UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

Aplicaciones del escalamiento multidimensional y el análisis de correspondencias a los seguros de salud y pensiones

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

NOMBRE DEL ALUMNO
Jeanette Castillo Balderas

TUTORA
M. en A. P. María del Pilar Alonso Reyes

2008

FACULTAD DE CIENCIAS
UNAM

Hoja de datos del jurado

1. Datos del alumno
Castillo
Balderas
Jeanette
56 51 20 63
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuario
09234509-0

2. Datos del tutor
M en AP
Alonso
Reyes
María del Pilar

3. Datos del sinodal 1
M en C
Flores
Díaz
José Antonio

4. Datos del sinodal 2
Mat
Chávez
Cano
Margarita Elvira

5. Datos del sinodal 3
Act
Trejo
González
Martha

6. Datos del sinodal 4
Act
Sánchez
Villareal
Francisco

7. Datos del trabajo escrito
Aplicaciones del escalamiento multidimensional y el análisis de correspondencias a los seguros de salud y pensiones
87 p.
2008

Indice

Capítulo 1. Análisis multivariado

1.1. Introducción 1
1.2. Definición 2
1.3. Clasificación de técnicas multivariadas 4
1.3.1. Técnicas de dependencia 4
1.3.2. Técnicas de interdependencia 4
1.3.3. Escalas no métricas 5
1.3.4. Escalas métricas 5
1.3.5. Principales técnicas multivariadas 6
1.3.5.1. Regresión múltiple 6
1.3.5.2. Análisis discriminante 7
1.3.5.3. Análisis de conglomerados 7
1.3.5.4. Análisis de factores 8
1.3.5.5. Análisis de correspondencias 8
1.3.5.6. Escalamiento multidimensional 8

Capítulo 2. Análisis de correspondencias

2.1. Introducción 10
2.2. Modelo 12
2.2.1. Datos de análisis 12
2.2.2. Las masas 13
2.2.3. Distancia 14
2.2.4. Objetivo 15
2.2.5. Solución al problema de optimización 17
2.2.6. Coordenadas factoriales 18
2.2.7. Relaciones de transición 20
2.2.8. Contribuciones 21
2.2.9. Filas y columnas suplementarias y valores de prueba 23

Capítulo 3. Escalamiento Multidimensional

3.1. Introducción 25
3.2. Medidas de proximidad 26
3.2.1. Similaridades 26
3.2.2. Disimilaridades 27
3.2.3. Distancias 27
3.3. Clasificación de modelos 28
3.3.1. Modelo clásico 28
3.3.1.1. Elección del número de dimensiones 33
3.3.2. Modelo métrico 33
3.3.3. Modelo no métrico 34
3.3.3.1. Medidas de bondad de ajuste 36
3.3.3.2. Elección del número de dimensiones 38

Capítulo 4. Aplicaciones del escalamiento multidimensional

4.1. Instituciones de seguros especializadas en salud (ISES) 40
4.2. Instituciones de seguros de pensiones derivadas de las Leyes de Seguridad Social (ISPDLSS) 42
4.3. Variables de análisis 43
4.4. Ejemplo para el mercado de las ISES 44
4.4.1. Análisis descriptivo de las variables 44
4.4.2. Segmentación del mercado de las ISES 49
4.5. Ejemplo para el mercado de las ISPDLSS 52
4.5.1. Análisis descriptivo de las variables 52
4.5.2. Segmentación del mercado de las ISPDLSS 58

Capítulo 5. Aplicaciones del análisis de correspondencias

5.1. Ejemplo para el mercado de las ISES 61
5.1.1. Notas de revelación 61
5.1.2. Análisis descriptivo de las variables 62
5.1.3. Segmentación del mercado de las ISES 65
5.2. Ejemplo para el mercado de las ISPDLSS 69
5.2.1. Análisis descriptivo de las variables 70
5.2.2. Segmentación del mercado de las ISPDLSS 72

Conclusiones 76

Anexos 80

Bibliografía 86

Capítulo 1

Análisis multivariado

1.1 Introducción

El progreso en la informática disponible en esta época ha hecho posible obtener avances notables en el análisis de datos de diversas índoles, lo cual era prácticamente inimaginable hace algunos añosa través de programas estadísticos básicos y especializados, permitiendo así el manejo de grandes y complejas bases de información.

Por otra parte, se ha podido observar un desarrollo continuo de técnicas estadísticas las cuales surgieron como respuesta a la creciente necesidad de obtener una capacidad analítica más profunda al abordar una gran diversidad de temas de investigación y para las cuales, si bien se disponía de los fundamentos teóricos, no fue sino hasta que los ordenadores tuvieron una capacidad de cálculo y memoria suficiente que los usuarios pudieron realizar ensayos de sus modelos, eliminando así las limitaciones de antaño.

Una gran parte del estudio de datos se ha venido desarrollando a través de la aplicación de estadística tradicional, sin embargo, alternativamente ha surgido un conjunto de técnicas conocidas como ***análisis multivariado***.

Debido a sus numerosas aplicaciones en la práctica, el análisis multivariado ha tenido un desarrollo creciente en los últimos años, convirtiéndose en una herramienta prácticamente imprescindible.

1.2 Definición de análisis multivariado

Se define como la rama de la estadística y del análisis de datos, que estudia, interpreta y elabora el material estadístico sobre la base de un conjunto de $n > 1$ variables, que pueden ser de tipo cuantitativo, cualitativo o una mezcla de ambos.[3]

El análisis multivariado entonces, se refiere a todos los métodos estadísticos que estudian simultáneamente múltiples medidas de cada individuo u objeto sometido a investigación. En muchos casos, estas técnicas son extensiones del análisis univariado y bivariado.

Sin embargo, una condición esencial es que las n variables sean dependientes, de naturaleza similar y que ninguna de ellas tenga una importancia superior a las demás.

El análisis multivariado es una metodología estadística compleja que requiere el uso continuo de conceptos de álgebra lineal, cálculo numérico y geometría, así como el conocimiento de tipos los de escalas que tiene la información objeto de estudio.

Sus principales objetivos son:

[3] Cuadras, C.M., Métodos de Análisis Multivariante, España, Publicaciones Universitarias S.A., 1991

1. Resumir datos a partir de un grupo de variables originales, obteniendo un conjunto pequeño de éstas pero con una pérdida mínima de información.

Permite identificar un conjunto de variables indicadoras necesarias para describir adecuadamente un fenómeno complejo, permitiendo su graficación y comparación de conjuntos de datos, lo que hace posible una interpretación más sencilla de la información, proporcionando así un mejor conocimiento del objeto de estudio.

2. Identificar grupos de datos en caso de que éstos existan.

Estas técnicas permiten visualizar relaciones que pueden existir en las observaciones de estudio, las cuales compartirán características similares y que no son identificables a simple vista.

3. Clasificar nuevas observaciones en grupos definidos.

Este objetivo se liga al anterior en el sentido de que una vez definidos dichos grupos, las nuevas observaciones pueden clasificarse de forma más sencilla en ellos.

4. Relacionar dos conjuntos de variables.

Permite determinar si existe algún tipo de correspondencia en un conjunto de datos, así como conocer el número de dimensiones de la misma.

El análisis multivariado tiene aplicaciones en diversos campos: en primera instancia se desarrollaron para resolver problemas de clasificación en biología, extendiéndose posteriormente en psicometría, marketing y ciencias sociales empleándolos para encontrar variables indicadoras y factores. Asimismo, han alcanzado una gran aplicación en ingeniería y ciencias de la

computación para resumir información y diseñar sistemas de reconocimiento de patrones.

1.3 Clasificación de técnicas multivariadas

En términos generales, éstas se diferencian unas de otras según su área de aplicación. Una forma de clasificarlas es estableciendo lo siguiente:

a) Si las variables son dependientes o independientes,
b) El número de variables involucradas en el estudio y
c) El tipo de escala de la información que se pretende estudiar.

1.3.1 Técnicas de dependencia

Estas técnicas se basan en el uso de variables independientes para predecir y explicar una o más variables dependientes, las cuales pueden contener información en escalas métrica o no métrica. Un ejemplo de técnica dependiente es el análisis de discriminante.

1.3.2 Técnicas de interdependencia

En el caso en que se relacionen las observaciones de una forma que no permita definir si son independientes o dependientes, sino que el análisis implica la utilización de todas las variables del conjunto de forma simultánea se habla de una técnica de interdependencia, cuyo objetivo es identificar la estructura de las mismas, lo que puede dar lugar al descubrimiento de nuevas relaciones.

1.3.3 Escalas no métricas

Las variables de este tipo pueden tener escalas nominal u ordinal. En el primer caso se asignan nombres o etiquetas a los objetos de estudio con base en la presencia o ausencia de una característica determinada. Para el procesamiento de estos datos, los nombres suelen ser remplazados por números, pero en ese caso el valor numérico es irrelevante. También se les conoce como escalas de categoría. El único tipo de comparaciones que se pueden hacer con este tipo de variables es el de igualdad o diferencia. Ejemplos de este tipo de escala son el estado civil, el género, la raza, la afiliación política.

Por su parte, en la escala ordinal los números asignados a los objetos representan el orden o rango en que se posee o carece de atributo. Permite hacer comparaciones como“mayor que” o “menor que”, igualdad o diferencia. Por ejemplo, preferencias sobre un número de opciones.

1.3.4 Escalas métricas

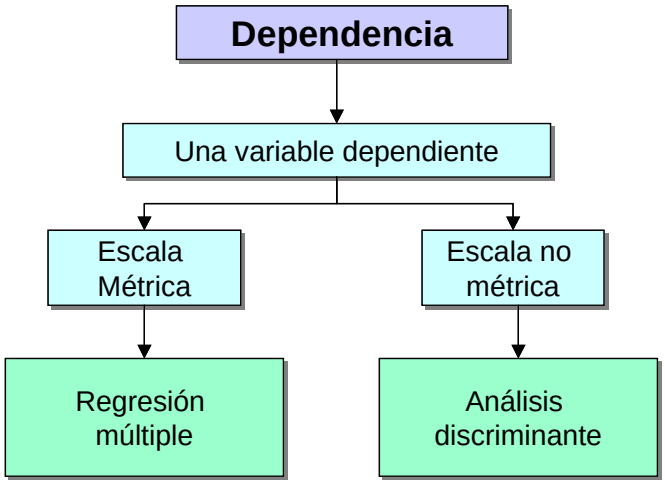
Se dividen en escalas de intervalo y de razón. En la primera, los números asignados a los objetos tienen todas las características de la escala ordinal y además las diferencias entre medidas representan intervalos equivalentes, por lo tanto, las operaciones de adición y sustracción tienen significado. El punto cero de esta escala es arbitrario, se pueden usar valores negativos y las diferencias se pueden expresar como razones. Ejemplos de este tipo de escala son la fecha y la temperatura.

En cuanto a la escala de razón, tiene todas las características de la escala de intervalo y además razones significativas entre pares arbitrarios de números, por lo que las operaciones de multiplicación y división tienen

significado. Asimismo, la posición del cero no es arbitraria, por ejemplo, la masa, la longitud, la edad.

1.3.5 Principales técnicas multivariadas

Existe una gran variedad de técnicas multivariadas disponibles para llevar a cabo el análisis de datos. No obstante, el uso de algunas ha sido más difundido que el de otras, por lo que a continuación se describen las más comunes.

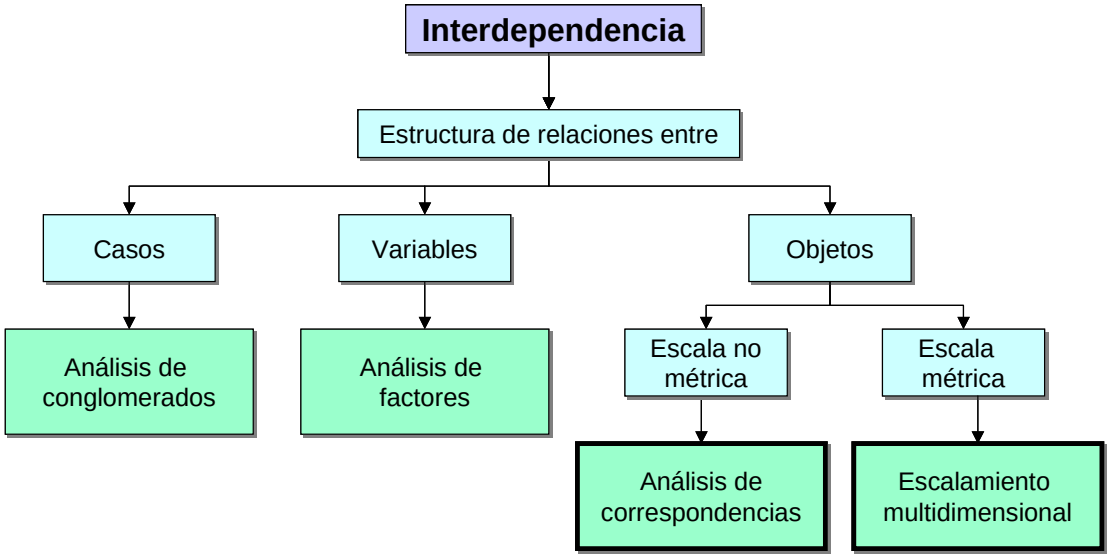


1.3.5.1 Regresión múltiple

La regresión múltiple se emplea para analizar la relación de una variable métrica dependiente con respecto a una o más independientes, al predecir los cambios de la primera en respuesta a los cambios en las variables independientes; cada una de estas últimas es ponderada y dicha asignación de peso indica su contribución relativa a la predicción conjunta.

1.3.5.2 Análisis discriminante

Es una técnica apropiada cuando la única variable dependiente es categórica y las independientes se suponen métricas. También es útil en situaciones donde la muestra total puede dividirse en grupos basándose en una variable dependiente caracterizada por varias clases conocidas. Su objetivo es entender las diferencias de los grupos y predecir la probabilidad de que un individuo u objeto pertenezca a una clase o grupo particular.



1.3.5.3 Análisis de conglomerados

Permite desarrollar subgrupos significativos de individuos u objetos clasificando una muestra de ellos en un número pequeño de grupos mutuamente excluyentes basados en similitudes. Por lo regular implica al menos dos etapas: la medida de alguna forma de similitud o asociación entre las entidades para determinar cuántos grupos existen en realidad. Posteriormente, describir las personas o variables para determinar su composición.

1.3.5.4 Análisis de factores

Suele usarse para analizar interrelaciones entre un gran número de variables y explicarlas en términos de sus dimensiones subyacentes comunes, denominados factores. El objetivo es encontrar un modo de condensar la información contenida en éstas en un conjunto más pequeño de factores con una pérdida mínima de información.

1.3.5.5 Análisis de correspondencias

Es una técnica de interdependencia para la reducción dimensional y la elaboración de mapas de percepciones, también conocidos como mapas preceptúales, entre objetos y un conjunto de características descriptivas o atributos especificados por el investigador.

Ésta es una técnica de composición debido a que el mapa de percepción se basa en la asociación entre objetos y un conjunto de características descriptivas o atributos especificados por el investigador. Es parecida al análisis de factores pero su aplicación más directa es la representación de la correspondencia de categorías de variables, particularmente de escala nominal.

Dicha correspondencia es la base del desarrollo de los mapas de percepción. Sus beneficios se basan en sus capacidades únicas para representar filas y columnas, por ejemplo, etiquetas y atributos en un mismo espacio.

1.3.5.6 Escalamiento multidimensional

Técnica cuyo objetivo es identificar las dimensiones subyacentes claves en las evaluaciones de los objetos de estudio, por ejemplo, se utiliza a menudo

en marketing para la evaluación de productos, servicios o compañías por parte de los clientes. También suele aplicarse en la comparación de cualidades físicas, como son gustos alimenticios, olores, percepciones de los asuntos o candidatos políticos, la evaluación de diferencias culturales entre diferentes grupos, entre otros.

El escalamiento multidimensional puede ayudar a determinar:

1. Las dimensiones que utilizan los encuestados cuando evalúan los objetos;
2. Cuántas dimensiones pueden utilizarse en una situación particular;
3. La importancia relativa de cada dimensión y
4. Cómo se relacionan visualmente los objetos.

En esta técnica también se elaboran mapas de percepción, con lo cual se obtiene la imagen recibida de un conjunto de objetos asociados a apreciaciones habituales, es decir, permite transformar juicios del consumidor de similitud o preferencia en distancias representadas en un espacio multidimensional.

El objeto de esta tesis es exponer la metodología del escalamiento multidimensional y el análisis de correspondencia, para que posteriormente se ejemplifique su aplicación al estudio de fenómenos reales, en este caso particular, la segmentación de dos de los ramos de seguros de México.

Capítulo 2

Análisis de correspondencias

2.1 Introducción

El análisis de correspondencias es una técnica de interdependencia empleada para la reducción de la dimensión y elaboración de mapas de percepción. Se basa en la asociación entre objetos y un conjunto de características descriptivas o atributos especificados. Una de las cualidades de esta técnica es su capacidad para utilizar tanto datos no métricos.

Se requiere que la información se represente en una matriz de datos rectangular, es decir, de tabulación cruzada, con entradas no negativas. Las filas y columnas no tienen significados predefinidos, y representan las respuestas de una o más variables categóricas.

Con la tabulación cruzada, para cualquier combinación de filas y columnas se relacionan con combinaciones basadas en frecuencias marginales, proporcionando una esperanza condicional, para lo cual a menudo se emplea la ji-cuadrada.
Una vez obtenidos los valores de la ji-cuadrada se estandarizan y se convierten en una distancia métrica, definiendo soluciones en dimensiones reducidas. Estos factores relacionan simultáneamente filas y columnas, los cuales son representados en un gráfico conocido como mapa de percepción.

Posteriormente debe evaluarse el ajuste en conjunto, para lo que se debe identificar el número apropiado de dimensiones y su importancia. Se obtiene un valore propio para cada dimensión con el fin de conseguir la contribución relativa en la explicación de la variación en las categorías. El número de éstas se elige basándose en el nivel conjunto de explicación deseada de la variación y el aumento de explicación ganado por la adición de otra dimensión.

Una vez establecida la dimensionalidad se puede identificar la asociación entre las categorías por su proximidad, para lo que se debe seleccionar un tipo de normalización y determinar si las comparaciones se van a hacer por filas, columnas o filas y columnas.

Entre las ventajas que esta técnica multivariada ofrece se encuentran las siguientes:

- Se puede representar en un espacio perceptual la tabulación cruzada simple de variables categóricas múltiples, analizando las respuestas existentes.

- Se representan las relaciones entre filas y columnas, así como entre las categorías tanto de filas como de las columnas, y

- Se dispone de una representación conjunta de categorías de filas y columnas en la misma dimensión, permitiendo identificar grupos caracterizados por atributos muy relacionados.

2.2 Modelo

2.2.1 Datos de análisis

Los datos forman una matriz de orden $I \times J$, en la que para cada elemento (i, j) se asocia un número no negativo que representa las frecuencias absolutas observadas de dos variables cualitativas de N elementos[3]. La primera variable se representa por filas, suponiendo que toma n valores posibles, mientras que la segunda se representa por columnas y toma p valores posibles:

J / I	1	2	j	…	p	
1	F_{11}	F_{12}	F_{1j}	…	F_{1p}	$F_{1.}$
2	F_{21}	F_{22}	F_{2j}	…	F_{2p}	$F_{2.}$
i	F_{i1}	F_{i2}	F_{ij}		F_{ip}	F_{2j}
…	…	…	…	…	…	…
n	F_{n1}	F_{n2}	F_{nj}	…	F_{np}	$F_{n.}$
	$F_{.1}$	$F_{.2}$	$F_{.j}$	…	$F_{.p}$	N

F_{ij} es la frecuencia absoluta observada del atributo *i* y el atributo *j*; $F_{n.}$ y $F_{.p}$ representan las distribuciones marginales por filas y columnas respectivamente, en donde

$$F_{n.} = \sum^{p} F_{1i}$$

[3] Si bien generalmente se manejan enteros, también pueden asociarse a los elementos (i,j) de estas tablas probabilidades, medidas, intensidades o preferencias.

$$N = \sum F_{ij}, i \in I, j \in J .$$

Asimismo, se define como perfiles de fila y columna a los cocientes dados por:

$$f_j^i = \frac{F_{ij}}{F_{n.}} \quad \text{y} \quad f_i^j = \frac{F_{ij}}{F_{.p}}$$

cada uno es una distribución condicionada de j a i y de i a j respectivamente. Estos proporcionarán las coordenadas de los puntos, es decir, la matriz de datos objeto del análisis no será la de frecuencias absolutas F_{ij} sino la de frecuencias relativas, a la cual se le denominará F y cuyo elemento genérico se denominará f_{ij} y se define como

$$f_{ij} = \frac{F_{ij}}{N} \quad \text{tal que} \sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij} = 1$$

La matriz F puede considerarse por filas o por columnas, lo que implica que el análisis de esta debe ser equivalente al aplicado a su transpuesta, lo anterior, en virtud de que la elección de la variable que se coloca en las filas o columnas es arbitraria, por lo que no debe influir en el análisis.

