreas urbanas y rurales se reflejan principalmente en el primer año de vida. En efecto, la probabilidad de morir en el primer año de vida para los nacidos en 1996 en el área rural es más del doble que en el área urbana, 109 y 44 por mil, respectivamente.

En materia de salud, la esperanza de vida es uno de los indicadores más importantes. Tomando en cuenta que el análisis abarca solo los primeros 12 años de vida, la cohorte de nacimientos en 1996 tiene una esperanza de vida de 10.6 años. A medida que transcurre el tiempo, la esperanza de vida de la cohorte se reduce gradualmente. Cabe notar, sin embargo, que la reducción más importante en la esperanza de vida de la cohorte se registra en el primer año de vida, esto como resultado de la más alta probabilidad de muerte en este tramo de edad. Por otra parte, la principal diferencia entre las esperanzas de vida de la cohorte urbana y la cohorte rural se marca en el primer año de vida. Una esperanza de vida de 11 años para la cohorte urbana frente a 10.1 años para la cohorte rural. A partir del segundo año de vida las esperanzas de vida de ambas cohortes son similares, lo que confirma que las diferencias de mortalidad entre ambas áreas residen básicamente en el primer año de vida.

El Cuadro 2 exhibe los límites de confianza para las esperanzas de vida de ambas cohortes, urbana y rural, considerando un nivel de confianza de 95%; mientras el Gráfico 2 muestra estos mismos límites, donde las líneas de color rojo representan los limites inferior y superior para las esperanzas de la cohorte urbana y las de color verde para la cohorte rural.

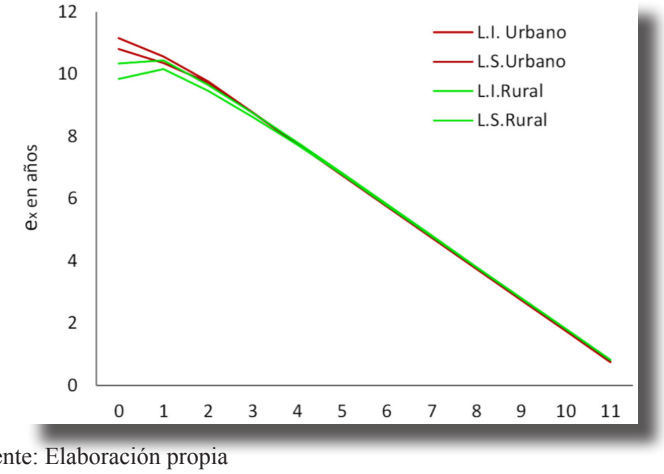
**Cuadro 2**
**Bolivia: Límites de confianza para las esperanzas de vida de las cohortes urbana y rural, a partir de la ENDSA 2008**

| Edad (años) | Cohorte Urbana | | Cohorte Rural | |
|---|---|---|---|---|
| | L.I. Urbano | L.S.Urbano | L.I.Rural | L.S.Rural |
| 0 | 10.8 | 11.2 | 9.8 | 10.3 |
| 1 | 10.4 | 10.6 | 10.2 | 10.4 |
| 2 | 9.7 | 9.8 | 9.5 | 9.6 |
| 3 | 8.7 | 8.8 | 8.6 | 8.7 |
| 4 | 7.7 | 7.8 | 7.7 | 7.8 |
| 5 | 6.7 | 6.8 | 6.8 | 6.8 |
| 6 | 5.7 | 5.8 | 5.8 | 5.8 |
| 7 | 4.7 | 4.8 | 4.8 | 4.8 |
| 8 | 3.7 | 3.8 | 3.8 | 3.8 |
| 9 | 2.7 | 2.8 | 2.8 | 2.8 |
| 10 | 1.7 | 1.8 | 1.8 | 1.8 |
| 11 | 0.7 | 0.8 | 0.8 | 0.8 |

Fuente: Elaboración propia

A partir de este análisis inferencial se constata que las diferencias de mortalidad entre ambas cohortes, urbana y rural, residen básicamente en el primer año de vida. En efecto, la esperanza de vida al nacer para la cohorte urbana es significativamente superior a la de la cohorte rural; en cambio, no hay evidencia estadística de que las esperanzas de vida de ambas cohortes, después del primer año de vida, sean distintas. Por ejemplo, se puede afirmar con 95% de seguridad que la esperanza de vida al nacer para la cohorte urbana está entre 10.8 y 11.2 años, intervalo de confianza superior al de la corte rural, el cual fluctúa entre 9.8 y 10.3 años. En consecuencia, se recomienda que las políticas de salud y otras políticas relacionadas con la salud, como son la nutrición y las condiciones sanitarias, tiene que estar orientadas fundamentalmente a mejorar las condiciones de salud en el primer año de vida y focalizadas principalmente en áreas rurales del país.

**Gráfico 2**
**Bolivia: Límites de confianza para las esperanzas de vida de las cohortes urbana y rural, a partir de la ENDSA de 2008**




Fuente: Elaboración propia

## BIBLIOGRAFÍA

Brass William (1964), "*Uses of Census and Surveys data for the Estimation of vital Rtaes*"

Keyfitz Nathan (1979), "*Introducción a las Matemáticas de Población*"

Preston Samuel (1993), "*Demographic Change in the United States, 1970-2025*"

United Nations (1982), "*Model Life Tables for Developing Countries*"

Índice

# CÁLCULO DE LAS PRIMAS TIPO POOLING Y PRIMAS DIFERENCIADAS EN EL SECTOR DE SEGUROS DE SALUD Y LA PROBABILIDAD DE OCURRENCIA, A PARTIR DE LA ENCUESTA A HOGARES 2009[1]

Lic. Crespo Chuquimia, Mercy

✉ *mercy_0074@hotmail.com*

## RESUMEN

Muchas veces nos preguntamos por qué las primas de los seguros de salud son tan elevadas y además diferenciadas por grupos de edad. En este documento se muestra la metodología de cálculo de las primas en seguros de salud tipo Pooling y diferenciadas, a partir de los datos de la Encuesta a Hogares 2009, realizado por el INE de Bolivia.

## PALABRAS CLAVE

*Seguros de salud, Primas, pooling.*

## 1. INTRODUCCIÓN

Existen diversos trabajos referentes al mercado de los seguros de salud entre los que destacan: El trabajo de Rothchild y Stiglitz, donde trata el tema de la incertidumbre por parte de las aseguradoras que no saben las condiciones de salud de los individuos. Existiendo dos tipos de individuos de alto y bajo riesgo. Un equilibrio agrupador beneficiaría a los primeros, pero perjudicaría a los segundos porque se estaría subsidiando a los individuos de alto riesgo. Finalmente llegan a la conclusión de que un equilibrio de "pooling" no es posible en este modelo. En cambio, sí hay las condiciones para un equilibrio separador, aunque no necesariamente se alcance.

El trabajo de Grossman analiza un equilibrio no-Nash con "pooling". Ya que las aseguradoras pueden negarse a aceptar determinados individuos riesgosos, entonces, éstos tendrán el incentivo necesario para comportarse como si fueran de bajo riesgo. Si, por ejemplo, las aseguradoras crearan un nuevo plan más conveniente para el grupo de bajo riesgo, los individuos de alto riesgo imitarían el comportamiento del otro grupo, aunque no les fuera conveniente, porque si no serían revelados como son realmente (alto riesgo). Entonces el equilibrio "pooling" es viable.

Neudeck y Podczek consideran que puede darse un equilibrio de "pooling", separador y mixto, pero bajo ciertas condiciones. Para la existencia del equilibrio separador la condición es un gran número de individuos de alto riesgo, de modo que el "pooling" es costoso; para el equilibrio "pooling" un pequeño número de individuos de alto riesgo y conducta no-Nash. También tratan el tema de los seguros privados y públicas, el problema del descreme por parte del sector privado.

1 Se trabajó con la encuesta a hogares 2009 porque después de esa encuesta cambia la periodicidad de las preguntas con respecto a gastos en salud (del último mes al último año), de tal forma que los gastos registrados en posteriores encuestas parecen estar subestimados.

## 2. ELEMENTOS DEL MERCADO DE SEGUROS DE SALUD

En el trabajo de García (1999) se desarrolla un modelo en base a los trabajos antes señalados. El problema de la oferta o mejor dicho de las empresas aseguradoras del mencionado modelo servirá de base para la determinación de la prima en el seguro de salud en el caso boliviano.

Los principales supuestos del modelo son:

- Hay dos tipos de individuos diferenciados en función de sus probabilidades de enfermedad y/o accidente: de alto riesgo ($P_H$) y de bajo riesgo ($P_L$).
- La probabilidad de ocurrencia de la enfermedad y/o accidente es P.
- Existe una proporción conocida de la población de alto riesgo ($\mu$) y el resto de bajo riesgo ($1-\mu$), cuando $0 < \mu < 1$.
- Los individuos que contraten un seguro de salud pagarán una fracción de sus ingresos.
- La prima que paguen los individuos será $\alpha_1$ y la cobertura de $\alpha_2$.
- Se maximiza la utilidad esperada según la función VNM (Von Neuman-Morgenstern).

### *2.1 DEMANDA*

Los seguros de salud pueden ser tanto públicos como privados y dependiendo de cuál se hable existe una diferencia fundamental en la afiliación. Mientras en los seguros privados la afiliación es voluntaria para los seguros públicos generalmente (lo cual se cumple en Bolivia) son obligatorios[2] .

En el caso privado (voluntario), como se mencionó anteriormente, se demandará un seguro de salud no tanto por la incertidumbre de su salud como por la aversión al riesgo financiero del mismo. La erogación en la que incurra por la enfermedad afectará la riqueza del individuo y esto es algo que desea evitar.

Ciertamente la demanda del seguro no solo dependerá de la aversión al riesgo, sino también (y quizá en mayor medida) del valor de la prima. Una prima de seguro se puede definir como el precio o contraprestación que cobra la empresa aseguradora al asegurado por el servicio prestado, es decir, por la cobertura de riesgo en un periodo de tiempo determinado.

Siguiendo con el modelo, el individuo que demande o compre un seguro de salud lo que buscará es maximizar su utilidad esperada que está en función a la probabilidad de ocurrencia del evento (enfermedad y/o accidente) además del contrato de seguro (prima y cobertura).

### *2.2 OFERTA*

Las empresas ofrecen sus servicios basándose en la prima que cobren. En un mercado exclusivamente privado esta oferta podría crecer y crear una competencia que redujera la prima. Pero en muchos países, incluido el nuestro, existe el seguro público obligatorio que modifica la situación.

2 La afiliación obligatoria del seguro público responde más que nada a razones prácticas, mientras mayor sea el número de asegurados menor (o al menos más próximo a la media) el riesgo de los individuos. Es decir, en cierto sentido se distribuye el riesgo en la población.

Con el fin de incrementar el beneficio las empresas aseguradoras pueden recurrir a diversos mecanismos: los deducibles, copagos, topes, alianzas, entre otras.

El problema principal de las empresas es maximizar su beneficio que está sujeto a lo que reciben por concepto de prima, la cobertura que ofrecen y, principalmente, de la probabilidad de ocurrencia del evento.

La determinación de esta probabilidad (P) representa gran parte del problema en los seguros de salud ya que es en este elemento que se encuentra la asimetría.

### *2.3 LA PRIMA*

Una forma de estimar la prima de salud mostrada en el trabajo de Baeza y Copetta (1999) es mediante la fórmula:

$$P\$ = Cac + e + Cadm + Tx + U$$

Donde:

$$Cac = \sum P_e * p * q$$

$$Cadm = Cf + Cm + CC + Ci$$

| | |
|---|---|
| P$: Precio del seguro | Cf: Costo financiero |
| Cac: Costo actuarial | Cm: Costo márketing |
| Pe: Probabilidad del evento | Cc: Costo de claim |
| q: Cantidad de los servicios | Ci: costo de información |
| p*q: costo del evento | U: Utilidades<br>Tx: Impuestos<br>e: Error (varianza) |

Ciertamente el componente principal de ésta fórmula es el costo actuarial o técnico que tiene la probabilidad de ocurrencia del evento. Con la determinación del costo actuarial en cierto sentido se tiene solucionado el problema de la prima.

Otra forma de determinar la prima es bajo el modelo propuesto anteriormente, mediante la solución del problema de la empresa aseguradora. Pero en este caso puede ofrecerse un contrato agrupador (*pooling*) o un contrato diferenciado a los individuos.

Cuando se habla del contrato agrupador la fórmula para el cálculo es:

$$\max \pi_{\alpha 1,\alpha 2} = \mu[(1-P_H)\alpha_1 - P_H\alpha_2] + (1-\mu)[(1-P_L)\alpha_1 - P_L\alpha_2]$$

Y para el contrato diferenciado la fórmula matemática es:

$$\max \pi_{\alpha 1,\alpha 2} = (1-P_i)\alpha_1 - P_i\alpha_2 \ ;$$

$$\forall i \ / \ i = 1,2,\dots,8$$

Usualmente la diferenciación en los contratos no solo depende de la probabilidad de ocurrencia del evento, sino también de características de los individuos que se pueden llamar índices[3] . Los ejemplos más comunes de estos índices son la edad y el género ya que se considera que los niños y ancianos son grupos más riesgosos en comparación al resto de la población, igualmente las mujeres en edad fértil serían más riesgosas que los hombres.

## 3. EL SEGURO DE SALUD EN BOLIVIA

Para aproximarnos a la estimación de una prima en el mercado de los seguros de salud tomamos en cuenta la Encuesta de Hogares del año 2009 (EH) realizada por el Instituto Nacional de Estadística.

Un primer elemento importante de la EH respecto a los seguros de salud es la cobertura

3 Spence en su estudio sobre la señalización en el mercado de trabajo hace una distinción entre una señal y un índice. Las características observables de un individuo que pueden ser modificadas por él mismo son las señales, mientras que los atributos observables pero inalterables son los índices.

en el área urbana mostrada en la tabla 1 así como también la distribución de la población asegurada de la tabla 2.

**Tabla 1**
**Tenencia de seguro en Bolivia - Urbano**

| Seguro | Número de personas | Participación (Porcentaje) |
|---|---|---|
| Público | 2.305,131 | 33,9 |
| Privado | 237,273 | 3,5 |
| Otro | 43,675 | 0,6 |
| Ninguno | 4.216,784 | 62,0 |
| **Total** | **6.802,863** | **100.0** |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

**Tabla 2**
**Bolivia – Urbano: Tipos de Seguros de Salud**

| Seguro | Número de personas | Participación (Porcentaje) |
|---|---|---|
| Caja de salud | 1.297,378 | 19,1 |
| SUMI | 508,581 | 7.5 |
| SSPAM (seguro de salud para el adulto mayor) | 119,765 | 1,8 |
| Otros seguros públicos | 379,407 | 5,6 |
| Seguro privado | 237,273 | 3,5 |
| Otro | 43,675 | 0,6 |
| Ninguno | 4.216,784 | 62,0 |
| **Total** | **6.802,863** | **100,0** |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

Del total de la población urbana de Bolivia en 2009, solamente el 38% contaba con algún tipo de seguro de salud mientras el restante 62% no contaba con ningún seguro. Este dato muestra que existen más de 4 millones de personas que estarían de alguna forma desprotegidas en cuanto a salud se trata. Si observamos la participación en base al seguro el 19,0% está afiliado a la caja de salud que forma parte del seguro público. Vale la pena recordad que este seguro público es de carácter obligatorio para los empleados públicos.

Tratando de encontrar la razón por qué las personas no optan por un seguro de salud privado de forma espontánea (voluntaria), podemos mencionar que un factor importante sería el alto monto de las primas que cobran las empresas aseguradoras en comparación al nivel de ingresos de nuestro país.

Veamos una desagregación de los ingresos en quintiles en el área urbana de Bolivia en la tabla 3.

**Tabla 3**
**Bolivia – Urbano: Ingresos de los hogares**

| Quintiles (Bolivianos) | Porcentaje |
|---|---|
| 20 – 1.350 | 20 |
| 1.350 – 2.177 | 20 |
| 2.177 – 3.294 | 20 |
| 3.294 – 5.044 | 20 |
| 5.044 y más | 20 |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

Lo más relevante de este cuadro, es el hecho que el 60% de los hogares en 2009 tenía un ingreso por debajo de Bs. 3.294 lo que nos muestra la realidad boliviana en cuanto a la disponibilidad de pagar por un seguro de salud privado.

## 4. DETERMINACIÓN DE LA PRIMA DE SALUD

El precio de la prima de salud se puede determinar mediante la expresión matemática mostrada en el marco teórico:

$$P\$ = Cac + e + Cadm + Tx + U$$

Por otro lado, como señalamos, un criterio separador usual es aquel que considera el género y edad. Pero en el presente trabajo, por el análisis de los datos desagregados de la EH, consideramos por conveniente el

análisis solo por edad.

Para calcular la probabilidad de ocurrencia del evento consideramos la respuesta a si gastó o no en salud de la EH como un aproximado. Entonces en la tabla 4 encontramos las probabilidades de enfermedad y/o accidente que se tiene en los diferentes grupos de edad. Los datos corresponden al mes de noviembre del año 2009, que seguramente se repiten (aproximadamente) en los demás meses del año.

Bajo el supuesto de que el gasto de salud reportado en la EH fuese el gasto que haría una aseguradora o, lo que es lo mismo, el costo del evento (p*q), podemos calcular la prima de salud para cada grupo de edad. Cabe aclarar que estos valores que se muestran en la tabla 5, pueden estar subestimados principalmente si se considera el gasto promedio. Por esta razón se toma el gasto máximo.

**Tabla 4**
**Bolivia – Urbano: Gastó en salud en el último mes según edad**

| Edad | Gasto en salud en el último mes | |
|---|---|---|
| | No (Porcentaje) | Si (Porcentaje) |
| 0-4 | 80,3 | 19,7 |
| 5-9 | 92,3 | 7,7 |
| 10-25 | 96,9 | 3,6 |
| 25-39 | 92,8 | 7,2 |
| 40-54 | 89,3 | 10,7 |
| 55-59 | 83,8 | 16,2 |
| 60-64 | 79,4 | 20,6 |
| 65 y más | 77,6 | 25,4 |
| **Total** | **90,9** | **9,1** |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

**Tabla 5**
**Bolivia – Urbano: Gasto máximo y promedio por edad**

| Edad | Gasto máximo (Bolivianos) | Gasto promedio (Bolivianos) |
|---|---|---|
| 0-4 | **900** | 136,5 |
| 5-9 | 3.800 | 223,2 |
| 10-25 | 5.200 | 311,2 |
| 25-39 | 4.080 | 344,4 |
| 40-54 | 8.000 | 571,3 |
| 55-59 | 7.000 | 717,6 |
| 60-64 | 18.000 | 1.078,5 |
| 65 y más | **5.000** | 420,1 |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

En los intervalos extremos de 0 a 4 años y de 65 años y más, podemos ver que los montos gastados son bajos en comparación a la tendencia que se tiene en los otros grupos de edad y, por lo tanto, son contradictorios con la teoría. La explicación a este fenómeno es que estos intervalos cuentan con un seguro público gratuito que muy probablemente reduzca el gasto en el que tienen que incurrir estos individuos. En específico hablamos del Seguro Universal Materno Infantil (SUMI) y el Seguro de Salud Para el Adulto Mayor (SSPAM).

### *4.1 COSTO ACTUARIAL*

Tomando en cuenta el gasto máximo en salud de los individuos juntamente con la probabilidad de enfermedad, podemos aproximarnos al costo actuarial o prima actuarialmente justa para dos casos: un contrato agrupador y un contrato diferenciado, el cual estará dividido por grupos de edad. En ambos casos el punto de vista que consideramos es el de la empresa aseguradora.

a) **La empresa ofrece un contrato agrupador**

Bajo el supuesto que la empresa no puede hacer ninguna diferencia entre los individuos y dado que en realidad tenemos los ocho grupos presentados en la tabla 4 con sus respectivas probabilidades, la empresa aseguradora ofrecerá un único contrato con una prima única para todos los individuos. Por lo tanto, el problema de la empresa aseguradora es el siguiente:

$$max\ \pi_{\alpha1,\alpha2} = \mu_1[(1-P_1)\alpha_1 - P_1\alpha_2] + \mu_2[(1-P_2)\alpha_1 - P_2\alpha_2] + \cdots + \mu_8[(1-P_8)\alpha_1 - P_8\alpha_2]$$

$$max\pi_{\alpha1,\alpha2} = \sum_{i=1}^{8} \mu_i[(1-P_i)\alpha_1 - P_i\alpha_2]$$

Donde:

$\alpha_1$ : La prima que paga el individuo único para todos

$\alpha_2$ : La cobertura del seguro

$P_i$ : La probabilidad de enfermedad y/o accidente en el grupo i

$\mu_i$ : La proporción de individuos en el grupo i

Bajo el supuesto de que nos encontramos en un mercado competitivo y entonces podemos esperar un beneficio igual a cero. Entonces tenemos la siguiente condición:

$$\frac{\alpha_2}{\alpha_1} = \frac{\sum_{i=1}^{8} \mu_i(1-P_i)}{\sum_{i=1}^{8} \mu_i P_i} = \varphi_p \quad (2)$$

Donde: $\sum_{i=1}^{8} \mu_i = 1$ ; $0 < P_i < 1$

Reemplazando los datos que se tienen en la tabla 4 y 5 tenemos como resultado los datos de la tabla 6.

**Tabla 6**
**Calculo de $\varphi_p$**

| Edad | Pi | $\%\mu_i$ | $\mu_i(1-P_i)$ | $\mu_i P_i$ |
|---|---|---|---|---|
| 0-4 | 0,197 | 10,64 | 8,55 | 2,10 |
| 5-9 | 0,077 | 11,93 | 11,01 | 0,92 |
| 10-25 | 0,036 | 30,69 | 29,57 | 1,11 |
| 25-39 | 0,072 | 19,80 | 18,38 | 1,43 |
| 40-54 | 0,107 | 14,52 | 12,96 | 1,55 |
| 55-59 | 0,162 | 3,40 | 2,85 | 0,55 |
| 60-64 | 0,206 | 3,05 | 2,42 | 0,63 |
| 65 y más | 0,254 | 5,98 | 4,46 | 1,52 |
| **Total** | **0,091** | **100,00** | **90,19** | **9,81** |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

Ya que los datos son mensuales y, además, contamos con un monto mensual máximo que se registra en la tabla 5 podemos tomar una cobertura propuesta *ad hoc* como el promedio ponderado del máximo valor por proporción de la población $\mu_i$ . Una ventaja de esta estimación es que en función a la cobertura propuesta se tendrá una prima diferente.

Con los datos:

$$\varphi_p = 9{,}2$$

$$Gasto\ máximo\ ponderado = 5.200 = \alpha_2$$

La prima para el contrato agrupador será:

$$\alpha_1 = 565{,}3 \text{ Bs.}$$

Esta prima será pagada de forma mensual, de tal forma que anualmente tendremos una cobertura de Bs 62.400.

b) **La empresa ofrece contrato "Diferenciado"**

Levantamos el supuesto de que la empresa no

puede distinguir entre los individuos de alto y bajo riesgo y, dado que podemos conocer esas probabilidades, se ofrece un contrato diferenciado con una prima distinta para cada grupo de edad. Realizaremos el cálculo de las primas diferenciando solamente los grupos de edad.

El problema de la empresa aseguradora es el siguiente:

$$\max \pi_{\alpha 1,\alpha 2} = (1 - P_i)\alpha_1 - P_i\alpha_2 \quad ;$$

$$\forall i = 1,2,\dots,8$$

Donde:

$\alpha_1$ : La prima que paga el individuo único para todos.

$\alpha_2$ : La cobertura del seguro.

$P_i$ : La probabilidad de enfermedad y/o accidente en el grupo *i*.

$1 - P_i$: La probabilidad de estar sano en el grupo *i*.

$$\frac{\alpha_2}{\alpha_1} = \frac{1 - P_i}{P_i} = \varphi_{si}$$

De hecho ocurre lo siguiente:

Si $P_i < P_j$ entonces se cumple que

$$\varphi_{sj} < \varphi_{si}$$

y como consecuencia inmediata la prima en el grupo *i* será menor a la prima en el grupo *j*.

Veamos con los datos tomando en cuenta el valor máximo registrado como gasto médico por grupo de edad.

**Tabla 7**
**Cálculo de $\alpha_1$ con cobertura total para cada grupo**

| Edad | Pi | $\varphi_{si}$ | $\alpha_2$ | $\alpha_1$ |
|---|---|---|---|---|
| 0-4 | 0,20 | 4,08 | **900** | 220,80 |
| 5-9 | 0,08 | 11,99 | 3.800 | 317,01 |
| 10-25 | 0,04 | 26,52 | 5.200 | 196,06 |
| 25-39 | 0,07 | 12,89 | 4.080 | 316,55 |
| 40-54 | 0,11 | 8,35 | 8.000 | 958,57 |
| 55-59 | 0,16 | 5,17 | 7.000 | 1.353,22 |
| 60-64 | 0,21 | 3,85 | 18.000 | 4.670,03 |
| 65 y más | 0,25 | 2,94 | **5.000** | 1.702,30 |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

Las primas del grupo 0 – 4 años y 65 y más años, podrían estar subestimadas por los bajos montos registrados en la encuesta. Sin embargo, podremos realizar una mejor comparación tomando un solo monto de cobertura.

Ahora veamos las primas por grupo de edad con un solo monto de cobertura igual a Bs. 5.200 mensuales. Esto lo hacemos con el fin de que sea comparable con el contrato agrupador.

**Tabla 8**
**Calculo de $\alpha_1$ con cobertura igual a 5200 mensual**

| Edad | Pi | $\varphi_{si}$ | $\alpha_2$ | $\alpha_1$ |
|---|---|---|---|---|
| 0-4 | 0,20 | 4,08 | 5.200 | 1.276 |
| 5-9 | 0,08 | 11,99 | 5.200 | **434** |
| 10-25 | 0,04 | 26,52 | 5.200 | **196** |
| 25-39 | 0,07 | 12,89 | 5,200 | **403** |
| 40-54 | 0,11 | 8,35 | 5.200 | 623 |
| 55-59 | 0,16 | 5,17 | 5.200 | 1.005 |
| 60-64 | 0,21 | 3,85 | 5.200 | 1.349 |
| 65 y más | 0,25 | 2,94 | 5.200 | 1.771 |

Fuente: Elaboración propia con la Encuesta a Hogares 2009. INE

En base a la información mostrada podemos constatar lo señalado en la teoría. Si se ofrece un contrato agrupador los individuos de alto riesgo se beneficiarán ya que poco a poco los seguros serán demandados por estos, dejando de lado al grupo de 5 – 39 años, pues la prima

que ellos pagarían en un contrato agrupador no refleja el riego de ese grupo.

## 5. CONCLUSIONES

- El Mercado de seguros de salud no va de acorde a la realidad y nivel de ingresos de nuestro país, ya que los costos de la prima son muy elevados comparados a los ingresos de los hogares bolivianos.

- El porcentaje de la población que no tiene seguro de salud es de 62%. Dato que podríamos comparar con la informalidad del empleo en Bolivia.

- El costo de las primas de salud ofrecidas por las empresas se puede reducir y ajustar teniendo mucha más información acerca de los individuos. De esta forma las personas con riesgo bajo estarán mucho más conformes con la prima que pagará a la empresa.

- Ofrecer una prima de tipo "Pooling", tendrá como consecuencia, un colapso en el mercado, de modo que solamente individuos de riesgo alto demandarán los seguros de salud y como consecuencia final el mercado de seguros de salud puede desaparecer.

Índice

## BIBLIOGRAFÍA


Akerlof, George. *The market for "lemons": quality uncertainty and the market mechanism*. En The Quarterly Journal of Economics, vol 84. 1970.

Baeza, Cristián y Copetta, Claudia. *Análisis conceptual de la necesidad y factibilidad de introducir mecanismos de ajuste de riesgo en el contexto de portabilidad de los subsidios públicos en el sistema de seguros de salud en Chile*. CLAISS. 1999.

Ferreiro, Alejandro; Saavedra, Eduardo y Zuleta, Gustavo. *Marco conceptual para la regulación de seguros de salud*. Banco Interamericano de Desarrollo. 2004.

García Núñez, Luis. *Seguros de salud públicos y privados: el caso chileno. Pontificia Universidad Católica del Perú*. 1999.

Instituto Nacional de Estadistica, *"Encuesta a Hogares 2009"* , Base de datos disponibles en sitio oficial INE, www.ine.gov.bo

Rothschild, Michael y Stiglitz, Joseph. *Equilibrium in competitive insurance markets: an essay on the economics of imperfect information*. En The Quarterly Journal of Economics, vol 90. 1976.

Spence, Michael. *Job market signaling*. En The Quarterly Journal of Economics, vol 87. 1973.

Villar, Antonio. *Lecciones de microeconomía*. Antoni Bosch editor. 1999.

# 2 CENSOS, 10 AÑOS DE ENCUESTAS A HOGARES EN BOLIVIA TIEMPO DE REPONDERAR Y DEFINIR UN DISEÑO MUESTRAL COMPARABLE PARA LOS INDICADORES DE BIENESTAR[1]

Lic. Chirino Gutiérrez, Álvaro
✉ achirino@aru.org.bo

## RESUMEN

Este documento propone un diseño muestral para el periodo 2002 - 2012 de la serie de encuestas a hogares que llevo a cabo el Instituto Nacional de Estadística de Bolivia, el diseño se orienta en lograr un rendimiento muestral optimo en los indicadores de pobreza moderada y pobreza extrema a nivel departamental (primera unidad sub-nacional) y nacional. Existen dos principales motivaciones para realizar este nuevo diseño; el primero, los diseños muestrales existentes en las encuestas de la serie difieren en varias características, como ser: la estratificación, la conglomeración, el marco muestral, la asignación de la muestra, el tamaño de muestra, et.al. Estas variaciones no permiten una lectura consistente de la evolución de los indicadores de bienestar a lo largo de estos años. La segunda motivación, es la pertinencia, el 2012 se realizó el último Censo de Población y Vivienda, los resultados de este censo muestran una estructura diferente a las proyecciones, que fueron la base de las estimaciones de las Encuestas a hogares. Y siguiendo las recomendaciones de la Naciones Unidas (2009) sobre las oportunidades de un nuevo censo, este es un momento idóneo para sugerir un nuevo diseño que traerá consigo la reponderación de la muestra.

## PALABRAS CLAVE

*Reponderación; Bolivia; Encuestas a hogares; Indicadores de Bienestar.*

## 1. INTRODUCCIÓN

El Instituto Nacional de Estadística de Bolivia (INE), al igual que la mayoría de los países de la región, desarrolló los últimos años, encuestas a hogares (EH) orientadas a medir bienestar, esta práctica tuvo el incentivo inicial del Programa MECOVI del Banco Mundial. El 2001 Bolivia llevó a cabo el Censo Nacional de Población y Vivienda (CNPV) y como uno de los resultados naturales del censo, la información obtenida por éste, se convirtió en la base para el desarrollo de las encuestas, en noviembre de 2012 se realizó nuevamente el CNPV. En el periodo intercensal (periodo entre censos) se realizaron un total de 10 rondas de encuestas, desde el 2002 al 2012, algunas particularidades en este periodo son; el 2003 y 2004 se realizó una encuesta de tipo contínua destinada a construir una nueva canasta familiar para el IPC[2], de 2002 a 2005 se empleó una nueva muestra como marco muestral, el 2010 no se realizó la encuesta, a partir del 2011 la encuesta casi duplicó el tamaño de muestra de viviendas.

Los cuadros 1 y 2 presentan un resumen acerca del diseño muestral del periodo intercensal, se verifica que en la serie no existe una relación respecto al diseño de un año a otro, es más, casi ningún año es similar en diseño al otro, esto motiva a pensar que las estimaciones que provienen de estas encuestas no son comparables completamente. Así también, los resultados del CNPV-2012 muestran que

1 Este documento fue presentado en el ***60th World Statistics Congress*** desarrollado en Rio de Janeiro, Brasil en Julio 2015.

2 Indice de Precios al Consumidor

la estructura del país no es la misma a la cual expanden las EH, por ejemplo, la EH-2012 expande a una población de alrededor 10,874,551 habitantes, sin embargo, el CNPV-2012 registro un total de 10,059,856 habitantes, más de 800 mil habitantes de diferencia.

**Cuadro 1**
**Diseños muestrales de las Encuestas a Hogares en Bolivia 2002 a 2006**

| Año | | 2002 | 2003-2004 | 2005 | 2006 |
|---|---|---|---|---|---|
| **Marco** | | Muestra Maestra, CNPV-2001 | Muestra Maestra, CNPV-2001 | Muestra Maestra, CNPV-2001 | CNPV-2001 |
| **Etapas** | **Urbano** | 2 | 2 | 2 | 2 |
| | **Rural** | 3 | 3 | 3 | 3 |
| **Cluster** | **Etapa 1** | Grupo de sectores censales | Grupo de sectores censales | Grupo de sectores censales | Grupo de sectores censales |
| | **Etapa 2** | Segmentos | Segmentos | Segmentos | Grupo de segmentos |
| **Estratificación, 1ra. Etapa** | | Población, 5 Niveles | Población, 5 Niveles | Población, 5 Niveles | Area y NBI, 8 levels |
| **Mecanismo de selección** | **Etapa 1** | $\pi ps$, Viviendas | $\pi ps$, Viviendas | $\pi ps$, Viviendas | $\pi ps$, Viviendas |
| | **Etapa 2** | $\pi ps$, Viviendas | $\pi ps$, Viviendas | $\pi ps$, Viviendas | $\pi ps$, Viviendas |
| | **Ultima etapa** | Sistemática | Sistemática | Sistemática | Sistemática |
| **Muestra** | **UPM** | 679 | 547 | 166 | 314 |
| | **Hogares** | 5746 | 8767 | 3865 | 4006 |
| | **Personas** | 24933 | 38500 | 16895 | 16511 |

Fuente: Elaboración propia

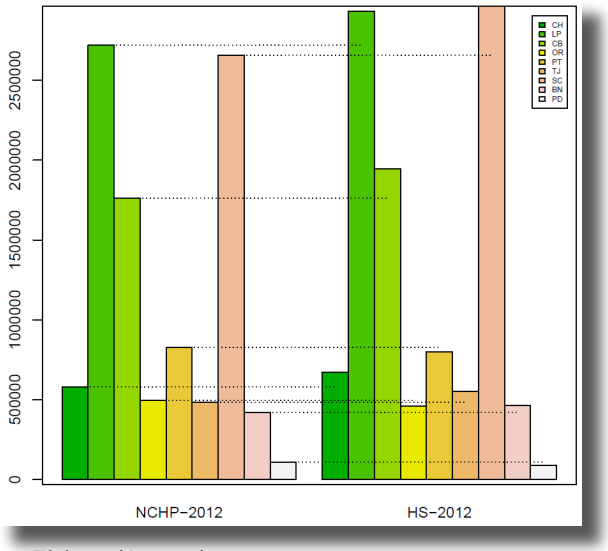
**Cuadro 2**
**Diseños muestrales de las Encuestas a Hogares en Bolivia 2007 a 2012**

| 2007 | 2008 | 2009 | 2011 | 2012 |
|---|---|---|---|---|
| CNPV-2001 | CNPV-2001 | CNPV-2001 | CNPV-2001 | CNPV-2001 |
| 2 | 2 | 2 | 2 | 2 |
| 3 | 3 | 3 | 2 | 2 |
| Grupo de sectores censales | Grupo de sectores censales | Grupo de sectores censales | sector censal | sector censal |
| Grupo de segmentos | Grupo de segmentos | Grupo de segmentos | Grupo de segmentos | Grupo de segmentos |
| Área y NBI, 8 niveles | población y consumo | Área y NBI, 8 niveles | Población y NBI, 20 niveles | Población y NBI, 20 niveles |
| Sistemática $\pi ps$, viviendas | Sistemática $\pi ps$, Viviendas | Sistemática $\pi ps$, viviendas | Sistemática $\pi ps$, viviendas | Sistemática $\pi ps$, hogares |
| Desconocido | Desconocido | Desconocido | | |
| Sistemática | Sistemática | Sistemática | Sistemática | Sistemática |
| 309 | 304 | 301 | 790 | 729 |
| 4069 | 3856 | 3932 | 8851 | 8360 |
| 16804 | 15030 | 15665 | 33821 | 31935 |

Fuente: Elaboración propia

Un resultado aún más importante que el nacional son los referidos a los departamentales, estas son las primeras unidades sub nacionales de Bolivia y son los primeros actores locales de la política pública, es por ello que es necesaria una información adecuada al nivel del departamento, la estructura reflejada por la EH-2012 a nivel departamental difiere de lo obtenido en el CNPV-2012, esta diferencia se la presenta en la Figura 1, donde se observa claramente que existen departamentos con sub y sobre cobertura.

**Figura 1**
**Población estimada de la EH-2012 vs Resultados del CNPV - 2012**




Fuente: Elaboración propia

Este documento plantea un diseño muestral expost, con el fin de poder generar estimaciones comparables y coherentes a los resultados del CNPV-2012 de toda la serie, el diseño muestral final es el resultado de un contraste en el rendimiento de muestreo de cinco propuestas alternativas de diseño y un ajuste en los factores de expansión para que se adecuen al ritmo de crecimiento del periodo intercensal 2001 a 2012, el análisis de rendimiento muestral se lo hace en base a los indicadores de pobreza moderada y extrema a nivel de los departamentos.

El documento presenta en la sección 2 la metodología empleada, en esta se describen las fuentes de datos empleadas, la definición de pobreza en las EH, el tratamiento para los componentes del muestreo como ser la estratificación y conglomeración, la construcción de las probabilidades teóricas y el factor de expansión y la proyección empleada para el periodo intercensal, en la sección 3 se muestran los resultados referentes al rendimiento de los diseños planteados, la evolución de la pobreza y el diseño muestral final. Finalmente, en la sección 4 se dan las conclusiones del trabajo.

### *1.1 OBJETIVO*

El objetivo general del trabajo es:

"Desarrollar un diseño muestral comparable para los indicadores de bienestar provenientes de la serie de encuestas a hogares 2002 a 2012, ajustado a los patrones de crecimiento poblacional de los Censos de Población y Vivienda 2001 y 2012" .

## 2. METODOLOGÍA

Con la finalidad de obtener un diseño muestral comparable para la serie de encuestas del 2002 a 2012 se optó por elaborar 5 distintos diseños muestrales y comparar el rendimiento de cada uno de ellos para los indicadores de pobreza moderada y extrema, para ello, se eligen los datos, se establece la definición de pobreza que seguirá el documento, se definen las características de conglomeración y estratificación, se realizan proyecciones para el periodo intercensal, se construyen las probabilidades teóricas y finalmente se elaboran los ponderadores muestrales ajustados a las proyecciones.

### *2.1 DATOS*

Las bases de datos y la documentación metodológica de las encuestas a hogares fueron obtenidas del banco de datos sociales del portal del INE[3], las bases de datos del Censo de Población y Vivienda 2001 que se emplea como la base del marco muestral de la serie, corresponde a la que distribuye el departamento de comunicación del INE . Finalmente, la información del CNPV-2012 fue obtenida del portal en versión Redatam del Censo[4] .

### *2.2 POBREZA*

Todas las encuestas a hogares del periodo de interés definen a la pobreza en base al ingreso y a su ubicación por encima o por debajo de la línea de pobreza, siguiendo la metodología de Foster, James, J. Greer, and Eric Thorbecke. 1984.

Donde la pobreza moderada para un individuo se define como:

$$p_i = I(y_i < z_k)$$

3 http://www.ine.gob.bo:8081/Webine10/enchogares1.aspx

4 http://datos.ine.gob.bo/binbol/RpWebEngine.exe/Portal?LANG=ESP

Donde $I(,)$ es una función indicatriz respecto el ingreso ($y_i$) del individuo $i$ y la línea de pobreza moderada ($z_k$) de la región $k$ del país.

De similar forma se define a la pobreza extrema como:

$$pe_i = I(y_i < ze_k)$$

Donde $I(,)$ es una función indicatriz respecto el ingreso ($y_i$) del individuo $i$ y la línea de pobreza extrema ($ze_k$) de la región $k$ del país.

El estudio emplea como un indicador agregado, a la incidencia de pobreza extrema y moderada, este se define dentro de un área $k$, como:

$$P_0 = \frac{\sum_k p_i}{N_k}$$

$$P_{0e} = \frac{\sum_k pe_i}{N_k}$$

Donde $P_0$ es la incidencia (headcount) de pobreza moderada y $P_{0e}$ la incidencia de pobreza extrema.

En el estudio se emplea a partir del 2006 las variables de pobreza reportadas en las bases de datos, para los años inferiores se emplea la información de Hernani and Villarroel (2012), esto a razón de la ausencia de las variables de pobreza en las bases de datos antes del 2006.

### *2.3 CLUSTER*

Durante el periodo de la serie de encuestas existieron 2 unidades primarias de muestreo (PSU), la primera estuvo desde el 2002 al 2009 y fue una agrupación de sectores censales que en promedio tenía 120 viviendas, en las encuestas a hogares de 2011 y 2012 las unidades primarias de muestreo fueron los sectores censales que en promedio tenían 90 viviendas, todas estas unidades en base al CNPV-2001, tal como se mostró en el cuadro 1, algunos años de encuesta existieron dos etapas en el área rural, sin embargo, no existe la documentación necesaria para conocer el criterio de formación y selección de estas áreas, es por ello que para fines de este trabajo se define la existencia de 2 etapas; la selección del cluster y la selección del hogar.

La PSU será la denominada UPM, esto debido a su presencia en la mayoría de las encuestas y que en las EH 2011 y 2012 se las puede identificar sin dificultad.

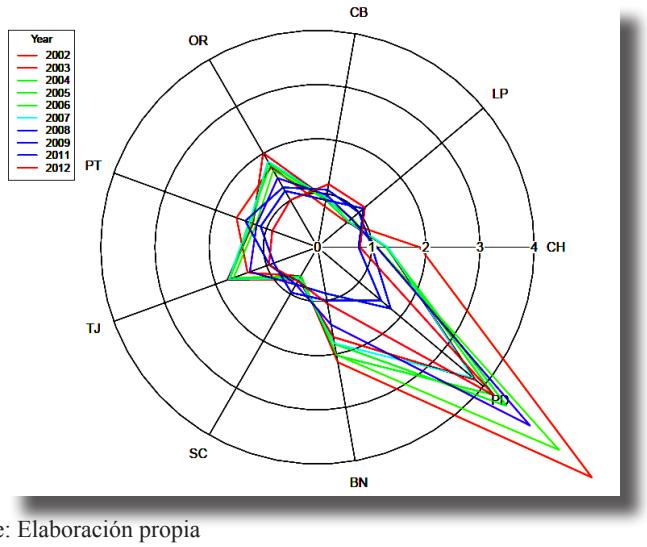
### *2.4 LA ESTRATIFICACIÓN*

La estratificación es la forma central de la cual depende la estructura de un diseño muestral complejo, a lo largo de la serie de encuestas se han empleado diversos estratos, todos los documentos metodológicos previos al 2011 definen los estratos a nivel nacional y no incluyen un tratamiento a nivel departamental, sin embargo, el analizar las bases de datos se encuentra que existió un criterio de selección basada en los departamentos, esto debido al patrón de sobre y sub representación de algunos departamentos respecto a una estructura proporcional del marco muestral, la figura 2 muestra el patrón, donde los anillos representan el cociente entre la muestra asignada y la necesaria, se observa claramente el caso del departamento de Pando, que casi siempre recibió más de 3 veces la muestra necesaria, esto pone en evidencia que existió una selección basada en los departamentos.

Para el trabajo se definen 5 tipos de estratos a nivel de las PSU, cada uno combinado con los 9 departamentos, esto a razón de la Figura 2, los estratos son:

- *Población:* 5 niveles al interior de los departamentos; ciudades capitales, ciudades intermedias, centros poblados mayores, centros poblados menores y área dispersa.
- *Geográfico:* 3 niveles; Altiplano, Valles y Llano.
- *Consumo:* 12 niveles al interior de los departamentos, fue construido en base a un modelo de regresión basado en las HS-1999 A 2002

**Figura 2**
**Sub y sobre muestreo en la serie EH 2002 - 2012**




Fuente: Elaboración propia

- *Componentes principales:* 5 niveles al interior de los departamentos, fue construido en basa a la técnica de los componentes principales empleando información del CNPV-2001.
- *NBI:* 4 niveles al interior de los departamentos, fue construido en base a la técnica del índice de necesidades básicas insatisfechas.

## 2.5 PROYECCIONES

Para las proyecciones de población en el periodo intercensal se empleó tasas de crecimiento exponencial, esta se define como:

$$r = \frac{\ln(P_{2012}) - \ln(P_{2001})}{Y_{2012-2001}}$$

Donde $P_t$ corresponde a la población censada el año $t$ y al tiempo $Y_{2012\text{-}2001}$ transcurrido

**Cuadro 3**
**Tasa de crecimiento anual y población del CNPV 2001 y 2012 por departamento y área**

| Departamento | NCPH-2001 | | | NCPH-2012 | | | Tasa de crecimiento anual (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Urbano | Rural | Total | Urbano | Rural | Total | Urbano | Rural | Departamento |
| Chuquisaca | 218126 | 313396 | 531522 | 283123 | 298224 | 581347 | 2.325 | -0.442 | 0.799 |
| La Paz | 1552146 | 797739 | 2349885 | 1814148 | 905196 | 2719344 | 1.390 | 1.126 | 1.302 |
| Cochabamba | 856409 | 599302 | 1455711 | 1200912 | 561849 | 1762761 | 3.013 | -0.575 | 1.706 |
| Oruro | 236110 | 156341 | 392451 | 316757 | 177830 | 494587 | 2.619 | 1.148 | 2.062 |
| Potosi | 239083 | 469930 | 709013 | 336412 | 491681 | 828093 | 3.044 | 0.403 | 1.384 |
| Tarija | 247736 | 143490 | 391226 | 314510 | 169008 | 483518 | 2.127 | 1.459 | 1.888 |
| Santa Cruz | 1545648 | 483823 | 2029471 | 2160579 | 497183 | 2657762 | 2.985 | 0.243 | 2.404 |
| Beni | 249152 | 113369 | 362521 | 308690 | 113318 | 422008 | 1.910 | -0.004 | 1.354 |
| Pando | 20820 | 31705 | 52525 | 53831 | 56605 | 110436 | 8.467 | 5.166 | 6.624 |
| Total | 5165230 | 3109095 | 8274325 | 6788962 | 3270894 | 10059856 | 2.436 | 0.452 | 1.742 |

Fuente: Elaboración propia

en años desde el 2001 al 2012, la tasa corresponde al 21 de Noviembre de 2012 (dia del CNPV-2012). La fórmula empleada para la proyección es:

$$\Pr_i = P_{2001} * e^{r*a}$$

Donde $P_ir$ corresponde a la proyección para el año $i$ y es el tiempo transcurrido en años desde la base 2001 al año de proyección, el cuadro 3 muestra las tasas de crecimiento anual por departamento y área.

### *2.6 PROBABILIDADES TEÓRICAS*

Con las consideraciones de los conglomerados y los estratos se elaboran las probabilidades teóricas asociadas a las dos etapas del diseño, para la primera etapa, se asume una probabilidad proporcional al tamaño sin reposición, empleando el total de vivienda de las PSU registradas en el marco muestral del CNPV-2001, este tiene la forma de:

$$\pi_{Ii} = n_{Is} * \frac{t_{is}}{\sum_S t_{is}}$$

Donde $\pi_{Ii}$ es la probabilidad de primera etapa para la PSU $i$, $n_{Is}$ el tamaño de muestra de PSU en el nivel $s$ de la estratificación y $t_{is}$ el total de viviendas en la PSU $i$ del nivel de la estratificación.

En general la probabilidad del hogar $k$, se define como el producto de la probabilidad de primera etapa y la última etapa.

$$\pi_{ki} = \pi_{Ii} * \frac{n_{IIi}}{t_{is}}$$

Donde $\pi_{ki}$ es la probabilidad para el individuo $k$ de la PSU $i$,$n_{IIi}$ el tamaño de muestra de la última etapa.

Para el análisis se eliminaron las PSU con $n_{IIi}=1$ para evitar el incremento de varianzas.

### *2.7 FACTORES DE EXPANSIÓN*

La construcción de los factores de expansión para los 5 distintos tipos de estratos se inició con el factor teórico, este es el inverso de la probabilidad de selección:

$$W_{ki} = \left(\pi_{ki} * \frac{n_{IIi}}{t_{is}}\right)^{-1}$$

Sobre este factor se incluyó un ajuste debido a las proyecciones $P_l$, este ajuste se lo hizo a nivel departamental y por área.

$$aW_{ki} = W_{ki} * \frac{P_l}{\sum_l W_{ki}}$$

Se acotó los factores finales en base al valor del percentil 99 de su distribución, esto siguiendo la recomendación de Naciones Unidas. (2009), con la finalidad de reducir la varianza de los estimadores de pobreza.

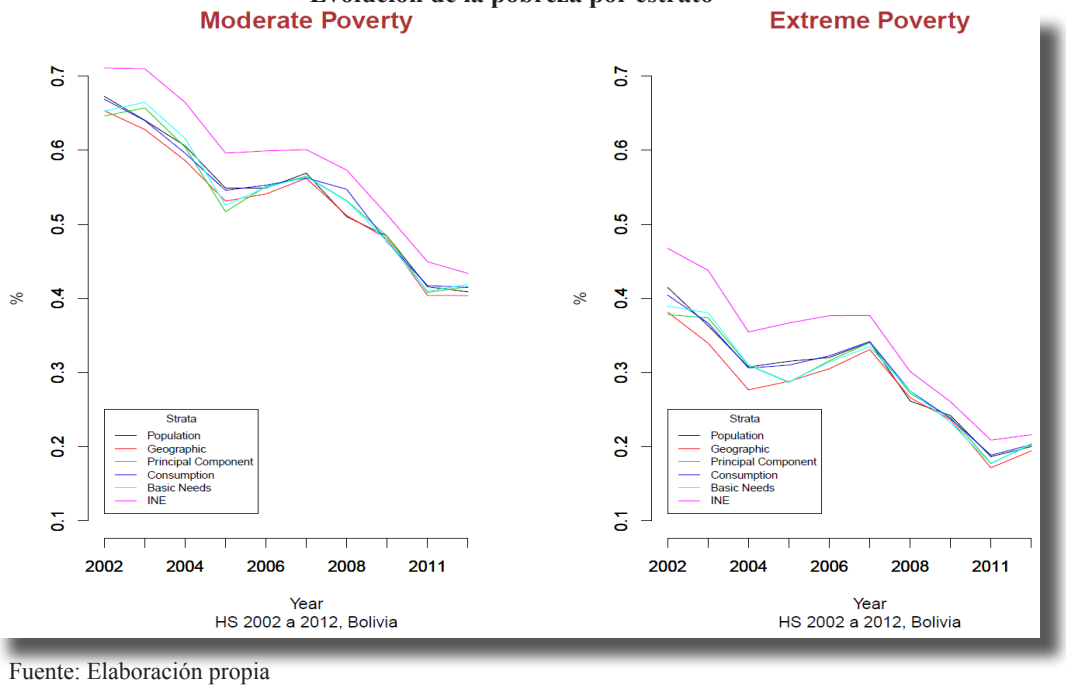
## 3. RESULTADOS

Esta sección presenta los resultados de los 5 tipos de diseño, mostrando la evolución de la pobreza, el rendimiento de muestreo y la elección del diseño final.

### 3.1 *EVOLUCIÓN DE LA POBREZA*

La figura 3, presenta la evolución de la pobreza para la serie de interés y considerando los 5 tipos de estratos definidos, se incluye la estimación oficial del INE. Se observa que todos los tipos de estratos estiman una incidencia menor a la oficial en ambos tipos de pobreza pero conservan la tendencia.

**Figura 3**
**Evolución de la pobreza por estrato**

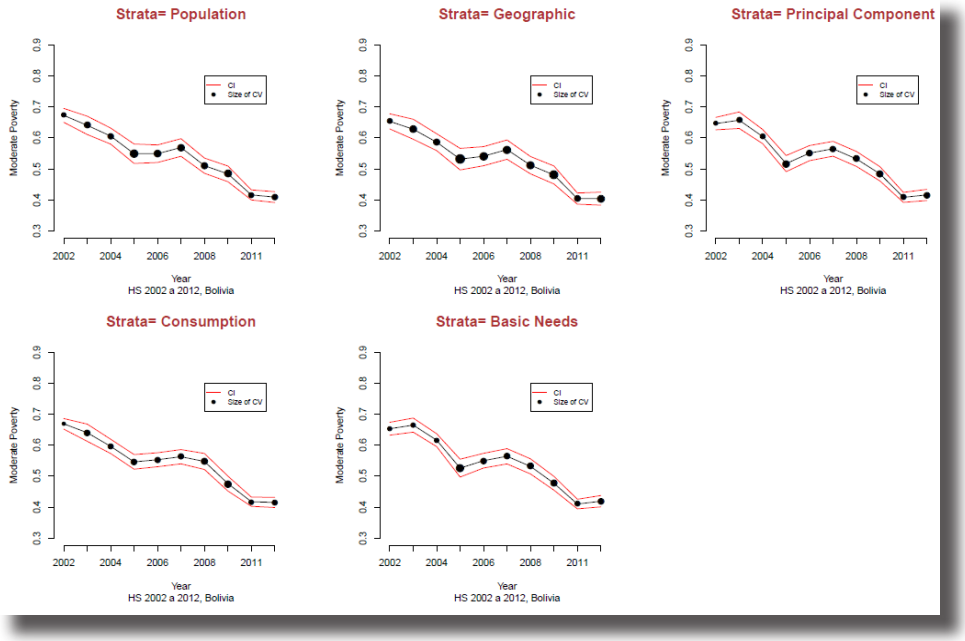



Fuente: Elaboración propia

### *3.2 DESEMPEÑO DEL DISEÑO MUESTRAL*

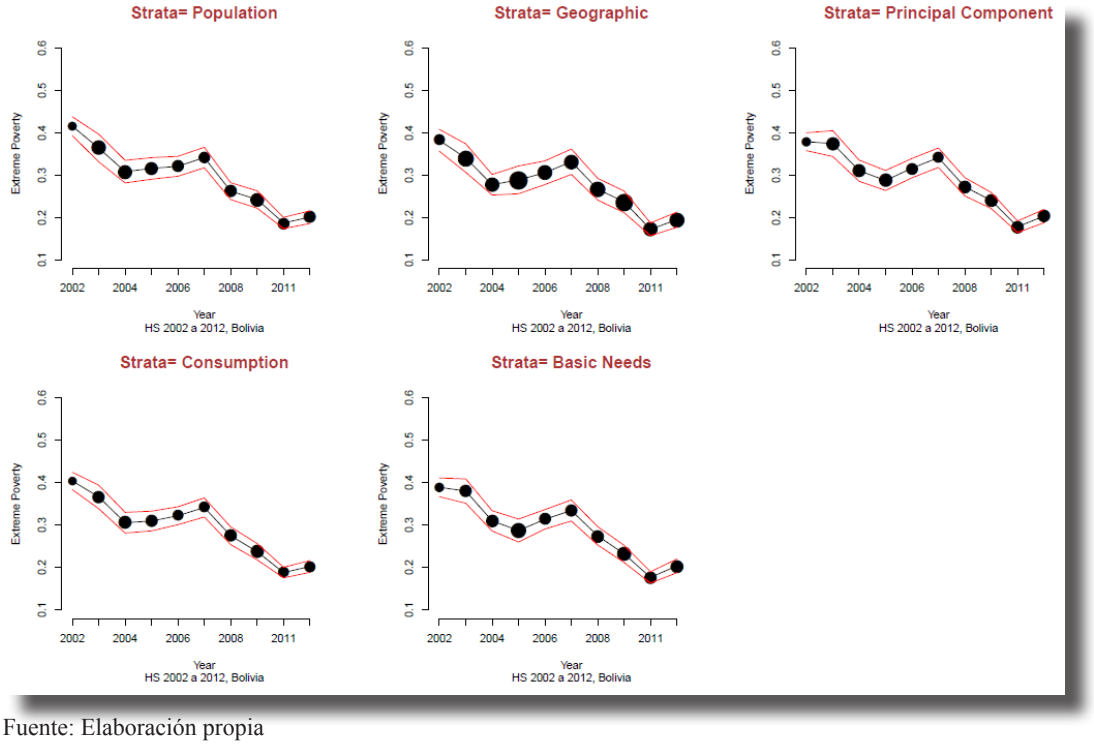
Para el rendimiento del muestreo se considera el coeficiente de variación muestral y el efecto de diseño para los dos tipos de pobreza a nivel nacional y departamental, las figuras 4, 5 muestran la evolución de la pobreza moderada y extrema por el tipo diseño e incluye el coeficiente de variación y los intervalos de confianza al 95 %.