2.2.2 Las masas

Un concepto importante en el análisis de correspondencias, es el de *las masas*, y las cuales representan la importancia de cada fila y columna, éstas se definen como:

$$f_i = \frac{f_{i.}}{N} \quad \forall\, i \in I \quad f_j = \frac{f_{.j}}{N} \quad \forall\, j \in J$$

en donde $f_{i.}$ y $f_{.j}$ son las distribuciones marginales por fila y columna respectivamente.

dando lugar así a dos sistemas de masas:

$$f_I = \{f_i, i \in I\} \quad \text{y} \quad f_J = \{f_j, j \in J\}$$

La relevancia de este concepto radica en que pondera a cada fila y columna con base en su importancia, la cual se pierde al trabajar únicamente con los perfiles.

Mediante el empleo de ambos conceptos se definen las nubes de puntos siguientes:

$$N(I) = \{(f_j^i, f_i), i \in I\}$$

$$N(J) = \{(f_i^j, f_j), j \in J\}$$

Asimismo, el perfil medio de las filas $\{i\}$ es el sistema de masas del conjunto J, y el de las columnas $\{j\}$ entonces, es el sistema de masas del conjunto I. También se denominan centros de gravedad o baricentros y se definen como

$$N(I) : \sum \{f_i . f_j^i, i \in I\} = f_J \equiv i_g$$

$$N(J) : \sum \{f_j . f_i^j, j \in J\} = f_I \equiv j_g$$

2.2.3 Distancia

A efecto de estudiar la relación entre $N(I)$ y $N(J)$ se definen las siguientes distancias:

- entre elementos de $N(J)$

$$d^2\left(i,i'\right)=\sum\left\{\left(f_j^i-f_j^{i'}\right)^2\frac{1}{f_j},j\in J\right\} \text{ con } i,i'\in I$$

- entre elementos de $N(I)$

$$d^2\left(j,j'\right)=\sum\left\{\left(f_j^i-f_j^{i'}\right)^2\frac{1}{f_i},i\in I\right\} \text{ con } j,j'\in J$$

El objetivo de dividir por f_j o f_i es dar a todas las columnas la misma importancia.

Al elegir la distancia adecuada para el estudio de la relación referida, es inmediato considerar la distancia euclídea debido a su sencillez, sin embargo, al aplicarla entre filas y columnas, ésta sólo refleja la diferencia que hay entre las frecuencias marginales $F_{n.}$ y $F_{.p}$, es decir, manifiesta el efecto de tamaño o talla mientras que entre los perfiles refleja semejanza.

Al elegir la distancia más apropiada debe tomarse en cuenta que:

- Todas las filas (puntos en $\mathbf{R}^n$) o columnas (puntos en $\mathbf{R}^p$) no tienen el mismo peso, ya que algunas contienen más datos que otras. Por esta razón debe darse más peso a aquellas que contengan más datos.
- La distancia euclídea entre puntos no es una buena medida de su proximidad.

Si a esta distancia se le incluye la ponderación por filas y columnas respectivamente, se obtiene una distancia que tiene la propiedad de la equivalencia distribucional y permite establecer relaciones de transición.

La propiedad de equivalencia distribucional consiste en que la distancia entre filas no se altera si se fusionan las columnas j y j' de perfil semejante. Esta propiedad garantiza robustez, al no perder ni ganar información mediante dichas fusiones o bien descomposiciones. La distancia que posee estas propiedades es la denominada ji-cuadrada y será la que se empleará en el análisis de correspondencias.

2.2.4 Objetivo

El objetivo del análisis de correspondencias consiste en encontrar los ejes principales o de máxima inercia (o varianza) de las nubes $N(I)$ y $N(J)$, es decir, la suma ponderada de la masa de cada punto por su distancia al punto de referencia elevada al cuadrado, por ejemplo con respecto a sus baricentros:

$$\text{Inercia}_{i_g[N(I)]} = \sum \left\{ f_i . d^2\left(i, i_g\right), i \in I \right\} = \sum \left\{ \left(f_{ij} - f_i f_j\right)^2 \frac{1}{f_i f_j}, i \in I, j \in J \right\} = \frac{\chi^2}{N}$$

$$\text{Inercia}_{j_g[N(J)]} = \sum \left\{ f_j . d^2\left(j, j_g\right), j \in J \right\} = \sum \left\{ \left(f_{ij} - f_i f_j\right)^2 \frac{1}{f_i f_j}, i \in I, j \in J \right\} = \frac{\chi^2}{N}$$

O respecto al origen:

$$\text{Inercia}_{O[N(I)]} = \sum \left\{ f_i . d^2\left(i, O\right), i \in I \right\} = \sum \left\{ f_i \left(f_j^i\right)^2 \frac{1}{f_j}, i \in I, j \in J \right\}$$

$$= \sum \left\{ \frac{f_{ij}^2}{f_i f_j}, i \in I, j \in J \right\}$$

$$\text{Inercia}_{O[N(J)]} = \sum \left\{ f_j . d^2\left(j, O\right), j \in J \right\} = \sum \left\{ f_j \left(f_i^j\right)^2 \frac{1}{f_i}, i \in I, j \in J \right\}$$

$$= \sum \left\{ \frac{f_{ij}^2}{f_i f_j}, i \in I, j \in J \right\}$$

2.2.5 Solución al problema de optimización

Para encontrar las direcciones que absorben el máximo de inercia de cada nube de puntos, es decir, los ejes principales de inercia, se empleará la siguiente notación, análoga para filas y columnas:

$F = \left[f_{ij} \right]$ es la matriz de frecuencias relativas y su dimensión es $n \times p$.

$D_n = \left[f_i \right]$ es la matriz diagonal de la distribución marginal de las filas.

$D_p = \left[f_j \right]$ es la matriz diagonal de la distribución marginal de las columnas.

$D_n^{-1} F = \left[f_j^i \right]$ es la matriz de los perfiles de fila; en $\mathbf{R}^p$ son las coordenadas f_j^i de las filas.

$D_p^{-1} F = \left[f_i^j \right]$ es la matriz de los perfiles de columna; en $\mathbf{R}^n$ son las coordenadas f_i^j de las columnas.

D_p^{-1} es la métrica bajo la cual $d^2\left(i, i'\right) = \sum \left\{ \left(f_j^i - f_j^{i'} \right)^2 \frac{1}{f_j}, j \in J \right\}$ con $i, i' \in I$.

D_n^{-1} es la métrica bajo la cual $d^2\left(j, j'\right) = \sum \left\{ \left(f_i^j - f_i^{j'} \right)^2 \frac{1}{f_i}, i \in I \right\}$ con $j, j' \in J$.

D_n es la matriz de masas $\left[f_i \right]$ de las filas.

D_p es la matriz de masas $\left[f_j \right]$ de las columnas.

Se tomarán en cuenta dos criterios de optimización:

a) De optimización en $\mathbf{R}^p$, con relación al origen O:

$$Max\left\{\sum f_i.d^2(i,O), i \in I\right\} = Max\left\{D_n\left[\left(D_n^{-1}F\right)D_p^{-1}u\right]^2\right\} = Max\left\{u'D_p^{-1}F'D_n^{-1}FD_p^{-1}u\right\}$$ con la restricción de normalización $u'D_p^{-1}u = 1$, en donde u es un vector propio ortonormal[4] de la matriz $F'D_n^{-1}FD_p^{-1}$, de valor propio[5] λ:

$$\left[F'D_n^{-1}FD_p^{-1}\right]u = \lambda u$$

b) Criterio de optimización en $\mathbf{R}^n$, con relación al origen O:

$$Max\left\{\sum f_j.d^2(j,O), j \in J\right\} = Max\left\{D_p\left[\left(D_p^{-1}F\right)D_n^{-1}v\right]^2\right\} = Max\left\{v'D_n^{-1}F'D_p^{-1}FD_n^{-1}v\right\}$$

con la restricción de normalización $v'D_n^{-1}v = 1$.

Cada v es un vector propio ortonormal de la matriz $F'D_p^{-1}FD_n^{-1}$ con el mismo valor propio asociado λ:

$$\left[F'D_p^{-1}FD_n^{-1}\right]v = \lambda v$$

[4] Ver definición en Anexo I
[5] Ver definición en Anexo I

2.2.6. Coordenadas factoriales

Anteriormente se mencionaron las relaciones de transición, esto es, poder representar las coordenadas de una fila i en función de las coordenadas de la columna j, implica que los dos conjuntos se pueden representar geométricamente de forma simultánea.

$$\psi = \left(D_n^{-1}F\right)D_p^{-1}u$$

$$\psi(i) = \sum\left\{\frac{f_j^i}{f_j}u_j\,,\, j\in J\right\} = \sum\left\{\frac{f_{ij}}{f_i f_j}u_j\,,\, j\in J\right\}$$

Con

$$E(\psi) = \sum\left\{f_i\,\psi(i), i\in I\right\} = 0$$

$$Var(\psi) = \sum\left\{f_i\psi^2(i), i\in I\right\} = \lambda$$

Con respecto a las coordenadas de las columnas:

$$\varphi = (D_p^{-1}F')D_n^{-1}v$$

$$\varphi(i) = \sum\left\{\frac{f_i^j}{f_i}v_i\,,i\in I\right\} = \sum\left\{\frac{f_{ij}}{f_i f_j}v_i\,,\, i\in I\right\}$$

Con

$$E(\varphi) = \sum\left\{f_j\,\varphi(j), j\in J\right\} = 0$$

$$Var(\varphi) = \sum\left\{f_j\varphi^2(j), j\in J\right\} = \lambda$$

Esta varianza se interpreta como la inercia de los distintos puntos respecto al origen de coordenadas, es decir, es la suma ponderada de las distancias de los puntos a otro que en este caso es el origen. El que ψ y φ tengan la misma varianza implica que están medidas en la misma escala

$$\sum\{\lambda_\alpha, \alpha = 1,2,...N\} = \frac{\chi^2}{K}$$

en donde N es el mínimo entre $n-1$ y $p-1$, en donde se omite el primer valor propio dado que su valor es 1 y representa la solución trivial; esta solución no se obtiene cuando se realiza el análisis respecto a los centros de gravedad i_g y j_g, obteniendo valores propios inferiores a 1.

A los conjuntos $(\psi_1, \varphi_1, \lambda_1)$, …, $(\psi_\alpha, \varphi_\alpha, \lambda_\alpha)$, …, $(\psi_N, \varphi_N, \lambda_N)$ se denominan el primer factor, α-ésimo factor, N-ésimo factor respectivamente. Prescindiendo del factor trivial, el orden de los valores propios es decreciente: $\lambda_1 \geq ... \geq \lambda_\alpha \geq ... \geq \lambda_N$. La fórmula de reconstrucción de datos es:

$$f_j^i = f_j \left[1 + \sum \left\{ \frac{1}{\sqrt{\lambda_\alpha}} \psi_\alpha(i) \varphi_\alpha(j), \alpha = 1,2,...,N \right\} \right]$$

$$f_i^j = f_i \left[1 + \sum \left\{ \frac{1}{\sqrt{\lambda_\alpha}} \psi_\alpha(i) \varphi_\alpha(j), \alpha = 1,2,...,N \right\} \right]$$

$$K(i,j) = K \cdot f_{ij} = K \frac{f_j^i}{f_i} = K \frac{f_i^j}{f_j}$$

A través de la fórmula de reconstitución de los datos es posible plantear la aplicación del postulado de parsimonia, es decir, el grado de calidad del ajuste

para cada coeficiente estimado. En otras palabras, la cantidad de ajuste por coeficiente estimado y editar o sobreajustar el modelo con coeficientes adicionales que consigan pequeñas ganancias en el ajuste del modelo. Lo anterior, siempre que el objetivo sea explorar con unas pocas dimensiones lo subyacente a una tabla de cantidades considerables.

2.2.7. Relaciones de transición

Una vez obtenidas las coordenadas factoriales es posible obtener para cada factor las relaciones de transición, las cuales de definen conforme a las siguientes expresiones:

$$\psi(i) = \frac{1}{\sqrt{\lambda}} \left\{ f_j^i \cdot \varphi(j), j \in J \right\}, \forall\ i \in I$$

$$\varphi(i) = \frac{1}{\sqrt{\lambda}} \left\{ f_i^j \cdot \psi(i), i \in I \right\}, \forall\ j \in J$$

De esta forma es posible representar los puntos de un espacio en función de las coordenadas de los puntos de otro, es decir, se puede llevar a cabo la representación simultánea de dos nubes de puntos. Las coordenadas de cada fila son el centro de gravedad de las columnas, ponderadas por los elementos de los perfiles de tal fila.

2.2.8. Contribuciones

Las modalidades que más inercia tienen son las que definen el factor, lo cual se encuentra condicionado a su masa. Por ejemplo, para el conjunto I de las filas:

$$\sum \left\{ f_i \cdot \psi_\alpha^2(i), i \in I \right\} = \lambda_\alpha$$

$f_i \cdot \psi_\alpha^2(i)$ es la parte de la inercia total del eje α correspondiente a la fila i. Se llamará contribución absoluta a todas las modalidades de I al factor α a:

$$C_\alpha(I) = \sum \left\{ f_i \cdot \psi_\alpha^2(i), i \in I \right\} = \lambda_\alpha$$

entonces $C_\alpha(i) = f_i \cdot \psi_\alpha^2(i)$ es la contribución absoluta de la fila i a la inercia del eje α. Análogamente, en lo que se refiere a las columnas, la contribución absoluta de las modalidades J al factor α es:

$$C_\alpha(J) = \sum \left\{ f_j \cdot \varphi_\alpha^2(j), j \in J \right\} = \lambda_\alpha$$

y $C_\alpha(j) = f_j \cdot \varphi_\alpha^2(j)$ es la contribución absoluta de la modalidad j a la inercia del eje α.

Dado que las contribuciones absolutas no representan adecuadamente la importancia de un punto en la construcción de un eje factorial; se recurre a las contribuciones relativas:

- contribución relativa de la modalidad i sobre

$$\alpha : R(i/\alpha) = \frac{C_\alpha(i)}{C_\alpha(I)} = \frac{f_i\, \psi_\alpha^2(i)}{\lambda_\alpha}$$

- contribución relativa de la modalidad j sobre

$$\alpha : R(j/\alpha) = \frac{C_\alpha(j)}{C_\alpha(J)} = \frac{f_j\, \varphi_\alpha^2(j)}{\lambda_\alpha}$$

Para evaluar cómo están representadas una fila o una columna por los distintos factores, se recurre a las siguientes contribuciones:

- contribución relativa de α sobre la modalidad de la fila i:

$$R(\alpha/i) = \frac{C_\alpha(i)}{C(i)} = \frac{f_i\, \psi_\alpha^2(i)}{\sum\{f_i\, \psi_\alpha^2(i), \alpha = 1,2,...,N\}} = \frac{\psi_\alpha^2(i)}{\sum\{\psi_\alpha^2(i), \alpha = 1,2,...,N\}}$$, y

- contribución relativa de α sobre la modalidad de la columna j:

$$R(\alpha/j) = \frac{C_\alpha(j)}{C(j)} = \frac{f_j\, \varphi_\alpha^2(j)}{\sum\{f_j\, \varphi_\alpha^2(j), \alpha = 1,2,...,N\}} = \frac{\varphi_\alpha^2(j)}{\sum\{\varphi_\alpha^2(j), \alpha = 1,2,...,N\}}$$

en donde $C(i) = \sum\{C_\alpha(i), \alpha = 1,2,...,N\}$ y $C(j) = \sum\{C_\alpha(j), \alpha = 1,2,...,N\}$, es decir, la suma de las contribuciones absolutas de la fila i y de la columna j respectivamente según todos los factores 1,2,…, α ,… N.

La suma de las contribuciones relativas de todos los factores a cada fila y a cada columna es 1:

$$\sum\{R(\alpha/i), \alpha = 1,2,...,N\} = 1$$

$$\sum \{R(\alpha / j), \alpha = 1, 2, ..., N\} = 1$$

Estas contribuciones relativas se pueden interpretar geométricamente como el componente sobre un eje de la distancia entre un punto y el origen de coordenadas ($\cos^2_\alpha$); es decir, el cuadrado del coseno del ángulo entre ambos segmentos (el cuadrado del cateto contiguo dividido por el cuadrado de la hipotenusa de un triángulo rectángulo).

Para evaluar la consistencia de una modalidad por medio de los primeros m ejes se deben sumar las contribuciones relativas de esos m ejes sobre tal modalidad.

Para interpretar un factor es conveniente elegir un reducido número de modalidades cuya contribución a la inercia sea fuerte, por lo que es deseable encontrar los puntos para los que la contribución relativa de modalidad a es elevada. Examinando las contribuciones relativas de los factores se puede saber si se encuentra alejado o no de la dirección del subespacio considerado.

2.2.9. Filas y columnas suplementarias y valores de prueba

Las filas y columnas utilizadas para calcular los planos factoriales se denominan elementos activos: deben formar un conjunto homogéneo y exhaustivo (describir completamente el tema) para que las distancias entre elementos puedan ser fácilmente interpretables.

Se analizan como elementos suplementarios observaciones obtenidas bajo condiciones pocas claras o distintas de las del resto, o elementos atípicos o de distinta naturaleza del resto.

Sea i^+ una fila suplementaria, $K(i^+, j)$ el valor del elemento de la fila i^+ columna j y $K(i^+)$ su marginal. Su perfil será:

$f_j^{i^+} = \left\{ f_j^{i^+}, j \in J \right\}$, siendo: $f_j^{i^+} = \dfrac{K(i^+, j)}{K(i^+)}$ y su masa f_{i^+}. La coordenada de esta fila suplementaria sobre el factor α es:

$$\psi_\alpha(i^+) = \frac{1}{\sqrt{\lambda_\alpha}} \left\{ f_j^{i^+} \cdot \varphi_\alpha(j), j \in J \right\}, \forall\, i \in I$$

La masa de la columna suplementaria j^+ es f_{j^+} y su perfil:

$$f_i^{j^+} = \left\{ f_i^{j^+}, i \in I \right\}, \text{ siendo } f_i^{j^+} = \frac{K(i, j^+)}{K(j^+)}$$

La coordenada de esta columna sobre el factor α es:

$$\varphi_\alpha(j^+) = \frac{1}{\sqrt{\lambda_\alpha}} \left\{ f_i^{j^+} \cdot \psi_\alpha(i), i \in I \right\}, \forall\, j \in J$$

Las contribuciones relativas de modalidad a factor son nulas, pero las contribuciones de factor a modalidad si pueden utilizar para interpretar más fácilmente dichos elementos.

Un resultado que tiene interés para los elementos suplementarios es el valor de prueba. Si los individuos caracterizados por una modalidad están aleatoriamente repartidos en la población, su centro de gravedad estará próximo al del conjunto de individuos; lo que mide el valor de prueba es la distancia entre los centros de gravedad, distancia que es convertida en variable normal típica. Para que el elemento esté significativamente relacionado con el factor su valor absoluto ha de ser mayor que 1.96.