**Figura 4**
**Desempeño del diseño para la Pobreza Moderada**




Fuente: Elaboración propia

**Figura 5**
**Desempeño del diseño para la Pobreza Extrema**




Fuente: Elaboración propia

Se observa que uno de los estratos con menor rendimiento es el geográfico y que el de consumo y componentes son los de mejor rendimiento. El cuadro 4 presenta un resumen del rendimiento a nivel departamental y nacional para la pobreza moderada y extrema.

Se observa que estrato de consumo es el que tienen un mejor rendimiento, seguido por el de componentes principales, luego el de necesidades básicas, el de población y el de más bajo rendimiento es el geográfico. El efecto de diseño es un indicador interesante de analizar, es superior a 7 en todos los casos, esto se debe a la alta homogenizad en las PSU.

**Cuadro 4**
**Rendimiento promedio del muestreo por indicador y estrato EH 2002 - 2012**

| Pobreza Moderada | | | | | |
|---|---|---|---|---|---|
| | Estrato | Nacional | | Departamento | |
| | | CV | DEFF | CV | DEFF |
| | Población | 2.364 | 13.180 | 8.567 | 10.295 |
| | Geográfico | 2.679 | 16.077 | 9.427 | 12.735 |
| | Componentes principales | 2.155 | 10.749 | 8.007 | 8.521 |
| | NBI | 2.120 | 10.460 | 7.880 | 8.667 |
| | Consumo | 2.070 | 10.028 | 7.719 | 7.918 |
| Pobreza extrema | | | | | |
| | Estrato | Nacional | | Departamento | |
| | | CV | DEFF | CV | DEFF |
| | Población | 3.883 | 12.512 | 15.032 | 10.505 |
| | Geográfico | 4.767 | 17.117 | 18.081 | 13.918 |
| | Componentes principales | 3.837 | 11.824 | 15.268 | 9.675 |
| | NBI | 3.937 | 12.353 | 15.356 | 9.995 |
| | Consumo | 3.595 | 10.593 | 14.336 | 8.626 |

Fuente: Elaboración propia

### *3.3 DISEÑO FINAL*

Por lo desarrollado se recomienda adoptar el diseño de muestra basado en el estrato de

consumo, puesto que es el que logra un mejor rendimiento en los indicadores de muestreo.

## 4. CONCLUSIÓN

Las conclusiones son:

- El diseño muestral que mejora los indicadores de rendimiento de la muestra es el diseño basado en el estrato de consumo.

- Existen en general efectos de diseños altos, es necesario reducir la muestra al interior de las PSU e incluir más PSU en la muestra.

- Los patrones de pobreza moderada y extrema están sobre estimados con el diseño oficial que no se ajusta al ritmo de crecimiento que existió en Bolivia en el periodo intercensal.

- Para el siguiente periodo intercensal es necesario tener una planificación previa del diseño muestral para toda la serie, logrando establecer efectivamente un sistema integrado de encuestas nacionales.

Índice

## BIBLIOGRAFÍA

Cochran, W. G. (1953). *Técnicas de Muestreo*. Mexico: John Wiley and Sons

Foster, James, J. Greer, and Eric Thorbecke. 1984. "*A Class of Decomposable Poverty Measures.*" Econometrica 52 (3): 761–65

Hernani W., Villarroel P. (2012) *Evolution of poverty in Bolivia: a multidementional approach*

INE, Bolivia. (2002-2012). *Documento Metodológico, Encuesta a Hogares 2002 - 2012*. La Paz, Bolivia.

Kish, L. (1955). *Survey Sampling. New York: John Wiley and Sons*. Lohr, S. L. (2000). Muestreo: Diseño y Análisis.

Naciones Unidas. (1986). *Sampling frames and sample design for integrated household survey programmers*.

Naciones Unidas. (2009). *Diseño de muestras para encuestas de hogares Directrices prácticas* (p. 252). Nueva York.

Rosén, B. (1997). *On sampling with probability proportional to size*. J. Statistics, Plann

Inference. Särndal, Swensson, y Wretman. (1992). *Model Assisted Survey Sampling*. Canada: Springer, Verlang

# ANÁLISIS CON MODELOS MULTINIVEL

Lic. Flores López Juan Carlos
✉ caarloslopez1@gmail.com

## RESUMEN

Esta investigación desarrolla una aplicación del modelo multinivel, a los datos simulados dado que en el medio no se pudo conseguir información. Por lo tanto, se toma en cuenta la información bajo las siguientes características, como variable respuesta se tomó en cuenta a las calificaciones obtenidas por los alumnos en Matemáticas y Lenguaje. Se asumió niños de 4 a 8 años, los cursos desde pre kínder, kínder, primero y segundo de primaria. Se consideró también 10 tipos de colegio y una muestra de 670 estudiantes

En este estudio, los estudiantes (nivel l) se encuentran anidados tanto en grados escolares como en escuelas (colegios), por lo que en los modelos que se llevaron a cabo, se puso a prueba si debían considerarse los grados o escuelas como un segundo nivel o incluso a los grados como un segundo nivel y las escuelas como tercer nivel.

Para el análisis se obtuvo la gráfica de dispersión de la variable nota de matemáticas la puntuación centrada donde se observa una relación con tendencia lineal positiva de magnitud media. Se realizó el estudio de las características residuales donde se observa la aproximación al supuesto de normalidad.

En la aplicación del modelo mismo se evidencia que existe una relación importante entre las diferentes dimensiones medidas por el instrumento el rendimiento académico particularmente en el caso de las Matemáticas. Los factores más relevantes son los de “Atención y persistencia” y “Desempeño académico”. Y para la comprobación de los resultados puede observar los resultados obtenidos en esta investigación.

## PALABRAS CLAVE

*Modelo multinivel, grados o escuelas, puntuación centrada, rendimiento académico.*

## 1. INTRODUCCIÓN

En la Universidad Mayor de San Andrés (U.M.S.A), la actividad de un docente tiempo completo está inmersa la investigación y como es de conocimiento de toda la comunidad universitaria de nuestro medio, esta actividad permite el avance y la actualización en el conocimiento científico, que exige nuestra realidad. A raíz de este hecho, se realiza la investigación para la gestión 2015 sobre la aplicación de los modelos multinivel. Este es con el propósito de contribuir con la comunidad estudiantil y todos lo que tengan interés en conocer estos métodos que es de gran importancia en las aplicaciones en el análisis de datos en las distintas disciplinas. Estos métodos pueden ser aplicados en todas las ciencias, como ser ingeniería, psicología, salud, sociología, económico, la industria, administración y centros de investigación de ámbito universitario, etc.

Estamos conscientes de que toda investigación Estadística, se basa en información recogida, estos pueden ser a través de encuestas, censos, simulación, etc. Una de las etapas que sigue después de la recolección de datos es la obtención de resultados de la información. Desde este punto de vista el proyecto tuvo el propósito de desarrollar el soporte teórico y la aplicación de los modelos multinivel.

Dentro los aspectos metodológicos, podemos mencionar que el tipo de estudio es exploratorio y descriptivo. Y desde el punto de vista de los métodos de investigación que se aplica en la presente investigación es Estadístico y de análisis.

## 2. LOS DATOS JERÁRQUICOS Y EL PROBLEMA MULTINIVEL

En la vida real existe una infinidad de clases o tipos de datos, incluyendo datos recolectados, en las ciencias pedagógica, exactas, económicas, ingeniería, etc. tienen una estructura jerárquica o en forma de clúster[1].

Por ejemplo, estudios de herencia realizados en seres humanos y en animales se tratan con niveles o jerarquías naturales, donde los hijos (descendientes) están agrupados dentro de las familias. Estos, a su vez, tienden a parecerse más a sus padres en sus características físicas y mentales que individuos escogidos al azar de la población. En segundo lugar, muchos experimentos también crean datos jerárquicos, por ejemplo, los diseños experimentales en las clínicas son llevados a cabo en centros y grupos de individuos elegidos aleatoriamente (Goldstein; 1995).

El termino jerarquía define unidades agrupadas en diferentes niveles. Así, los hijos serían las unidades del nivel 1 en una estructura de datos de 2 niveles, donde el nivel 2 está dado por las familias; los estudiantes (las unidades del nivel 1) están agrupados dentro de las escuelas (las unidades del nivel 2).

Los modelos multinivel o jerárquicos se presentan en distintos tipos de aplicaciones, por ejemplo: La aplicación en el área de la salud y la educación, en las que los datos se presentan de una forma anidada o jerárquica; por ejemplo, pacientes dentro de hospitales, o estudiantes dentro de escuelas. Las aplicaciones de los modelos jerárquicos también se enmarcan perfectamente en una amplia gama de aplicaciones del gobierno y los negocios, etc. Donde muestran clústers que son efectuadas en una o varias etapas, y ofrecen una aproximación unificadora por medio del análisis de los modelos de efectos aleatorios (random-effects), de varianza-covarianza (variance-components o ANOVA) y de los modelos mixtos (mixedmodels). Así sucesivamente.

Respecto a los métodos de estimación los modelos multinivel, se dividen en:

1) Aquellos que utilizan métodos de máxima verosimilitud.

2) Los que se basan en la Estadística Bayesiana.

A su vez, dentro de cada uno de estos grupos, se encuentran distintas formas de estimación. El cuadro 1 sintetiza las formas de estimación existentes para estructuras de datos multinivel.

1 Se entiende por "clúster" un agrupamiento que contiene elementos de un menor nivel. Por ejemplo, en una muestra, el conjunto de familias en un vecindario. El "nivel", por otra parte, es un componente de los datos jerárquicos. El nivel 1 es el menor nivel; por ejemplo, estudiantes dentro de las escuelas o medidas repetidas para un mismo individuo.

**Cuadro 1**
**Métodos de estimación para los modelos multinivel**

| | *Métodos* | *Abreviatura* |
|---|---|---|
| ***Máxima verosimilitud*** | Mínimos Cuadrados Generalizados | IGLS |
| | Mínimos Cuadrados Generalizados Iterativos | RIGLS |
| | Cuasi-verosimilitud Marginal | MQL |
| | Cuasi-verosimilitud Penalizada | PQL |
| ***Estadística Bayesiana*** | Full Bayes estimación | FB |
| | Empirical Bayes estimation | EM |
| | Cadenas de Markov-Monte Carlo | MCMC |

A continuación, mencionamos un modelo jerárquico lineal de dos niveles (HLM). Este modelo, está dado por:

$$\underline{y}_j = X_j \beta + Z_j \underline{\delta}_j + \underline{\in}_j$$

$$j = 1 \quad m$$

Con

$$\begin{pmatrix} \underline{\epsilon}_j \\ \underline{\delta}_j \end{pmatrix} \sim N\left( \begin{pmatrix} \phi \\ \phi \end{pmatrix}, \begin{pmatrix} \sum_j \theta & \phi \\ \phi & \Omega(\varepsilon) \end{pmatrix} \right)$$

Y $\left(\underline{\epsilon}_j, \underline{\delta}_j\right) \perp \left(\underline{\epsilon}_\ell, \underline{\delta}_\ell\right) \ \forall\, j \neq \ell$

Las dimensiones de los vectores, β y δj respectivamente, son $n_j$ ,$r$ y $s$. Como en todos los modelos de tipo regresión, las variables explicativas $X$ y $Z$ son consideradas como variables fijas, que también pueden ser expresadas en las distribuciones de las variables aleatorias $\varepsilon$ y $\delta$ son condicionales en $X$ y $Z$. Las variables aleatorias $\varepsilon$ y $\delta$ también se llaman vectores residuales en los niveles 1 y 2, respectivamente. Las variables $\delta$ también se denominan pendientes aleatorias. Las unidades de nivel dos también son llamadas clusters.

La especificación estándar y más frecuentemente usada de las matrices de covarianza son los residuales de nivel uno, que son independientes e idénticamente distribuidos (i.i.d), por ejemplo:

$$\sum(\theta) = \sigma^2 I_{nj}$$

Donde $I_{nj}$ es la matriz identidad $n_j$-*dimensional*; y que todos los elementos de la matriz de covarianza de nivel dos Ω son parámetros libres (lo que se podría identificar Ω con ξ), o algunos de ellos están limitados a 0 y los otros son parámetros libres.

El cuestionamiento de la especificación de este modelo puede ser dirigido a diversos aspectos: la elección de variables incluidas en $X$, la elección de variables para $Z$, los residuos después de haber esperado el valor 0, la homogeneidad de las matrices de covarianza a través de los clústers, la especificación de las matrices de covarianza, y las distribuciones normales multivariable. Hay que tener en cuenta que las variables explicativas $X$ y $Z$ son consideradas como determinísticas; la

suposición de que los valores esperados de los residuos (para las variables explicativas fijas) son cero, es análoga a la suposición de que los residuos no están correlacionados con las variables explicativas en un modelo con variables explicativas aleatorias.

## 3. GENERALIDADES Y CARACTERÍSTICAS

Después de una ardua búsqueda de información, no se pudo conseguir información para la aplicación de los modelos multinivel, toda información era incompleta y muy difícil de aplicar con este tipo de información. Dado el escenario incierto de la disponibilidad de una base de datos para su aplicación de este tipo de modelos, se tomó la decisión de utilizar información simulada.

## 4. RESULTADOS

En este estudio se utilizaron como *variables respuesta* las calificaciones obtenidas por los alumnos en Matemáticas y Lenguaje. Se asumió niños de 4 a 8 años, los cursos de pre kínder, kínder, primero y segundo de primaria. Se consideró también 10 tipos de colegios y una muestra de 670 estudiantes.

Se realizó el modelamiento como variable dependiente nota de la materia Matemáticas como variable respuesta, y variables independientes: grado y colegios, cuyos resultados fueron los siguientes: el intercepto cambió de 7.282 a 20.595. La variable "grado" demuestra efectos significativos, y el ajuste del modelo también mejora en forma significativa, disminuyendo la medida de desviación de 194.518 a 33.429.

Lo anterior señala que al haber controlado las diferencias individuales, se manifiestan las diferencias entre los grados. Señala también que la diferencia en el rendimiento entre alumnos no es muy amplia, y que varía significativamente, hay características homogéneas en las escuelas. EI modelo indica que existe una estructura que justifique un modelamiento multinivel.

Bajo este criterio también es necesario realizar el análisis de residuos, para la comprobación de la normalidad de la información.

Lo que se observa es la información de los residuos que se comporta aproximadamente normal, cumpliendo los requerimientos del modelo. En este estudio, los estudiantes (nivel l) se encuentran anidados tanto en grados escolares como en escuelas (colegios), por lo que en los modelos que se llevaron a cabo, se puso aprueba si debían considerarse los grados o escuelas como un segundo nivel, o incluso a los grados como un segundo nivel y las escuelas como un tercer nivel.

Los siguientes apartados se organizan entorno a las dos **variables de respuesta** analizadas: Calificaciones de Matemáticas y Calificaciones de Lenguaje.

En cada apartado se ponen aprueba modelos con base en las siguientes estructuras de datos:

| | |
|---|---|
| Nivel 1: | Grados |
| Nivel 2: | Colegios |

Y posteriormente:

| | |
|---|---|
| Nivel 1: | Alumnos. |
| Nivel 2: | Grado escolar. |

Por último, se pone a prueba:

| | |
|---|---|
| Nivel 1: | Alumnos |
| Nivel 2: | Grados escolar |
| Nivel 3: | Colegios. |

Después de realizar un estudio de los modelos multinivel, tratando de responder el rendimiento en la asignatura de matemáticas y lenguaje considerando como variable de pendiente y el grado y colegio como niveles de primer y segundo orden, podemos decir lo siguiente:

- Se estableció que existe una relación importante entre las diferentes dimensiones medidas por el instrumento el rendimiento académico particularmente en el caso de las Matemáticas. Los factores más relevantes son los de "Atención *y* persistencia" *y* "Desempeño académico".

- A pesar de que los resultados obtenidos en secciones anteriores señalaban que la relación entre las variables explicativas y las variables respuesta es distinta en las diferentes escuelas, los resultados del análisis multinivel no apoyan esto. Sin embargo, debe interpretarse este resultado con cuidado, ya que como se mencionó al inicio, pueden no haberse manifestado estas diferencias debido a que el número de colegios incluidos en la muestra es reducido.

- EI grado escolar que cursa el niño cuando se lleve a cabo el seguimiento es un aspecto que debe ser incluido tanto en los modelos de regresión jerárquicos como en los tradicionales, ya que incrementa la precisión de las estimaciones. En los grados avanzados existe mayor exactitud en las predicciones.

En conclusión, los análisis presentados en esta sección señalan que no es importante la inclusión de un segundo nivel correspondiente al *grado* que cursaban los alumnos al momento de la evaluación. Esto es verdad para Matemáticas, *y* aunque en Lenguaje no fue necesario *utilizar* un modelo multinivel, se confirma a través del uso de modelos de regresión lineal tradicional la necesidad de tomar en cuenta el *grado* que cursan. Entre más avanzado sea éste, se *alcanzará* mayor exactitud en la predicción.

De *igual* manera, aquellos modelos que utilizan las dimensiones medidas por el instrumento en lugar de la puntación global, logran un mejor *ajuste.* Se determinó que los factores más significativos son el "Desempeño académico" y la "Atención".

Los resultados de este estudio indican que el género no tiene efectos sobre el rendimiento.

Asimismo, el uso del modelamiento multinivel ha servido para comprobar que las diferencias que existen entre los colegios, no requieren una interpretación diferente para cada escuela de las puntuaciones obtenidas con el instrumento. Puede utilizarse la misma transformación de puntajes en todos los colegios incluidos en el estudio.

Tanto los análisis de regresión OLS como los modelos de regresión jerárquica, han permitido conocer la magnitud de la relación entre las puntuaciones obtenidas en la prueba de seguimiento y las calificaciones escolares posteriores. Han permitido, asimismo, establecer cuál de los factores medidos por la prueba tiene mayor efecto sobre el rendimiento. También han logrado establecer la relación que existe entre el momento en que se realiza aplicación del instrumento (grado escolar en que se aplica) y su capacidad

predictiva, así como la importancia de los factores varia en los diferentes grados, y no en las diferentes escuelas (en forma significativa). Sin embargo, aún no se ha realizado la estimación de la exactitud de la prueba para detectar niños en riesgo de presentar problemas de aprendizaje. Es decir, no se tuvo información acerca de los porcentajes de éxito que se tienen en la identificación y clasificación de los niños.

## BIBLIOGRAFÍA


Gelman Andrew "*Data Analysis Using Regression and Multilevel/Hierarchical Models*".

Goldstein Harvey "*Handbook of Multilevel Analysis*"

Hox J.J. "*Applied Multinivel Analysis*"

Martínez-Garrido Cynthia y F. Javier Murillo "*Programas para la realización de Modelos Multinivel. Un análisis comparativo entre MLwiN, HLM*".

Min Yang, JonRasbash, Harvey Goldstein, Maria Barbosa "*MLwiN Macros for advanced Multilevel modeling*".

Snijders Tom A. B. "*Multilevel Analysis"*


Índice

# MODELACIÓN POISSON CON ENFOQUE BAYESIANO PARA EXPLICAR EL EFECTO DE LA EDUCACIÓN Y DEL ÁREA DE RESIDENCIA SOBRE LA MORTALIDAD DURANTE LOS PRIMEROS AÑOS DE VIDA

Lic. Loza Cruz, Patricia
✉ *lcruzpatricia@gmail.com*

## RESUMEN

El presente trabajo de investigación es parte de un estudio mucho más amplio en el que se analiza, con el uso de modelos tipo Poisson, los factores determinantes de la mortalidad durante los primeros años de vida a partir de dos enfoques estadísticos: el bayesiano y el frecuentista. En este trabajo sólo se presentan algunos resultados basados en el enfoque bayesiano para determinar el efecto de la educación de las madres y el lugar de su residencia sobre los niveles de mortalidad neonatal (MNN), post-neonatal (MPNN) y post-infantil (MPI). La información usada proviene de la encuesta nacional de demografía y salud realizada en 2008. Se concluye que la educación de las madres tiene un alto impacto positivo sobre los tres tipos de mortalidad, luego de controlar las demás covariables. La tasa de MNN en hijos de madres con educación alta es 38.1% menos que en hijos de madres con educación baja o sin educación. Esta diferencia se incrementa a 47.8% para la MPNN y a 60.4% en el caso de la MPI. El problema de sobredispersión fue superado a través de dos mecanismos: una especificación adecuada del modelo y el uso de un modelo binomial negativo. Se recomienda impulsar políticas orientadas a incrementar la cobertura de la educación principalmente secundaria, de mujeres y focalizada en áreas rurales del país.

## PALABRAS CLAVE

*Enfoque bayesiano, mortalidad neo-natal, mortalidad post-neonatal, mortalidad post-infantil.*

## 1. INTRODUCCIÓN

Es de gran importancia para los planificadores del sector de la salud del país conocer tanto el nivel como la evolución de la mortalidad en los primeros años de vida, pues ello les permite evaluar la eficacia de las acciones pasadas en el área, así como delinear nuevas políticas y programas. Empero, también es de vital importancia conocer los factores socioeconómicos que determinaron tanto esos niveles de mortalidad como su tendencia.

En el orden metodológico, un modelo de regresión Poisson es adecuado para abordar el análisis de los factores determinantes de la mortalidad en los primeros años de vida. El problema de sobredispersión, típico en datos tipo Poisson, puede superarse recurriendo o bien a un "mejor" tratamiento de la información en la etapa de modelación o bien a modelos alternativos.

## 2. OBJETIVO

Determinar el efecto de la educación y lugar de residencia sobre el nivel de la mortalidad neonatal, post-neonatal y post-infantil, durante el periodo de análisis (1993-2007).

## 3. INFORMACIÓN

La Encuesta Nacional de Demografía y Salud, principal fuente de información del País con relación al sector salud, es la más adecuada para realizar la presente investigación. En el

módulo de "historia de nacimientos" de la ENDSA 2008 se obtiene información acerca de la fecha de nacimiento, edad actual y edad al morir si es el caso, entre otros datos, para cada uno de los hijos nacidos vivos de las mujeres en edad fértil entrevistadas. A partir de esta información se elaboró el Cuadro 1, información que se usó en el análisis posterior.

En el Cuadro 1 se muestran las tasas de mortalidad clasificadas por grupo de edad del niño, cohorte de nacimiento del niño, área de residencia y nivel de educación de las madres. Se han definido tres grupos de edad: 0 meses, 1-11 meses y 12-59 meses. Estos grupos de edad corresponden a las edades consideradas para calcular las tasas de mortalidad neonatal (MNN), mortalidad post-neonatal (MPNN) y mortalidad pos-infantil (MPI), respectivamente. Por esta razón, cuando por ejemplo se hace referencia a la tasa de MNN se estará haciendo mención a la mortalidad en el primer mes de vida. La variable cohorte también tiene tres categorías: la cohorte de nacidos vivos en el periodo 1993-1997, la cohorte de 1998-2002 y la cohorte de nacimientos en el periodo 2003-2007.

**Cuadro 1**
**Bolivia: Tasas de mortalidad (por mil) por grupo de edad, según cohorte de nacimiento, área de residencia y nivel de educación, ENDSA 2008**

| Área de residencia | Nivel de educación | Grupo de edad (en meses) | Cohorte de Nacimiento 1993-1997 | 1998-2002 | 2003-2007 |
|---|---|---|---|---|---|
| Urbana | Baja | 0 | 418.2 | 362.2 | 295.0 |
| | | 1-11 | 50.3 | 30.0 | 27.0 |
| | | 12-59 | 7.1 | 5.6 | 1.7 |
| Urbana | Alta | 0 | 197.0 | 265.7 | 210.0 |
| | | 1-11 | 23.0 | 19.6 | 11.2 |
| | | 12-59 | 3.7 | 1.5 | 1.7 |
| Rural | Baja | 0 | 570.1 | 573.1 | 465.5 |
| | | 1-11 | 57.7 | 47.1 | 36.8 |
| | | 12-59 | 10.7 | 7.4 | 6.4 |
| Rural | Alta | 0 | 215.5 | 305.2 | 393.4 |
| | | 1-11 | 24.0 | 21.9 | 22.7 |
| | | 12-59 | 3.7 | 2.5 | 1.7 |

Fuente: Elaboración propia

Con relación a las variables área de residencia y nivel de educación, ambas tienen dos categorías: residencia urbana y rural para la variable área de residencia y educación baja y alta para la variable nivel de educación. A madres sin educación formal o con educación primaria se las ha denominado con educación baja, mientras a las madres con educación secundaria o superior se las ha denominado con educación alta.

## 4. EL MÉTODO BAYESIANO

A diferencia del enfoque clásico, el enfoque bayesiano considera al parámetro como una

variable aleatoria e interpreta la probabilidad desde el punto de vista subjetivo. Para realizar el proceso de inferencia bayesiana es necesario la especificación de una distribución previa o *apriori* de probabilidad, la cual representa el conocimiento acerca del parámetro antes de obtener cualquier información respecto a los datos. Este enfoque tiene como punto central el Teorema de Bayes el cual se detalla a continuación:

Sea $y = y_1, y_2, \dots\dots, y_n$ un vector $n$ de observaciones cuya distribución de probabilidad $f(y, \theta)$ depende de $k$ parámetros involucrados en el vector $\theta = \theta_1, \theta_2, \dots\dots, \theta_k$. Se supone también que $\theta$ tiene una distribución de probabilidad previa $f(\theta)$. Entonces la distribución conjunta de $\theta$ e $y$ es:

$$f(y, \theta) = f(y|\theta) * f(\theta)$$

$$= f(\theta|y) * f(y)$$

donde la distribución de probabilidad condicional de $\theta$ dado el vector de observaciones y tiene la forma

$$f(\theta|y) = \frac{f(y|\theta) * f(\theta)}{f(y)}$$

$$= c * f(y|\theta) * f(\theta)$$

$$= l(\theta|y) * f(\theta)$$

Siendo $c = \frac{1}{f(y)}$ la constante normalizadora, $l(\theta|y)$ representa la función de verosimilitud, $f(\theta)$ la distribución previa y $f(\theta|y)$ la distribución posterior del parámetro.

La distribución de probabilidad previa $f(\theta)$ puede tener información específica acerca del parámetro, así como también puede tener información general y ser considerado como una distribución previa no informativa.

Una selección común para una previa no informativa es la previa propuesta por Jeffreys (1961) quien definió una distribución previa no informativa considerando tranformaciones uno a uno del parametro $\phi = h(\theta)$. Mediante transformacion de las variables se puede ver que la densidad previa $f(\theta)$ es proporcional a la densidad previa para $\phi$, esto es

$$f(\theta) \propto \overline{J(\phi)}$$

donde $\overline{J(\phi)}$ es la matriz de información de Fisher.

Para una variable aleatoria y con una distribución Poisson ($\theta$), la previa no informativa de Jeffreys está dada por:

$$f(\theta) = \overline{J(\phi)} = \frac{1}{\theta}$$

La función de verosimilitud es la función a través de la cual los datos modifican el conocimiento previo de $\theta$. El modelo lineal generalizado Poisson asume que se distribuye Poisson con media $\theta$ y con una función de enlace logarítmico, tal que $\ln \theta = X\beta$. En esta situación, la distribución muestral para los datos $y = y_1, y_2, \dots\dots, y_n$ queda expresada como:

$$f(y, \beta) = \prod_{i=1}^{n} \frac{1}{y_1!} e^{\exp(\eta_i)} \exp(\eta_i)^{y_i}$$

donde $\eta_i = X\beta$ es el predictor lineal para el i-ésimo caso.

La distribución posterior contiene toda la información concerniente al parámetro de interés $\theta$. En consecuencia, cualquier inferencia con respecto a $\theta$ se hará a partir de dicha distribución. En el caso del modelo Poisson, la distribución posterior llega a ser

$$f(\theta|y) \propto \theta^{\sum_{i=1}^{n} y_i - \frac{1}{2}} e^{-n\theta}$$

la cual es el kernel de una distribución Gamma con parámetros

$$\left(\sum_{i=1}^{n} y_i - \frac{1}{2}, n\right)$$

En el contexto bayesiano, las estimaciones de los parámetros del modelo se obtienen generalmente con el procedimiento denominado muestreador de Gibbs, un procedimiento específico del denominado método MCMC. En términos generales, el procedimiento consiste en extraer muestras en forma sucesiva de las probabilidades condicionales de los parámetros del modelo y consta de los siguientes pasos:

1. Definir $t = 0$ , y elegir un valor inicial arbitrariamente de
$$\theta^0 = \theta_1^0, \theta_2^0, \ldots\ldots, \theta_k^0$$
2. Generar cada componente de $\theta$ como sigue:

- Mostrar $\theta_1^{t+1}$ desde
$$f(\theta_1|\theta_2^{(t)}, \ldots\ldots, \theta_k^{(t)}, y)$$
- Mostrar $\theta_2^{t+1}$ desde
$$f(\theta_2|\theta_1^{t+1}, \theta_3^{(t)} \ldots\ldots, \theta_k^{(t)}, y)$$
- ……..
- Mostrar $\theta_k^{t+1}$ desde
$$f(\theta_k|\theta_1^{t+1}, \ldots\ldots, \theta_{k-1}^{t+1}, y)$$

3. Generar $t = t+1$. Si $t<T$ el número de muestras deseadas, retornar al paso 2. En otro caso, parar.

## 5. ALGUNOS RESULTADOS

Con base en criterios estadísticos para la comparación de modelos, el modelo que mejor se ajusta a la información y que, a la vez, supera el problema de sobredispersión es el que en la parte sistemática toma en cuenta, además de los efectos individuales, los efectos de interacción entre el grupo de edad y la cohorte de nacimientos, y la interacción entre el grupo de edad y el nivel de educación. El área de residencia no interactúa con ninguna de las variables analizadas. El modelo final para analizar las tasas de mortalidad, entonces, es representado como:

$$Edad * Cohorte + Edad * Educación + Residencia$$

**Cuadro 2**
**Estimación Bayesiana de los Parámetros para el Modelo Elegido "EDA*COH+EDA*EDU+RESI"**

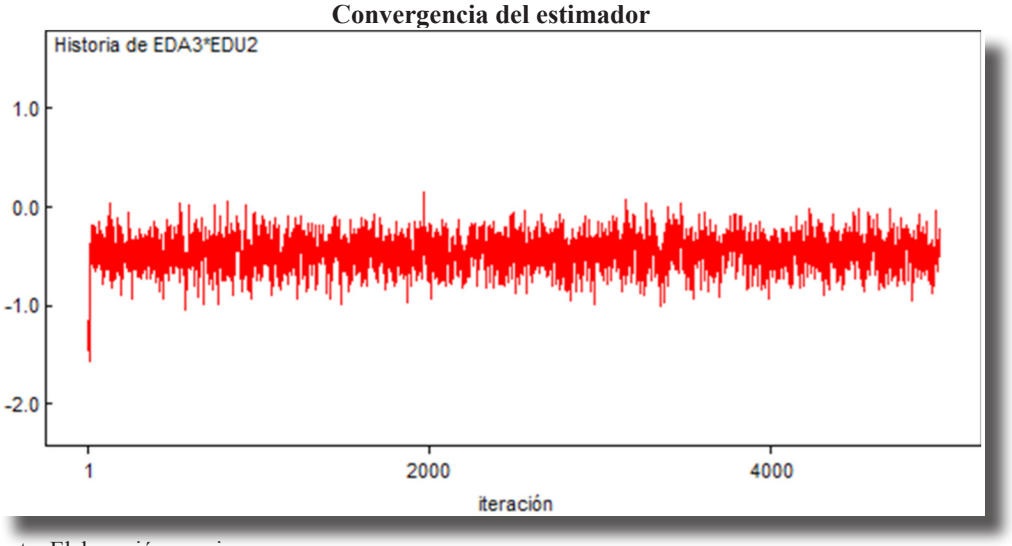
| node | mean | sd | MC error | 2.50% | median | 97.50% | start | sample |
|---|---|---|---|---|---|---|---|---|
| eda2 | -2.195 | 0.077 | 0.005 | -2.348 | -2.196 | -2.040 | 1001 | 4000 |
| eda3 | -3.946 | 0.084 | 0.004 | -4.108 | -3.946 | -3.785 | 1001 | 4000 |
| coh2 | 0.011 | 0.076 | 0.004 | -0.141 | 0.011 | 0.162 | 1001 | 4000 |
| coh3 | -0.168 | 0.083 | 0.004 | -0.326 | -0.168 | 0.000 | 1001 | 4000 |
| edu2 | -0.481 | 0.088 | 0.003 | -0.654 | -0.482 | -0.311 | 1001 | 4000 |
| res2 | -0.356 | 0.048 | 0.001 | -0.449 | -0.357 | -0.261 | 1001 | 4000 |
| eda2coh2 | -0.298 | 0.108 | 0.006 | -0.504 | -0.297 | -0.084 | 1001 | 4000 |
| eda2coh3 | -0.361 | 0.121 | 0.006 | -0.608 | -0.359 | -0.129 | 1001 | 4000 |
| eda3coh2 | -0.399 | 0.124 | 0.005 | -0.637 | -0.401 | -0.151 | 1001 | 4000 |
| eda3coh3 | -0.546 | 0.169 | 0.005 | -0.879 | -0.545 | -0.217 | 1001 | 4000 |
| eda2edu2 | -0.171 | 0.125 | 0.004 | -0.420 | -0.172 | 0.071 | 1001 | 4000 |
| eda3edu2 | -0.456 | 0.162 | 0.005 | -0.786 | -0.452 | -0.151 | 1001 | 4000 |
| cons | -0.587 | 0.057 | 0.004 | -0.698 | -0.587 | -0.473 | 1001 | 4000 |

Fuente: Elaboración propia

Las estimaciones de los parámetros en el modelo elegido fueron realizadas tanto con el software Winbugs como con programas propios elaborados por la autora. Los resultados son expuesto en el Cuadro 2.

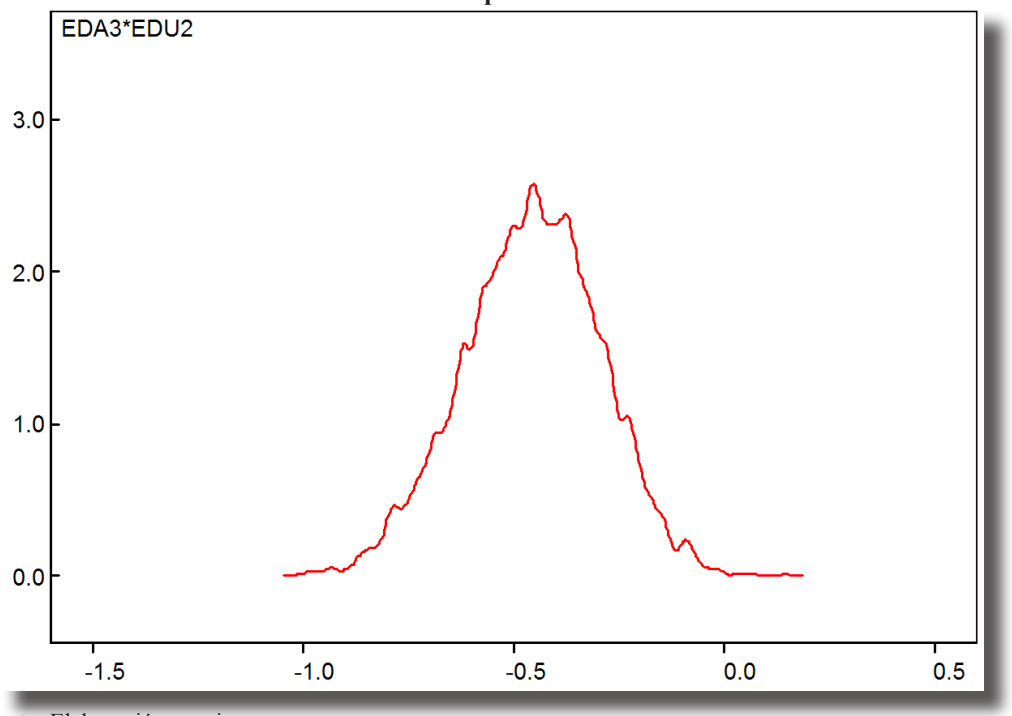
El Gráfico 1 muestra la convergencia del estimador para el coeficiente de la interacción entre el tercer grupo de edad (12-59 meses) y el segundo nivel de educación (educación alta), simbolizado por *EDA3*EDU2*, luego de eliminar ("burn") las primeras 1000 simulaciones. El valor estimado de tal coeficiente es -0.456. La distribución suavizada para el mismo estimador se muestra

**Gráfico 1**
**Convergencia del estimador**



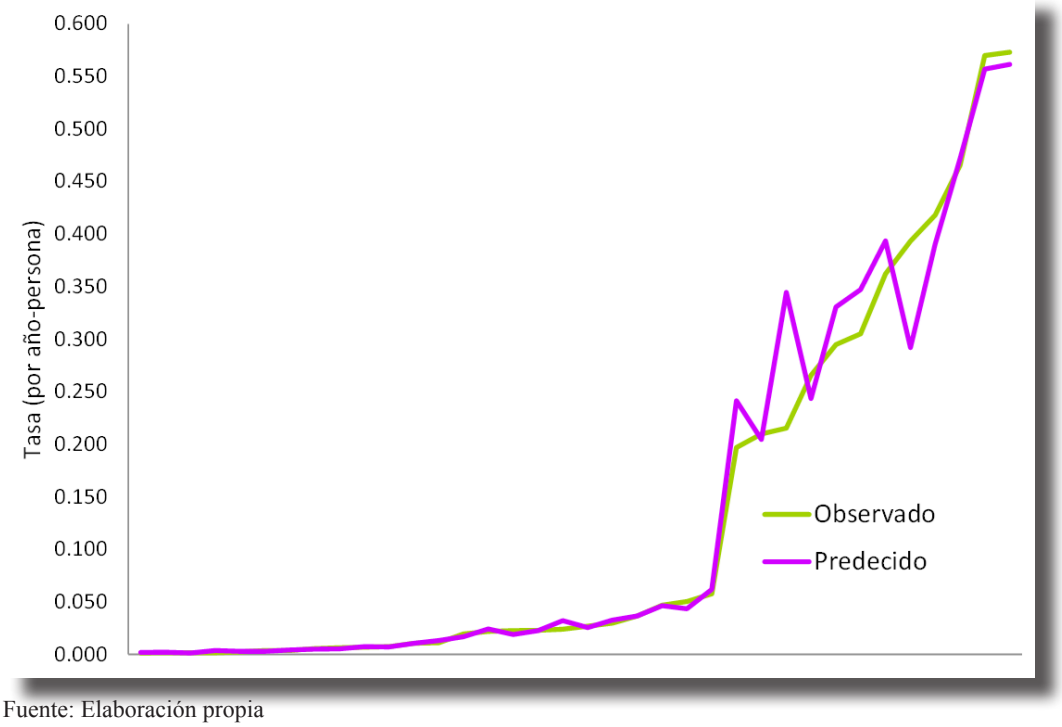

Fuente: Elaboración propia

**Gráfico 2**
**Distribución suavizada para el mismo estimador**




Fuente: Elaboración propia

**Gráfico 3**
**Tasas de mortalidad observadas y tasas predecidas con el modelo elegido "EDA*COH+EDA+RESI**




Fuente: Elaboración propia

en el Gráfico 2. Similares convergencias y distribuciones suavizadas se tienen para cada uno de los estimadores de parámetros.

Tasas de mortalidad observadas y predecidas con el modelo elegido son comparadas en el Gráfico 3. Si bien se aprecia algunas diferencias en niveles intermedios de mortalidad, el ajuste del modelo es casi perfecto en niveles extremos.

A diferencia del informe completo, en este documento se resaltan sólo dos de los resultados de la investigación. Por una parte, la educación tiene un efecto altamente significativo sobre la tasa de mortalidad en los primeros años de vida. En efecto, la tasa de mortalidad neonatal (MNN) en hijos de madres con educación alta es 38.1% menos que en hijos de madres con educación baja, controlado el efecto de cohorte de nacimiento y de área de residencia. Esta diferencia porcentual se incrementa a 47.8% en el caso de la mortalidad post-neonatal (MPNN) y a 60.4% en el caso de la mortalidad post-infantil (MPI).

En consecuencia, la educación tiene un alto impacto positivo en los tres tipos de mortalidad, principalmente en la tasa de mortalidad post-infantil (ver Cuadro 3).

**Cuadro 3**
**Efecto de la educación sobre el nivel de mortalidad, por tipo de mortalidad**

| Tipo de mortalidad | Educación | |
|---|---|---|
| | Baja | Alta (%)* |
| MNN | - | -38.1 (0.0000) |
| MPNN | - | -47.8 (0.0000) |
| MPI | - | -60.4 (0.0000) |

Fuente: Elaboración propia
*En paréntesis el valor-p del efecto

El segundo resultado tiene que ver con el efecto del área de residencia. Si bien su efecto es inferior en comparación con el

efecto de la educación, también es altamente significativo. Esto es, la mortalidad en el área urbana es 29.9% menos que en el área rural, controlado el efecto de cohorte de nacimiento y de educación. Tal diferencia porcentual es la misma en cada uno de los tres tipos de mortalidad (MNN, MPNN y MPI), en cada nivel de educación y en cada cohorte de nacimientos (ver Cuadro 4).

**Cuadro 4**
**Efecto de área de residencia sobre el nivel de mortalidad**

| Tipo de mortalidad | Educación | |
|---|---|---|
| | Baja | Alta (%)* |
| MNN | - | -38.1 (0.0000) |
| MPNN | - | -47.8 (0.0000) |
| MPI | - | -60.4 (0.0000) |

Fuente: Elaboración propia
*En paréntesis el valor-p del efecto

Índice

## BIBLIOGRAFÍA

Box, G.E.P and Tiao, G. C.(1973), "*Bayesian Inference in Statistical Analisys*".Bretthorst,

G. Larry (1988), "*Bayesian Spectrum Analisys and Parameter Estimation*"

Lee, Peter (1997), "*Bayesian Statistics: An Introduction*".

Winkler, R. (2003) "*Introduction to Bayesian Inference and Decision*".

# Psychometric behaviour of the strengths and difficulties questionnaire (SDQ) in the Spanish national health survey 2006

Manuel Gómez-Beneyto[1,2], Andreu Nolasco[3*], Joaquín Moncho[3], Pamela Pereyra-Zamora[3], Nayara Tamayo-Fonseca[3], Mikel Munarriz[4], José Salazar[5,6], Rafael Tabarés-Seisdedos[1,2] and Manuel Girón[7]

**Abstract**

**Background:** The Strengths and Difficulties Questionnaire (SDQ) is a tool to measure the risk for mental disorders in children. The aim of this study is to describe the diagnostic efficiency and internal structure of the SDQ in the sample of children studied in the Spanish National Health Survey 2006.

**Methods:** A representative sample of 6,773 children aged 4 to 15 years was studied. The data were obtained using the Minors Questionnaire in the Spanish National Health Survey 2006. The ROC curve was constructed and calculations made of the area under the curve, sensitivity, specificity and the Youden J indices. The factorial structure was studied using models of exploratory factorial analysis (EFA) and confirmatory factorial analysis (CFA).

**Results:** The prevalence of behavioural disorders varied between 0.47% and 1.18% according to the requisites of the diagnostic definition. The area under the ROC curve varied from 0.84 to 0.91 according to the diagnosis. Factor models were cross-validated by means of two different random subsamples for EFA and CFA. An EFA suggested a three correlated factor model. CFA confirmed this model. A five-factor model according to EFA and the theoretical five-factor model described in the bibliography were also confirmed. The reliabilities of the factors of the different models were acceptable (>0.70, except for one factor with reliability 0.62).

**Conclusions:** The diagnostic behaviour of the SDQ in the Spanish population is within the working limits described in other countries. According to the results obtained in this study, the diagnostic efficiency of the questionnaire is adequate to identify probable cases of psychiatric disorders in low prevalence populations. Regarding the factorial structure we found that both the five and the three factor models fit the data with acceptable goodness of fit indexes, the latter including an externalization and internalization dimension and perhaps a meaningful positive social dimension.
Accordingly, we recommend studying whether these differences depend on sociocultural factors or are, in fact, due to methodological questions.

**Keywords:** Psychometrics, Mental disorders diagnosed in childhood, Health survey, Strengths and difficulties questionnaire, Spain

* Correspondence: nolasco@ua.es
[3]Research Unit for the Analysis of Mortality and Health Statistics, University of Alicante, Alicante, Spain
Full list of author information is available at the end of the article

## Background

In its 2006 edition, the Spanish National Health Survey (SNHS) used for the first time the Strengths and Difficulties Questionnaire (SDQ) to measure the risk for a mental disorder in children aged 4 to 15 years [1,2]. The SDQ provides separate scores for very important clinical and epidemiological dimensions, such as hyperactivity, emotional symptoms, behavioural problems and difficulties with peers. It also includes a prosocial behaviour, meant to measure the child positive social skills. In addition there are three versions to be used by the parents, the teachers and a self-report questionnaire for 11–16 16 years old, as well as an extended version which includes an estimation of the impact on functioning, distress and burden on others. This study will focus only on the parent version.

The SDQ was originally designed as a screening tool for population-based surveys [3,4] and it has been used in national health surveys in several countries [5-7]. It has also been used successfully for clinical evaluation in clinical settings and as a research tool. Studies undertaken in different cultures have shown that it possesses fair reliability and good criterion and convergent validity [8-10]. Regarding the internal structure there are a large number of studies confirming the existence of the aforementioned five theoretical dimensions, using both exploratory (EFA) and confirmatory (CFA) factor analysis [10,11]. However there are also discrepancies, some authors reporting three [12-15] and four factor solutions [16], and a few others who could not even find a clinically meaningful solution. A recent British study has confirmed both the five factor and the three factor solution [15] and in a study covering five European countries it is argued that the number of factors in the model may be country-dependent [14]. The three factor solution validated in some studies is particularly interesting since it gathers hyperactivity and behavioural problems in one factor, emotional symptoms and difficulties with peers in another factor and prosocial behaviour as a third factor. The first two factors constitute the well known dimensions of externalization and internalization. This is compatible with a hierarchical model of psychopathology. However the value of the prosocial behaviour dimension is not so clear. In constructing the questionnaire Goodman [17] added ten items reflecting traits of strength (half of them reverse-scored to reflect difficulties) to make it more acceptable to parents by enquiring about strengths as well as weaknesses. Out of ten strength items, the five directly-scored constitute the prosocial behaviour dimension, two reverse-scored items are included in hyperactivity, another two in peer problems and one in conduct problems. The introduction of strength items and directly/reversed-scored items has complicated the exploration of the factorial structure of the SDQ. A sixth factor including some of the strength items has been reported in previous studies and discarded as a methodological artefact [11,18].

The Spanish version of the SDQ [19] used in this study has been validated in a sample population of the Canary Isles [20] by a semi structured diagnostic interview [21] administered and scored by specialists. The diagnostic parameters obtained were acceptable and similar to those of the original study [3], but the cut point to identify probable cases was higher. Analysis of the dimensionality of the questionnaire using EFA showed a similar structure, though not equivalent to that expected from the theoretical structure.

The reported discrepancies in the structure of the SDQ and the uncertainties surrounding the Spanish version warrant a further examination of the psychometric behaviour of the SDQ. Thus, the aim of this study is to describe the diagnostic efficiency and internal structure of the SDQ in the sample of children studied in the SNHS 2006.

## Method

### Sample

The study data were obtained using the Minors Questionnaire of the SNHS 2006 [22]. The survey has a cross-sectional design, and contemplates a sample of children aged 0 to 15 years, distributed throughout Spain. Details of the methodology (sample design, sample size and sampling procedure) have been published elsewhere [2]. In brief, the number of children surveyed was 9,122, of whom 6,773 were aged 4 or over (51.2% men, 48.8% women). This latter was the size of the study sample, and was representative of the corresponding population. As a result of the complex sample design of the SNHS, the analysis used the weightings corresponding to the sample subjects. Applied to the children studied, these weightings enabled the number of children represented by each sample child to be established. The original weightings ($\lambda$) were calculated according to the sample design and included in the database supplied by the Spanish Ministry of Health, Social Politics and Equality, and were transformed to adjust the weights to the actual sample size studied. The estimations thus obtained were unbiased and coincide with those obtained using the methods incorporated in the sample design, although the random error of the estimations should be considered approximate. Among the various different solutions for use of the weightings we selected the method that consists of the transformation of the weightings under the normalized form:

Normalized weight sample unit i = $\omega = \frac{n}{N}\lambda_i$

where n = number of sample minors

$\lambda_i$ = original weight unit i.

N = population sample represented = $\sum_i \lambda_i$

With these weights, a sample of the same size as that studied is reproduced, thus avoiding the problem of artificially reducing the random errors that would be estimated with the original weights, as these would reproduce a sample size similar to the study population, i.e., very large.

### Measures

To evaluate the presence of mental health problems, the survey included the Spanish version of the SDQ. The SDQ is composed of 25 questions grouped in five dimensions, four relating to psychopathology (emotional symptoms, conduct problems, symptoms of hyperactivity/inattention and peer problems) and prosocial behaviour. Each dimension has 5 items that are each scored between 0 and 2 according to their frequency, obtaining a score of 0–10 for each dimension. The total difficulty is obtained by adding the 20 items for difficulties (excluding prosocial behaviour).

The SNHS also includes questions directed to the informants of the child, aimed to detect cases of disease, with five of these questions being included in this study: Does the child suffer or has he/she ever suffered from conduct problems (including hyperactivity), Does the child suffer or has he/she ever suffered from mental disorder (depression, anxiety,. . .). If the answer to either of these two questions is "Yes", then: Has he/she had them during the last 12 months? Has a physician ever said he/she has them? During the last 12 months, have these disorders or health problems limited the child in any of his/her usual activities in any way?

### Data Analysis

To estimate the efficiency of the questionnaire as a screening tool the ROC curve was drawn and calculations made of the area under the curve, sensitivity, specificity and the Youden J indices. The total difficulty score was used for the calculation and the children were considered to be cases if the informant had answered positively either of the first two questions and the remaining three designed to determine the presence of a disorder.

In order to estimate factor analysis models those cases in which a value was lost in any of the items studied were not included. The initial 6,773 minors fell to 6,506 who had complete information for all 25 items on the questionnaire. The factorial structure of the questionnaire was studied using models of exploratory factorial analysis and confirmatory factorial analysis, using the software FACTOR v8.1 [23,24] and LISREL v8.80 [25] respectively. The variables (items of the questionnaire) were defined as ordinal. The polychoric correlation matrixes between the items, obtained using the weightings corresponding to the sample subjects, were used as an element to reproduce both for the EFA and for the CFA.

In order to cross-validate factor models, the initial sample (n = 6,506) was divided in two random subsamples of the same size (n = 3,253). An EFA was performed on one subsample, using Parallel analysis based on 500 replications [26] as a test to establish the number of factors to retain. The estimation method was Unweighted Least Squares, and in order to obtain a simple factor solution we used Promin. This rotation method allows factors to be oblique in order to maximize factor simplicity [27]. The reliability of each of the factor construct was calculated after the model analyses.

The other subsample was used to validate the factor structure, previously obtained, by means of CFA. Four CFA models were adjusted. The first one with the factors obtained in the EFA (3 factors), without correlation structure. The second one with the same factors including their correlation structure. The third one with five factors obtained by EFA and with correlation structure. And the last one with the five-factor theoretical model described in the bibliography. As a general rule each item was assigned to an only factor, the one with the higher factor loading in EFA. The estimation method was Diagonally Weighted Least Squares. The goodness of fit of the CFA models was done with the usual indicators (Chi-Squared, RMSEA, ECVI, GFI, CFI, AGFI, NFI). Additionally, the reliability of each of the factor constructs was calculated in each factor as the proportion represented by the square sum of standardized factor loadings of its items with respect to the square sum of standardized factor loadings plus the sum of measurement errors associated with each item (McDonald's Omega index) [28].

## Results

The prevalence of conduct problems (including hyperactivity) diagnosed by a physician, present in the past 12 months and limiting activities of daily living, was 0.93% (CI: 0.70-1.16, n = 63), the prevalence of emotional symptoms (depression, anxiety) with the same characteristics was 0.47% (CI: 0.31-0.64, n = 32) and that of any disorder was 1.18% (CI: 0.92-1.44, n = 80). The area under the ROC curve for each of these diagnoses was 0.91 (CI: 0.88-0.94), 0.84 (CI: 0.77-0.91) and 0.88 (CI: 0.84-0.92), respectively. The diagnostic parameters for the presence of any disorder, for different cut points, are shown in Table 1.