Capítulo 3

Escalamiento multidimensional

3.1 Introducción

Esta técnica por objetivo describir e interpretar las relaciones existentes entre un conjunto de datos. Inicia considerando las distancias o disimilaridades entre un conjunto de n individuos u objetos las cuales conforman una matriz D, cuadrada de $n \times n$. Esta puede representar las diferencias entre n productos fabricados por una empresa, las percepciones entre n candidatos políticos, diferencias entre n sectores industriales, etc. Estas distancias pueden generarse a partir de ciertas variables, o pueden ser resultado de una estimación directa, por ejemplo preguntando a un grupo de jueces por sus opiniones sobre las similitudes entre los elementos considerados.

Dado un conjunto de coordenadas pertenecientes a n individuos, es fácil calcular la distancia euclidiana o cualquier otro tipo de medida de disimilaridad entre cada par de objetos, sin embrago, los métodos de escalamiento trabajan de manera inversa, dada la información correspondiente a las distancias entre individuos, se trata de encontrar las coordenadas de los puntos que tengan una disimilaridad lo más pequeña posible respecto a la configuración original, preferentemente en un número pequeño de dimensiones, usualmente requerido de dimensión 2 ó 3, a fin

de que puedan ser reconocidos patrones asociados al comportamiento de los datos originales.

Las técnicas de escalamiento se clasifican en dos: clásico y por otra parte métrico y no métrico. La primera consiste en un método algebraico que reconstruye las coordenadas considerando que las disimilaridades se encuentran representadas por distancias euclidianas; fue creado por Torgerson en los años cincuentas y fue popularizado por Gower en 1966 bajo el nombre de análisis de coordenadas principales.

El segundo basa su construcción a partir de la idea de minimizar la varianza entre las disimilaridades observadas y las encontradas en una dimensión específica. Se llama métrico, cuando la matriz inicial es propiamente de distancias, y no métrico, cuando la matriz es de similaridades. Este procedimiento involucra métodos de optimización numérica y requiere la utilización de algoritmos que sean implementados en una computadora.

3.2 Medidas de proximidad

En virtud de que el escalamiento multidimensional se basa en la comparación de objetos, es necesario definir los criterios con los que se llevará a cabo dicha comparación.

3.2.1 Similaridades

Se denomina similaridad a la medida que indica el grado de semejanza o similitud entre dos objetos i y j, por ejemplo $s(i,j)$. Entre mayor sea la semejanza, mayor

será el valor de la similaridad. Por lo tanto, un valor pequeño indicará poca semejanza entre los objetos en comparación.

3.2.2 Disimilaridades

Se llama disimilaridad a la medida que indica el grado de diferencia entre los objetos i y j, cuya notación será $\delta(i,j)$. A mayor diferencia entre dos objetos mayor será el valor de $\delta(i,j)$.

3.2.3 Distancias

Se denomina distancia a la función denotada por $d(r,s)$, que cumple las siguientes propiedades:

1) $d_{ij} \geq 0$

2) $d_{ii} = 0$

3) $d_{ij} = d_{ji}$

4) $d_{ij} \leq d_{it} + d_{jt}$

5) $d_{ij} = 0 \Leftrightarrow i \equiv j$

6) $d_{ij} \leq \max\{d_{it}, d_{jt}\}$

7) d_{ij} es euclidiana, significa que existe un espacio euclidiano R^m y dos puntos $P_i, P_j \in R^m$, de coordenadas $P_i = (x_{i1},...,x_{im})$, $P_j = (x_{j1},...,x_{jm})$

en donde $d_{ij} = d_{2(P_i,P_j)} = \left[\sum_{h=1}^{m}\left(x_{ih} - x_{jh}\right)^2\right]^{\frac{1}{2}}$

Conforme a las propiedades citadas, las distancias se clasifican así:

Denominación	P1	P2	P3	P4	P5	P6	P7
Disimilaridad	✓	✓	✓				
Distancia métrica	✓	✓	✓	✓	✓		
Distancia ultramétrica	✓	✓	✓			✓	
Distancia euclídiana	✓	✓	✓	✓			✓

3.3 Clasificación de modelos

3.3.1 Escalamiento clásico

Se denotará $d(i,j)=d_{ij}$ a la distancia o disimilaridad entre los objetos i y j de un conjunto de n objetos; con base en éstas se genera una matriz que se denominará $D=\left[d_{ij}\right]$ de dimensión $n\times n$.

Se llama solución clásica al modelo basado en una matriz D siempre que ésta sea conformada por distancias euclidianas. Lo anterior puede comprobarse si la matriz B $b_{ij}=-\frac{1}{2}\left[d_{ij}^2-d_{i.}^2-d_{.j}^2+d_{..}^2\right]$, es semidefinida positiva[3].

Para los casos en los cuales no se conoce la distancia entre los puntos sino sus similaridades, las cuales se denotarán $C=(c_{ij})$, también es aplicable el escalamiento multidimensional una vez que se hayan transformado dichas similaridades en distancias a través del siguiente cálculo:

[3] Ver la definición en el anexo I y el teorema correspondiente en el Anexo II

$$d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{\frac{1}{2}}$$

Ahora bien, si la matriz de similaridades es semidefinida positiva, la matriz de distancias que se obtiene de su transformación es euclidiana.

Cuando se dispone de las coordenadas de los n individuos en un espacio euclidiano p-dimensional, se puede calcular fácilmente la distancia euclidiana entre cada par de puntos. Una forma de hacerlo es a través de la matriz B de $(n \times n)$, en donde $B = XX^T$, la cual representa la matriz de las sumas de cuadrados y productos entre individuos.

El (i, j)-ésimo término de la matriz B se define de la siguiente forma:

$$b_{ij} = \sum_{k=1}^{p} x_{ik} x_{jk}$$

A partir de esta expresión se puede construir la matriz D de distancias, en donde d_{ij}^2 representa el cuadrado de la distancia euclidiana entre los puntos i y j:

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = \sum_{k=1}^{p} (x_{ik}^2 - 2x_{ik}x_{jk} + x_{jk}^2) = \sum_{k=1}^{p} x_{ik}^2 - 2\sum_{k=1}^{p} x_{ik}x_{jk} + \sum_{k=1}^{p} x_{jk}^2$$

$$d_{ij}^2 = b_{ii} - 2b_{ij} + b_{jj} \text{ .....(1)}$$

Por otro lado, si se considera el problema inverso y se supone que se conocen las distancias pero no las coordenadas de los puntos que les dan origen, primero hay que encontrar la matriz B y posteriormente factorizarla de la forma $B = XX^T$.

Dado que se quiere encontrar b_{ij} en términos de d_{ij}^2, y no existe una solución única, se impone la restricción de que el centro de gravedad $\bar{x}$, se sitúe en el origen de coordenadas, de esta forma $\sum_{i=1}^{n} x_{ik} = 0$, $\sum_{j=1}^{n} x_{jk} = 0$, para toda k.

Sumando la expresión (1) sobre todos los valores de i y j, se obtienen las siguientes desigualdades:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} \left[b_{ii} - 2b_{ij} + b_{jj} \right] = T - 2\sum_{i=1}^{n} b_{ij} + nb_{jj} \text{ ; donde } T = \sum b_{ii}$$

$$\sum_{i=1}^{n} d_{ij}^2 = T - 2\sum_{i=1}^{n}\sum_{k=1}^{p} x_{ik} x_{jk} + nb_{jj}$$

$$= T - 2\sum_{k=1}^{p} x_{jk} \sum_{i=1}^{n} x_{ik} + nb_{jj}$$

$$= T + nb_{jj}$$

$$\sum_{j=1}^{p} d_{ij}^2 = \sum \left[b_{ii} - 2b_{ij} + b_{jj} \right]$$

$$= T - 2\sum_{j=1}^{p} b_{ij} + nb_{ii}$$

$$= T - 2\sum_{i=1}^{n}\sum_{k=1}^{p} x_{ik} x_{jk} + nb_{ii}$$

$$= T - 2\sum_{k=1}^{p} x_{jk} \sum_{j=1}^{n} x_{ik} + nb_{ii}$$

$$= T + nb_{ii}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \left[b_{ii} - 2b_{ij} + b_{jj} \right]$$

$$= \sum_{i=1}^{n} [T + nb_{ii}]$$

$$= nT + n\sum_{i=1}^{n} b_{ii}$$

$$= nT + nT$$

$$= 2nT$$

donde $T = \sum_{i=1}^{n} b_{ii} = \sum_{j=1}^{n} b_{jj}$ *es la traza de la matriz* B

$$d_{ij}^2 = b_{ii} - 2b_{ij} + b_{jj} \quad \text{..... (1)}$$

$$\sum_i d_{ij}^2 = T + nb_{jj} \quad \text{..... (2)}$$

$$\sum_j d_{ij}^2 = T + nb_{ii} \quad \text{..... (3)}$$

$$\sum_{ij} d_{ij}^2 = 2nT \quad \text{..... (4)}$$

De (1) se obtiene lo siguiente:

$$b_{ij} = \frac{b_{ii} + b_{jj} - d_{ij}^2}{2} = -\frac{1}{2}(d_{ij}^2 - b_{ii} - b_{jj}) \quad \text{..... (5)}$$

De (2) se obtiene:

$$b_{jj} = \frac{1}{n}\left[\sum_i d_{ij}^2 - \sum_i b_{ii}\right] \quad \text{..... (6)}$$

De (3) se obtiene:

$$b_{ii} = \frac{1}{n}\left[\sum_j d_{ij}^2 - \sum_i b_{jj}\right] \quad \text{.....… (7)}$$

Sustituyendo (6) y (7) en (5) se llega a que:

$$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - \frac{1}{n}\left(\sum_j d_{ij}^2 - \sum_i b_{ii} \right) - \frac{1}{n}\left(\sum_j d_{ij}^2 - \sum_i b_{ii} \right) \right]$$

$$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - \frac{1}{n}\sum_j d_{ij}^2 + \frac{2}{n}\sum_i b_{ii} - \frac{1}{n}\sum_i d_{ij}^2 \right]$$

$$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - \frac{1}{n}\sum_j d_{ij}^2 + \frac{2}{n}T - \frac{1}{n}\sum_i d_{ij}^2 \right]$$

Por (4) $\sum_{ij} d_{ij}^2 = 2nT \Rightarrow \frac{\sum_{ij} d_{ij}^2}{n^2} = \frac{2T}{n}$

$$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - \frac{1}{n}\sum_j d_{ij}^2 + \frac{\sum_{ij} d_{ij}^2}{n^2} - \frac{1}{n}\sum_i d_{ij}^2 \right]$$

Denotando por $d_{i.}^2 = \frac{1}{n}\sum_i d_{ij}^2, d_{.j}^2 = \frac{1}{n}\sum_j d_{ij}^2, d_{..}^2 = \frac{\sum_{ij} d_{ij}^2}{n^2}$, se llega a la siguiente expresión:

$$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2 \right]$$

Por medio de la ecuación anterior se puede reconstruir la matriz B a partir de las distancias dadas entre cada par de individuos.

Posteriormente se factoriza la matriz B en la forma $B = XX^T$ obteniéndose así las coordenadas de los puntos a partir de la matriz X, para lo se debe descomponer en valores propios a la matriz B.

Por la forma en que se construyó la matriz B, se observa que es una matriz simétrica y semidefinida positiva, además si B es de rango k entonces va a tener k eigenvalores distintos de cero. Por el teorema de factorización de Young-Householder[4] es posible encontrar una matriz X tal que $B = XX^T$.

Como B es simétrica se puede descomponer como $B = Q\Lambda Q^T$, con Q la matriz ortogonal que contiene los eigenvectores de la matriz B y Λ la matriz diagonal tal que en su diagonal principal se encuentran los eigenvectores de B, es decir, $\Lambda_{ii} = \lambda_i$.

La matriz B se puede escribir como:

$B = Q\Lambda Q^T = Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q^T$ con $\Lambda_{ii}^{\frac{1}{2}} = \sqrt{\lambda_i}$, de tal forma que $\Lambda = \Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}$

Si $X = Q\Lambda^{\frac{1}{2}}$, entonces $X^T = (Q\Lambda^{\frac{1}{2}})^T = (\Lambda^{\frac{1}{2}})^T Q^T = \Lambda^{\frac{1}{2}}Q^T$

$\therefore B = Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q^T = XX^T$, con $X = Q\Lambda^{\frac{1}{2}}$

Entonces para encontrar las coordenadas de los puntos se necesita calcular la matriz $X = Q\Lambda^{\frac{1}{2}}$.

3.3.1.1 Elección del número de dimensiones

Para elegir el número de dimensiones $p^* \leq p$ para el cual se desea obtener una configuración se sugiere tomar en cuenta a los eigenvectores asociados con los p^* eigenvalores más grandes y calcular la proporción de información que es

[4] Ver Anexo I

explicada por la configuración elegida en el número de dimensiones p^* de la siguiente forma:

$$T = \left[\frac{\sum_{k=1}^{p^*} \lambda_k}{\sum_{k=1}^{p} |\lambda_k|} \right] \times 100$$

donde los λ_i son todos los eigenvalores asociados a la matriz B.

3.3.2 Escalamiento métrico

El supuesto principal es que se emplean variables cuantitativas medidas en escalas de intervalo o razón por lo que existe una función (lineal, cuadrática, etc.) que relaciona linealmente las medidas de proximidad con las distancias.

En 1952 Torgenson trabajó en el primer método de escalamiento métrico basándose en el trabajo de Young y Householder (1938, 1914); se obtenía la solución del modelo clásico pero con la suposición fundamental de que las medidas de disimilaridad son en realidad distancias euclidianas, es decir que

$$\delta_{ij} \approx d_{ij} = d(i,j)) = \left[\sum_{k=1}^{n} \left(x_{ik} - x_{jk} \right)^2 \right]^{\frac{1}{2}}$$

El método del modelo métrico parte de la matriz $D = \left[d_{ij} \right]$ de distancias euclidianas entre los objetos que se pretende analizar. Entonces se construye la matriz B, cuyos elementos están definidos por $b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2 \right]$.

Posteriormente, se obtienen los k eigenvalores positivos más grandes de B

$$\lambda_1, \lambda_2, ..., \lambda_k \quad (k \leq n)$$

con sus correspondientes eigenvectores

$$X = \left[x_{(1)}, x_{(2)}, ..., x_{(n)} \right] \text{ donde } x_{(i)}^T x_{(i)} = 1$$

Se obtiene la configuración buscada, dada por los renglones de la matriz X:

$$x_1^T, x_2^T, \ldots, x_k^T$$

3.3.3 Escalamiento no métrico

Su fundamento teórico fue elaborado por Kruskal (1964) y su hipótesis fundamental es que las medidas de proximidad, generalmente disimilaridades, están relacionadas con las distancias entre los puntos mediante una función monótona (ordenada) a diferencia del modelo métrico que supone una función exacta. Esto es congruente con la suposición de que los datos están medidos en una escala ordinal ya que sólo se utiliza su relación de orden.

En los problemas de escalamiento no métrico se parte de una matriz de diferencias o disimilitudes entre objetos, los cuales se ha obtenido generalmente por consultas o partir de procedimientos de ordenación de elementos. Por ejemplo, se ha aplicado para estudiar las semejanzas entre actitudes, preferencias o percepciones de personas sobre asuntos políticos o sociales o para evaluar preferencias respecto a productos y servicios en marketing y en calidad.

Se supone que la matriz de similaridades está relacionada con una de distancias, pero de una manera compleja. Es decir, se acepta que los encuestados utilizan en sus valoraciones ciertas variables o dimensiones, pero que, además, los datos incluyen elementos de error y variabilidad personal. Por lo tanto, las variables que explican las similitudes entre los elementos comparados determinarán una

distancia euclidiana entre ellos, d_{ij}, que están relacionadas con las similitudes dadas, δ_{ij}, mediante una función desconocida

$$\delta_{ij} = f(d_{ij})$$

donde la única condición que se impone es que f es una función monótona, es decir, si

$$\delta_{ij} > \delta_{ik} \Leftrightarrow d_{ij} > d_{ik}$$

La mayoría de los algoritmos para encontrar la solución a este modelo consiste en cuatro fases: en la primera se obtiene una configuración inicial. Después se estandarizan las coordenadas de los puntos y las distancias entre ellos. Posteriormente, se calculan las disparidades, en una fase denominada no métrica. Por último, en la llamada fase métrica se calculan las nuevas coordenadas de los puntos.

Después de calcular la configuración inicial, cada iteración consiste en una estandarización, una fase métrica y una no métrica. Al finalizar se calcula la función S-Stress, la cual nos indica la bondad de ajuste de la solución obtenida. Las iteraciones continúan hasta que el cambio en su valor de una iteración a la siguiente sea menor que algún valor establecido.

3.3.3.1 Medidas de bondad de ajuste

Se tiene la función de sumas de cuadrados, denotada como SS, en la cual si las distancias son iguales a las disimilaridades observadas se toma el valor cero y una discrepancia entre δ_{ij} y d_{ij} aumenta el valor de SS.

$$SS = \sum_{i \neq j}^{n} (\delta_{ij} - d_{ij})^2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (\delta_{ij} - d_{ij})^2$$

Dado que el ajuste es mejor cuando el valor de SS disminuye, es necesario encontrar una configuración de puntos de tal forma que SS sea mínimo.

La medida de SS es invariante bajo transformaciones tales como rotaciones, traslaciones y reflexiones, lo cual es una característica muy deseable, desafortunadamente no pasa lo mismo para cambios de escala ya que el valor de SS no es invariante ante este tipo de transformaciones.

Se introduce un factor de escala definido como:

$$SC = \sum_{i \neq j}^{n} d_{ij}^2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^2$$

de tal forma que $S = \frac{SS}{SC}$ represente una medida de bondad de ajuste que conserva todas las características de SS con la ventaja de que es invariante ante los cambios de escala.

En el escalamiento multidimensional se utiliza una estandarización de la medida S, conocida como S-Stress, que conserva igualmente todas las características de S.

$$STRESS = \sqrt{\frac{SS}{SC}} = \left[\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (\delta_{ij} - d_{ij})^2}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^2} \right]^{\frac{1}{2}}$$

Al elegir una medida de bondad de ajuste basada en el cálculo de SS, se supone implícitamente que $d_{ij} = \delta_{ij} + \varepsilon_{ij}$, donde ε_{ij} representa el error que se genera debido a que las disimilaridades observadas no ajustan perfectamente a las distancias encontradas en la dimensión elegida.

Se espera que $d_{ij} = a + b\delta_{ij} + \varepsilon_{ij}$, por lo que $SS = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij} - a - b\delta_{ij})^2$, esto último implica que dado un conjunto de puntos en una dimensión determinada y un conjunto de valores d_{ij}, se necesitan determinar los valores a y b que minimicen la suma de cuadrados anterior, lo cual se podría realizar implementando una regresión lineal de d_{ij} sobre δ_{ij}; una vez determinados los valores a y b se debe implementar un algoritmo de optimización que encuentre los puntos $X_1, X_2, ..., X_n$ que minimicen la función de Stress. El procedimiento anterior se tendría que repetir hasta que se alcanzara algún criterio de convergencia.

Generalizando, se puede definir la relación entre δ_{ij} y d_{ij} de la siguiente manera:

$$d_{ij} = f(\delta_{ij}) + \varepsilon_{ij} \quad \text{.....} \quad (8)$$

por lo que la suma de cuadrados residual SS queda:

$$SS = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij} - f(\delta_{ij}))^2$$

Una variante de la función de Stress es aquella que considera las disimilaridades al cuadrado, a esta expresión se le conoce con el nombre de S-Stress.

$$S - Stress = \left(\frac{\sum_{i \neq j}^{n} (d_{ij}^2 - \hat{d}_{ij}^2)^2}{\sum_{i \neq j}^{n} d_{ij}^2} \right)^{\frac{1}{2}}$$

Los valores $\hat{d}_{ij}$ se obtienen a partir de un método conocido con el nombre de regresión monótona.