Before performing the EFA on the first random subsample, we calculated the Barlett's sphericity test, which was significant ($p < 0.00001$), and the Kaiser-Meyer-Olkin measure was 0.855, and so the data showed a good sampling adequacy for the factor analysis. Results of Parallel analysis suggested a three-factor model, since these are the only ones that explained variability above the mean of random replications. The three-factor model (F1, F2, F3)

**Table 1 Diagnostic characteristics and Youden J index to diagnose the presence of any disorder**

| SDQ cut point [a] | Sensitivity | Specificity | Youden J |
|---|---|---|---|
| 12.50 | 0,898 | 0,740 | 0,638 |
| 13.50 | 0,833 | 0,787 | 0,620 |
| 14.50 | 0,773 | 0,821 | 0,594 |
| 15.50 | 0,713 | 0,859 | 0,572 |
| 16.50 [b] | 0,652 | 0,887 | 0,539 |
| 17.50 | 0,648 | 0,905 | 0,553 |
| 18.50 | 0,645 | 0,926 | 0,571 |
| 19.50 [c] | 0,613 | 0,941 | 0,554 |

[a] Positive result if the score is ≥ at the specified cut point.
[b] Cut point proposed by Rodriguez Hernández J [20].
[c] Cut point proposed by Goodman [4].

explained 50.0% of variability (26.1%, 15.0% and 8.8% variability explained by the respective factors) and the rotated loading matrix of which is given in Table 2. Only 4 items had a factor loading above 0.30 in more than one factor, and so the interpretation of the factors is rather clear. The construct reliability of the factors proposed for the model was 0.825, 0.908 and 0.880 for the respective factors F1, F2 and F3. A five-factor model was also built with these data in order to have a model with the same number of factors than the original theoretical model described in the bibliography. This model explained 59.4% of variability, 26.1%, 15.0%, 8.8%, 4.9% and 4.5% variability explained by the respective factors. These factors corresponded to the 5 eigenvalues above 1. The construct reliability of the factors proposed for the model was 0.786, 0.832, 0.908, 0.743, and 0.810 for the respective factors F1, F2, F3, F4 and F5. The rotated loading matrix for this model is given in Table 3, and only 2 items had a factor loading above 0.30 in more than one factor.

The CFA's were performed on the second random subsample. Four different models were built. According to the results of the EFA, two three-factor models were first adjusted, one of them without correlation and the other one with correlation between the factors. Secondly, two five-factor models were built. The first one according to the result of the five-factor model of the EFA and the second one according to the theoretical structure of the questionnaire, each factor comprising 5 items, such that each of the items on the questionnaire was assigned to just one of the 5 latent factors (according to the 5 subscales on the questionnaire). The adjusted models included the possible correlation structure between the latent factors.

Table 4 shows that all the correlated factor models had good indices of goodness of fit.

Figures 1, 2 and 3 show the results of the standardized factor loadings, correlations between the factors, reliabilities of factors and error term of variables (items of the questionnaire) of the CFA models with correlated factors. All the standardized factor loadings were above 0.40 (except for item 6, which was 0.39 in the three-factor model and 0.34 in the five-factor model according to EFA, and for item 23, which was 0.30 in the five-factor model according to EFA.

The study of modification indices suggests the possible presence of some high correlations between certain items, which could improve the fit of CFA models.

## Discussion

The overall prevalence of cases detected in this study was 1.18%, being 0.93% for conduct problems (including hyperactivity) and 0.47% for emotional symptoms (depression, anxiety). These figures relate to the prevalence in minors who, according to their parents, had been diagnosed by a physician and who had also presented limitations in their activities of daily living during the previous 12 months; this prevalence of cases, therefore, was relatively severe. Considering that the prevalence of

**Table 2 Rotated loading matrix* of EFA with 3 factors (n = 3,253)**

| Nº | Item | F1 | F2 | F3 |
|---|---|---|---|---|
| 3 | Somatic | **0.472** | 0.127 | 0.006 |
| 6 | Solitary | **0.536** | −0.148 | −0.101 |
| 8 | Worries | **0.644** | 0.025 | −0.018 |
| 13 | Unhappy | **0.716** | −0.093 | 0.057 |
| 19 | Bullied | **0.619** | −0.026 | 0.126 |
| 23 | Adults | **0.425** | 0.021 | 0.107 |
| 24 | Fears | **0.386** | 0.140 | 0.189 |
| 1 | Considerate | 0.043 | **0.633** | −0.197 |
| 4 | Shares | 0.020 | **0.617** | −0.016 |
| 9 | Caring | 0.170 | **0.793** | −0.014 |
| 11 | Good friend | 0.246 | **−0.755** | −0.148 |
| 14 | Popular | 0.217 | **−0.736** | −0.092 |
| 17 | Kind to kids | −0.033 | **0.797** | 0.010 |
| 20 | Often volunteers to help | 0.167 | **0.668** | −0.103 |
| 22 | Steals | 0.276 | −0.292 | 0.211 |
| 2 | Restless | −0.044 | 0.138 | **0.798** |
| 5 | Tempers | 0.166 | 0.064 | **0.591** |
| 7 | Obedient | −0.149 | **−0.470** | **0.503** |
| 10 | Fidgety | −0.049 | 0.106 | **0.850** |
| 12 | Fights | 0.245 | −0.134 | **0.408** |
| 15 | Distractible | 0.146 | 0.099 | **0.506** |
| 16 | Clingy | **0.349** | 0.011 | **0.384** |
| 18 | Lies | 0.101 | −0.082 | **0.514** |
| 21 | Thinks before acting | −0.132 | **−0.383** | **0.518** |
| 25 | Persistent | −0.011 | **−0.380** | **0.455** |

(*) Factor loadings above 0.30 in bold.

**Table 3 Rotated loading matrix* of EFA with 5 factors (n = 3,253)**

| Nº | Item | F1 | F2 | F3 | F4 | F5 |
|---|---|---|---|---|---|---|
| 12 | Fights | **0.443** | 0.163 | −0.124 | 0.143 | 0.007 |
| 18 | Lies | **0.554** | −0.038 | −0.012 | 0.080 | 0.145 |
| 22 | Steals | **0.802** | −0.060 | −0.189 | −0.223 | −0.081 |
| 23 | Adults | **0.433** | 0.229 | 0.092 | −0.203 | −0.016 |
| 3 | Somatic | 0.034 | **0.473** | 0.095 | −0.015 | −0.066 |
| 6 | Solitary | 0.025 | **0.480** | −0.108 | −0.274 | 0.063 |
| 8 | Worries | −0.011 | **0.679** | −0.024 | −0.022 | −0.085 |
| 13 | Unhappy | 0.019 | **0.778** | −0.156 | 0.034 | −0.089 |
| 16 | Clingy | −0.191 | **0.604** | −0.048 | 0.269 | 0.238 |
| 19 | Bullied | **0.336** | **0.475** | 0.026 | −0.195 | 0.034 |
| 24 | Fears | 0.093 | **0.402** | 0.126 | 0.044 | 0.052 |
| 1 | Considerate | −0.102 | 0.016 | **0.627** | −0.103 | −0.066 |
| 4 | Shares | −0.069 | 0.033 | **0.606** | 0.003 | 0.010 |
| 7 | Obedient | 0.138 | −0.065 | **−0.445** | 0.218 | 0.277 |
| 9 | Caring | 0.056 | 0.112 | **0.802** | −0.072 | −0.013 |
| 11 | Good friend | 0.082 | 0.186 | **−0.752** | −0.116 | −0.122 |
| 14 | Popular | 0.048 | 0.165 | **−0.694** | −0.190 | 0.035 |
| 17 | Kind to kids | −0.043 | −0.053 | **0.835** | −0.085 | 0.116 |
| 20 | Often volunteers to help | 0.133 | 0.072 | **0.644** | −0.014 | −0.211 |
| 2 | Restless | 0.231 | 0.099 | 0.104 | **0.505** | 0.257 |
| 5 | Tempers | 0.234 | 0.258 | 0.021 | **0.380** | 0.116 |
| 10 | Fidgety | **0.305** | 0.069 | 0.084 | **0.500** | 0.271 |
| 15 | Distractible | 0.101 | 0.162 | 0.262 | −0.180 | **0.675** |
| 21 | Thinks before acting | 0.017 | −0.036 | −0.295 | 0.041 | **0.548** |
| 25 | Persistent | −0.012 | 0.019 | −0.217 | −0.243 | **0.790** |

(*) Factor loadings above 0.30 in bold.

**Table 4 Goodness of fit indexes for CFA models built on a random subsample of data (n = 3,253)**

| Factor model | Chi-Square (df) | RMSEA (IC 90%) | GFI | AGFI | CFI | NFI | ECVI |
|---|---|---|---|---|---|---|---|
| Three uncorrelated factors according EFA analysis | 2736.8 (275) | 0.0758 (0.0740–0.0775) | 0.880 | 0.859 | 0.963 | 0.959 | 0.872 |
| Three correlated factors according to the EFA analysis | 2572.1 (272) | 0.0597 (0.0580–0.0615) | 0.933 | 0.920 | 0.965 | 0.961 | 0.823 |
| Five correlated factors according to the EFA analysis | 2255.0 (265) | 0.0589 (0.0571–0.0608) | 0.948 | 0.937 | 0.970 | 0.966 | 0.730 |
| Five correlated factors according to the theoretical model | 2693.3 (265) | 0.0571 (0.0553–0.0589) | 0.927 | 0.911 | 0.963 | 0.959 | 0.865 |

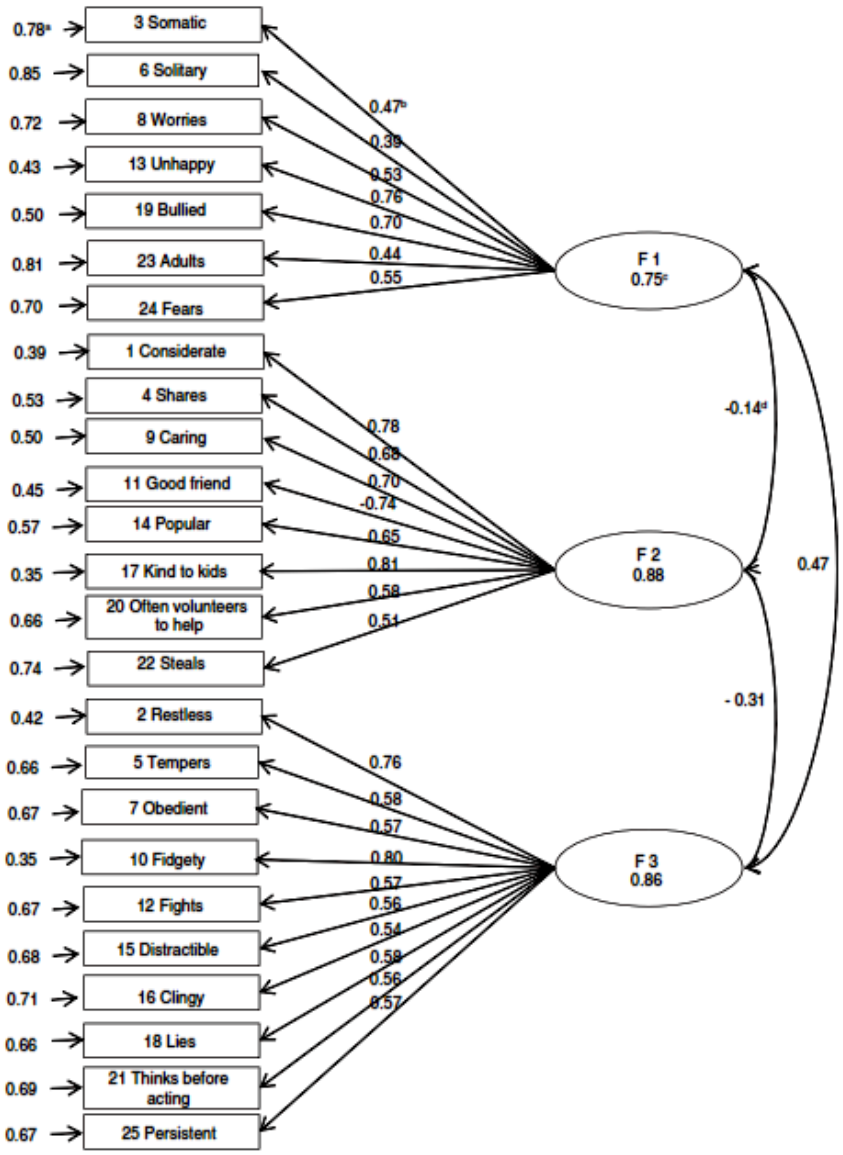



**Figure 1 Result of the CFA with 3 factors according EFA using the second random subsample (n = 3,253). a** Error term of variables; **b** Standardized factor loadings; **c** Reliability; **d** Correlation between factors.

children clinically attended seen in this age range is around one sixth of that in the general population, the corresponding prevalence in the general population would be 5%, a value that agrees with that found in previous epidemiological studies in Spain [29].

The area under the ROC curve of 0.88 for the total difficulties is similar to the mean of 0.87 mentioned by Stone et al. [8] in a review of seven studies.

Considering the Youden J index as an indicator of the efficiency of the questionnaire [30], the cut point corresponding to the highest index (0.64) is 12/13, which indicates a sensitivity of 0.90 and a specificity of 0.74. This cut point is near that proposed by Goodman for the English population (15/16) and much below that proposed by Rodríguez for the population of the Canary Isles (19/20). Several earlier studies provide results on the sensitivity and specificity of the SDQ [8], but none of these surpasses those obtained in the present work.

All the factor reliability coefficients were acceptable (>0.70) except for factor 1 of the five factor solution (0.62). Reliabilities in previous studies are reported as Cronbach's alpha and are generally low, particularly in

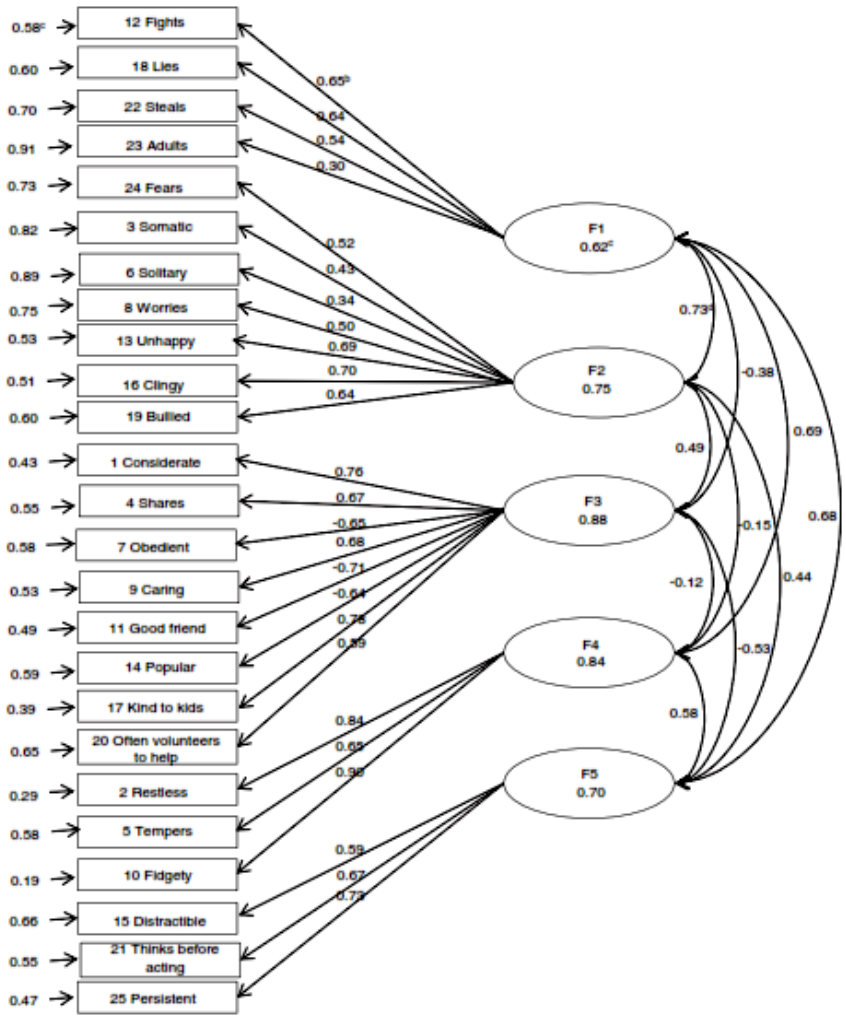



**Figure 2 Result of the CFA with 5 factors according EFA using the second random subsample (n = 3,253). a** Error term of variables; **b** Standardized factor loadings; **c** Reliability; **d** Correlation between factors.

conduct problems and problems with peers [8]. We have not found studies reporting on factor reliabilities. Concerning the use of Cronbach's Alpha, unidimensionality of each scale is not entirely clear and the value of the Cronbach's alpha could not be a good indicator of the internal consistency [31] and this why we use model-based reliabilities after the factor analysis.

Regarding the EFA three factor solution, only "steals" (item 22) loaded < 0.30. There were four items loading >0.30 in more than one factor ("obedient, item 7; clingy, item 16; thinks before acting, item 21 and persistent (item 25). CFA analysis in a different subsample confirmed the validity of this structure, including high factor reliabilities for the three factors.

The first and the third factor could be conceptualized as internalizing and externalizing dimensions respectively, and the third factor is clearly a social dimension. The internalizing dimension consists of four emotional symptoms plus "bullied" (item 19), "gets better with adults" (item 23) and "solitary" (item 6). This combination makes full sense from the clinical point of view. The externalizing dimension comprises all the hyperactivity and conduct problems plus "clingy" (item 16). This cluster is also clinically acceptable except for "clingy" which should belong to the internalizing factor.

The clustering of emotional symptoms under an internalizing dimension, and hyperactivity and behavioural problems under an externalizing one is in keeping with clinical and psychopathological knowledge and it has

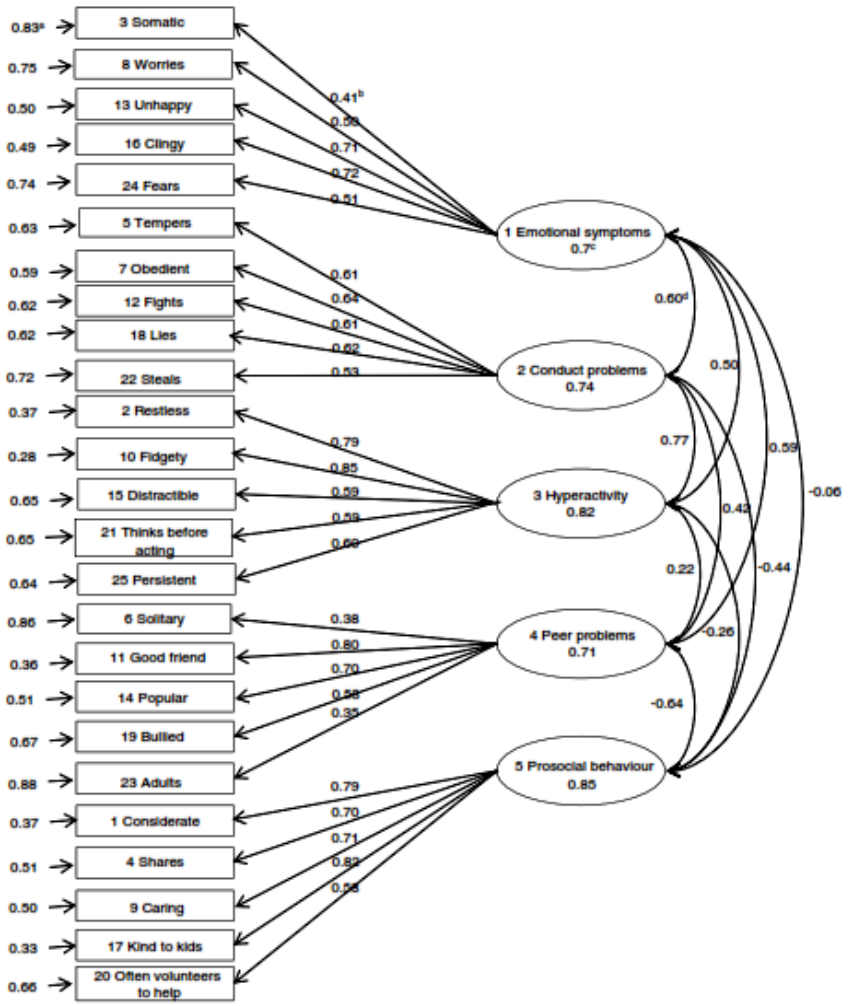



**Figure 3 Result of the CFA with 5 theoretical factors using the second random subsample (n = 3,253). a** Error term of variables; **b** Standardized factor loadings; **c** Reliability; **d** Correlation between factors.

also been verified using other questionnaires. Thus, this is not a new finding but reinforces the validity of the SDQ by proving that it is in line with established psychopathological knowledge.

The second factor covers the five prosocial behaviour items, plus "good friend" (item 11) and "popular" (item 14), which in theory should belong to the peer problems dimension. These seven items constitute a meaningful combination of social items. However it should not be overlooked that three items which have higher factor loadings in the first and third factor, also load over 0.30 on this second factor. Taken together these ten items represent those added by Goodman to reflect strengths. Therefore it may well be that this factor is a method artefact, as noted by some authors [11]. Nevertheless, from our point of view there is not enough evidence to discard the social factor as an artefact. Prosocial behaviour and peer relationships are the bases of social capital and social capital plays an important role in social cohesion and in individual and public health [32]. Being such an important issue we think further research is warranted in establishing the validity of this dimension.

A three factor solution has been described in four previous studies [12-15]. Three of these studies reported a distribution of items identical to ours. However, Goodman [17] confirmed the validity of a somewhat different model: an internalizing dimension including all emotional and peer difficulties, an externalizing dimension including hyperactivity and conduct problems, and the prosocial behaviour.

In the EFA five factor model all items loaded >0.3 in only one factor, except for "bullied" (item 19) and "fidgety" (item 10). Factor reliabilities were acceptable. This model provides a solution similar to the theoretical structure originally proposed in several respects. The first and second factors include behavioural problems and emotional symptoms respectively as expected, except that "tempers" (item 5) is not included in behavioural problems and "bullied" and "solitary" (items 19 and 6) load in emotional symptoms. There are also important differences. "Good friend" and "popular" from the peer difficulties dimension (items 11 and 14) and prosocial behaviour cluster together, as in the three factor solution, making up again a meaningful social dimension. Finally the hyperactivity scale splits in two factors, hyperactivity and inattention, which is compatible with our current psychopathological understanding of Attention Deficit Hyperactivity Disorder. However, in spite of these discrepancies between the expected and the empirical model, the CFA confirmed the validity of the EFA five factor structure as well as the theoretical structure with good goodness of fit indexes.

The validity of the five factor model has been supported by the majority of previous SDQ studies. Those using EFA report different but closely similar distribution of items within the five factor structure. This is not surprising considering the different cultures where it has been tested and the use of parent, teacher or self-report questionnaires. Out of 18 studies reviewed by Stone [8], eight applied CFA and in five of them the five factor structure was supported using the parent version.

Finally, we may ask which of the two factor models is better. According to our estimations both models fit the data. Only two previous studies have also confirmed both models [14,15]. Whether these two models may have different applications in different circumstances or whether they reflect culture-dependent solutions is an open question. Goodman [15] gives some evidence to support the use of the externalization/internalization dimensions to screen for difficulties when surveying low prevalence populations. Essau [14] finds that the number of factors is dependent on the country where the survey has been carried out.

This study has some strengths and limitations. First, it is necessary to bear in mind the diagnostic criteria to define the result variable (case/non case) is very demanding, and it could not be comparable to a diagnostic interview.

The use of weightings corresponding to the sample subjects guarantee that the estimations are unbiased, although the random error could be underestimated. However, the size of the samples used both for the EFA and for the CFA was very high, and so we estimated models based on much evidence.

On the other hand, the use of polychoric correlation matrixes for the estimation of the factor analysis models resulted efficient and made it possible to incorporate both the weightings of the sample subjects and the ordinal metric of the items.

Even though the criteria to estimate CFA models was based on the assignment of each factor to the item with the highest factor loading in the corresponding EFA (three or five-factor models), the CFA made it possible to qualify the factor structures proposed as acceptable. The modification indices suggest the possible presence of some high correlations between certain items that could improve the fit of CFA models.

## Conclusions

The diagnostic behaviour of the SDQ in the Spanish population is within the working limits described in other countries. According to the results obtained in this study, the diagnostic efficiency of the questionnaire is adequate to identify probable cases of psychiatric disorders in low prevalence populations. Regarding the factorial structure we found that both the five and the three factor models fit the data with acceptable goodness of fit indexes, the latter including an externalization and internalization dimension and perhaps a meaningful positive social dimension.

Accordingly, we recommend studying whether these differences depend on sociocultural factors or are, in fact, due to methodological questions.

### Abbreviations

SDQ: Strengths and difficulties questionnaire; EFA: Exploratory factorial analysis; CFA: Confirmatory factorial analysis; SNHS: Spanish national health survey.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

The study was conceived and designed collectively by all the authors. MGB and AN coordinated and conducted the analysis and wrote the first version of the manuscript. All the authors contributed equally to the analysis of the data, interpretation, and discussion of results. All the authors have read and approved the final version.

### Acknowledgements

We would like to thank Dr. Urbano Lorenzo-Seva (Department of Psychology at Rovira i Virgili University, Spain) for his comments and methodological suggestions.

### Author details

[1]Teaching Unit of Psychiatry and Psychological Medicine, Department of Medicine, University of Valencia, Valencia, Spain. [2]CIBERSAM, Instituto de Salud Carlos III, Madrid, Spain. [3]Research Unit for the Analysis of Mortality and Health Statistics, University of Alicante, Alicante, Spain. [4]Burriana Center of Mental Health, Burriana, Spain. [5]Paterna Center of Mental Health, Paterna, Spain. [6]CIBERSAM, Instituto de Salud Carlos III, Madrid, Spain. [7]Department of Clinical Medicine, Universitat Miguel Hernández, Sant Joan d'Alacant, Spain.

**Received: 4 May 2012 Accepted: 15 March 2013**
**Published: 22 March 2013**

**References**

1. Ministerio de Sanidad Política Social e Igualdad: *Encuesta Nacional de Salud 2006. Descripción de variables compuestas*. http://www.msssi.gob.es/estadEstudios/estadisticas/encuestaNacional/encuesta2006.htm.
2. Ministerio de Sanidad Política Social e Igualdad: *Encuesta Nacional de Salud 2006. Metodología detallada*. http://www.msssi.gob.es/estadEstudios/estadisticas/encuestaNacional/encuestaNac2006/metodENS2006.pdf.
3. Goodman R, Ford T, Simmons H, Gatward R, Meltzer H: **Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample.** *Br J Psychiatry* 2000, **177:**534–539.
4. Goodman R: **Psychometric Properties of the Strengths and Difficulties Questionnaire.** *J Am Acad Child Adolesc Psychiatry* 2001, **40:**1337–1345.
5. Becker A, Steinhausen H, Baldursson G, Dalsgaard S, Lorenzo M, Ralston S, Döpfner M, Rothenberger A: **Psychopathological screening of children with ADHD: Strengths and Difficulties Questionnaire in a pan- European study.** *Eur Child Adolesc Psychiatry* 2006, **15:**56–62.
6. Public Health Institute in Finland, Scientific Institute of Public Health Brussels: *European Health Survey Information Database (EUHSID project)*. https://hishes.iph.fgov.be/index.php?hishes=home.
7. Goodman A, Heiervang E, Fleitlich-Bilyk B, Alyahri A, Patel V, Mullick M, Slobodskaya H, Neves Dos Santos D, Goodman R: **Cross-national differences in questionnaires do not necessarily reflect comparable differences in disorder prevalence.** *Soc Psychiatry Psychiatr Epidemiol* 2012, **47:**1321–1331.
8. Stone L, Otten R, Engels R, Vermulst A, Janssens J: **Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4 to 12 years-old: A review.** *Clin Child Fam Psychol Rev* 2010, **13:**254–274.
9. Hill C, Hughes J: **An examination of the convergent and discriminant validity of the Strengths and Difficulties Questionnaire.** *School Psychol Quart* 2007, **22:**380–406.
10. Tsang K, Wong P, Lo SK: **Assessing psychosocial well-being of adolescents: a systematic review of measuring instruments.** *Child Care Health Dev* 2012, **38:**629–646.
11. Palmieri P, Smith G: **Examining the structural validity of the Strengths and Difficulties Questionnaire (SDQ) in a U.S. sample of custodial grandmothers.** *Psychol Assess* 2007, **19:**189–98.
12. Dickey W, Blumberg S: **Revisiting the factor structure of the Strengths and Difficulties Questionnaire.** *J Am Acad Child Adolesc Psychiatry* 2004, **43:**1159–1167.
13. Di Riso D, Salcuni S, Chessa D, Raudino A, Lis A, Altoè G: **The Strengths and Difficulties Questionnaire (SDQ). Early evidence of its reliability and validity in a community sample of Italian children.** *Pers Indiv Differ* 2010, **49:**570–575.
14. Essau C, Olaya B, Anastassiou-Hadjicharalambous X, Pauli G, Gilvarry D, Bray D, O'callaghan J, Ollendick T: **Psychometric properties of the Strength and Difficulties Questionnaire from five European countries.** *Int J Methods Psychiatr Res* 2012, **21:**232–245.
15. Goodman A, Lamping DL, Ploubidis GB: **When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British parents, teachers and children.** *J Abnorm Child Psychol* 2010, **38:**1179–1191.
16. Muris P, Meesters C, Eijkelenboom A, Vincken M: **The self-report version of the Strengths and Difficulties Questionnaire: Its psychometric properties in 8- to 13-year-old non-clinical children.** *Br J Clin Psychol* 2004, **43:**437–448.
17. Goodman R: **A modified version of the Rutter Parent Questionnaire including extra items on children's strengths: A research note.** *J Child Psychol Psychiatry* 1994, **35:**1483–1494.
18. Van Roy B, Veenstra M, Clench-Aas J: **Construct validity of the five-factor Strengths and Difficulties Questionnaire (SDQ) in pre-, early, and late adolescence.** *J Child Psychol Psychiatry* 2008, **49:**1304–1312.
19. García P, Goodman R, Mazaria J, Torres A, Rodríguez-Sacristán J, Hervás A: **El Cuestionario de Capacidades y Dificultades.** *Rev Psiquiatr Infanto-Juv* 2000, **1:**12–17.
20. Rodríguez Hernández PJ: **Estudio de la prevalencia de los trastornos psiquiátricos infantiles en la Comunidad Autónoma Canaria**. In *PhD thesis*. Universidad de la Laguna, Departamento de Medicina Interna, Dermatología y Psiquiatría; 2006.
21. Rodríguez Hernández PJ, Herreros O: *Historia clínica, evaluación y diagnóstico en Psiquiatría Infantil. Curso de Formación Continuada en Psiquiatría Infantil.* Sociedad Española de Pediatría Extrahospitalaria y Atención Primaria-SEPEAP. http://www.sepeap.org/imagenes/secciones/Image/_USER_/Ps_inf_Historia_clinica_evaluacion_diagnostico.pdf.
22. Ministerio de Sanidad Política Social e Igualdad: *Encuesta Nacional de Salud 2006.* Cuestionario de Menores. http://www.msssi.gob.es/estadEstudios/estadisticas/encuestaNacional/encuestaNac2006/ENS_06_Menores_definitivo.pdf.
23. Lorenzo-Seva U, Ferrando PJ: **FACTOR: A computer program to fit the exploratory factor analysis model.** *Behav Res Methods Instrum Comput* 2006, **38:**88–91.
24. Lorenzo-Seva U, Ferrando PJ: *FACTOR v 8.1*. Rovira i Virgili University; 2012. http://psico.fcep.urv.es/utilitats/factor/Download.html.
25. Jöreskog KG, Sörbom D: *LISREL 8.80 for Windows*. Lincolnwood, IL: Scientific Software International, Inc; 2006.
26. Timmerman M, Lorenzo-Seva U: **Dimensionality assessment of Ordered Polytomous Items with Parallel Analysis.** *Psychol Methods* 2011, **16:**209–220.
27. Lorenzo-Seva U: **Promin: A method for oblique factor rotation.** *Multivariate Behav Res* 1999, **34:**347–356.
28. McDonald R: *Test theory. A unified treatment*. Erlbaum Associates: Mahwah, NJ; 1999.
29. Gomez-Beneyto M, Bonet A, Catala M, Puche E, Vila V: **Prevalence of mental disorders among children in Valencia, Spain.** *Acta Psychiatr Scand* 1994, **89:**352–357.
30. Indrayan A: *ROC Curve. Biostatistics Resource for Medical, Health and Allied Research*. http://www.medicalbiostatistics.com/ROCCurve.pdf.
31. Cortina J: **What is coefficient Alpha? An examination of theory and applications.** *J Appl Psychol* 1993, **78:**98–104.
32. Rocco L, Suhrcke M: *Is social capital good for health? A European perspective*. Copenhagen: WHO Regional Office for Europe; 2012.



**doi:10.1186/1471-244X-13-95**
**Cite this article as:** Gómez-Beneyto *et al.*: **Psychometric behaviour of the strengths and difficulties questionnaire (SDQ) in the Spanish national health survey 2006.** *BMC Psychiatry* 2013 **13**:95.




Índice

# ANÁLISIS FACTORIAL

Dr. Cs. Ruiz Aranibar, Gustavo

✉ gustavoruiz432@hotmail.com

## RESUMEN

El Análisis Factorial (AF) es una técnica de reducción de datos, que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables. Estos grupos homogéneos se forman con las variables que se correlacionan mucho entre si y procurando, inicialmente, que unos grupos sean independientes de otros.

El AF es una técnica de reducción de dimensionalidad de los datos; su propósito final, consiste en buscar un número mínimo de dimensiones capaces de explicar el máximo de información contenida en los datos. A diferencia de lo que ocurre en otras técnicas como el análisis de varianza o el de regresión, en el AF todas las variables del análisis cumplen el mismo papel, todas ellas son independientes en el sentido de que no existe a priori una dependencia conceptual de unas variables sobre otras.

## PALABRAS CLAVE

*Análisis factorial, factores, correlación, desviación estándar, valor propio, vector propio, variable aleatoria, varianza común, denormalizacion, iterar, rotación.*

## 1. INTRODUCCIÓN

El Análisis Factorial (AF) es usada en ciencias sociales, mercadotecnia, gestión de producto, investigación operativa, medicina y otras ciencias aplicadas que tratan con grandes cantidades de datos. El objetivo, es explicar la mayor parte de la variabilidad entre varias variables aleatorias observables, en términos de un número menor de variables aleatorias no observables llamadas factores. Las variables aleatorias no observables se modelan como combinaciones lineales de los factores más términos de errores.

El AF es un método del análisis estadístico múltiple, utilizado para el estudio e interpretación de las correlaciones entre un grupo de variables, partiendo de la idea de que dichas correlaciones no son aleatorias, sino que se deben a la existencia de factores comunes entre ellas y el objetivo del AF es la identificación y cuantificación de dichos factores comunes.

De una manera general, el AF se puede aplicar cuando existen relaciones entre las variables, es decir para explicar las relaciones con la ayuda de un número limitado de factores independientes, estando basada en el análisis de correlaciones existentes entre las variables medidas sobre un grupo de “n” individuos u objetos.

Las variables observables deben ser homogéneas y correlacionadas, sin que alguna predomine sobre las demás. La información estadística en el AF es de carácter multidimensional, por lo tanto, la geometría, el cálculo matricial, el álgebra lineal y la estadística son fundamentales en el desarrollo del presente trabajo.

## 2. BREVE HISTORIA DE LOS PRECURSORES DEL AF.

**Jacobi Carl Gustav**, (Potsdam, actual Alemania, 1804-Berlín, 1851) Matemático

alemán. Hijo de una familia de banqueros de origen judío, estudió en la Universidad de Berlín, donde se doctoró en 1825. Convertido al cristianismo, tuvo oportunidad de acceder a un puesto de profesor en la Universidad de Königsberg. Destacadísimo pedagogo, influyó en numerosas generaciones posteriores de matemáticos alemanes. Sus trabajos más relevantes se produjeron en el campo del álgebra matricial y lineal, en el que introdujo y desarrolló el concepto de determinante, aplicándolo asimismo al estudio de las funciones de variables múltiples. Entre 1826 y 1827 estableció, independientemente del noruego Niels Henrik Abel, los principios fundamentales de la teoría de las funciones elípticas. En el ámbito de la teoría de números, demostró el teorema de Bachet sobre el total de las descomposiciones posibles de un entero, y en el de la mecánica física, trató con profundidad y rigor el problema de los tres cuerpos. Su obra más notable es sobre la formación y propiedades de los determinantes (1841).

**Charles Edward Spearman**. (Londres, nació el día 10 de septiembre de 1863, falleció el 7 de septiembre de 1945). Psicólogo de profesión, estudió Estadística y logró desarrollar notables aplicaciones de la estadística en el campo de la Psicología, desarrollando el AF. Propuso la existencia de un factor general de inteligencia (Factor g) que subyace a las habilidades para la ejecución de las tareas intelectuales. Estudió en las universidades de Leipzig, donde alcanzó el título cuando tenía 40 años. Además, cursó estudios en Wurzburgo y Göttingen y enseñó e investigó en la Universidad de Londres (1907 – 1931). El AF tiene su origen en las investigaciones realizadas con los trabajos en Psicología, (1904).

Creó y desarrollo la metodología de los llamados experimentos factoriales para la estadística, que son aquellos experimentos en los que se estudia simultáneamente dos o más factores, y donde los tratamientos se forman por la combinación de los diferentes niveles de cada uno de los factores.

Por todo esto, es considerado uno de los grandes estadisticos de todos los tiempos. Su método, inscrito en las matemáticas experimentales, estudia las dimensiones del campo empírico. Sus aportes metodológicos, no sólo se han constituido en herramientas fundamentales para algunos ámbitos de la psicología, sino que son instrumentos para la ciencia estadística. El desarrollo de Spearman es útil en todas las ciencias sociales que requieran de técnicas de estadística correlacional para poder interpretar la información.

**Thurstone Louis León.** (Chicago, 1887 - Chapel Hill, 1955) Psicólogo estadounidense. Se doctoró en la Universidad de Chicago, donde dio clases la mayor parte de su vida. Especializado en Psicometría, desarrolló nuevas técnicas para medir las cualidades mentales. Realizó y publicó varias escalas de actitud que pretendían medir la influencia de la propaganda en los prejuicios del hombre; también se interesó por la medición del aprendizaje e intentó expresar a través de unidades absolutas el desarrollo del aprendizaje.

Fue uno de los primeros en utilizar el AF para medir la inteligencia; es por su trabajo sobre AF por lo que es más conocido Thurstone, ya que él y sus seguidores lo aplicaron a múltiples problemas, como el análisis de las capacidades perceptivas humanas o el desarrollo de nuevos tests de aptitudes. Se interesó también por las características de la personalidad y publicó un test de tendencias Psiconeuróticas.

En la evolución de la Psicometría, y en particular en el campo de la medición de la inteligencia, hay que destacar las aportaciones de Charles Spearman, que propuso en 1927 la distinción entre un "factor g" (factor general), común a todas las pruebas de medición y presente en cualquier tarea intelectual, y un "factor s" (específico), asociado a cada operación en particular. Thurstone dio un paso más en la distinción de Spearman al identificar en 1934, con la ayuda de las técnicas estadísticas de AF, siete aptitudes primarias incluidas en la inteligencia: comprensión verbal, fluidez verbal, aptitud numérica, visualización espacial, velocidad perceptiva, memoria y razonamiento. Ello llevó a Thurstone a concebir la inteligencia como una combinación de varias capacidades distintivas; de este modo, el factor general "g" debe entenderse como factor secundario, detectable únicamente gracias a las correlaciones entre las aptitudes primarias.

La técnica para la medición conocida como AF múltiple, que puede manejar varios factores de capacidad simultáneamente, se convirtió a partir del impulso que le dio Thurstone en un potente instrumento de análisis estadístico aplicado a la investigación psicológica, cuyas repercusiones se extienden hasta nuestros días. Sus trabajos sobre AF pudieron aplicarse a múltiples problemas, como el análisis de las capacidades perceptivas humanas o el desarrollo de nuevos tests de aptitudes. Se interesó por las características de la personalidad y elaboró un test de tendencias psiconeuróticas. Fundador y director de la revista Psicometrika, entre sus obras destacan The nature of intelligence (1924) y Vectors of the mind (1935).

**Karl Pearson.** (27 de marzo de 1857, 27 de abril de 1936). Es uno de los fundadores de la estadística, desarrollo en el ámbito de la investigación, las ideas de Galton F. sobre regresión y correlación. Hijo de una familia acomodada estudió en la University College School y posteriormente en Cambridge especializándose en matemáticas. Años más tarde viajó a Berlín y Heidelberg donde estudió Literatura alemana medieval, destacando en esta área y colaborando en Cambridge en los estudios germánicos. Estudió también abogacía en la asociación profesional de Inner Temple.

Karl Pearson fue uno de los miembros fundadores de la "Escuela Biométrica", que estudiaba la aplicación de la Estadística en materias como la biología, campo en el que Pearson poseía diversos conocimientos que sirvieron en futuras generaciones de investigadores y siempre será recordado por sus teorías positivistas radicales, su investigación estadística en materia de biología y por ser el fundador de la bioestadística.

En Julio de 1900, una de las más importantes contribuciones de Pearson a la Estadística fue presentada en la publicación de un artículo. Esta contribución era el Test de la $\chi 2$. Pearson usó esta fórmula para obtener la distribución muestral de $\chi 2$ en grandes muestras, las cuales estaba particularmente interesado en estudiar, como una función de k, la cual resultó ser una forma especial de la distribución de Pearson tipo 3, ahora conocida como "distribución $\chi 2$ para K-1 grados de libertad". Además, daba una pequeña tabla de la integral de la distribución de $\chi 2$ desde 1 a 70 y para k desde 3 a 20. Este test de la $\chi 2$ de bondad del ajuste es una de las mayores y más útiles contribuciones de Pearson a los tests estadísticos.

Tras el reconocimiento por su trabajo sobre curvas de distribución, Pearson continuó recibiendo reconocimientos y honores. En 1893, comenzó su serie de 18 artículos

titulados "Mathematical Contributions to the Theory of Evolution", que contendrían parte de su trabajo más valioso. El mismo año que empezó estos artículos, Pearson acuñó el término "desviación estándar". Entre 1906 y 1914 Pearson estuvo consagrado al desarrollo de un centro de postgrado para promover el desarrollo de la estadística como una rama de las matemáticas aplicadas. En el verano de 1933, tras una larga vida consagrada al avance estadístico, Pearson abandonó su trabajo en la Universidad.

El hecho de que tras la retirada de Pearson el departamento de estadística aplicada fuera dividido en dos unidades independientes muestra el importante trabajo soportado por Pearson. Incluso después de la muerte de Karl Pearson en 1936, su apellido continúa siendo uno de los más destacados en el campo de las matemáticas. No hay duda de que las contribuciones de Pearson a lo largo de su vida consolidaron la Estadística como una disciplina por derecho propio.

## 3. CONCEPTOS MATEMÁTICO-ESTADÍSTICOS NECESARIOS EN EL AF.

El análisis estadístico requiere del instrumento matemático, por esto los datos deben presentarse en una forma susceptible de tratamiento matemático de ahí que la aplicación del AF, requiere en principio que se disponga de una matriz de información de datos cuantitativos *X*, de orden *n*m*, (*n*=observaciones y *m*=variables), *n* > *m*:

Matriz de datos observados:

$$X = X_{ij} = \begin{bmatrix} x_{11} & \cdots & x_{1j} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} \end{bmatrix} \quad \begin{matrix} i = 1,2,\dots,n \\ \mathrm{j} = 1,2,\dots,\mathrm{m} \end{matrix}$$

A partir de esta información, se tiene que aplicar las siguientes ecuaciones que son utilizadas para determinar los estadísticos de: los promedios, desviaciones estándar, suma de los productos cruzados de desviaciones de las medias y los coeficientes de correlación. Promedios de cada columna:

$$\overline{X_j} = \frac{\sum_{i=1}^{m} X_{ij}}{n} \qquad j = 1,2,\dots,m$$

Media temporal para obtener un cálculo más exacto de :

$$T_j = \frac{\sum_{i=1}^{m} X_{ij}}{m}$$

Suma de los productos cruzados de las desviaciones:

$$S_{jk} = \sum_{i=1}^{n} (X_{ij} - T_j)(X_{ik} - T_k) - \frac{\sum_{i=1}^{n}(X_{ij} - T_j)\sum_{i=1}^{n}(X_{ik} - T_k)}{n}$$

$$i = 1,2,\dots,n \quad , \quad j = 1,2,\dots,m \quad , \quad k = 1,2,\dots,m$$

Coeficientes de correlación y su dominio:

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}} \qquad -1 \le r_{jk} \le +1$$

$$j = 1,2,\dots,m \quad y \quad k = 1,2,\dots,m$$

Desviaciones estándar:

$$S_j = \frac{\sqrt{S_{jj}}}{\sqrt{n-1}} \qquad j = 1,2,\dots,m$$

**Matriz identidad.**

Una matriz de identidad es usada como una aproximación a R (matriz de correlación).

$$V_{ij} = I \qquad \begin{matrix} \forall\, i = j \;\; V_{ij} = 1 \\ \forall\, i \neq j \; V_{ij} = 0 \end{matrix} \quad y$$

**Valores propios**[1].

Se llama valor propio de una matriz simétrica A de orden n, las n soluciones de la ecuación:

$$|A - \lambda I| = 0 \;\; \lambda_1, \lambda_2, \ldots, \lambda_n$$

El desarrollo de este determinante, da lugar en efecto al nacimiento a una ecuación de grado *n* en $\lambda$, llamada ecuación característica:

$$|A - \lambda I| = (-\lambda)^n + c_{n-1}(-\lambda)^{n-1} + \cdots + c_1(-\lambda) + c_0 = 0$$

Esta ecuación posee de una manera general *n* raíces. En particular, el coeficiente $c_{n-1}$ es la suma de los elementos de la diagonal de la matriz inicial A, llamada traza de la matriz, mientras que el término independiente $c_0$ es el valor del determinante de A:

$$c_{n-1} = tr\, A = \sum_{i=1}^{n} a_{ii} \qquad c_0 = |A|$$

De una manera general, las raíces de la ecuación característica pueden ser reales o complejas, positivas, nulas o negativas, distintas o diferentes. En el caso de matrices simétricas, que interesan en forma particular, deben ser todas reales, teniéndose entonces:

$$\sum_{i=1}^{n} \lambda_i = tr\, A \qquad \prod_{i=1}^{n} \lambda_i = |A|$$

La determinación de los valores propios puede estar basado en el algoritmo de Rutishauer, que consiste en descomponer la matriz A dada, como producto de dos matrices B y H, tal que B es una matriz triangular inferior con todos los elementos de la diagonal principal igual a 1, y H es una matriz triangular superior. Por un proceso iterativo se consigue transformar la matriz A en una matriz triangular superior tal que, los valores propios buscados son los elementos de la diagonal principal, existiendo para esta determinación también otros métodos.

La determinación de los valores propios puede estar basado en el algoritmo de Rutishauer, que consiste en descomponer la matriz A dada, como producto de dos matrices B y H, tal que B es una matriz triangular inferior con todos los elementos de la diagonal principal igual a 1, y H es una matriz triangular superior. Por un proceso iterativo se consigue transformar la matriz A en una matriz triangular superior tal que, los valores propios buscados son los elementos de la diagonal principal, existiendo para esta determinación también otros métodos.

**Vectores propios**[2]

Sea A una matriz de orden n sobre un cuerpo K. Un escalar λ se denomina un valor propio de A si existe un vector (columna) no nulo v, es decir:

$$\lambda \in K \;\; v \in K^n \;\; Av = \lambda v$$

Todo vector que satisfaga esta relación se llama un vector propio de A perteneciente al valor propio $\lambda_i$.

Norma inicial:

$$v_I = 0$$

$$v_I = \left( \sum_{i \le k} 2 * A_{ik}^2 \right)^{1/2}$$

Norma final calculada en cada etapa de ese dominio:

$$v_F = \frac{v_I * 10^{-6}}{N}$$

1 Llamados, autovalores o valores característicos. En ingles: laten root, characteristic root o eigenvalue. En francés: valeurs propres o valeurs caractéristiques.

2 Llamados, autovectores o vectores característicos. En ingles: laten vector, characteristic vector o eigenvector. En francés: vecteurs propes o vecteurs caractéristiques.

Esta norma final es un conjunto suficientemente pequeña que requiere que cualquier elemento fuera de la diagonal debe ser más pequeño que $V_F$ en magnitud absoluta define la convergencia del proceso.

Indicador inicializado. Este indicador posteriormente es utilizado para determinar cuándo cualquier elemento fuera de la diagonal ha sido encontrado, que son más grandes que en el presente inicio.

Cada elemento fuera de la diagonal es seleccionado en etapas y una transformación es ejecutada para eliminar el elemento fuera de la diagonal (pivote), como se muestra en las siguientes ecuaciones:

$$\lambda = -A_{1m}$$

$$\mu = \frac{A_{11} - A_{mm}}{2}$$

$$\omega = signo\ (\mu)\frac{\lambda}{\sqrt{\lambda^2 + \mu^2}}$$

$$sen\ \theta = \frac{\omega}{\sqrt{2*(1+\sqrt{1-\omega^2})}}$$

$$\cos\ \theta = \sqrt{1 - sen^2\theta}$$

$$B = A_{i1} * \cos\ \theta - A_{im} * sen\ \theta$$

$$C = A_{i1} * sen\ \theta - A_{im} * \cos\ \theta$$

$$B = R_{i1} * \cos\ \theta - R_{im} * sen\ \theta$$

$$R_{im} = R_{i1} * sen\ \theta + R_{im} * \cos\ \theta$$

$$R_{i1} = B$$

$$A_{11} = A_{i1} * \cos^2\theta + A_{mm} * sen^2\theta - 2 * A_{im} * sen\theta * \cos\theta$$

$$A_{mm} = A_{11} * sen^2\theta + A_{mm} * \cos^2\theta + 2 * A_{im} * sen\theta * \cos\theta$$

$$A_{im} = (A_{11} - A_{mm}) * \sin\theta * \cos\theta + A_{im} * (\cos^2\theta - sen^2\theta)$$

Los cálculos anteriores son repetidos hasta que todos los elementos del pivote sean menores que en el comienzo.

El método de diagonalización fue realizado por Jacobi y adaptado por von Neumann para las computadoras.

Se encuentra el valor de *k*, el número de valores propios que son más grandes o iguales que la constante especificada los valores propios determinados son arreglados en orden descendiente.

El porcentaje acumulado de esos *k* valores propios son:

$$d_j = \sum_{i=1}^{j} \frac{\lambda_i}{m} \qquad j = 1,\ 2,.....,\ k$$

*m* = número de valores propios (o variables) y $k \le m$

Se calcula los coeficientes de cada factor, multiplicando los elementos de los vectores propios normalizados por la raíz cuadrada de los correspondientes valores propios, es decir se determina una matriz factor de los valores propios asociada a los vectores propios:

$$a_{ij} = v_{ij} * \sqrt{\lambda_j}$$

$i = 1,\ 2,.....,\ m$ son las variables

$j = 1,\ 2,.....,\ k$ son los valores propios retenidos $k \le m$

Posteriormente se ejecuta la rotación ortogonal de la matriz factor de orden m por k, tal que:

$$\sum_{j}^{k}\left( m\sum_{i}^{m}\left(\frac{a_{ij}^2}{h_i^2}\right)^2 - \left(\sum_{i}^{m}\left(\frac{a_{ij}^2}{h_i^2}\right)\right)\right)^2$$

$i = 1,\ 2,.....,\ m$ son las variables

$j = 1,\ 2,.....,\ k$ son los factores

$a_{ij}$ es la matriz retenida o saturada, para la i-iem variable del j-ime factor

$h_i^2$ es la varianza común j-ime variable definida a continuación

**Varianza común**[3].

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \qquad i = 1,\ 2,.....,\ m$$

Matriz factor normalizada:

$$b_{ij} = \frac{a_{ij}}{\sqrt{h_i^2}} \qquad i = 1,\ 2,.....,\ m$$
$$j = 1,\ 2,.....,\ k$$

Varianza para la matriz factor:

$$V_c = \sum_{j=1}^{k} \frac{m\sum_{i=1}^{m}\left(b_{ij}^2\right)^2 - \sum_{i=1}^{m}\left(b_{ij}^2\right)^2}{m^2}$$

$c = 1,\ 2,.....$ (ciclo de iteración)

La prueba de convergencia se ejecuta para cuatro operaciones sucesivas, se detiene la rotación y se ejecuta la ecuación de denormalización:

Si $V_c - V_{c-1} \le 10^{-7}$

De otra manera se repite la rotación de factores hasta que la prueba de convergencia sea satisfecha.

Rotación de dos factores: Se rota dos factores normalizados $b_{ij}$ al mismo tiempo:

1 con 2, 1 con 3, …., 1 con k, 2 con 3,…..,2 con k,….,k-1 con k. Esto constituye un ciclo de iteración.

Asumiendo que $x$ y $y$ son factores a ser rotados, donde $x$ es el número más bajo o el factor del lado izquierda, y siguiendo la notación para la rotación estos dos factores son usados:

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ ..... & \\ ..... & \\ x_m & y_m \end{bmatrix} * \begin{bmatrix} \cos\ \phi & -sen\ \phi \\ sen\ \phi & \cos\ \phi \end{bmatrix} = \begin{bmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ ..... & \\ ..... & \\ X_m & Y_m \end{bmatrix}$$

Donde $x_i$ y $y_i$ son los valores almacenados que serán utilizables, y $X_i$ y $Y_i$ son los valores normalizados deseados, que son funciones de $\phi$ , el ángulo de rotación.