Ahora bien, minimizar la función de Stress implica el desarrollo de dos etapas: la primera involucra algún método de regresión y la segunda se ocupa de minimizar la función a partir de un algoritmo de optimización como podría ser el método de Newton-Raphson o Stepest Descent, los cuales se implementan mediante métodos numéricos cuya explicación queda fuera del alcance del presente trabajo, sin embargo, en términos generales estos algoritmos trabajan a partir de una configuración inicial arbitraria de coordenadas en la dimensión elegida y por medio de un proceso iterativo van moviendo poco a poco los puntos de manera que la función de Stress sea más pequeña cada vez, este proceso se repite hasta que ya no sea posible mover los puntos para obtener una configuración que minimice la medida de ajuste o bien hasta que haya realizado un número específico de iteraciones previamente establecido.

3.3.3.2 Elección del número de dimensiones

Se busca obtener una solución en dos o tres dimensiones de manera que se cuente con una imagen que facilite la interpretación de los resultados, por lo que la solución obtenida con el número de dimensiones elegido deberá ser evaluada a fin de encontrar la mejor posible; un criterio que se utiliza para determinar el grado de ajuste a partir de la configuración obtenida para una dimensión específica, es a través de la función de S-Stress.

En 1964, Kruskal publicó algunos criterios de bondad de ajuste para la función de S-Stress, los cuales se formularon a partir de la experiencia propia, de esta forma se tienen los siguientes parámetros:

S-Stress %	**Ajuste**
20	Pobre
10	Regular
5	Bueno
2.5	Excelente
0	Perfecto

Capítulo 4

Aplicaciones del escalamiento multidimensional

En este capítulo se realizan dos ejemplos para el escalamiento multidimensional, mediante los cuales se segmentarán los mercados de los seguros de salud y seguros de pensiones derivados de las leyes de seguridad social, por considerar que las condiciones actuales de la población mexicana con una tendencia al envejecimiento en su pirámide poblacional, aunado a una concientización de la importancia del uso de estas modalidades del seguro, incidirá en el incremento en la demanda de este tipo de seguros.

4.1 Instituciones de Seguros Especializadas en Salud (ISES)

En los decretos publicados en el Diario Oficial de la Federación del 3 de enero de 1997 y 31 de diciembre de 1999, el Congreso de la Unión aprobó importantes reformas a la Ley General de Instituciones y Sociedades Mutualistas de Seguros, en particular, la operación de seguros de accidentes personales se dividió en los ramos de: accidentes personales, gastos médicos y salud.

Conforme a las Reglas de Operación para el Seguro de Salud, “El ramo de salud constituye la base para que las sociedades u organizaciones conocidas como

entidades administradoras de medicina prepagada, se transformen en Instituciones de Seguros Especializadas en Salud (ISES)". Su finalidad es la prestación de servicios dirigidos a prevenir o restaurar la salud a través de acciones que se realicen en beneficio del asegurado, comercializándolos a futuro y el cumplimiento de la obligación de prestarlos depende de un acontecimiento futuro e incierto.

Al mes de marzo de 2007, el mercado de las ISES se encontraba conformado por 11 instituciones:

INSTITUCIONES AUTORIZADAS	
701	Plan Seguro, S.A. de C.V., Compañía de Seguros
702	Médica Integral GNP, S.A de C.V.
703	Seguros del Sanatorio Durango, S.A. de C.V.
704	Preventis, S.A. de C.V., Grupo Financiero BBVA Bancomer
705	ING Salud, S.A. de C.V.
706	Salud Inbursa, S.A.
707	General De Salud, Compañía De Seguros, S.A.
708	Novamedic Seguros de Salud, S.A. de C.V.
710	Vitamédica, S.A. de C.V.
711	Servicios Integrales de Salud Nova, S.A. de C.V.
712	Seguros Centauro, Salud Especializada, S.A. de C.V.
713	Saludcoop México, S.A. de C.V.

4.2 Instituciones del seguro de pensiones, derivadas de las Leyes de Seguridad Social (ISPDLSS)

Con escrito del 10 de febrero de 1997, publicado en el Diario Oficial de la Federación el 26 de febrero del mismo año, se dieron a conocer las Reglas de operación para los seguros de pensiones derivados de las leyes de seguridad social.

Conforme a la Ley General de Instituciones y Sociedades Mutualistas de Seguros en su artículo 8°, fracción II, para los Seguros de Pensiones derivados de las leyes de seguridad social, se considera el pago de las rentas periódicas durante la vida del asegurado o las que correspondan a sus beneficiarios de acuerdo con los contratos de seguro celebrados en los términos de la ley aplicable.

Al mes de marzo de 2007, el mercado de las ISPDLSS se conformaba por las siguientes instituciones:

INSTITUCIONES AUTORIZADAS	
901	ING Pensiones, S.A. de C.V.
902	Pensiones Banorte Generali, S.A. de C.V., Grupo Financiero Banorte
903	HSBC Pensiones, S.A.
905	Pensiones BBVA Bancomer, S.A. de C.V., Grupo Financiero BBVA Bancomer
906	Profuturo GNP Pensiones, S.A. de C.V.

INSTITUCIONES AUTORIZADAS	
907	Principal Pensiones, S.A. de C.V.
908	Pensiones Banamex, S.A. de C.V., Grupo Financiero Banamex
909	Metlife Pensiones México, S.A.
910	Pensiones Inbursa, S.A., Grupo Financiero Inbursa
911	Royal and Sunalliance Pensiones (México), S.A. de C.V.

4.3 Variables de análisis

Para llevar a cabo la segmentación de estos mercados, se utilizó la información de los estados financieros de las instituciones con corte a la fecha antes señalada, del cual se eligieron las siguientes variables:

RUBRO DEL ESTADO DE RESULTADOS	VARIABLE
Primas emitidas	PE
Incremento neto de la reserva de riesgos en curso	IRRC
Primas de retención devengadas	PRD
Costo neto de adquisición	CNA
Costo neto de siniestralidad, reclamaciones y otras obligaciones contractuales	CNS
Resultado técnico (utilidad o pérdida)	RT
Gastos de operación netos	GON
Resultado de operación (utilidad o pérdida)	RO

RUBRO DEL ESTADO DE RESULTADOS	VARIABLE
Resultado integral de financiamiento o productos financieros	RIF
Resultado del ejercicio (utilidad o pérdida)	RE

4.4 Ejemplo para el mercado de las ISES

Para el mercado de las ISES, la información es la siguiente:

ISPDSS	VARIABLES*									
	PE	IRCC	PRD	CNA	CNS	RT	GON	RO	RIF	RE
701	94.7	14.3	80.3	21.7	47.5	11.1	13.2	-2.0	3.1	1.1
702	58.3	24.3	34.1	9.9	33.3	-9.1	12.4	-21.3	-2.8	-24.9
703	23.1	3.5	19.6	1.8	15.9	1.8	2.0	-0.2	0.5	0.3
704	15.9	-12.8	28.8	1.5	19.9	7.4	5.6	1.8	1.0	5.6
705	30.1	-5.7	35.9	5.5	19.3	11.2	7.6	3.6	3.2	4.7
706	0.0	-0.6	0.6	0.0	0.6	0.0	0.6	-0.6	0.0	-0.7
707	13.4	-2.2	15.7	3.4	7.2	5.0	3.2	1.8	1.0	2.0
708	45.6	30.2	15.4	4.7	3.8	6.9	4.7	2.2	-0.1	2.2
710	0.1	-0.3	0.4	0.4	0.2	-0.2	38.5	-39.0	45.0	5.4
711	27.2	-6.1	33.3	2.7	24.3	6.4	3.2	3.2	-0.4	2.0
712	14.6	2.4	12.2	1.0	6.8	4.4	3.8	0.6	0.8	1.4
713	1.6	0.4	1.3	0.2	1.8	-0.7	3.1	-3.8	0.1	-3.7

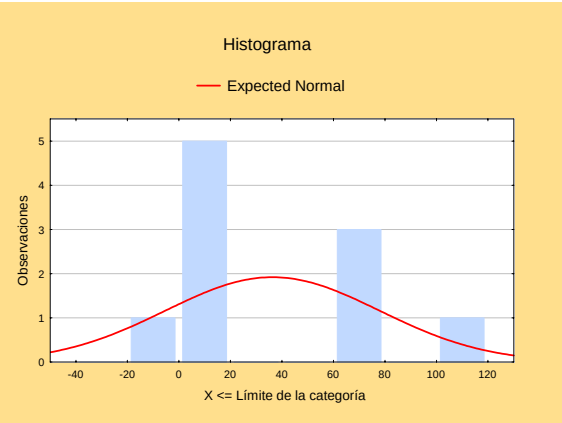
*Cifras en millones de pesos

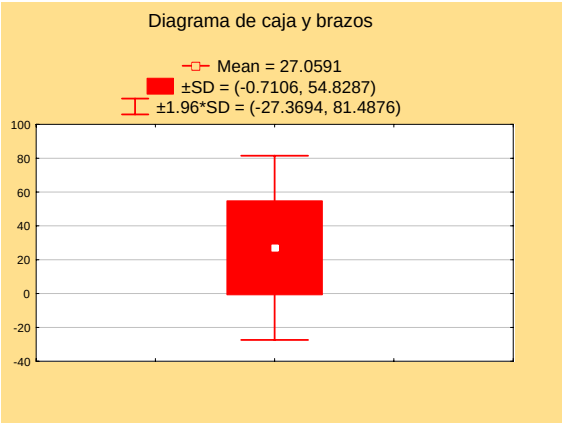
4.4.1 Análisis descriptivo de las variables

Una vez elegidas las variables, se lleva a cabo un análisis exploratorio de la información, de tal forma que se pueda detectar alguna inconsistencia o caso particular que deba ser tomado en cuenta antes de aplicar el modelo.

	Casos	Media	Mínimo	Máximo	Desviación Estándar
PE	12	27.06	0.00	94.66	27.77
IRRC	12	3.93	-12.84	30.16	12.71
PRD	12	23.13	0.44	80.33	22.23
CNA	12	4.39	0.04	21.68	6.14
CNS	12	15.06	0.21	47.51	14.68
RT	12	3.68	-9.15	11.16	5.73
GON	12	8.16	0.61	38.53	10.32
RO	12	-4.48	-39.00	3.59	12.74
RIF	12	4.29	-2.78	45.05	12.93
RE	12	-0.39	-24.93	5.61	8.16

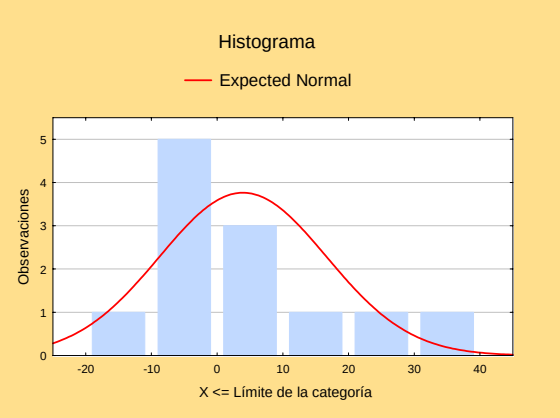
Primas emitidas

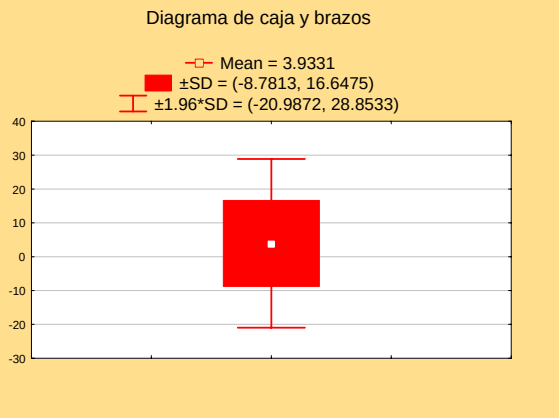




Las primas emitidas presentan un promedio de $27.06[1], observándose una emisión nula o escasa para las instituciones 706, 710 y 713, no obstante la emisión de $94.7 que reporta la institución 701. Lo anterior, general una dispersión considerable en los datos, al reportar una desviación estándar de $27.7.

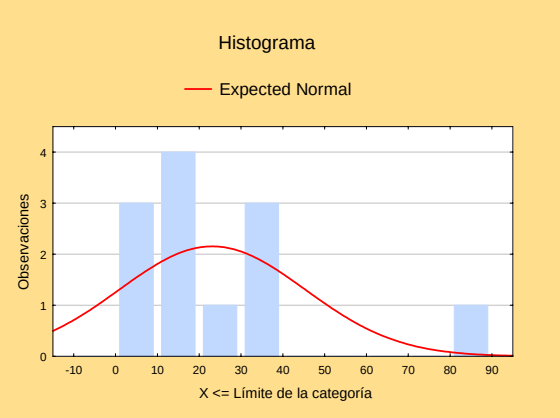
Incremento de la reserva de riesgos en curso

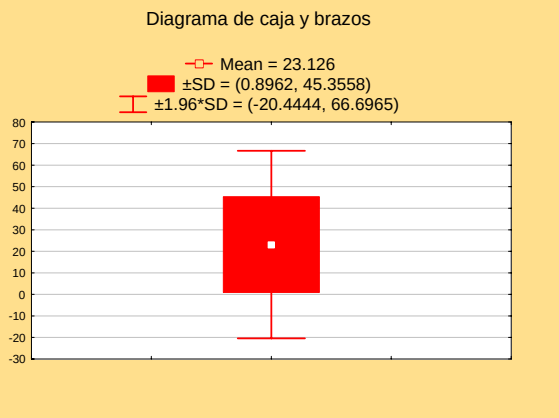




Esta variable reporta valores más bien bajos para el mercado en general, sólo las instituciones 702 y 708 presentan cifras elevadas. Su desviación es de $12.7.

Prima de riesgo no devengada

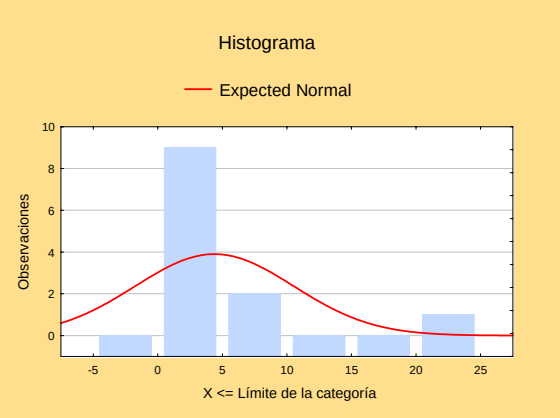


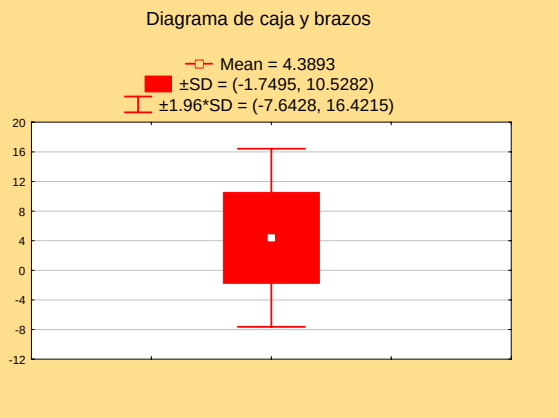


[1] Todas las cifras utilizadas en esta sección están expresadas en millones de pesos

La prima de riesgo no devengada reporta una desviación de $22.3, es decir, se observa dispersión en los datos, los cuales van desde $0.4 hasta $80.3, presentando una media por $23.1.

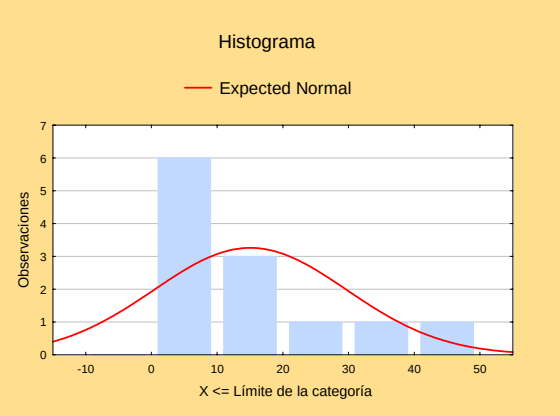
Costo neto de adquisición

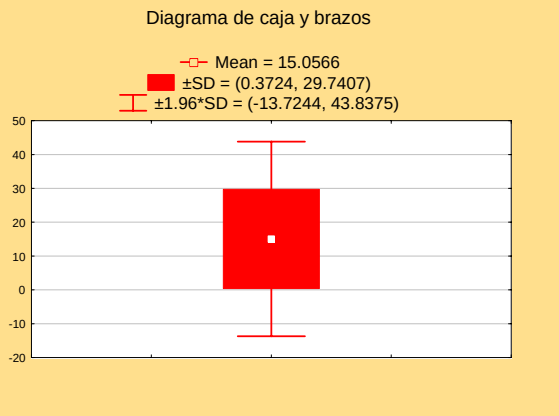




De las 12 ISES, 11 reportan en promedio costos por adquisición de $2.8, sin embargo, la institución 701 reporta un monto de $21.7 lo cual eleva el promedio del mercado a $4.4.

Costo neto de siniestralidad

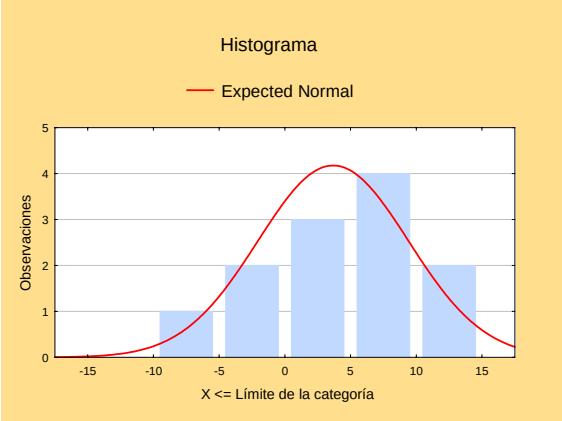


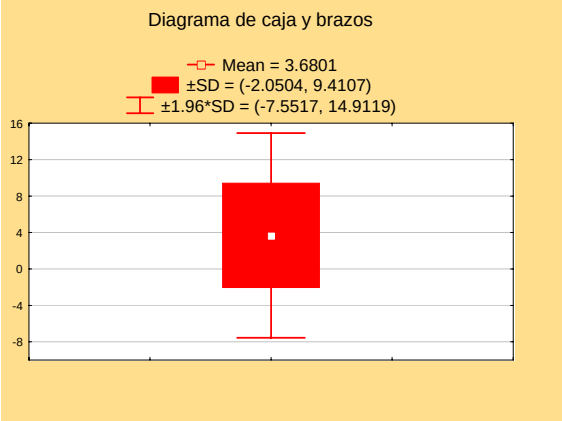


El costo de siniestralidad es presenta una división marcada de dos grupos de instituciones: la primera reporta montos superiores a $15.0 y hasta $47.5, mientras

que el segundo grupo va desde $0.2 hasta $6.8, reportando una desviación estándar por $14.7.