Las etapas de cálculo son las siguientes:

Cálculo de NUM y DEN

$$A = \sum_{i=1}^{m}(x_i + y_i)(x_i - y_i)$$

$$B = 2\sum_{i=1}^{m} x_i y_i$$

$$C = \sum_{i=1}^{m}\left((x_i + y_i)(x_i - y_i) + 2x_i y_i\right)\left((x_i + y_i)(x_i - y_i) - 2x_i y_i\right)$$

$$D = 4\sum_{i=1}^{m}(x_i + y_i)(x_i - y_i)x_i y_i$$

$$NUM = D - \frac{2AB}{m}$$

$$DEN = C - \frac{(A+B)(A-B)}{m}$$

Comparación de NUM y DEN:

Los cuatro casos siguientes se producen:

NUM < DEN determinar B1

NUM > DEN determinar B2

3 En inglés: communality. En Francés: varianza commune.

(NUM + DEN) $\geq \varepsilon^*$ determinar B3

(NUM + DEN) $= \varepsilon$ saltar a la próxima rotación

$\varepsilon^*$ es un factor de tolerancia pequeño

B1: $tg\ 4\theta = |\ NUM\ |/|\ DEN\ |$

$si\ tg\ 4\theta$

i) Si tg $4\theta < \varepsilon$ y DEN es positivo, salta a la próxima rotación

ii) DEN es negativo entonces:

$$\cos\ \theta = sen\ \theta = \frac{\sqrt{2}}{2}$$

Luego pasa a calcular E

$si\ tg\ 4\theta \geq \varepsilon$

Se calcula:

$$\cos\theta = \frac{1}{\sqrt{1+tg^2 4\theta}}$$

$$\text{sen}\,\theta = tg4\theta^* \cos 4\theta$$

y se calcula C

$B2: ctg\ 4\theta = |NUM|/|DEN|$

i) Si $ctg\ 4\theta < \varepsilon$
$\cos 4\theta = 0\ \ y\ \ sen\ 4\theta = 1$
se pasa a calcular

Si $ctg\ 4\theta \geq \varepsilon$ se calcula:

$$sen\ 4\theta = \frac{1}{\sqrt{1+ctg^2 4\theta}}$$

$$\cos 4\theta = ctg4\theta * sen\ 4\theta$$

se calcula C

$$B3: cos\ 4\theta = sen\ 4\theta = \frac{\sqrt{2}}{2}$$

se calcula C

C: Determinación de $\cos\theta$ y $\text{sen}\,\theta$

$$\cos 2\theta = \sqrt{\frac{1+\cos 4\theta}{2}}$$

$$\text{sen}\,2\theta = \frac{\text{sen}\,4\theta}{2\ \cos 2\theta}$$

$$\cos\theta = \sqrt{\frac{1+\cos 2\theta}{2}}$$

$$\text{sen}\,\theta = \frac{\text{sen}\,2\theta}{2\ \cos\theta}$$

D: Determinación de $\cos\phi$ y $\text{sen}\,\phi$

D1. Si DEN es positivo:

$$\cos\phi = \cos\theta$$

$$\text{sen}\,\phi = \text{sen}\,\theta$$

Ir a D2

Si DEN es negativo, calcular:

$$\cos\phi = \sqrt{\frac{2}{2}}\cos\theta + \sqrt{\frac{2}{2}}\,\text{sen}\,\theta$$

$$\text{sen}\,\phi = \left|\sqrt{\frac{2}{2}}\cos\theta - \sqrt{\frac{2}{2}}\,\text{sen}\,\theta\right|$$

Ir a D2

D2.si NUM es positivo:

$$\cos\phi = |\cos\phi|$$

$$\text{sen}\,\phi = |\text{sen}\,\phi|$$

Ir a E

Si NUM es negativo:

$\cos\phi = |\cos\phi|$

$\text{sen}\,\phi = -|\text{sen}\,\phi|$

E: Rotación:

Posteriormente, un componente principal en la solución, es la rotación de la matriz factor denominada varimax, desarrollada por Kaiser J. B. Carrol (1958). Transforma la matriz factorial hasta conseguir la solución que verifique que la suma de las simplicidades de los factores sea máxima. (Simplicidad = varianza de los cuadrados de las saturaciones). Rotan los factores forzando a que unas saturaciones se aproximen más a uno y las otras a cero, para facilitar así su interpretación.

$$X_i = x_i \cos\ \phi + y_i\ sen\ \phi$$

$$Y_i = x_i\ sen\ \phi + y_i \cos\ \phi$$

$$i = 1,\ 2,.....,\ m$$

Después de ser completado el ciclo *k(k-1)/2* las rotaciones son completadas, se vuelve a calcular la varianza para la matriz factor.

Denormalización:

$$a_{ij} = b_{ij} * h_i$$

$i = 1,\ 2,.....,\ m \qquad j = 1,\ 2,.....,\ k$

Prueba de la matriz común:

$f_i^2 = \sum_{j=1}^{k} a_{ij}^2$ Matriz común final

$d_i = h_i^2 - f_i^2$ Diferencias $i = 1,\ 2,.....,\ m$

En este cálculo, se guarda en orden el número de dimensiones independientes tan pequeño como sea posible, solamente los valores propios (o coeficientes de correlación) más grandes o igual a 1 son retenidos en el análisis.

Es de hacer notar la diferencia entre: el análisis del componente principal que es usado para determinar el número mínimo de dimensiones independientes necesarias para contar al menos de la varianza en el conjunto original de variables; en cambio, la rotación varimax es usada para reducir las columnas (factores) y no las filas (variables) de la matriz factor.

## 4. APLICACIÓN

La técnica del AF es aplicada a cualquier rama del saber humano, ya sea a la Psicología, minería, geología, metalurgia, medicina, odontología, agronomía, etc., donde se tenga un determinado número de muestras con sus respectivas variables, todas las que deberán estar expresadas en forma cuantitativa. Por ejemplo, en educación para llegar a determinar el rendimiento académico de los estudiantes, se llegará a considerar las notas obtenidas en cada una de las asignaturas por el estudiante, las cuales constituyen las variables. En la Tabla 1., se tiene una determinada información base, y las tablas siguientes son las obtenidas de acuerdo al programa desarrollado, conforme a la teoría señalada anteriormente, en la cual se posee 25 muestras, cada una constituida por 7 variables:

**Tabla 1**
**Valores de las 7 variables para las 25 muestras**

| $X_{ij}$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|
| 1 | 3,760 | 3,660 | 0,540 | 5,275 | 9,768 | 13,741 | 4,782 |
| 2 | 8,590 | 4,490 | 1,340 | 10,022 | 7,500 | 10,162 | 2,130 |
| 3 | 6,220 | 6,140 | 4,520 | 9,842 | 2,175 | 2,732 | 1,089 |
| 4 | 7,570 | 7,280 | 7,070 | 12,662 | 1,791 | 2,101 | 0,822 |
| 5 | 9,030 | 7,080 | 2,590 | 11,762 | 4,539 | 6,217 | 1,276 |
| 6 | 5,510 | 3,980 | 1,300 | 6,924 | 5,236 | 7,304 | 2,403 |
| 7 | 3,270 | 0,620 | 0,440 | 3,357 | 7,629 | 8,838 | 8,389 |
| 8 | 8,740 | 7,000 | 3,310 | 11,675 | 3,529 | 4,757 | 1,119 |
| 9 | 9,640 | 9,490 | 1,030 | 13,567 | 13,133 | 18,519 | 2,354 |
| 10 | 9,730 | 1,330 | 1,000 | 9,871 | 9,871 | 11,064 | 3,704 |
| 11 | 8,590 | 2,980 | 1,170 | 9,170 | 7,851 | 9,909 | 2,616 |
| 12 | 7,120 | 5,490 | 3,680 | 9,716 | 2,642 | 3,430 | 1,189 |
| 13 | 4,690 | 3,010 | 2,170 | 5,983 | 2,760 | 3,554 | 2,013 |
| 14 | 5,510 | 1,340 | 1,270 | 5,808 | 4,566 | 5,382 | 3,427 |
| 15 | 1,660 | 1,610 | 1,570 | 2,799 | 1,783 | 2,087 | 3,716 |
| 16 | 5,900 | 5,760 | 1,550 | 8,388 | 5,395 | 7,497 | 1,973 |
| 17 | 9,840 | 9,270 | 1,510 | 13,604 | 9,017 | 12,668 | 1,745 |
| 18 | 8,390 | 4,920 | 2,540 | 10,053 | 3,956 | 5,237 | 1,432 |
| 19 | 4,940 | 4,380 | 1,030 | 6,678 | 6,494 | 9,059 | 2,807 |
| 20 | 7,230 | 2,300 | 1,770 | 7,790 | 4,393 | 5,374 | 2,274 |
| 21 | 9,460 | 7,310 | 1,040 | 11,999 | 11,579 | 16,182 | 2,415 |
| 22 | 9,550 | 5,350 | 4,250 | 11,742 | 2,766 | 3,509 | 1,054 |
| 23 | 4,940 | 4,520 | 4,500 | 8,067 | 1,793 | 2,103 | 1,292 |
| 24 | 8,210 | 3,080 | 2,420 | 9,097 | 3,753 | 4,657 | 1,719 |
| 25 | 9,410 | 6,440 | 5,110 | 12,495 | 2,446 | 3,103 | 0,914 |

Fuente: Elaboración propia

En base a esta matriz de datos iníciales, aplicando el programa de AF desarrollado: se llega a obtener las siguientes tablas:

**Tabla 2**
**Total de las columnas**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\sum X_j$ | 177,500 | 119,330 | 58,720 | 228,346 | 136,455 | 179,186 | 58,654 |

Fuente: Elaboración propia

**Tabla 3**
**Promedios**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\overline{X}_j$ | 7,10000 | 4,77320 | 2,34880 | 9,13384 | 5,45820 | 7,16744 | 2,34616 |

Fuente: Elaboración propia

**Tabla 4**
**Suma de cuadrados y productos cruzados**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $SCP_{ij}$ | 750,097 | 492,280 | 231,787 | 959,858 | 604,324 | 784,220 | 235,410 |
| | 492,280 | 388,642 | 180,395 | 669,409 | 375,293 | 505,024 | 130,657 |
| | 231,787 | 180,395 | 117,809 | 324,953 | 128,429 | 165,650 | 58,114 |
| | 959,858 | 669,409 | 324,953 | 1256,672 | 749,150 | 979,738 | 286,960 |
| | 604,324 | 375,293 | 128,429 | 749,150 | 649,402 | 848,859 | 272,051 |
| | 784,220 | 505,024 | 165,650 | 979,738 | 848,859 | 1117,969 | 344,935 |
| | 235,410 | 130,657 | 58,114 | 286,960 | 272,051 | 344,935 | 165,419 |

Fuente: Elaboración propia

**Tabla 5**
**Desviaciones estándar**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $S_j$ | 2,3238 | 2,4178 | 1,6656 | 3,0178 | 3,2733 | 4,5581 | 1,6105 |

Fuente: Elaboración propia

**Tabla 6**
**Coeficientes de correlación**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $R_{ij}$ | 1,0000 | 0,5803 | 0,2011 | 0,9113 | 0,2833 | 0,2865 | -0,5332 |
| | 0,5803 | 1,0000 | 0,3638 | 0,8337 | 0,1658 | 0,2611 | -0,6087 |
| | 0,2011 | 0,3638 | 1,0000 | 0,4386 | -0,7042 | -0,6805 | -0,6488 |
| | 0,9113 | 0,8337 | 0,4386 | 1,0000 | 0,1630 | 0,2023 | -0,6755 |
| | 0,2833 | 0,1658 | -0,7042 | 0,1630 | 1,0000 | 0,9902 | 0,4272 |
| | 0,2865 | 0,2611 | -0,6805 | 0,2023 | 0,9902 | 1,0000 | 0,3571 |
| | -0,5332 | -0,6087 | -0,6488 | -0,6755 | 0,4272 | 0,3571 | 1,0000 |

Fuente: Elaboración propia

**Tabla 7**
**Valores propios**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\lambda_{i=j}$ | 3,3946 | 2,8055 | 0,4373 | 0,2779 | 0,0810 | 0,0034 | 0,0003 |

Fuente: Elaboración propia

**Tabla 8**
**Porcentajes acumulativos de valores propios**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\%\sum \lambda_{i=j}$ | 0,4849 | 0,8857 | 0,9482 | 0,9879 | 0,9995 | 1,0000 | 1,0000 |

Fuente: Elaboración propia

| $V_{ij}$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0,4053 | 0,4316 | 0,3854 | 0,4939 | -0,1277 | -0,0968 | -0,4809 |
| | -0,2929 | -0,2224 | 0,3559 | -0,2323 | -0,5751 | -0,5800 | -0,1303 |
| | -0,6674 | 0,6980 | 0,1477 | -0,1186 | 0,0294 | 0,1743 | 0,0176 |
| | 0,0888 | -0,0338 | 0,6276 | 0,2103 | 0,1108 | -0,0061 | 0,7353 |
| | -0,2267 | -0,4366 | 0,5121 | -0,1054 | 0,3890 | 0,3549 | -0,4553 |
| | 0,4098 | 0,1443 | 0,1875 | -0,5878 | -0,4232 | 0,5003 | 0,0332 |
| | -0,2782 | -0,2540 | -0,1081 | -0,5359 | -0,5562 | 0,4975 | 0,0489 |

Fuente: Elaboración propia

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Var. 1 | 0,7467 | -0,4906 | -0,4413 | 0,0468 | -0,0645 | 0,0239 | -0,0045 |
| Var. 2 | 0,7952 | .0,3726 | 0,4616 | -0,0178 | -0,1242 | 0,0084 | -0,0041 |
| Var. 3 | 0,7102 | 0,5961 | 0,0977 | 0,3309 | 0,1457 | 0,0109 | -0,0018 |
| Var. 4 | 0,9100 | -0,3890 | -0,0785 | 0,1109 | -0,0300 | -0,0343 | 0,0087 |
| Var. 5 | -0,2353 | -0,9633 | 0,0195 | 0,0584 | 0,1107 | -0,0247 | -0,0090 |
| Var. 6 | -0,1784 | -0,9715 | 0,1153 | -0,0032 | 0,1010 | 0,0292 | 0,0081 |
| Var. 7 | -0,8861 | -0,2182 | -0,0116 | 0,3876 | -0,1295 | 0,0019 | 0,0008 |

Fuente: Elaboración propia

**Varianza d** **ración**

| Iteración Ciclo | Varianzas |
|---|---|
| 0 | 0,2231 |
| 1 | 0,3592 |
| 2 | 0,3727 |
| 3 | 0,3730 |
| 4 | 0,3730 |
| 5 | 0,3730 |
| 6 | 0,3730 |
| 7 | 0,3730 |
| 8 | 0,3730 |
| 9 | 0,3730 |
| 10 | 0,3730 |

Fuente: Elaboración propia

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Var. 1 | 0,9587 | -0,1477 | 0,2018 | -0,1333 | 0,0085 | 0,0208 | 0,0006 |
| Var. 2 | 0,3763 | -0,0976 | 0,9032 | -0,1729 | 0,0655 | 0,0056 | 0,0007 |
| Var. 3 | 0,2434 | 0,7707 | 0,3285 | -0,0845 | 0,4813 | -0,0011 | -0,0002 |
| Var. 4 | 0,8091 | -0,0321 | 0,5415 | -0,1579 | 0,1561 | -0,0422 | -0,0011 |
| Var. 5 | 0,1555 | -0,9760 | 0,0461 | 0,1419 | 0,0145 | -0,0235 | -0,0157 |
| Var. 6 | 0,1271 | -0,9762 | 0,1528 | 0,0812 | -0,0049 | 0,0247 | 0,0157 |
| Var. 7 | -0,4439 | -0,4298 | -0,4002 | 0,6754 | -0,0437 | 0,0006 | -0,0002 |

Fuente: Elaboración propia

## 5. CONCLUSIONES

Los conceptos y definiciones matemático-estadísticos dados en el presente trabajo, fueron utilizados en el desarrollo del programa computacional, en la misma secuencia mencionada de operaciones, cuya aplicación y obtención de resultados se muestra mediante una aplicación numérica; observándose desde la matriz de datos que se ha utilizado, y los resultados parciales que se determinan, como: sumas, promedios, desviaciones estándar, suma de cuadrados y productos, coeficientes de la matriz de correlación, valores propios, porcentaje acumulativo de los valores propios, vectores propios, matriz factor, varianza de la matriz factor para cada ciclo de iteración, rotación de la matriz factor y verificación de las varianzas comunes.

La finalidad de esta investigación es llegar a que el interesado tenga un acceso inicial al conocimiento y comprensión del AF sin dificultad, pudiendo profundizar a futuro este tema con mayor facilidad, recomendándose los textos especializados 5, 7 y 8 de la bibliografía para su mejor comprensión.

## 6. COLABORACIÓN

Prof. Mary Nilda Avilés de Ruiz. Licenciada en Idiomas. Universidad Autónoma Gabriel René Moreno Santa Cruz – Bolivia (agosto, 2005)

## BIBLIOGRAFÍA

Balestra Pietro, *Calcul Matriciel pour Économistes.* Éditions Castella, Lausanne, Suisse, 1972, pp. 242 – X.

Dagnelie Pierre, *Analyse Statistique á Plusieurs Variables. Les Presses Agronomiques de Gembloux*, Gembloux, Belgique, 1975 (2º éddition), pp. 362 – XIV.

Davis C. John, *Statistics and Data Analysis in Geology*. John Wiley & Sons, New York, USA, 1973, pp. 550 – VII.

International Business Machines Corporation, *System Reference Library, Scientific Subroutines,* New York, USA, 1967, pp. 191.

Joreskog j.k., klovan J. E. Reyment R. A., *Geological Factor Analysis. Elsevier Scientific Publishing Company*. Amsterdam, Holanda, 1976, pp 178 – VII.

Kendall S. Maurice, *Multivariate Analysis*. Charles Griffin & Co. Ltd., Londres, Inglaterra, 1975, pp. 210 – XI.

Harman Harry H., *Modern Factor Analysis*, The University of Chicago Press, Chicago, USA, 1967, pp. 474 – XVI.

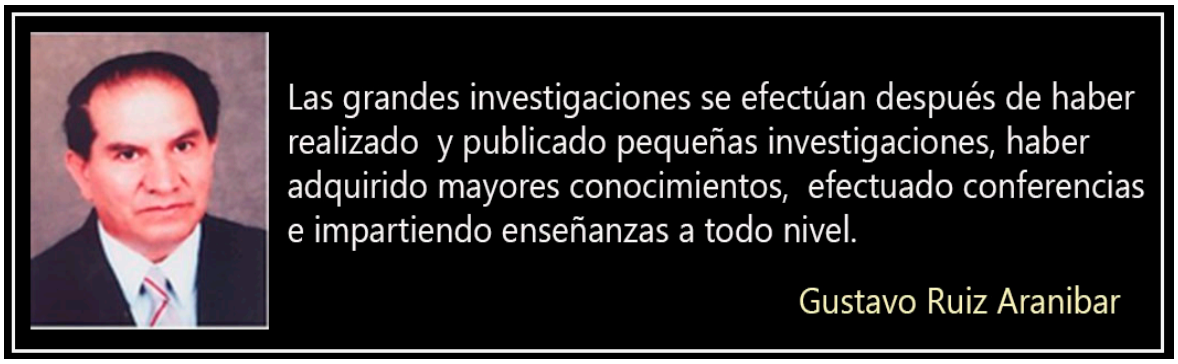
Horst Paul, *Factor Analysis of Data Matrices*. Copyright by Hold Rinehart and Wiston, Inc, Seattle, Washington, USA, 1965, pp. 730 – XXIV.

Lipschutz Seymour, *Algebra Lineal*. Editorial McGraw-Hill, Madrid, España, 1998, pp. 553 – XV.

Shafiro S. Miriam et al. *Courant Institute of Mathematical Sciences*, New York University, New York, USA, 1965, pp. 395.

Morice E., Bertrand M., *Penglaou C. Dictionnaire de Statistique*. Editorial Dunod, Paris, Francia, 1968, pp. 196.

Ruiz Aranibar Gustavo[4]. *Libreria Científica de Programas Informáticos*, La Paz –Bolivia.




Índice

4 Calle 20 y Av. Ballivian, Nº 8035, Calacoto, La Paz – Bolivia, Tel. 591-22772162 Cel. 67111778 gustavoruiz432@hotmail.com.bo ruizaranibargustavo@gmail.com.bo Blog: Gustavo Ruiz Aranibar

# ESTIMADORES ROBUSTOS DE TENDENCIA CENTRAL

Lic. Valdez Blanco, Dindo

✉ dindovaldez@hotmail.com

## RESUMEN

El presente artículo tiene por objeto, dar a conocer algunos estimadores robustos de tendencia central, y mostrar su aplicabilidad. En vista que el estimador de tendencia central más utilizado es la media muestral $\bar{x}$ , es necesario considerar la presencia de datos atípicos en la muestra, ya que estos pueden distorsionar la estimación que se realiza con la media aritmética

## PALABRAS CLAVE

*Estimación robusta, mediana de Hodges-Lehmann, media de Takashi, trimedia de Tukey, Media de Huber.*

## 1. INTRODUCCIÓN

Los estimadores robustos denominados también estimadores no paramétricos tienen la ventaja de disminuir la influencia de los valores extremos o de alguna manera ponderar los datos de tal forma que el estimador de posición central sea lo más representativo posible y el margen de error se minimice.

Los Estimadores Robustos, son estimadores libres de la suposición de la forma de distribución de la población de la cual se extrae la muestra. Al contrario de los estimadores clásicos que tienen asociada un tipo de distribución de la población. Así, por ejemplo, a la media aritmética se le asocia la distribución normal o mesocurtica.

Es más, a los Estimadores Clásicos se les asocia un criterio de óptimo prefijado, expresado por medio de las llamadas Normas Mínimas, basado en la distancia de las observaciones respecto el estimador. Las principales Normas Mínimas son las siguientes:

Norma $L_1$ : "la suma absoluta de los residuales es mínima."

$$Minimo\ L_1 = \sum_{i=1}^{n} |x_i - M|$$

La Norma $L_1$ está asociada a la Mediana y es conocida como la Norma de Laplace.

Norma $L_2$: " La suma de los cuadrados de los residuales es mínima."

$$Mínimo\ L_2 = \sum_{i=1}^{n} (x_i - M)$$

La norma $L_2$ está asociada a la media aritmética y es conocida como el principio de mínimos cuadrados

En contraposición los estimadores robustos no tiene asociados ninguna distribución y ninguna norma óptima. Los principales objetivos de usar los estimadores Robustos se pueden resumir en los siguientes puntos:

a. Construir una estimación segura ante una cantidad apreciable de datos atípicos.

b. Poner un límite a la influencia del sesgo escondido debido a la presencia de datos atípicos (los que se salen de una tolerancia).

c. Aislar de manera clara los datos atípicos para un tratamiento por separado.

d. Seguir cercanamente el sentido estricto del modelo Paramétrico.

## 2. ESTIMACION NO PARAMÉTRICA O ROBUSTA DE TENDENCIA CENTRAL

Los estimadores no paramétricos de tendencia central son los llamados estimadores de orden o estadísticos de orden, puesto que las observaciones o valores de la variable aleatoria $X$ deben ser ordenados: $x_{(1)}$ , $x_{(2)}$,..., $x_{(n)}$. Tal que debe cumplirse que: $x_{(1)} < x_{(2)} <... < x_{(n)}$.

De los diversos estimadores no paramétricos o robustos existentes, sólo se indicarán algunos de ellos.

### *2.1 LA MEDIANA DE HODGES – LEHMANN*

Este estimador fue desarrollado por Joseph L. Hodges y Erich L. Lehmann en 1960, el mismo se basa en un algoritmo muy sencillo, es la mediana de los promedios de todos los pares sucesivos de observaciones de una muestra de n observaciones ordenadas.

Sea la serie de datos ordenados: $x_{(1)}$ , $x_{(2)}$ ,..., $x_{(n)}$. En base a estos se definen los promedios sucesivos denominados: $Y_1$ , $Y_2$,..., $Y_{n-1}$ tal que:

$$Y_1 = \frac{x_{(1)} + x_{(2)}}{2} \; ; \; Y_2 = \frac{x_{(2)} + x_{(3)}}{2} \; ; \; \dots \; ;$$

$$Y_{n-1} = \frac{x_{(n-1)} + x_{(n)}}{2}$$

De tal forma que se obtiene una nueva serie ordenada: $Y_1 < Y_2 <... < Y_{n-1}$. Siendo la mediana de Hodges – Lehmann la mediana de esta nueva serie.

***Ejemplo:*** Suponer que 5 turistas se registran en un hotel, sus edades son: 18, 17, 18, 19 y 60 años. Para calcular el estimador de Hodges – Lehmann, se tiene la muestra ordenada:

$$x_{(1)} = 17 \quad x_{(2)} = 18 \quad x_{(3)} = 18$$

$$x_{(4)} = 19 \quad x_{(5)} = 60$$

En base a los que se calcula la serie de promedios:

$$Y_1 = 17{,}5 \quad Y_2 = 18$$

$$Y_3 = 18{,}5 \quad Y_4 = 39{,}5$$

Hallamos la mediana de los cuatro promedios y resulta:

$$M_{H-L} = \frac{18 + 18{,}5}{2} = 18{,}25$$

### *2.2 LA MEDIA DE TAKASHI*

El Estimador de Takashi, presentado en 1969 por Takashi Yamagawa[1] toma la mediana sucesiva de las observaciones o mediciones y luego a esa nueva serie originada le aplica la media aritmética.

Sea la serie de datos ordenados: $x_{(1)}$ , $x_{(2)}$,..., $x_{(n)}$. En base a estos se definen las medianas sucesivas denominadas: $Y_1$ , $Y_2$,..., $Y_{n-1}$ tal que:

$$Y_1 = \frac{x_{(1)} + x_{(2)}}{2} \; ;$$

$$Y_2 = \frac{x_{(2)} + x_{(3)}}{2} \; ; \; \dots ;$$

$$Y_{n-1} = \frac{x_{(n-1)} + x_{(n)}}{2}$$

De tal forma que se obtiene una nueva serie ordenada: $Y_1 < Y_2 <... < Y_{n-1}$. Siendo la media de Takashi la media aritmética de esta nueva serie.

***Ejemplo:*** El estimador de Takashi para los

1 Huber (1964) "Robust estimation of a location parameter"

datos de las edades de los turistas del ejemplo anterior se basa en la muestra ordenada de las edades:

$$x_{(1)} = 17 \quad x_{(2)} = 18 \quad x_{(3)} = 18$$

$$x_{(4)} = 19 \quad x_{(5)} = 60$$

En base a los que se calcula la serie de medianas:

$$Y_1 = 17{,}5 \quad Y_2 = 18$$

$$Y_3 = 18{,}5 \quad Y_4 = 39{,}5$$

Finalmente se calcula la media aritmética de las medianas:

$$M_T = \frac{17{,}5 + 18 + 18{,}5 + 39{,}5}{4} = 23{,}37$$

*2.3 LA TRIMEDIA DE TUKEY*

Este estimador fue desarrollado por John Tukey en 1960[2] y es un promedio ponderado del primer, segundo y tercer cuartil. Sean los cuartiles de una muesra aleatoria de X, entonces el estimador de Tukey se define como:

$$T = \frac{1}{4}Q_1 + \frac{1}{2}Q_2 + \frac{1}{4}Q_3$$

***Ejemplo:*** Los cuartiles de los datos: 18, 17, 18, 19 y 60 son:

$$Q_1 = 17.5 \quad Q_2 = 18 \quad Q_3 = 39{,}5$$

Por tanto la trimedia de Tukey es:

$$T = \frac{1}{4}17{,}5 + \frac{1}{2}18 + \frac{1}{4}39{,}5 = 23{,}25$$

*2.4 LA MEDIA ITERATIVA DE HUBER*

Este estimador fue desarrollado por Peter J. Huber[3] en el año 1964. El estimador Huber se desarrolla en base a las funciones:

$$Min \sum_{i=1}^{n} (x_i - M)^2$$

Si se cumple:

$$|x_i - M| \leq K\sigma \;\; ; \;\; i = 1,2,\dots,n$$

$$Min \sum_{i=1}^{n} K\sigma(2|x_i - M| - K\sigma)$$

Si cumple $|x_i - M| \geq K\sigma \;\; ; \;\; i = 1,2,\dots,n$

Generalmente *K* adopta valores de 2 ó 3. Pues $K\sigma$ representa la tolerancia de la medición. El estimador de Huber utiliza una función de peso *P* de la siguiente forma:

$$P_i = 1 \;\; si \;\; |x_i - M| \leq K\sigma$$

$$P_i = \frac{K\sigma}{|x_i - M|} \;\; si \;\; |x_i - M| \geq K\sigma$$

De esta manera se otorga una ponderación más baja a las observaciones que se encuentran con mayor desviación de la deseada, lo que influirá directamente en la estimación final.

***Ejemplo:*** Consideremos las observaciones $x_i$: 17, 18, 18, 19, 60. A simple vista la observación 60 aparenta ser un dato atípico. Tomando como desviación máxima σ = 10 y usando K = 2 se tiene que Kσ = 20 se procede a utilizar el proceso iterativo para la estimación.

Recordemos que la media aritmética ponderada está dada por:

$$\bar{X}_P = \frac{\sum_{i=1}^{n} P_i x_i}{\sum_{i=1}^{n} P_i}$$

Utilizaremos la media aritmética de los datos como M=26,4 para la asignación de las ponderaciones, los residuales de las muestras son:

2 Huber (1964) Robust estimation of a location parameter.

3 Huber (1981) Robust Statistic.

$$|x_1 - M| = 6{,}44$$

$$|x_2 - M| = 5{,}44$$

$$|x_3 - M| = 5{,}44$$

$$|x_4 - M| = 4{,}44$$

$$|x_5 - M| = 36{,}56$$

De tal forma las ponderaciones de las observaciones resultan ser:

$$P_1 = 1 \quad P_2 = 1 \quad P_3 = 1$$

$$P_4 = 1 \quad P_5 = \frac{20}{33.6} = 0{,}5952$$

Así la media aritmética ponderada resulta:

$$\bar{X}_P = \frac{\sum_{i=1}^{n} P_i x_i}{\sum_{i=1}^{n} P_i} = 23{,}44$$

Con esta media ponderada calculamos los nuevos residuales:

$$|x_1 - M| = 6{,}44$$

$$|x_2 - M| = 5{,}44$$

$$|x_3 - M| = 5{,}44$$

$$|x_4 - M| = 4{,}44$$

Y se calculan los nuevos pesos tomando en cuenta que si $|x_i\text{-}M| \leq 20$ tendrá un peso igual a la unidad y si es $\geq 20$ se le calcula el peso tal como se indicó antes:

$$P_1 = 1 \quad P_2 = 1 \quad P_3 = 1$$

$$P_4 = 1 \quad P_5 = \frac{20}{36{,}56} = 0{,}547$$

La nueva media ponderada resulta:

$$\bar{X}_P = \frac{\sum_{i=1}^{n} P_i x_i}{\sum_{i=1}^{n} P_i} = 23{,}05$$

Este proceso iterativo se repite hasta que la media ponderada converja a un número el cuál será la media estimada de Huber.

Índice

## 3. CONCLUSIONES

Los estimadores robustos de tendencia central ofrecen la ventaja de que evitan el "uso y abuso" que se ha hecho de la media aritmética, puesto que "a todo" le aplicamos la estimación de la media aritmética. Por otra parte, dichos estimadores eliminan el uso arbitrario del rechazo de observaciones atípicas, en vista que minimizan su efecto en el cálculo del estimador de posición. Por último, asumir una distribución normal para la población de donde se extrae la muestra (lo cual implica usar la media aritmética como estimador) no es conveniente cuando el número de observaciones es muy pequeño.

## BIBLIOGRAFÍA

Huber, P.J. 1964: *Robust estimation of a location parameter*. Annals of Mathematical Statistics. Vol.35; pág. 73-101.

Huber, P.J. 1981: *Robust Statistics*. New York. Wiley & Sons.

# SOBRE LA RIQUEZA Y LA POBREZA DE UNA NACIÓN EN TÉRMINOS DE PRODUCTIVIDAD "UNA SENCILLA EXPLICACIÓN ESTADÍSTICA"

Mg. Sc. Vargas Salazar, F. Rodrigo

✉ *cipacohc@gmail.com*

## RESUMEN

En los países pobres el trabajo o empleo productivo es inútil, ya que se caracteriza por el empleo redundante, elevado mercado informal, elevado desempleo, elevado subempleo vs los países ricos que generan trabajo o empleo productivo útil, el cual se caracteriza por la elevada alfabetización y larga escolarización; ¡clave para el progreso de una nación!

## PALABRAS CLAVE

*Riqueza, pobreza, productividad, trabajo o empleo productivo.*

## 1. INTRODUCCIÓN

¿Seremos capaces de percatarnos de la enorme consecuencia ecológica, económica, social y política que conlleva la explosión demográfica? Pues lo gravitacional no es el terrible problema que conlleva alimentar a tantísima gente –indigente- sino más bien el de procurarles un trabajo o empleo productivo que les saque de la miseria en que viven. Las religiones, el socialismo y el comunismo o dicho de otro modo los natalistas no se ponen de acuerdo con los antinatalistas sobre cuánto poder tiene (nada, un poco, mucho o muchísimo) la famosa expresión estadística "la explosión demográfica" sobre la riqueza y pobreza de una nación.

## 2. LA RIQUEZA Y POBREZA DE UNA NACIÓN

Una persona es pobre como nos recuerda el Doctor (Tejero, J., 2004) cuando es analfabeto y cuando puede efectuar únicamente trabajos manuales primarios o serviles o de artesanía, ya que genera pobreza debido a su baja productividad. Sin embargo, todo converge como lo veremos más delante de manera lógica en la frase "altas tasas de natalidad", esto para nosotros será "el circulo vicioso de la pobreza". Por otro lado, también consideraremos un país pobre a toda nación que tiene un PIB per cápita inferior a 5500 US$, y nación rica a aquella que tiene un PIB per cápita superior a 5500 US$.

## 3. LA PRODUCCIÓN

Entenderemos por producción o fabricación, la actividad que transforma determinados bienes en otros distintos que poseen una utilidad mayor. Lo que tratamos de decir es que debemos tener en cuenta que hay actividad productiva del "bien A en bien B" si y solo si el bien B representa mayor utilidad económica que la del bien A, para lo cual también debemos considerar que el bien A y el bien B son el mismo objeto en distintos momentos. De esta manera, inferimos que el bien A (materia prima) y el bien B (producto) son bienes económicamente distintos. En consecuencia, las transformaciones de bienes en otros bienes que pueden dar lugar a actividades productivas son de tres especies: tecnológicas, espaciales y temporales. Ahora bien, esta sección del trabajo quedaría vaga

si no relacionásemos la eficiencia con la producción. Y si leemos al Doctor Wilfredo Pareto, comprenderemos enseguida que a + eficiencia productiva que posea una determinada empresa + será la capacidad productiva de dicha empresa.

Entonces, para llegar a la eficiencia productiva[1], diremos: 1) que una empresa es una unidad de producción que produce un determinado bien; 2) una empresa se encuentra en condiciones de máxima eficiencia productiva cuando su producto es máximo respecto a las cantidades globales de los factores de producción empleados. En consecuencia, para que un producto de una empresa sea máximo dadas las cantidades de los factores de producción, es necesario que los factores se distribuyan racionalmente entre las secciones de la empresa de tal modo que cada sección tenga el mismo nivel de productividad.

## 4. LA PRODUCTIVIDAD

Entenderemos por productividad[2], la capacidad productiva o capacidad de producción, es decir, como el máximo nivel de producción que puede alcanzar una empresa por medio de una estructura productiva en un tiempo dado. El incremento o no de la capacidad productiva viene dada por la toma de decisiones en por ejemplo la inversión para adquirir tecnología, mayor cantidad de mano de obra, más expertos en un área X, etc. Un ejemplo que podemos describir para que quede claro lo anterior, sería el referido a una industria láctea en el que 100 trabajadores poseen una producción en volumen de 2000 litros de leche evaporada/día de acuerdo a la tecnología y mano de obra cualificada que posee. Entonces diremos que la productividad es de 20 litros/trabajador. De lo anterior, deducimos que la producción y la productividad de una empresa X debe ser capaz de satisfacer la demanda de un mercado. Cuando no es capaz de hacerlo hablamos de demanda insatisfecha y cuando sobrepasa dicha oferta entonces decimos que la empresa X está perdiendo clientes. En términos concretos notamos que: la productividad no es una medida de la producción ni de la cantidad que se ha fabricado (Bain, 1985), más bien es una medida de lo bien que se ha combinado y empleado los recursos para cumplir con los resultados deseados: Productividad = [resultados/recursos empleados]; o escrito de otra manera: Productividad = [producción de bienes o servicios/insumos].

Finalmente, otro Doctor en Economía (Wannacott; 1993: 329-348), indica que la clave del crecimiento económico o rentabilidad es la productividad. Esta proposición es evidente si por ejemplo esperamos que aumente la producción de un producto X en un Laboratorio entonces el profesional debería producir más de ese producto X/hora, o dicho de otra manera, si aumentase la productividad media de su trabajo el profesional entonces tendríamos un lógico incremento en la producción del producto X. Por tanto, el crecimiento del nivel de producción depende del efecto combinado de los incrementos en el número de horas trabajadas, de la producción y cuánto se produzca en un periodo dado. Así, solo una mayor productividad supone mayor

1 *En otras palabras, la eficiencia productiva [EP] responde a la ecuación: EP = [ciencia + tecnología + recurso humano + recurso material].*

2 *Revisando nuestra Historia, empezamos con la revolución agrícola y tuvimos mundialmente el primer indicio de productividad en la que se redujo el número de hambrientos, luego con la revolución industrial dimos curso a una intensa productividad, finalmente con la actual revolución tecnológica, biológica, científica globalizada vivimos el mayor periodo de productividad vista por el hombre. Y todo tiene una sola causa: "elevada alfabetización y larga escolarización, es decir educación, el cual es un problema catastrófico en los países pobres y subdesarrollados, pues en éstos importa más lo político y lo popular".*

rentabilidad.

## 5. LOS GENIOS DEL CAPITALISMO

*5.1.* ***DAVID HUME*** *(1711-1776)*, británico, filósofo, profesor de moral, fue el iniciador de las divergencias respecto a las opiniones dominantes en la época. Con Hume la filosofía británica se apartó de la metafísica. Su filosofía que consistía en ver las cosas de otra manera influyó en Immanuel Kant y en Adam Smith; junto con Smith son considerados los precursores del capitalismo. Modelo económico practicado por los países industrializados.

*5.2.* ***ADAM SMITH*** *(1723-1790)*, escocés, filósofo, profesor de moral, un genio de la observación, del análisis, síntesis y exposición, padre del liberalismo económico y de la economía moderna, nos proporciona dos reglas valiosísimas: 1) la riqueza de las naciones es (independientemente de su extensión territorial, población, suelo y clima) a) directamente proporcional al crecimiento económico e b) inversamente proporcional al crecimiento de la población [3]; 2) la riqueza de las naciones es c) directamente proporcional a la tasa de rentabilidad e d) inversamente proporcional a la tasa de interés.

Cuando decimos matemáticamente que la riqueza de las naciones es directamente proporcional al crecimiento económico queremos decir que una nación es tanto más rica, más desarrollada social, económica y políticamente, cuanto más crezca económicamente. Por su parte, cuando decimos que la riqueza de las naciones es inversamente proporcional al crecimiento de la población queremos decir que una nación es tanto más pobre, más subdesarrollada social, económica y políticamente, cuanto más crezca poblacionalmente.

## 6. LOS GENIOS DEL SOCIALISMO-COMUNISMO

*6.1.* ***KARL MARX*** *(1818-1883)*, judío alemán, quien en octubre del 1843 una vez en la capital francesa se relacionó con el movimiento obrero-socialista. Si bien a menudo miró con desagrado las ideas paupérrimas de los grupúsculos socialistas de la época, siempre se emocionó con la fraternidad que unía a los obreros durante las reuniones políticas a las que asistía. En París conoció en Engels y juntos escribieron el "manifiesto comunista" la que señala la idea de que la lucha entre las clases sociales constituye el motor de la evolución histórica y que las clases sociales, los obreros son todo y el capitalismo es nada. En 1900 Lenin se convirtió en el artífice del socialismo ruso, sucediéndole a éste Stalin, entre otros.

## 7. LOS GENIOS IMPARCIALES

*7.1.* ***MAX WEBER*** *(1864-1920)*, judío alemán, jurista, sociólogo, economista, filósofo, políglota, de un saber enciclopédico anonadante, es el primero en escribir sobre la influencia de las religiones-culturas en la consecución de logros materiales y en la elevada alfabetización y larga escolarización [4]de manera organizada y a gran escala. Weber concede gran relevancia a la racionalización de Occidente, al racionalismo occidental, racionalismo económico, conducta racional, ética económica, sociología de las religiones,

3 Entenderemos en este trabajo la expresión "crecimiento de la población" como sinónimo de "tasa de natalidad" o de "crecimiento demográfico".

4 *La expresión "larga escolarización" implica pasar por el Pre kínder, Kínder, Primaria, Secundaria, Universidad, Diplomado, Especialidad, Maestría hasta llegar al Doctorado y/o post Doctorado en cualquier área científica. En consecuencia, ser científicos con la habilidad de investigar, crear, innovar o inventar es la cúspide de la larga escolarización en los países industrializados.*

valor de la herencia biológica y el modo de reaccionar ante el destino y el medio para la consecución de esos logros éticos, materiales, científicos organizados y en gran escala,…, en otras palabras "el capitalismo genuino". Nos proporciona una regla valiosísima: 1) el capitalismo organizado, la riqueza de las naciones, la cultura occidental, el desarrollo económico, social y político es a) directamente proporcional a la ética, a la conducta racional y a la cultura racional (elevada alfabetización y larga escolarización creciente) e b) inversamente proporcional a la corrupción, al analfabetismo y a la cultura irracional.

Cuando decimos matemáticamente que la riqueza de las naciones es directamente proporcional a la ética, a la conducta racional y a la cultura racional queremos decir que una nación es tanto más rica, más desarrollada económica, social y políticamente, cuanto más ética, conducta racional y cultura racional posea. Por su parte, cuando decimos que la riqueza de las naciones es inversamente proporcional a la corrupción, al analfabetismo y a la cultura irracional queremos decir que una nación es tanto más pobre, cuanto más corrupción, analfabetismo y cultura irracional posea[5].

Weber se preguntaba ¿en qué medida las actitudes religiosas y culturales inciden en la pobreza? Él observó y comprobó que determinadas comunidades, determinadas actitudes religiosas y culturales, eran más exitosas que otras en la consecución de bienes materiales e intelectuales. Es así que estableció dos reglas: 1) son comunidades pobres las que sufren las más altas tasas de natalidad, corrupción, analfabetismo y han padecido o padecen un sistema político socialista o comunista o populista, 2) son comunidades ricas las que gozan las más bajas tasas de "natalidad, corrupción y analfabetismo".

Weber nos sugiere en su investigación que en los países ricos "la cultura de los valores democráticos se impone sobre la religión; poseen una elevada alfabetización y larga escolarización y una amplia racionalización de la religión y de la cultura [6]". En los países pobres islámicos "la religión no racionalizada sigue siendo la base de su identidad pobre. La religión es su cultura y aquella es el principio y fin de su existencia". En los países pobres latinos "sus tradiciones paganas, religiones y culturas no racionalizadas, etc., son la base de su identidad pobre". Weber también nos señala que "la responsabilidad y la ética deberían ser los dos valores más imprescindibles y trascendentes de cualquier grupo humano, comunidad, sociedad o Estado desde la más pobre hasta la más rica con el fin de garantizar la permanencia, supervivencia y buen desarrollo y progreso de éstos". Con esto tenemos que, los países pobres se caracterizan por sus bajos niveles éticos. Por ende, a + bajo nivel de ética + corrupción. Y los países ricos se caracterizan por sus altos niveles morales y éticos. Por ende, a + alto nivel ético – corrupción.

*7.2.* ***ALBERT EINSTEIN*** *(1879-1955)*, judío alemán, que junto con Arquímedes e Isaac Newton forman el trio de científicos más brillantes que jamás haya conocido la Historia de la humanidad. Ellos fueron, son y seguirán siendo los tres héroes de la Ciencia. Su fórmula E=mc2, nos habla también de economía, cuando nos dice: + energía,

5 *Esto nos resulta lógico, pues solamente una conducta y cultura racionales nos conduce a una elevada alfabetización y larga escolarización, a la procreación muy bien controlada, a procrear hijos que se puedan mantener y educar dignamente hasta que ellos lleguen a su Doctorado o mínimamente culminen con una Carrera Universitaria y sean el orgullo de su familia, de sí mismos y de su nación.*

6 *Entendemos por religión racional a la religión actualizada, es decir la razón por sobre la fe; y por cultura racional a la cultura progresista, es decir la razón por sobre las tradiciones festivas ridículas periódicas, paganas y retrogradas, etc.*

supone + civilización, y con + civilización se requiere + energía. Einstein también nos recuerda que la riqueza y pobreza de las naciones también está determinada por el sistema político-económico que rija en cada país. Esto tiene que ver con los más de 2800 millones de personas pobres en nuestro planeta, y son pobres porque no disponen de un trabajo o empleo productivo. Sin embargo, ¿Cuál es la causa de esa pobreza? Según (Tejero, J., 2004) los países que más padecen de explosión demográfica son los países más pobres del planeta: África posee (1.91% de tasa anual de crecimiento demográfico) en relación al Total mundial (1.33%); EEUU y Canadá (1.34%); América Latina (4.13%); Asia (1.38%); Europa (0.60%); Oceanía (1.66%). Por ende, un país es pobre por sus elevados incrementos demográficos, y su pobreza es tanto mayor cuanto más elevadas son sus tasas de corrupción y analfabetismo. Contrariamente, los países con bajos incrementos demográficos son los más ricos y prósperos a nivel mundial. Luego, por sus bajos incrementos demográficos, la riqueza es tanto mayor cuanto más elevadas son las tasas de alfabetismo. Empero, EEUU es un caso especial, ya que, si bien es la nación con el mayor nivel económico, cultural, social, político, científico, tecnológico, empresarial y laboral a nivel mundial, ésta padece las más altas tasas de natalidad y analfabetismo entre las naciones más ricas, como consecuencia de sus muy prolíferas reproductoras etnias "negra e hispana".

*7.3.* ***GERALD M. MEIER*** *(1923-2011)*, un renombrado Doctor en economía nos recuerda una regla económica conocida como el cociente o índice de Meier, el cual es muy relevante, es la más representativa y trascendental de las ciencias económicas sociales y políticas que dice: "el crecimiento económico debe duplicar o triplicar dependiendo del nivel tecnológico de la nación al crecimiento demográfico para la creación de trabajo o empleo productivo".

Por medio del índice, razón o proporción matemática de Meier (Crecimiento PIB anual dividido entre el Crecimiento poblacional anual) podemos argumentar o determinar o diferenciar la riqueza y pobreza entre naciones o entre regiones. O en otras palabras, el índice de Meier dice: a) en las naciones ricas el índice de Meier es mayor que 2.5 o 3, b) en las naciones pobres el índice de Meier es inferior a 2.5, c) si el índice de Meier es inferior a 1 hablamos entonces de un desastre económico para esa nación.

Los países pobres (ricos o muy ricos en recursos naturales[7]) suelen tener un crecimiento económico (PIB) mayor que el de los países industrializados. Empero, lo estadísticamente significativo es que las tasas más elevadas del crecimiento PIB por habitante o PIB per cápita (que son las más trascendentales económicamente) corresponden a países industrializados. ¿Cómo es esto posible, parece algo incoherente o no? La explicación está dada a través del siguiente ejemplo: América Latina creció poblacionalmente 2.9 veces más que los países industrializados (crecimiento poblacional anual de América Latina 2.42 dividido entre crecimiento poblacional anual de países industrializados 0.83). Por ende inferimos que América Latina (un conglomerado de países pobres y subdesarrollados) necesita un ritmo de crecimiento económico sostenible anual mayor al 7.26% en su PIB para salir del subdesarrollo y pobreza; racionalmente

7 *Por incomprensible e increíble que parezca son los países que no poseen recursos naturales los más productivos del planeta, ejemplo de ello son: Indonesia, Malasia, Filipinas y Tailandia (llamados los tigres o dragones menores); Hong Kong, Taiwán, Singapur y Corea del sur (llamados los tigres o dragones mayores). Otros casos dignos de mencionar son Israel, Japón, Australia, Finlandia, Irlanda, etc. Finalmente están aquellas naciones que están escalando a pasos gigantes la montaña de la productividad tales como Brunéi, Vietnam, Laos, Birmania, Camboya y Chile.*

imposible de alcanzar[8]. Luego, tenemos que el PIB anual de América Latina es de tan solo 4.13, su Crecimiento poblacional anual es el terrorífico 2.42, por ende el PIB per cápita es el absurdo 1.71, esto nos lleva a deducir lógicamente que haga lo que se haga, América Latina está condenada al abismo tercermundista, pues recordemos que para salir de la pobreza el índice de Meier debe ser superior a 2.5 o 3. Hasta aquí es preciso aclarar que el éxito socioeconómico y político de los países ricos no se debe a su elevado crecimiento económico a costa de los países pobres (como nos hacen creer los políticos de gobiernos de turno), sino a su racional crecimiento demográfico.

*7.4.* ***MICHAEL P. TODARO***, otro renombrado Doctor en economía nos recuerda otra regla económica sobre la Tasa Bruta de Natalidad (TBN), que establece después de muchísimas investigaciones mundiales que: 1) la TBN de 15 mil nacidos vivos por cada 1000 habitantes es la barrera que separa a los países ricos de los países pobres. Es decir, una TBN si es superior a 15 por mil nacidos vivos es un país pobre; e inferior a 15 por mil nacidos vivos es un país rico.

## 8. LA SECUENCIA ESTADÍSTICA

i. A + tasa de crecimiento demográfico + tasa de desempleo, y a + tasa de desempleo + pobreza, y a + pobreza + tasa de hambre[9].

ii. El desarrollo económico es directamente proporcional al crecimiento económico. E inversamente proporcional al crecimiento de la población.

iii. El buen desarrollo y progreso de las naciones es directamente proporcional a las tasas de responsabilidad y ética e inversamente proporcional a las tasas de irresponsabilidad y corrupción.

iv. A + elevada alfabetización y larga escolarización + eficiencia productiva, y a + eficiencia productiva + productividad, y a + productividad + crecimiento económico, y a + crecimiento económico + desarrollo económico de una persona, de una nación o región, y a + desarrollo económico + progreso de una nación.

## 9. DISCUSIÓN

¿Hay algo más irracional, no solidario, irresponsable, inhumano y fuera de moral que procrear hijos como conejos destinados a la desdicha, a la explotación económica, sexual, política y religiosa?, ¿Hay algo en cualquier país tan indigno y fuera de moral cómo el político y funcionario público corrupto que asegura su poder?, ¿Alfabetizar plenamente a las familias que se reproducen como conejos con carencias de la más imprescindible infraestructura es o no es estadísticamente un imposible? Los datos estadísticos propuestos de las diversas fuentes consultadas nos dejan ver que la tremenda explosión demográfica es la causa del desastre ecológico, desempleo masivo, pobreza masiva, bajos salarios masivos, analfabetismo masivo, emigraciones masivas, agitaciones sociales, corrupción, delincuencia, políticas masivas, terrorismo masivo, explotación de niños, narcotráfico, prostitución, tráfico de órganos y de personas, etc., realidades o males rancios, endógenos o internos que no son aceptadas por las

8 *Japón (comparando territorialmente con nuestro país, es del tamaño de la ciudad de Tarija), ha sido el único en el mundo que tuvo la mayor tasa de crecimiento PIB anual en el periodo de 1913 a 1998 y fue de 4.22%, según nos señala el Profesor PhD Guisán.*

9 *Todos los países subdesarrollados del planeta, especialmente en África, Latinoamérica y países islámicos poseen altas tasas de crecimiento, desempleo y pobreza. Paradójicamente son ricos y otros muy ricos en recurso naturales. La corrupción muy bien disimulada y lo voraces que pueden ser los políticos y/o gobernantes modelo de dichos países lo explican todo.*

religiones ni por los latifundistas ni por los gobiernos seguidores del judío alemán Karl Marx: "socialistas-comunistas, populistas, fascistas, corporativistas, autárquicas, neo marxistas, marxistas leninistas, marxistas maoístas".

## 10. CONCLUSIÓN

Hemos develado cómo la estadística nos auxilia en comprender de manera única, simple y elegante la riqueza y la pobreza de una nación en términos de productividad que nos traslada a 6 conclusiones: i) La tasa de rentabilidad debería ser dos veces la tasa de interés para conseguir un desarrollo económico equilibrado con suficiente trabajo o empleo productivo; ii) Un país es pobre porque es incapaz de crear un trabajo o empleo productivo que demanda su desmedido crecimiento demográfico (Esto no lo aceptan los natalistas, ya sea por su ignorancia, creencias religiosas, convicciones políticas o sociales o por espurios intereses políticos, económicos y sociales). Los gobiernos de dichos países son ajenos a la democracia pluripartidista y ética y a la economía liberal de mercado; iii) Los inconvenientes principales para la creación de trabajo o empleo productivo en países pobres o subdesarrollados son: la explosión demográfica incontrolable, la corrupción, el analfabetismo, la fuga de capitales, actitudes religiosas culturales irracionales; iv) La tasa anual de crecimiento poblacional debería ser más importante que la tasa anual de crecimiento económico con el único objetivo mayor de lograr suficiente trabajo o empleo productivo; v) El control de natalidad, los incentivos laborales y económicos, salarios muy bien remunerados, la elevada alfabetización y larga escolarización y la mejora sanitaria deberían ser el único fin de todo sistema político en países subdesarrollados; vi) La riqueza y pobreza de las naciones depende básicamente de la productividad de sus ciudadanos, de su constitución política de Estado, de su larga escolarización que a su vez es consecuencia de su genotipo (herencia genética), fenotipo (clima, religión, cultura) y estado nutricional, del poder que posean los sindicatos, y especialmente de los políticos dirigentes de una nación. Y dado que son los políticos quienes fabrican un sistema político, está en las manos de éstos últimos el destino (éxito o fracaso) de su país.

## BIBLIOGRAFÍA

Apuleyo Mendoza, Plinio; Montaner Carlos Alberto; Vargas Llosa, Álvaro. 1999. *Fabricantes de miseria*. Barcelona, España: Plaza & Janés.

Apuleyo Mendoza, Plinio; Montaner Carlos Alberto; Vargas Llosa, Álvaro. 1996. *Manual del perfecto idiota latinoamericano*. Barcelona, España: Plaza & Janés.

Harrison, Lawrence. 1985. *El subdesarrollo está en la mente*. Madrid, España: Playor.

Hume, David. 2008. *Ensayos económicos: los orígenes del capitalismo moderno*. España: Biblioteca nueva.

Marx, Karl. 2010. *Capital, trabajo, plusvalía.* Buenos Aires, Argentina: Longseller.

Smith, Adam. 2011. *La riqueza de las naciones*. Barcelona, España: Brontes.

Wannacott et al. 1993. *Economía*. (4ª ed.). Madrid, España: Mc Graw Hill.

Weber, Max. *Economía y Sociedad*. Fondo de Cultura económica. México 1944.

Weber, Max. 2012. *El político y el científico*. Madrid, España: Alianza.

Weber, Max. 2012. *La ética protestante y el espíritu del capitalismo*. Barcelona, España: Brontes.

Índice