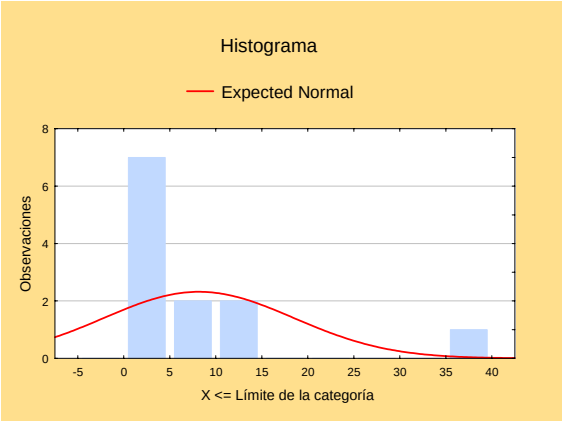
Resultado técnico

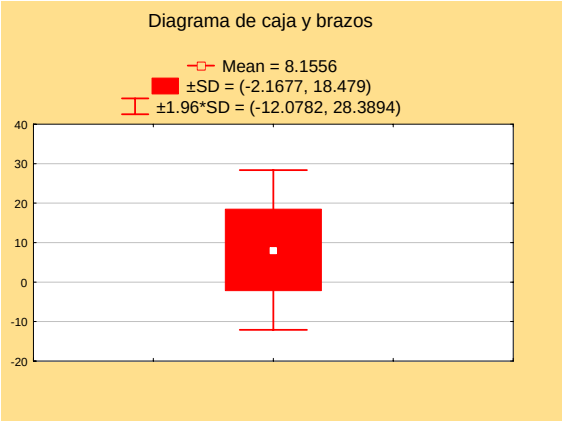




No obstante los costos de siniestralidad, el mercado en promedio reporta un resultado técnico positivo de $3.7. Es la variable con menor dispersión en el análisis.

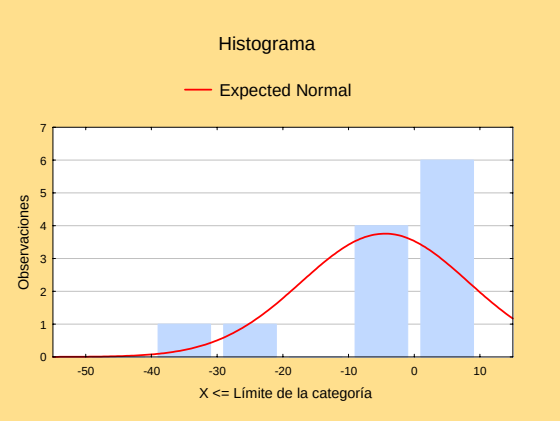
Gastos operación netos

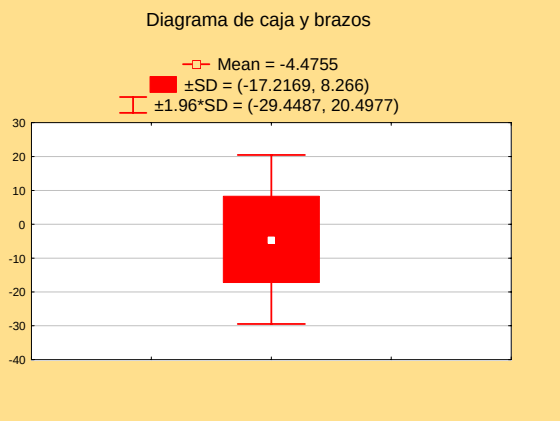




Todas las instituciones reportan saldos positivos para este rubro, los cuales van desde $0.6 para la institución 706, hasta $38.5 para la 710, lo cual no concuerda con la emisión que esta reporta.

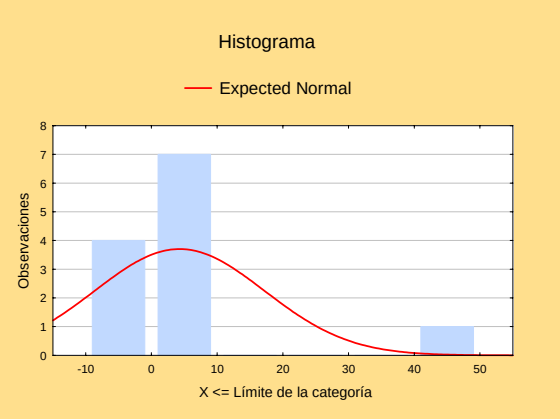
Resultado de la operación

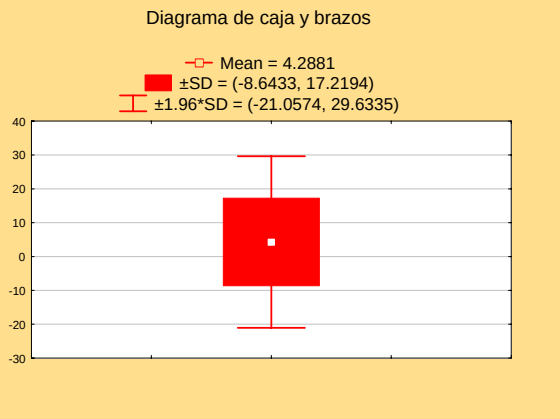




Como consecuencia de las altos gastos de operación, siniestralidad y costos de adquisición, las instituciones 701, 702, 706, 710 y 713, reportan un resultado de operación negativo. La media del mercado es de -$4.5

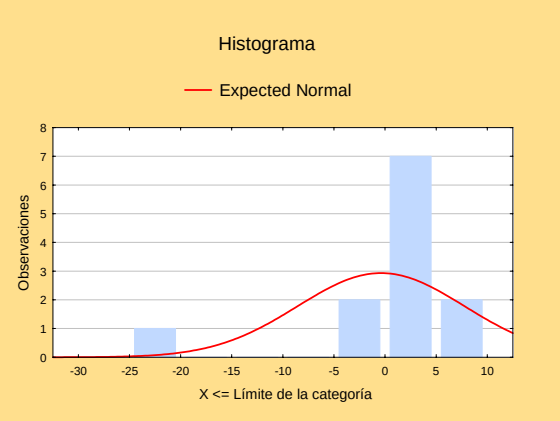
Resultado integral de financiamiento

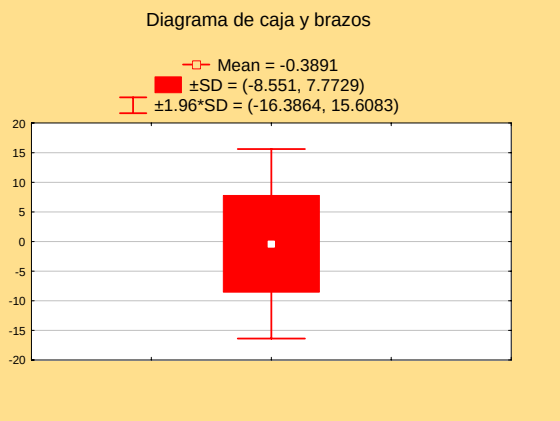




Esta variable reporta una desviación estándar alta por $12.9, dado que sus valores máximo y mínimo son de $45.1 y -$2.8 respectivamente. No obstante el promedio del mercado es positivo por $4.3

Resultado del ejercicio





A nivel mercado se observa un resultado del ejercicio negativo por -$0.4. No obstante, el 75% de las ISES reportan utilidades aunque por cifras bajas.

4.4.2 Segmentación del mercado de las ISES

Para obtener el análisis anterior, se empleó el programa ASCAL del paquete estadístico SPSS (Statistical Package for Social Sciences) en su versión 11.0.0.

El primer paso consiste en agrupar toda la información antes señala en una tabla de información a partir de la cual se obtiene una matriz de disimilitudes:

ISES	MATRIZ DE DISIMILITUDES											
	701	702	703	704	705	706	707	708	710	711	712	713
701	0.0	73.4	102.7	104.3	87.6	136.0	114.8	95.9	148.8	90.8	115.9	133.9
702	73.4	0.0	59.7	72.5	62.3	86.9	73.0	55.2	102.9	60.7	71.4	82.9
703	**102.7**	59.7	0.0	22.2	24.2	34.0	15.5	37.9	77.5	20.1	15.0	32.3
704	**104.3**	72.5	22.2	0.0	18.6	41.2	21.9	56.5	79.6	15.4	26.7	40.7
705	87.6	62.3	24.2	18.6	0.0	52.8	30.4	47.3	85.0	10.5	33.6	51.7
706	**136.0**	86.9	34.0	41.2	52.8	0.0	22.6	58.0	70.5	49.8	20.8	5.6

707	**114.8**	73.0	15.5	21.9	30.4	22.6	0.0	45.9	73.2	28.5	6.6	22.1
708	95.9	55.2	37.9	56.5	47.3	58.0	45.9	0.0	90.5	49.0	42.2	56.4
710	**148.8**	102.9	77.5	79.6	85.0	70.5	73.2	90.5	0.0	87.1	71.9	67.8
711	90.8	60.7	20.1	15.4	10.5	49.8	28.5	49.0	87.1	0.0	31.6	48.6
712	**115.9**	71.4	15.0	26.7	33.6	20.8	6.6	42.2	71.9	31.6	0.0	19.7
713	**133.9**	82.9	32.3	40.7	51.7	5.6	22.1	56.4	67.8	48.6	19.7	0.0

Se observa que la institución 701 reporta mayores diferencias con respecto a la demás ISES, lo anterior se confirmará en el mapa de percepción.

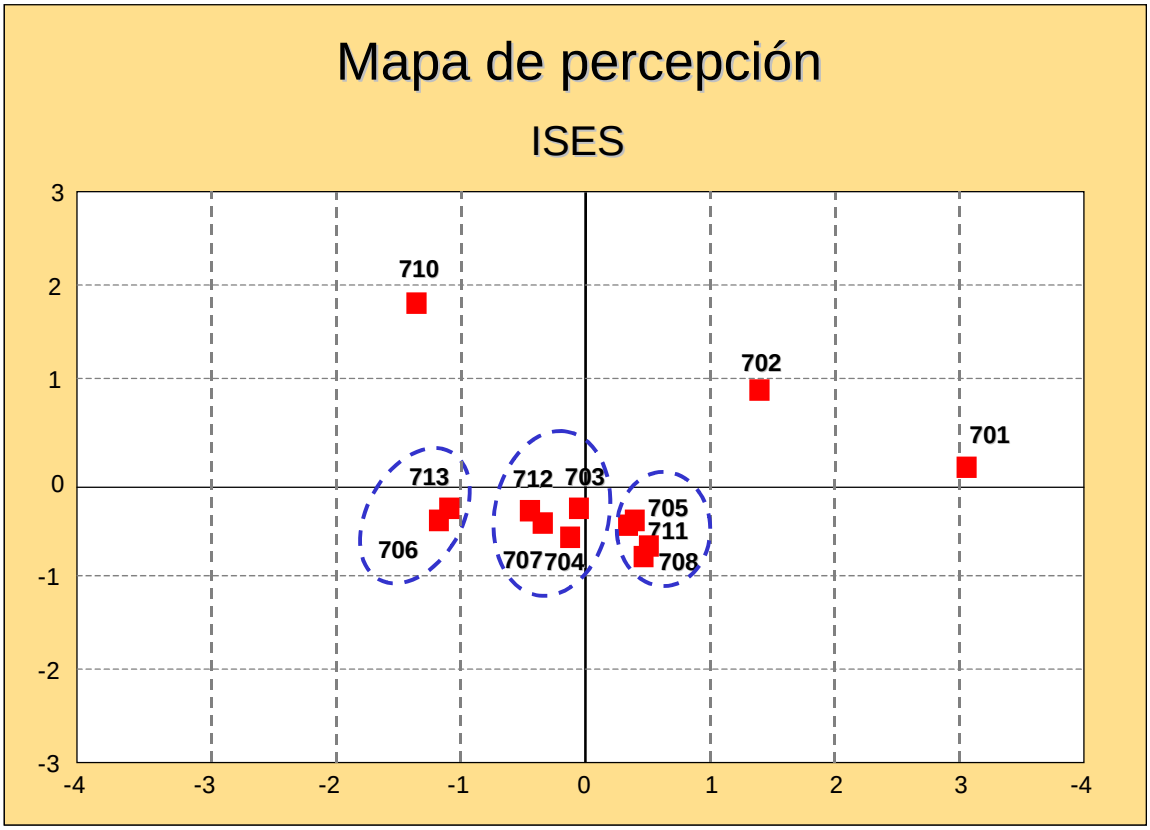
A continuación se presentan los resultados que el paquete SPSS genera al aplicar el programa ASCAL a la matriz de disimilitudes antes exhibida, para un máximo de 30 iteraciones posibles, un criterio de convergencia de 0.00100 y un valor mínimo de S-Stress de 0.00500:

El proceso sólo generó 3 iteraciones deteniéndose porque la mejora en la función S-Stress alcanzó un nivel menor a 0.001000.

Iteración	S-Stress	Mejoramiento
1	0.11935	
2	0.10331	0.01603
3	0.10271	0.00060

La función S-Stress resultante fue de 0.13149 es decir, 13.14% el cual conforme a los criterios de bondad de ajuste para la función de Stress de Kruskal[2] se clasifica como regular.

[2] Ver capítulo III



Conforme al escalamiento multidimensional, se obtienen 3 grupos de ISES integrados por:

Grupo A: Salud Inbursa y SaludCoop.

Grupo B: Seguros Centauro, Durango, General de Salud y Preventis.

Grupo C: ING Salud, Salud Nova y Novamedic.

Las instituciones Plan Seguro, Médica Integral y Vitamédica, no pertenecen a ningún grupo en particular.

4.5 Ejemplo para el mercado de las ISPDLSS

Para el mercado de las ISPDLSS, la información es la siguiente:

ISPDLSS	VARIABLES*									
	PE	IRCC	PRD	CNA	CNS	RT	GON	RO	RIF	RE
901	7.4	-23.5	30.9	0.0	71.9	-41.1	2.0	-44.7	142.8	70.3
902	324.7	263.3	61.4	15.3	120.4	-74.3	12.2	-96.0	287.2	113.2
903	121.3	78.1	43.2	10.4	112.5	-79.7	7.9	-89.0	116.0	15.7
905	424.2	317.6	106.6	18.8	259.7	-171.9	16.8	-208.7	315.4	75.8
906	273.9	225.4	48.4	31.6	185.5	-168.6	4.3	-181.9	209.9	18.1
907	21.0	-46.2	67.3	0.3	46.5	20.4	14.1	4.6	2.4	7.0
908	221.8	174.0	47.8	8.9	133.6	-94.7	10.2	-113.7	207.0	64.6
909	2.1	-12.9	15.1	0.0	67.8	-52.8	4.1	-65.4	84.1	13.3
910	2.6	-39.2	41.9	0.1	192.7	-150.9	4.8	-158.2	298.7	-15.1
911	0.0	-0.5	0.5	0.0	3.3	-2.8	0.7	-3.6	3.8	0.2

*Cifras en millones de pesos

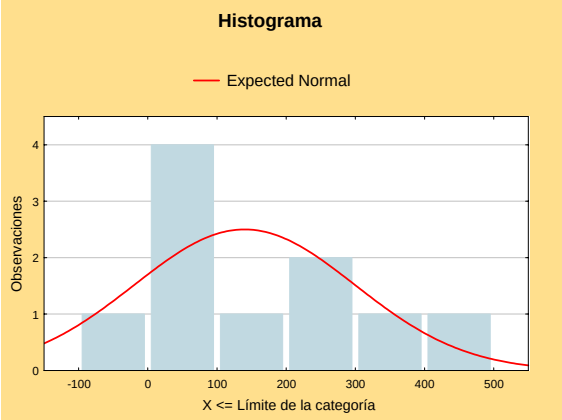
4.5.1 Análisis descriptivo de las variables

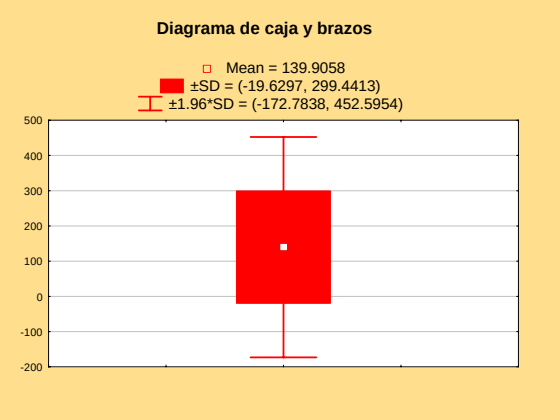
Con base en las variables anteriores, se lleva a cabo un análisis exploratorio de la información, de tal forma que se pueda detectar alguna inconsistencia o caso particular que deba ser tomado en cuenta antes de aplicar el modelo.

	Casos	Media	Mínimo	Máximo	Desviación Estándar
PE	10	139.91	0.00	424.20	159.54
IRRC	10	93.60	-46.21	317.58	139.11
PRD	10	46.30	0.55	106.62	29.15
CNA	10	8.56	0.00	31.56	10.75
CNS	10	119.40	3.31	259.74	77.02

RT	10	-81.65	-171.94	20.40	66.55
GON	10	7.71	0.74	16.79	5.41
RO	10	-95.66	-208.69	4.60	71.90
RIF	10	166.74	2.43	315.38	115.96
RE	10	36.32	-15.09	113.15	41.52

Primas emitidas

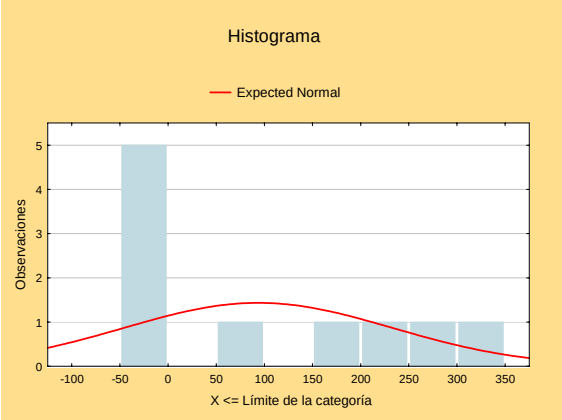


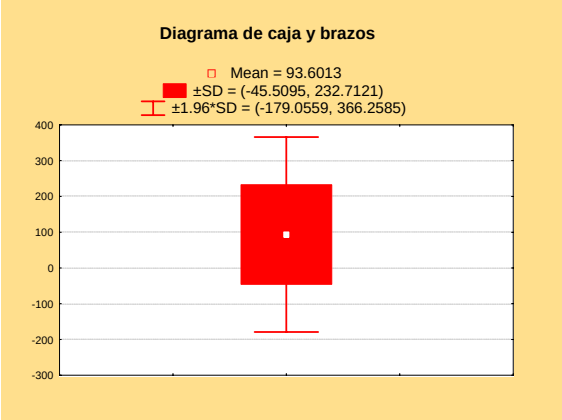


Presenta la mayor desviación estándar de las variables elegidas para el análisis, en virtud de presentar una emisión nula para la institución 911, en comparación con los $424.2[3] millones que reporta la institución 905.

Incremento en la reserva de riesgos en curso

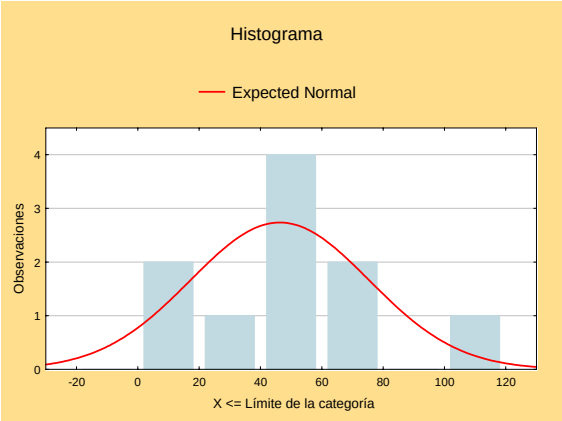
[3] Todas las cifras utilizadas en esta sección están expresadas en millones de pesos

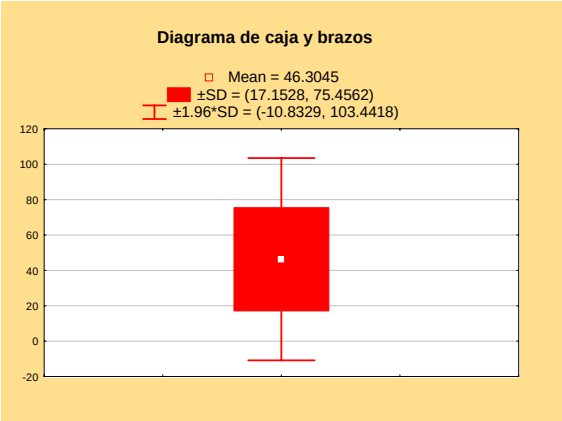




Se puede observar una gran dispersión en sus valores, al alcanzar saldos por $317.6 millones para la institución 905, en comparación con los saldos negativos que reportan la 901, 910, 909 y 907, siendo esta última en la que la variable alcanza su mínimo ($46.4).

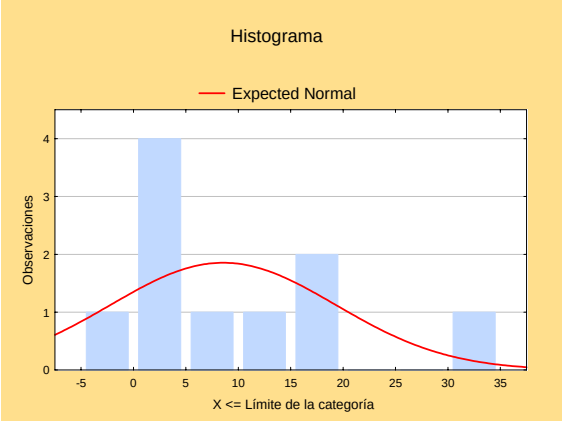
Primas retenidas devengadas

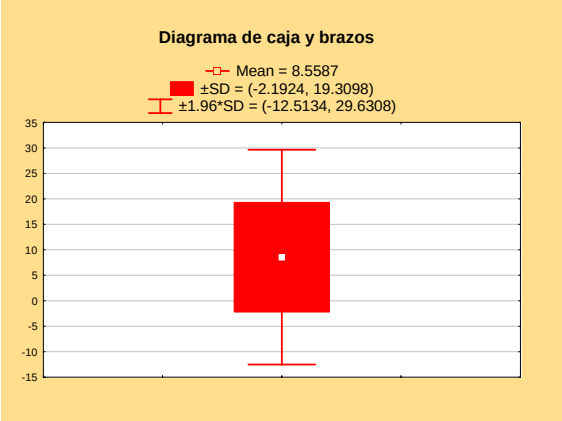




Todas las cifras utilizadas en esta sección están expresadas en millones de pesos utilizadas en esta sección están expresadas en millones de pesos $29.2 observándose nuevamente, que el máximo lo reporta la institución 905.

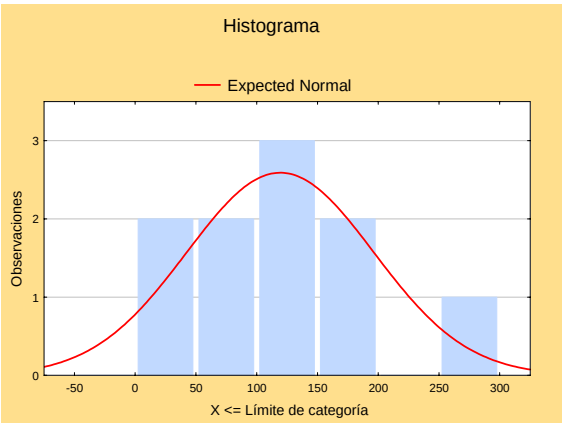
Costo neto de adquisición

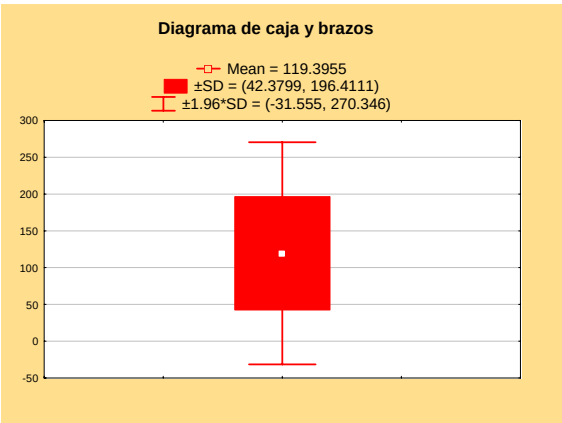




El costo neto de adquisición reporta una media de $8.6, no obstante su máximo valor de $31.6 que reporta la institución 906; lo anterior debido a que el 50% de las instituciones reportan valores por nulos o por debajo de los $500 pesos.

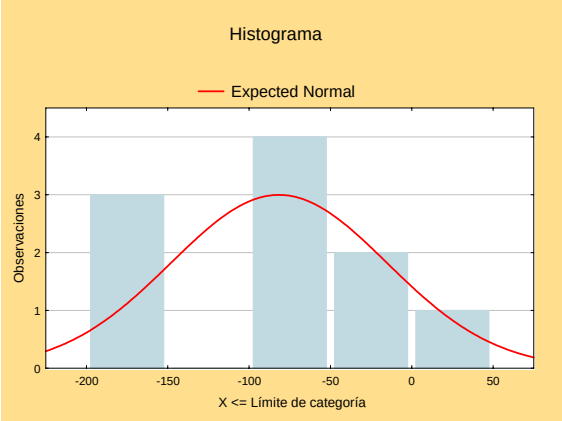
Costo neto de la siniestralidad

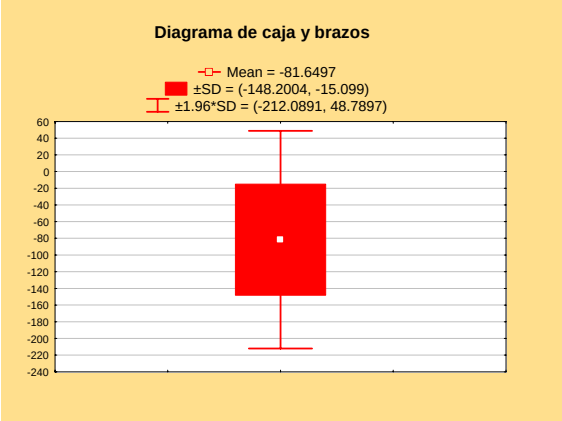




La siniestralidad reporta valores elevados en proporción a las primas retenidas devengadas de todo el mercado, ubicando su media en $119.4.

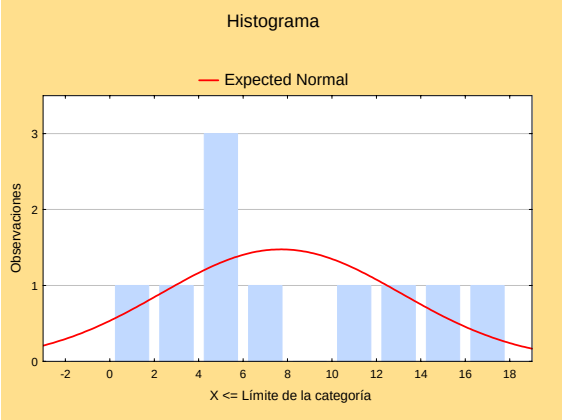
Resultado técnico

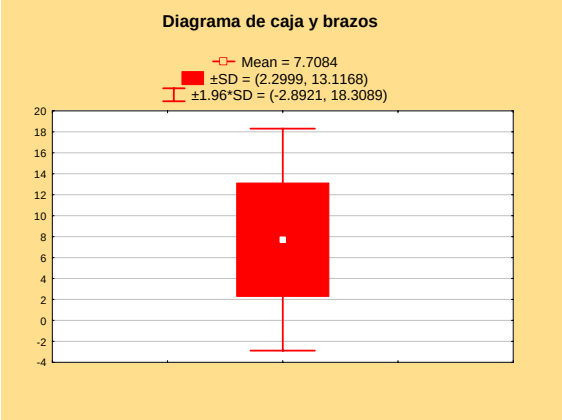




En virtud de la alta siniestralidad reportada, aunado a los costos de adquisición, el resultado técnico del mercado es negativo, excepto para la institución 907.

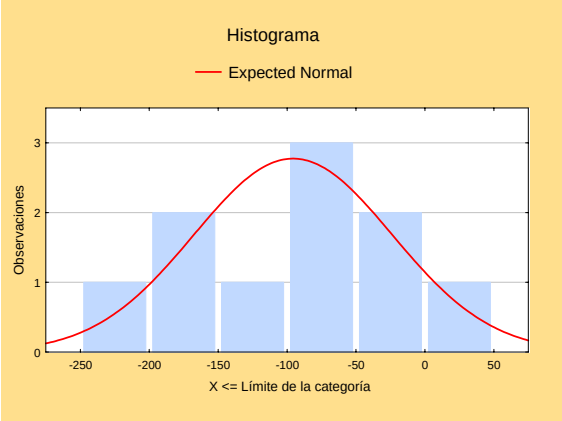
Gastos de operación netos

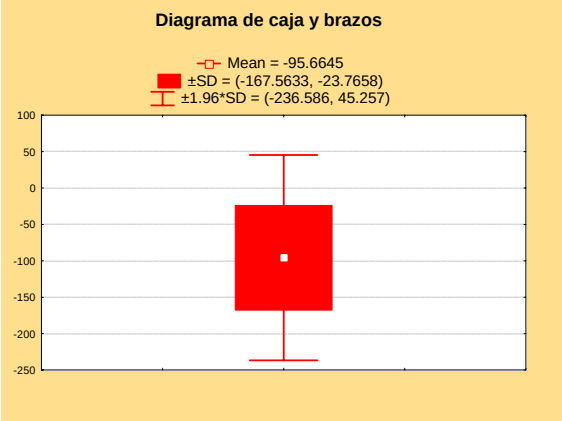




Esta variable presenta el comportamiento más homogéneo en el mercado, con una desviación estándar de apenas $5.4. La media de los gastos del mercado es de $7.7, siendo de nueva cuenta la institución 905 la que reporta el valor máximo.

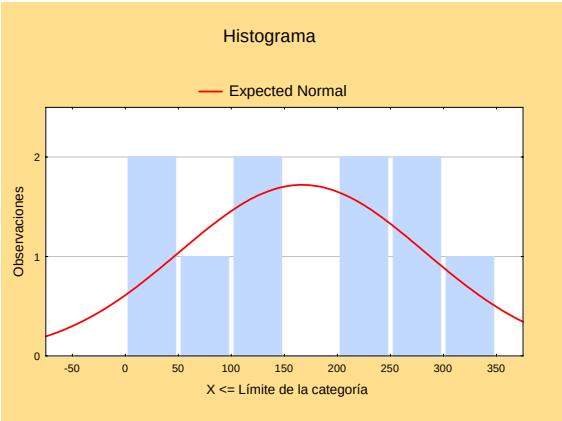
Resultado de la operación

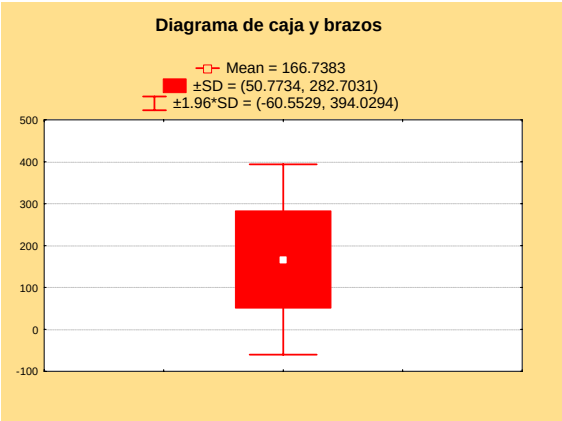




Aunado al resultado técnico negativo el cual decrece al sumar los costos de operación, el mercado reporta pérdida en la operación por $95.7 en promedio; sólo la institución 907 reporta utilidades en este rubro por $4.6.

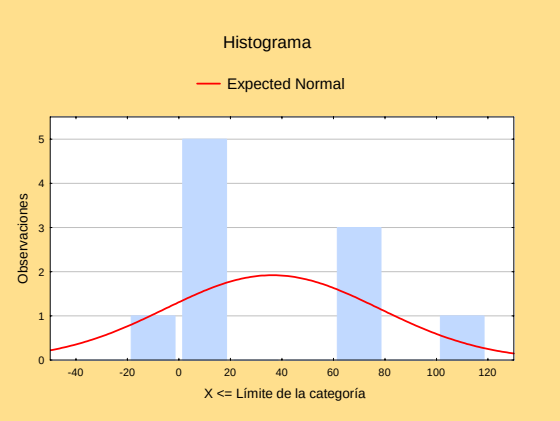
Resultado integral de financiamiento

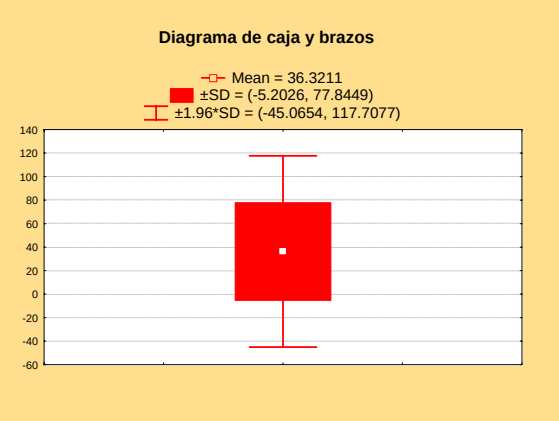




El resultado integral de financiamiento presenta un gran dispersión en sus saldos, los cuales van desde $2.4 hasta $315.4, reportando una desviación estándar de $116.0.

Resultado del ejercicio





En virtud del comportamiento de la variable anterior, la 9 de las 10 instituciones logran sobreponerse a los resultados técnicos y de operación negativos, para reportar utilidad en el ejercicio con un promedio de $36.3. La única institución que reporta pérdida es la 910, por $15.1.

4.5.2 Segmentación del mercado de ISPDLSS

Para obtener el análisis anterior, se empleó nuevamente el programa ASCAL del paquete estadístico antes mencionado.

Se agrupó toda la información antes señala en una tabla de información a partir de la cual se obtiene una matriz de disimilitudes:

ISPDLSS	MATRIZ DE DISIMILITUDES									
	901	902	903	905	906	907	908	909	910	911
901	0.0	461.5	180.0	**636.8**	435.3	181.0	317.8	87.6	267.4	183.3
902	461.5	0.0	339.1	242.4	200.1	552.5	168.8	488.1	476.9	547.3
903	180.0	339.1	0.0	491.1	276.0	249.7	176.9	166.4	280.3	246.6
905	636.8	242.4	491.1	0.0	235.6	726.3	328.6	652.1	570.7	729.5
906	435.3	200.1	276.0	235.6	0.0	521.7	144.5	435.3	393.1	514.4

907	181.0	552.5	249.7	**726.3**	521.7	0.0	411.4	147.4	408.2	98.4
908	317.8	168.8	176.9	328.6	144.5	411.4	0.0	332.5	342.1	406.2
909	87.6	488.1	166.4	**652.1**	435.3	147.4	332.5	0.0	286.5	132.2
910	267.4	476.9	280.3	570.7	393.1	408.2	342.1	286.5	0.0	414.9
911	183.3	547.3	246.6	**729.5**	514.4	98.4	406.2	132.2	414.9	0.0

De la matriz anterior se confirma que la institución que guarda mayores diferencias con respecto a las demás es la 905, cuyas distancias son significativamente mayores a las demás instituciones. Sin embargo, las demás instituciones parecen presentar cierta homogeneidad.

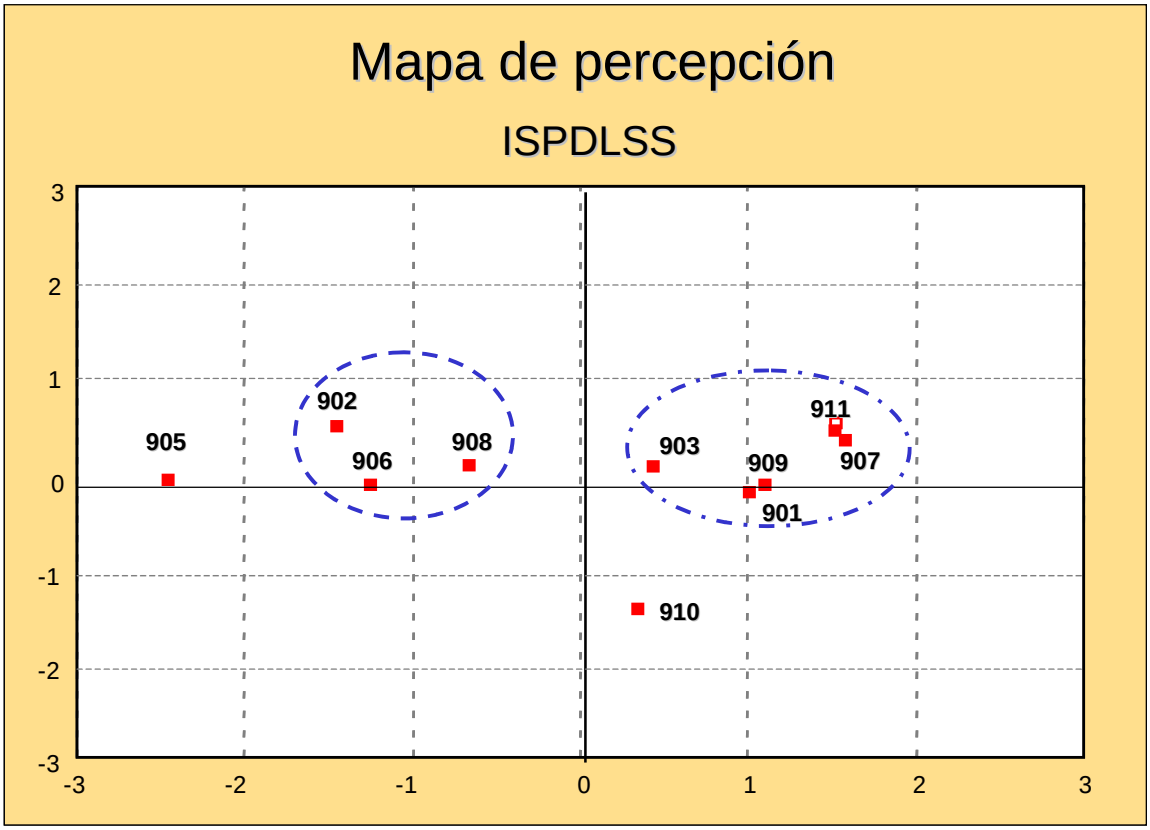
A continuación se presentan los resultados que el paquete SPSS genera al aplicar el programa ASCAL a la matriz de disimilitudes antes exhibida, para un máximo de 30 iteraciones posibles, un criterio de convergencia de 0.00100 y un valor mínimo de S-Stress de 0.00500:

El proceso sólo generó 3 iteraciones deteniéndose porque la mejora en la función S-Stress alcanzo un nivel menor a 0.001000.

Iteración	S-Stress	Mejoramiento
1	0.03205	
2	0.02779	0.00426
3	0.02768	0.00011

La función S-Stress resultante fue de 0.04618 es decir, 4.62%, el cual conforme a los criterios de bondad de ajuste para la función de Stress de Kruskal[4] se clasifica como bueno.

[4] Ver capítulo III



Conforme al escalamiento multidimensional, se obtuvieron 2 grupos de ISPDLSS integrados por:

Grupo A: HSBC Pensiones, ING pensiones, Principal Pensiones, Metlife Pensiones y Royal and Sunalliance Pensiones.

Grupo B: Pensiones Banorte, Profuturo y Pensiones Banamex.

Las instituciones Pensiones Bancomer y Pensiones Inbursa, no pertenecen a ningún grupo, como pudimos observar en la descripción de las variables de análisis, ambas presentaron algunos valores muy por encima de la media de las demás instituciones, aunque entre si, de acuerdo al análisis no guardan similitudes.

Capítulo 5

Aplicaciones del análisis de correspondencias

Se utilizará la técnica del análisis de correspondencias para segmentar nuevamente los mercados anteriores, no obstante por la naturaleza del análisis, deberán llevarse a cabo algunas consideraciones y adecuaciones en la información de las instituciones de seguros.

El análisis de correspondencias es una técnica empleada con el fin de reducir dimensiones, o bien, encontrar relaciones entre variables cuya información se caracteriza por tener una escala nominal.

5.1 Ejemplo para el mercado de las ISES

En el caso del mercado de las instituciones de Salud, la información que servirá de base para este análisis se obtuvo de las denominadas Notas de revelación de dichas aseguradoras. Para lo referente a las instituciones de pensiones, se utilizó la información estadística que publica la Comisión Nacional de Seguros y Fianzas, al cierre de 2006, la cual se encuentra disponible en su página web.

5.1.1 Notas de revelación

A través de la Circular S-18.2.2 del 7 de noviembre de 2006, publicada en el Diario Oficial de la Federación el 4 de diciembre de 2006, se dieron a conocer las

disposiciones de carácter general sobre notas a los estados financieros anuales de las instituciones de seguros.

En dichas notas, las instituciones tienen la obligatoriedad de hacer pública información cuantitativa y cualitativa relativa a su operación, situación técnico-financiera y riesgos inherentes a sus actividades, mediante la inclusión de notas a sus estados financieros anuales.

Este informe debió realizarse a partir del cierre del ejercicio 2006.

De las notas de revelación de las compañías referidas, se eligieron las siguientes variables para la segmentación del mercado, considerando únicamente variables nominales:

RUBRO DE LAS NOTAS DE REVELACIÓN	VARIABLE
Concentración geográfica del riesgo	CGR
Tipo de inversiones	INV

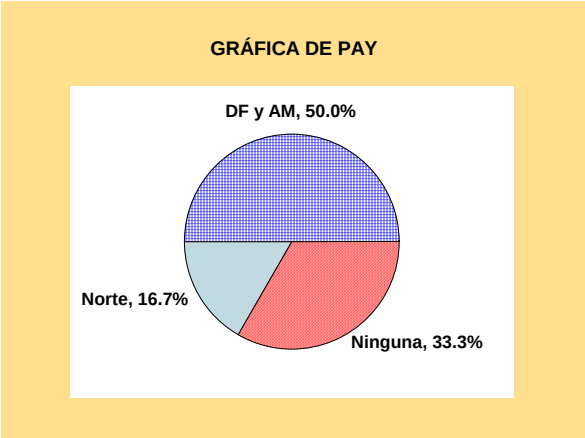
5.1.2 Análisis descriptivo de las variables

El análisis inicia construyendo una tabla de frecuencias la cual resulta del cruce de dos variables de tipo categórico, en este caso se eligió la variable "Concentración geográfica del riesgo" para las filas y "Tipo de inversiones" para las columnas.

Concentración geográfica del riesgo

En la información correspondiente al numeral I, Disposición Vigésima Novena de la Nota de Revelación 12, se solicita que las instituciones señalen "*cuando sea factible, de la distribución geográfica de sus primas emitidas, considerando que la*

concentración geográfica del riesgo asegurado se refiere a la ubicación geográfica en donde se localiza dicho riesgo, no en donde fue emitido el contrato".



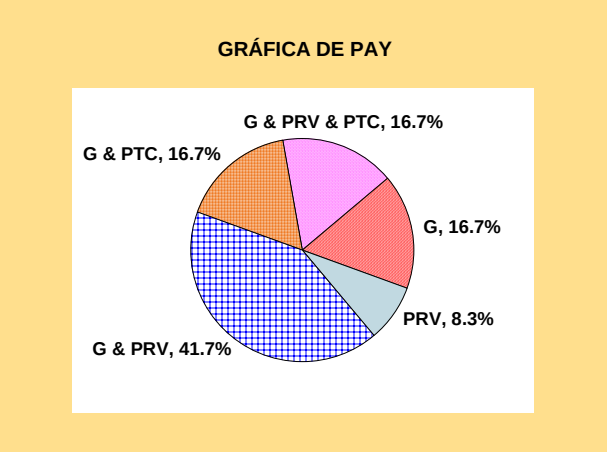
El 50.0% de las ISES reportan una concentración de primas en el Distrito Federal y el Área Metropolitana; el 16.7% indica que es en el Norte del país su concentración y el 33.3% señala que no tiene una concentración específica:

ISES	CONCENTRACIÓN GEOGRÁFICA DEL RIESGO		
	DF Y ÁREA MATROPOLITANA	NORTE	NINGUNA
701	X		
702	X		
703	X		
704	X		
705			X
706			X
707		X	
708	X		

ISES	CONCENTRACIÓN GEOGRÁFICA DEL RIESGO		
	DF Y ÁREA MATROPOLITANA	NORTE	NINGUNA
710			X
711		X	
712			X
713	X		

Tipo de inversiones

La Nota de Revelación 4, en su Disposición Séptima requiere a las instituciones presentar información referente a su portafolio de inversiones al cierre del ejercicio, presentando el detalle del valor de las inversiones, que en el caso de las ISES resultó ser para todas en Moneda Nacional, para los siguientes rubros[1]:



En donde:

G Valores Gubernamentales

PRV Valores Privados de renta variable

[1] En total son nueve tipos de valores, pero las ISES sólo reportan información en los rubros presentados.

PTC Valores Privados de tasa conocida.

ISES	TIPO DE INVERSIONES				
	G	PRV	G & PRV	G & PTC	G & PRV & PTC
701			X		
702			X		
703				X	
704	X				
705					X
706			X		
707					X
708		X			
710				X	
711	X				
712			X		
713			X		

5.1.3 Segmentación del mercado de las ISES

Con la información anterior, se generó la siguiente tabla de frecuencias:

CGR	TIPO DE INVERSIONES					
	G	PRV	G & PRV	G & PTC	G & PRV & PTC	TOTAL
DF y AM	1	1	3	1	0	6
NORTE	1	0	0	0	1	2
NINGUNA	0	0	2	1	1	4

TOTAL	2	1	5	2	2	12

En primera instancia se obtienen los perfiles de renglones y columnas, los cuales indican la presencia de un tipo de inversión en cada una de las zonas de concentración geográfica del riesgo y viceversa:

Perfiles por fila

	G	PRV	G & PRV	G & PTC	G & PRV & PTC	TOTAL
DF y AM	16.67	16.67	50.00	16.67	-	100.00
NORTE	50.00	-	-	-	50.00	100.00
NINGUNA	-	-	50.00	25.00	25.00	100.00

Perfiles por columna

	G	PRV	G & PRV	G & PTC	G & PRV & PTC
DF y AM	50.00	100.00	60.00	50.00	0.00
NORTE	50.00	0.00	0.00	0.00	50.00
NINGUNA	0.00	0.00	40.00	50.00	50.00
TOTALES	100.00	100.00	100.00	100.00	100.00

A partir de estos renglones se busca obtener una representación geométrica (un mapa de percepción) de las zonas de concentración geográfica del riesgo con base en la forma en que se distribuyen las frecuencias relativas según el tipo de inversión y viceversa, empleando para este fin la distancia χ^2.

Posteriormente mediante la distancia euclidiana cada zona de concentración geográfica del riesgo queda representada por nuevos valores los cuales consideran la forma en que se distribuyen las frecuencias de los tipos de inversión y la reducción de dimensiones para las zonas se realiza a partir de estos valores

transformados. En caso de los tipos de inversión con respecto a las zonas es análogo.

A continuación se obtiene la matriz de covarianzas, la cual genera una serie de indicadores del modelo obtenido para dar una interpretación final del mismo:

Dimensiones	Valores propios	Proporción de inercia	Inercia acumulada	χ^2
1	0.41	70.45	70.45	4.93
2	0.17	29.55	100.00	2.07

Se genera una solución en dos dimensiones, recordando que el máximo número de dimensiones posibles es el mínimo entre el número de filas en la tabla menos uno y el número de columnas de la tabla menos uno.

Los valores propios indican el porcentaje de variabilidad total que representa cada dimensión y por medio de ellos se observa el porcentaje de variabilidad que se conserva con la representación de una determinado número de dimensiones. Su suma, aporta la inercia total del modelo, es decir, la variabilidad total de los datos, que en este modelo es 0.58.

En cuanto a la proporción de inercia, se observa que la primera dimensión aporta el 70.5% de la variabilidad, mientras que la segunda dimensión aporta sólo el 29.6% de esta.

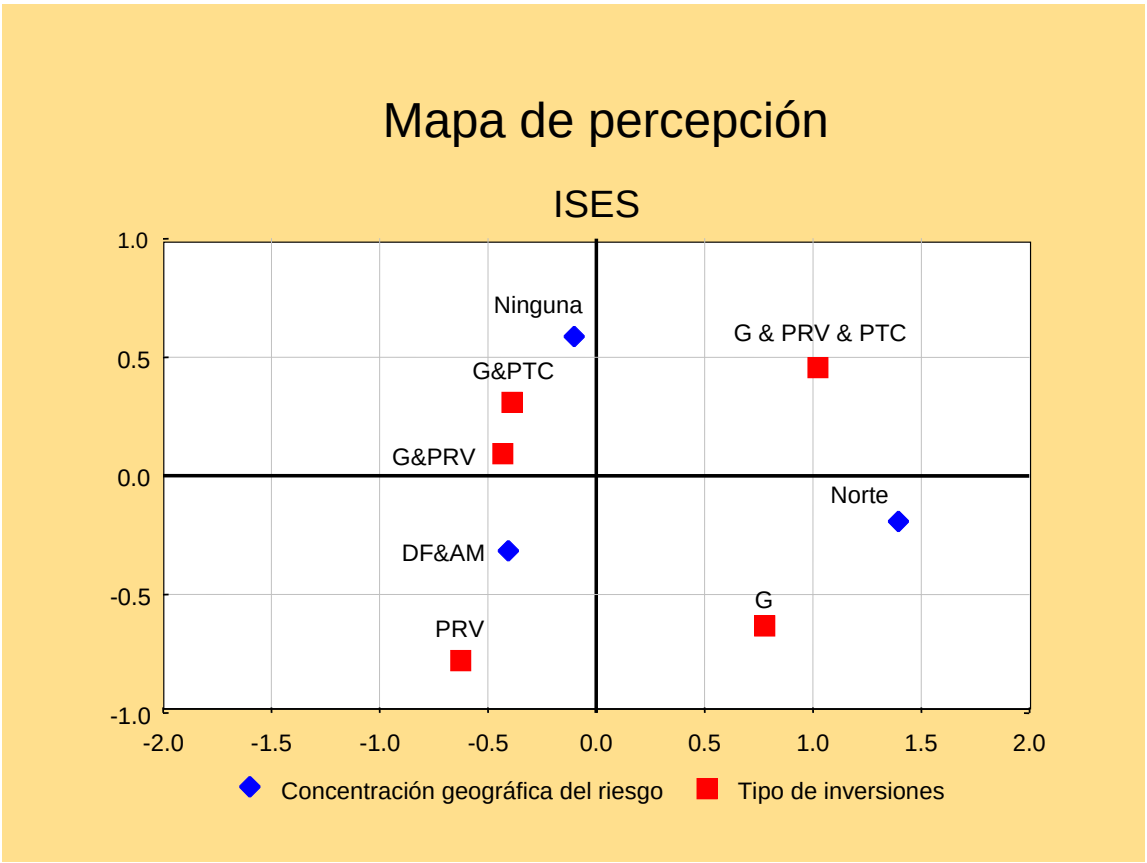
Al analizar la inercia acumulada es deseable obtener el mayor valor posible para las dimensiones elegidas, ya que este porcentaje indica la calidad de la solución obtenida y se pretende representar todas las zonas de concentración de riesgo y tipos de inversión en menos dimensiones que las originales, pero perdiendo la mínima cantidad de información. Como se observa en el párrafo anterior, la inercia

acumulada es el 100.0% para el modelo, lo cual indica que la calidad de la solución es óptima.

Finalmente, se generan las coordenadas que representarán las zonas de concentración del riesgo y los tipos de inversión en un mapa de percepción, lo cual es posible debido a que existe una correspondencia entre dichas variables. Por lo tanto, mientras más asociados estén una zona y un tipo de inversión, mayor será la presencia de ese tipo de inversión en la zona, mayor será su aportación a las coordenadas de dicha zona y ambos estarán más cerca en el plano:

	Coordenadas Dimensión 1	Coordenadas Dimensión 2
DF y AM	-0.40	-0.32
NORTE	1.40	-0.20
NINGUNA	-0.10	0.58

	Coordenadas Dimensión 1	Coordenadas Dimensión 2
G	0.78	-0.63
PRV	-0.63	-0.78
G & PRV	-0.44	0.10
G & PTC	-0.39	0.31
G & PRV & PTC	1.02	0.46



Analizando el mapa de percepciones obtenido se observa que las ISES con concentración geográfica del riesgo en la zona norte están asociadas a inversiones de tipo gubernamental, mientras que las inversiones en valores privados de renta variable se asocian a la zona D.F. y Área metropolitana. Finalmente, las instituciones que no reportan concentración geográfica de riesgo se relacionan con inversiones en valores de tipo gubernamental y privados de tasa conocida.

5.2 Ejemplo para el mercado de las ISPDLSS

De la información estadística disponible el cierre de diciembre de 2006, publicada por la Comisión Nacional de Seguros y Fianzas en su portal de internet, se eligió realizar el análisis con base en las variables “tipo de pensión” y “región” de México:

TABLA DE INFORMACIÓN ESTADÍSTICA	VARIABLE
Región	REG
Tipo de pensión	TP

5.2.1 Análisis descriptivo de las variables

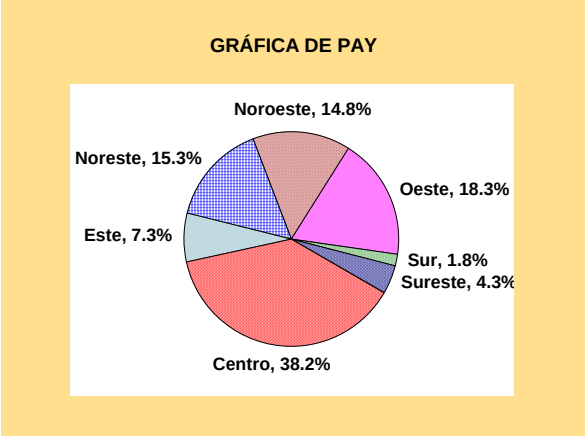
Para construir la tabla de frecuencias, en primera instancia se dividió al país en regiones:

REGIÓN	ESTADO	REGIÓN	ESTADO
CENTRO	Distrito Federal	NOROESTE	Sinaloa
CENTRO	Hidalgo	NOROESTE	Sonora
CENTRO	México	OESTE	Aguascalientes
CENTRO	Morelos	OESTE	Colima
CENTRO	Querétaro	OESTE	Guanajuato
CENTRO	Tlaxcala	OESTE	Jalisco
ESTE	Puebla	OESTE	Michoacán
ESTE	Veracruz	OESTE	Nayarit
NORESTE	Coahuila	OESTE	Zacatecas

REGIÓN	ESTADO	REGIÓN	ESTADO
NORESTE	Nuevo León	SUR	Guerrero
NORESTE	San Luis Potosí	SUR	Oaxaca
NORESTE	Tamaulipas	SURESTE	Campeche
NOROESTE	Baja California Norte	SURESTE	Chiapas
NOROESTE	Baja California Sur	SURESTE	Quintana Roo
NOROESTE	Chihuahua	SURESTE	Tabasco
NOROESTE	Durango	SURESTE	Yucatán

Región

El país se dividió en siete regiones, dependiendo del estado en donde se paga la pensión, éstas son: centro, noreste, noreste, este, oeste, sur y sureste:



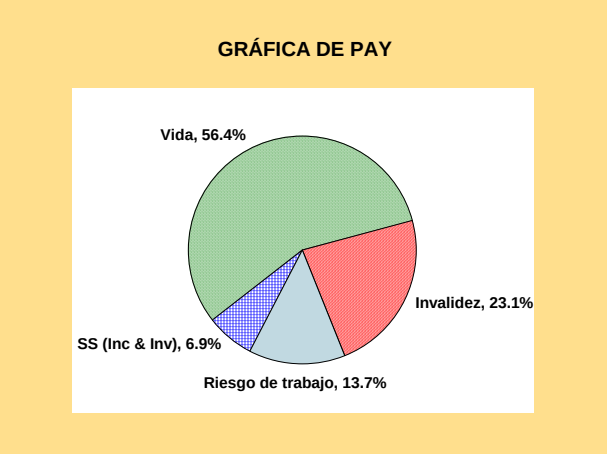
Como se observa, la región centro es predominante en el total de pensiones que se pagan en sus estados, principalmente por las cifras en el D.F.. Por otro lado, la región sur es la que menos datos aporta al análisis al representar sólo el 1.8% del total.

Tipo de pensión

Los tipos de pensión de clasifican en cinco rubros: invalidez, riesgo de trabajo, seguro de supervivencia (por incapacidad o invalidez) y vida.

TIPO DE PENSIÓN	VARIABLE
Invalidez	INV
Riesgo de trabajo	RT
Seguro de Supervivencia (Incapacidad o Invalidez)	SS (INC & INV)
Vida	VIDA

Su distribución es la siguiente:



5.2.2 Segmentación del mercado de las ISPDLSS

A continuación se muestra la tabla de frecuencias que servirá de base para el análisis:

	TIPO DE PENSIÓN				
REGIÓN	INV	RT	SS (INC& INV)	VIDA	TOTAL
CENTRO	25,520	15,929	7,008	65,664	114,121
ESTE	5,381	3,478	1,643	11,328	21,830
NORESTE	12,847	6,157	3,799	22,992	45,795

NOROESTE	9,029	5,000	3,279	26,801	44,109
OESTE	13,614	7,542	3,705	29,935	54,796
SUR	605	823	354	3,485	5,267
SURESTE	1,898	1,927	757	8,170	12,752
TOTALES	68,894	40,856	20,545	168,375	298,670

El proceso de análisis es similar al exhibido en el ejemplo de las ISES, por esta razón se omitirán algunos detalles del procedimiento, dando paso a la interpretación de la solución obtenida.

Perfiles por fila

	TIPO DE PENSIÓN				
REGIÓN	INV	RT	SS (INC& INV)	VIDA	TOTAL
CENTRO	22.36	13.96	6.14	57.54	100.00
ESTE	24.65	15.93	7.53	51.89	100.00
NORESTE	28.05	13.44	8.30	50.21	100.00
NOROESTE	20.47	11.34	7.43	60.76	100.00
OESTE	24.84	13.76	6.76	54.63	100.00
SUR	11.49	15.63	6.72	66.17	100.00
SURESTE	14.88	15.11	5.94	64.07	100.00

Perfiles por columna

	TIPO DE INVERSIÓN			
REGIÓN	INV	RT	SS (INC& INV)	VIDA
CENTRO	37.04	38.99	34.11	39.00
ESTE	7.81	8.51	8.00	6.73
NORESTE	18.65	15.07	18.49	13.66
NOROESTE	13.11	12.24	15.96	15.92
OESTE	19.76	18.46	18.03	17.78
SUR	0.88	2.01	1.72	2.07
SURESTE	2.75	4.72	3.68	4.85
TOTALES	100.00	100.00	100.00	100.00

Indicadores obtenidos de la matriz de covarianzas:

Dimensiones	Valores propios	Proporción de inercia	Inercia acumulada	χ^2
1	0.01	81.58	81.58	2296.85
2	0.00	13.27	94.85	373.71
3	0.00	5.15	100.00	144.97

Si bien la aplicación del modelo genera tres dimensiones, se puede observar que con las dos primeras se obtiene un 94.85% de inercia acumulada, lo que da un grado de calidad razonable a la solución. Por lo tanto, para efectos de la representación geométrica se utilizarán sólo dos dimensiones.

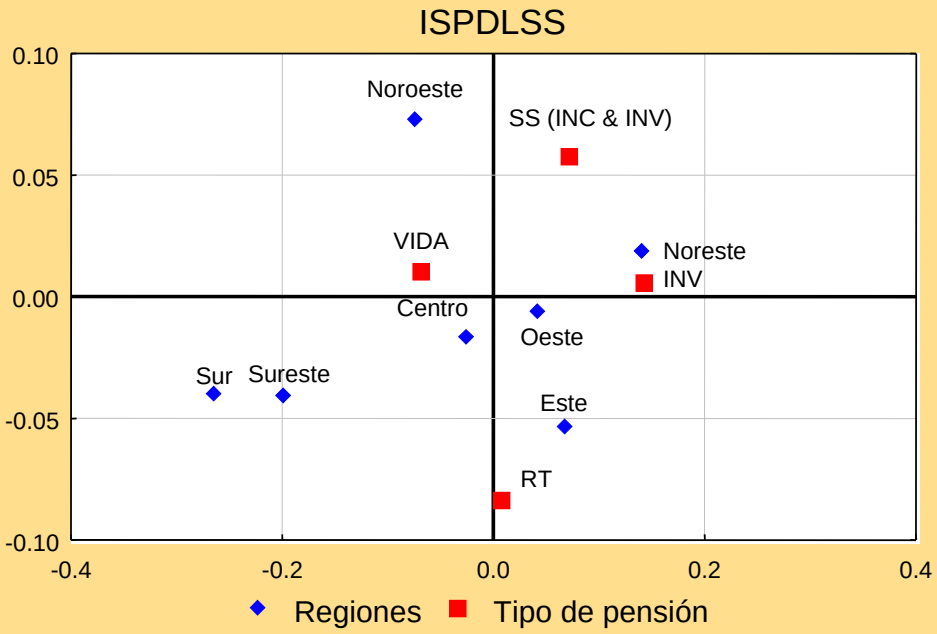
La primer dimensión aporta el 81.58% de la variabilidad, mientras que la segunda el 13.27%, quedando la tercera con tan sólo 5.15%.

Las coordenadas obtenidas para la representación de las asociaciones entre los tipos de pensiones y las regiones del país en que éstas se pagan son las siguientes:

	Coordenadas Dimensión 1	Coordenadas Dimensión 2
CENTRO	-0.03	-0.02
ESTE	0.07	-0.05
NORESTE	0.14	0.02
NOROESTE	-0.07	0.07
OESTE	0.04	-0.01
SUR	-0.26	-0.04
SURESTE	-0.20	-0.04
TOTALES	-0.03	-0.02

	Coordenadas Dimensión 1	Coordenadas Dimensión 2
INV	0.14	0.01
RT	0.01	-0.08
SS (INC&INV)	0.07	0.06
VIDA	-0.07	0.01

Mapa de percepciones



Analizando el mapa de percepciones se observa que las pensiones pagadas por invalidez están relacionadas con la región noreste, mientras que las pensiones de vida lo están con las regiones centro y oeste.

Se pagan en la región noroeste un mayor número de pensiones de seguros de supervivencia (por incapacidad e invalidez) que en otras regiones. En la zona este hay más pagos por pensiones por riesgo de trabajo.

Las regiones sur y sureste no se asocian con ningún tipo de pensiones en particular, lo anterior se explica por que el número de casos de esta región es considerablemente menor a las demás.

Conclusiones

En el trabajo desarrollado se expusieron dos técnicas de estadística multivariada de interdependencia, es decir, en las cuales se pretende identificar la estructura de la información empleada, como pueden ser variables, observaciones u objetos, de tal forma que se puedan clasificar a dichos objetos dentro de grupos como un total, en lugar de tener que analizarlos en forma individual: el análisis de correspondencias y el escalamiento multidimensional.

El análisis de correspondencias es utilizado para describir las relaciones existentes entre dos o más variables nominales sobre un espacio con el menor número de dimensiones posibles de tal forma que puedan ser representados gráficamente a través de los denominados mapas de percepción.

Dichas relaciones se reflejan a través de las distancias en el gráfico, en donde las categorías similares son representadas próximas unas de otras.

Una de las bondades de esta metodología es que puede ser aplicada a variables nominales, lo que permite segmentar con base en características físicas, de percepción, geográficas, etc., lo cual no es posible bajo otro tipo de metodologías, incluso multivariadas.

Por su parte, el escalamiento multidimensional también pretende encontrar la estructura existente pero, a diferencia del análisis de correspondencias, en un conjunto de medidas de proximidades entre objetos.

Lo anterior se logra asignando las observaciones a posiciones específicas en un espacio de pocas dimensiones, de modo que las distancias entre los

puntos en el espacio concuerden al máximo con las similaridades (o disimilaridades) dadas.

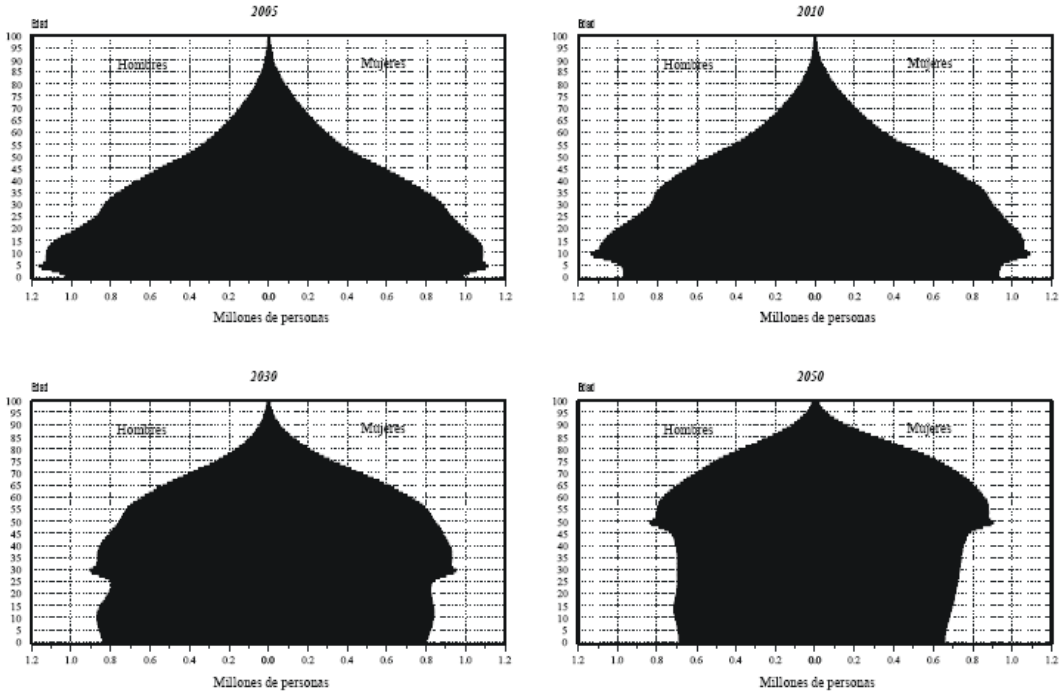
La información empleada puede ser arreglada en forma de matrices de proximidades o como variables que posteriormente se convierten a estas matrices. Dichas proximidades pueden ser en escala de razón o de intervalo.

Es necesario especificar al menos tres variables y el número de dimensiones no puede superar al número de objetos menos uno.

Las aplicaciones de estas metodologías son muy extensas debido a la flexibilidad en el tipo de variables y datos a emplear. En particular, en este trabajo se aplicó a dos tipos de seguros del mercado mexicano: los seguros de pensiones derivados de las leyes de la seguridad social y los seguros de salud.

Ambos mercados cobran cada día mayor relevancia a nivel mundial, en virtud del comportamiento que presenta la población, en el caso de México, la cual presenta una esperanza de vida mayor y “para las siguientes décadas la población de “adultos mayores” constituirá el grueso de la pirámide poblacional.

Gráfica 10. Pirámides de población, 2005-2050



Fuente: Estimaciones del CONAPO.

La situación anterior implica un gran reto para el mercado asegurador mexicano al demandar un mayor número de recursos económicos para afrontar los gastos derivados del pago de pensiones y servicios médicos que brinden una protección adecuada y suficiente a la población.

Es de suma importancia crear una conciencia colectiva sobre el impacto que estos gastos tienen sobre el patrimonio personal, lo que hace necesario adoptar un sistema de previsión personal que permita conservar la estabilidad económica ante una contingencia en la salud, así como formar un soporte financiero para la vejez.

Se analizó información financiera y estadística derivada de las notas de revelación de los estados financieros a diciembre de 2006, así como información estadística disponible en la página de la Comisión Nacional de Seguros y Fianzas para segmentar los mercados, ubicándose similitudes entre HSBC Pensiones, ING pensiones, Principal Pensiones, Metlife

Pensiones y Royal and Sunalliance Pensiones en un primero grupo y Pensiones Banorte, Profuturo y Pensiones Banamex en un segundo grupo, Pensiones Bancomer y Pensiones Inbursa no se pudieron ubicar dentro de ningún grupo.

En cuanto a las ISES, se identificaron tres grupos: el primero conformado por Salud Inbursa y SaludCoop; el segundo por Seguros Centauro, Durango, General de Salud y Preventis y finalmente ING Salud, Salud Nova y Novamedic en otro. No obstante, Plan Seguro, Médica Integral y Vitamédica no pertenecen a ninguno de los grupos citados.

No obstante que estos resultados dan un panorama muy preliminar de la segmentación de estos mercados, los grupos identificados pueden ser objeto de estudios en forma separada con el fin de obtener descripción a mayor detalle a fin de delimitar sus características particulares, con tanto nivel de detalle como la información disponible lo permita.

Actualmente no se cuenta con suficiente información sobre estas modalidades del seguro, en virtud de que se encuentran conformados por un pequeño número de aseguradoras, 12 en el caso de salud y 10 en pensiones, sin embargo, las reformas a las leyes de seguridad social, cuya penetración a partir de 2008 incluirá a los trabajadores del estado, la reversión de cuotas en el caso de los servicios de salud y la política de revelación de información por parte de las instituciones a través de modificaciones en la normativa que las regula, hará posible tener acceso a un mayor conocimiento de las características particulares de estos negocios así como de su desarrollo en el tiempo.

Anexos

Anexo I

El álgebra matricial es indispensable para describir los métodos por lo que se citan definiciones comunes que se encontrarán a través del presente trabajo y cuyo conocimiento es esencial para la comprensión de lo expuesto.

- Matriz **A**
- Renglones y columnas $(R \times C)$
- Una matriz **A** de $(m \times n)$ es un arreglo rectangular de m renglones y n columnas. El elemento (i, j) correspondiente al i-ésimo renglón de la j-ésima columna de la matriz **A** se denotará $\left(a_{i,j}\right)$
- Matriz cuadrada

Si $m = n$, **A** es una matriz cuadrada de orden n y los números $a_{1,1}, a_{2,2}, ..., a_{nn}$ forman la diagonal principal de **A**.

- Producto de matrices

Si A = $\left(a_{i,j}\right)$ es una matriz $m \times p$ y B = $\left(b_{i,j}\right)$ es una matriz $p \times n$, entonces el producto de A y B, denotada AB, es la matriz $(m \times n)$ C = $\left(c_{i,j}\right)$, definida por

$$c_{i,j} = a_{i,1}b_{1,j} + a_{i,2}b_{2,j} + ... + a_{i,p}b_{p,j} = \sum_{k=1}^{p} a_{i,k}b_{k,j} \quad (1 \le i \le m, 1 \le j \le n)$$

- Transpuesta de una matriz

Si **A** = $\left(a_{i,j}\right)$es una matriz $(m\times n)$ entonces la matriz $(n\times m)$ A^T = $\left(a^T{}_{i,j}\right)$, donde $a^T_{i,j}=a_{i,j}\quad\left(1\le i\le n,1\le j\le m\right)$es llamada transpuesta de A. Entonces la transpuesta de

A es obtenida por el intercambio de renglones y columnas de A.

- Matriz simétrica

Una matriz **A** se denomina simétrica si $A=A^T$. De igual forma **A** = $\left(a_{i,j}\right)$es simétrica si los elementos simétricos (mirror images en la diagonal) son iguales, es decir, si cada

$a_{i,j}=a_{j,i}$.

- Matriz diagonal

Una matriz diagonal es una matriz cuadrada **A** = $\left(a_{i,j}\right)$cuyos términos diferentes de la diagonal son iguales a cero, es decir, $a_{i,j}$ = 0 para $\forall\ i\neq j$

- Traza de una matriz cuadrada

Sea **A** = $\left(a_{i,j}\right)$ una matriz cuadrada. La diagonal (o diagonal principal) de A consiste de elementos $a_{1,1},a_{2,2},...,a_{n,n}$. La traza de **A** , escrita *tr **A*** , es la suma de los elementos de la dia gonal, es decir, $tr\ A=a_{1,1}+a_{2,2}+...+a_{n,n}=\sum_{i=1}^{n}a_{i,i}$

- Matriz identidad

La matriz cuadrada con 1's en la diagonal y 0´s en los demás lugares, denotada por I_n o simplemenrte I es llamada la matriz identidad.

- Determinante de una matriz cuadrada

Sea **A** = $\left(a_{i,j}\right)$ una matriz cuadrada. Se define el determinante de A (denotado $\det(A)$ o $|A|$ $\det(A)$ o $|A| = \sum(\pm) a_{1,j_1}, a_{2,j_2}, ..., a_{n,j_n}$

en donde la sumatoria sobre de los rangos de todas las permutaciones $j_1, j_2, ..., j_n$ del conjunto $S = \{1, 2, ..., n\}$ .El signo + o – se toma de acuerdo a la permutación $j_1, j_2, ..., j_n$ correspondiente.

- Inversa de una matriz y Matriz no singular

  Una matriz **A** $(n \times n)$ se denomina *no singular* (o *invertible*) si existe una matriz **B** $(n \times n)$ tal que **AB** = **BA** = I_n .

  La matriz **B** es llamada *inversa* de **A**. Si no existe tal matriz **B**, entonces **A** es llamada

  singular (o no *invertible*).

- Vectores

Un vector en el plano es un vector 2-vector

u = $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ donde x_1 y y_1 son números reales, llamados los componentes de u.

Si u = (x_1, y_1) y c es un escalar (un número real), entonces el múltiplo escalar c u de u por c es el vector (cx_1, cy_1). Es decir, el múltiplo escalar c u de u por c es obtenido multiplicando cada componente de u por c.

Si $c > 0$, entonces c u está en la misma dirección de u, sin embargo, si $d > 0$, entonces d u está en la dirección opuesta.

- Vectores linealmente independientes

Sea $S = \{v_1, v_{2,}..., v_r\}$ un conjunto no vacío de vectores, entonces la ecuación vector

$k_1 v_1 + k_2 v_2 + ... + k_r v_r = 0$ tiene una única solución, a saber $k_1 = 0, k_2 = 0, ..., k_r = 0$

Si esta es la única solución, entonces S es un conjunto linealmente independiente. Si existen otras soluciones, entonces S es un conjunto dependiente.

- Rango de una matriz

La dimensión común del espacio de renglones y el espacio de columnas de una matriz A es llamado el rango de A y es denotado por Rango (A); la dimensión de del espacio nulo de A es denominado nulidad de A y es denotado por nulidad (A).

- Vectores ortogonales

Dos vectores distintos de cero son ortogonales si y sólo si su producto punto* es igual a 0. Si se considera u y v son perpendiculares entonces alguno o ambos

vectores son 0, entonces se puede establecer sin excepción que dos vectores u y v son ortogonales (perpendiculares) si y sólo si $u \cdot v = 0$. Para indicar que u y v son vectores ortogonales escribimos $u \perp v$.

- Si u y v son vectores en el espacio de dos o tres dimensiones y θ es el ángulo entre u y v, entonces el producto punto o producto interior euclidiano $u \cdot v$ es definido por

$$u \cdot v = \begin{cases} \|u\| \ \|v\| \ \cos\theta & si\ u \neq 0\ y\ v \neq 0 \\ 0 & si\ u = 0\ y\ v = 0 \end{cases}$$

La norma del vector u es denotada por $\|u\|$. Siguiendo la forma del teorema de Pitágoras la norma del vector $u = (u_1, u_2)$ en el espacio bidimensional o tridimensional

$\|u\| = \sqrt{u_1^2 + u_2^2}$ o $\|u\| = \sqrt{u_1^2 + u_2^2 + u_3^2}$

- Vectores ortonormales

Un conjunto de vectores en el interior de un producto espacio se denomina ortogonal si todos los pares de distintos vectores son ortogonales. Un conjunto ortogonal en el cual cada vector tiene norma 1 es llamado ortonormal.

- Matriz cuadrada ortogonal

Una matriz real **A** es ortogonal si $AA^T = A^T A = I$. Observe que un matriz ortogonal A es necesariamente cuadrada e invertible, con inversa $A^{-1} = A^T$.

- Eigenvalores

Sea A una matriz $n \times n$. El número real λ es llamado un eigenvalor de A si existe un vector diferente de cero x en R^n tal que

$Ax = \lambda x$.

Cada vector x diferente de cero que satisface esta expresión es denominado eigenvector de A asociado con el eigenvalor λ. Se debe mencionar que la palabra "eigenvalor" es sólo un híbrido ("eigen" en Alemán significa "propio"). Los Eigenvalores son también llamados valores propios, valores característicos y valores latent; y los eigenvectores también son llamados vectores propios.

Anexo II

Teorema 1. Sea $D = \left[d_{ij} \right]$ una matriz de distancia y defínanse las matrices $A = \left[a_{ij} \right]$ con $a_{ij} = -\frac{1}{2} d_{ij}^2$ y $B = HAH$ donde $H = I_n - n^{-1} 1_n 1_n^t$ con $1_n = (1,1,...,1)^t$, de modo que $b_{ij} = a_{ij} - \overline{a}_{.j} - \overline{a}_{i.} + \overline{a}_{..}$. Entonces D es eculideana si y sólo si B es semidefinida positiva.

Teorema 2. Si $C \geq 0$ entonces $D = (d_{ij})$ definida por $d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{\frac{1}{2}}$ es euclideana con matriz de producto interno centrado $B = HCH$.

Bibliografía

- Ardanuy, Ramón, Cuadernos de estadística, España, La Muralla, S.A.,2000.
- Benzecri J.P., Correspondence analysis handbook, New York, Dekker, Inc, 1992.
- Borg, Ingwer, Modern multidimensional scaling, theory and applications, Nueva York, Springer-Verlag, 1997.
- Colman, Bernard, Introductory linear algebra with aplications. 6a edición, Prentice Hall, 1997.
- Comisión Nacional de Seguros y Fianzas. Circular S-18.2.2 “mediante la cual se emiten las disposiciones de carácter general sobre notas a los estados financieros anuales de las instituciones de seguros”. México, 2006.
- Comisión Nacional de Seguros y Fianzas. “Reglas para la Operación del ramo de Salud”, México, 2000.
- Comisión Nacional de Seguros y Fianzas. “Reglas de operación para los seguros de pensiones derivados de las leyes de seguridad social, México, 1997.
- Cuadras, C.M., Métodos de análisis multivariante, España, Publicaciones Universitarias S.A., 1991.
- Davison, Mark L., Multidimensional scaling, Estados Unidos, John Wiley & Sons, 1983.
- Dillon, William R., Multivariate analysis methods and applications, Estados Unidos, John Wiley & Sons, 1984.
- Friedberg, Stephen H. Álgebra lineal, México, Publicaciones Cultural, 1982.
- Hair, J.F., et al.,Analisis multivariante, 5ª edición, Prentice Hall, Madrid 1999.
- Johnson, Richard A., Applied mutivariate statistical analysis, Estados Unidos, Prentice Hall, 1988.

- Ley General de Instituciones y Sociedades Mutualistas de Seguros.
- Mardia, K. V. Multivariate analysis, Academic Press Inc. 1982
- Revista “Datos diagnósticos tendencias”, Número 36, Abril 2003, Publicación trimestral de la Asociación Mexicana de Agencias de Investigación de Mercado y Opinión Pública, A.C.
- www.cnsf.gob.mx